Embedding Glyph Identifiers in XML Documents

Christian Wittern

Introduction: What is EGIX

On December 20, 2002, a group of Japanese experts led by KAWAMATA Akira, submitted a technical note entitled Embedding Glyph Identifiers in XML Documents and now known as EGIX. This document was previously published as a technical report of the Japanese Standards Association in Japanese as JIS TR X 0047:2001[1], which in turn is partly based on the work of the ‘Extended Kanji Processing Council’ and its XXP GAIJI Exchange Specification, which outlined a way to transparently use the private use area available in Shift_JIS.

The document, which has among its contributors such well-known names as MURATA Makoto or KOMACHI Yushi, in essence identifies a namespace "http://www.xml.gr.jp/xmlns/PRE/Reference" and an attribute name that is to be used as a reference to glyphs. In slightly more detail, the content of the element, if existing, is intended to be interpreted as another possible form (in most cases the orthographic form) of the glyph that is indicated by the attribute with the name name. The value of the attribute is to be a glyph identifier as assigned by the Registration Authority in accordance with ISO 10036:1996, Cor.1:2001 *Information technology -- Font information interchange -- Procedures for registration of font-related identifiers*[2]. The document also gives three non-normative examples in an appendix. Since these examples shade some light on the intended use, I will quote them here in full[3]:

A.1 Example 1

```
<html xmlns="http://www.w3.org/1999/
xhtml">
<body xmlns:glyph="http://
www.xml.gr.jp/PRE/Reference">
<p><span glyph:name="ISO/IEC
10036/RA//Glyphs:10003290"
>吉</span>田茂</p>
</body>
</html>
```

A.2 Example 2

```
<html xmlns="http://www.w3.org/1999/
xhtml">
<body xmlns:glyph="http://
www.xml.gr.jp/PRE/Reference">
<p><span glyph:name="ISO/IEC
10036/RA//Glyphs:10003290"
>吉 (The version of Short Upper
Line)</span>田茂</p>
</body>
</html>
```

A.3 Example 3

```
<html xmlns="http://www.w3.org/1999/
```

```
xhtml">
<body xmlns:glyph="http://
www.xml.gr.jp/PRE/Reference">
<p><img glyph:name="ISO/IEC
10036/RA//Glyphs:10003290"
src="http://www.mojikyo.gr.jp/gif/
003/003290.gif"
alt="吉 (The version of Short Upper
Line)" />田茂</p>
</body>
</html>
```

In the following, I will attempt a review of this proposed standard for glyph reference exchanges, considering its appropriateness for the problem at hand in comparison to other proposals, its effectiveness, completeness, appropriateness and overall usability.

The problem: Not encoded glyphs (or characters)

The problem this proposal tries to solve is the fact that with the ever expanding use of computers and information technology, there is a constantly rising need for more characters or a more finegrained selection of glyphs for existing characters. Because of the long history of the writing systems based on sinitic logographs, and the varying orthographic standards in different times and places, a great variety of styles and forms of glyphs has developed. This situation is complicated by the fact, that some characters can be used interchangeably in certain context, while they could not in others. On the other hand, the development of encoded character sets for information processing has started in an era where resources were scarce and therefore tried to minimize the number of characters encoded.

The field of character studies, which has a long tradition in China has received important new impulses from the efforts of digitization of premodern materials and especially from the encoding and standardization efforts. Since encoding of characters for the purpose of using them in information processing started in the seventies of the last century, there has also been a considerable progress in terms of the purpose and theory of character standardization. One important distinction which has been arrived at is the distinction between a character, that is an abstract entity of information transmission and the different instantiations of such an entity in graphical forms, that might differ in style, size, stroke counts and even components, but nevertheless refer to the same abstract unit. While for the purpose of information exchange itself and many information processing purposes the graphical instantiation can be discarded as of no significance, there are on the other hand situations where the graphical appearance is an inseparatable part of the information, which might be the case where a specific glyph is used in a proper name.

In the late eighties of the last century, when efforts began to encode characters for all modern languages in one single character repertoire, it was decided that Chinese Characters should be unified where possible to avoid the situation were semantically identical characters would be encoded multiple times. A set of unification rules had been established, and the first set of more than 20000 characters was released in the early 1990. At this time, it was also recognized that there would be a need to be able to refer more specifically to a glyph, for example in selecting glyphs from fonts for typesetting. To take care of this, a separate registration for glyphs was initiated through ISO 9541:1991 (Information technology -- Font information Interchange -- Part 1. Architecture) and the above mentioned ISO 10036. The task of maintaining a glyph registry was assigned to the Association of Font Information Interchange (AFII).

The need for yet more characters was however not stilled and work began to encode more characters. With every new release of a batch of characters (at the time of this writing, there are more than 70000 graphical forms encoded and another 30,000 or more are in the pipeline), the unification rules became less stringent and more glyphs that had not been encoded because of the unification rules became encoded. And not surprisingly, the process of registration glyphs through the registry, which had be seen as an alternative to character encoding, was hardly used. By accepting so many characters into the Universal Character Set, there was not much left to do for AFII and consequently, the President of AFII asked the committees governing its conduct of business to withdraw the standard[4].

It should be mentioned here in passing, that to extend the size of coded character sets even more, as seems to be the current fashion can not be an answer to the problems here, since the character set is open, and the necessity to distinguish visual distinct forms in information processing is depending on the context, thus clearly belongs to a different layer.

Proposed solutions for the problem

As mentioned above, there is an important conceptual difference between an abstract character and a glyph instance. In information processing, it is crucial to be able to address these at different levels. In the context of markup languages, as is the case with EGIX, there is the additional layer of markup which has much more expressive power and has the potential to significantly distinguish fine shades of glyph differences. The proposal under review uses markup for this purpose and this is a significant achievement. There are other proposals with a similar purpose and it is well worth making a comparison. The proposals I have in mind are:

- SVG: altGlyph (http://www.w3.org/Graphics/SVG)
- MathML: mglyph (http://www.w3.org/Math)
- Rick Jelliffe (1999): Elements for Non-Unicode Characters in XML (http://www.ascc.net/~ricko/xcs/missing_chars.html)
- TEI: Writing System Declaration (http://www.tei-c.org/Activities/CE)

All these proposals are attempts to either allow markup to refer to completely new, hitherto unencoded glyphs, or to allow annotation of existing characters with more specific glyphs or both. The former two of these schemes allow specification of a graphical entity through a markup element, while the latter two do also introduce new elements, but in addition also provide markup constructs for additional information about the character. Compared with these proposals, the EGIX is unique in that it uses just a single attribute value to refer to the glyph. Similar to SVG and MathML, there is no mechanism in EGIX, that would allow to attach further information about the character or glyph, although the fact that in the examples of the EGIX proposal there are additional prose comments illustrate the fact that the editors were aware that there might be a need to attach further information[5]. A further main difference of EGIX to all the other proposals listed above is the fact that a normative reference to the glyph is included. In fact, this is the central aspect of the proposal, it provides a firm and stable base for exchanging non-encoded characters by allowing only references to glyphs that have been previously registered with the glyph registry.

Information exchange depends on the fact that references to identical information items can be recognized and accordingly processed. If the same glyph is referenced using the five schemes under discussing here, there would be no hope of a processing system being able to recognize the fact that they all refer to the same fact. On the other hand, a global reference system only works in practice, if it is actually used for the purpose it was intended.

The proposals by Rick Jelliffe and the TEI (which is still work in progress) do not rely on a universal identifier for a glyph. Instead they use markup constructs to associate additional information with a given glyph. This will allow a system receiving files encoded according to these proposals, to process the glyphs according to the processing context (for example rendering for display, indexing for search) as required. The association with the standard glyph, which is at best implicit in the EGIX proposal, could be expressed explicitly in the TEI proposal. Besides providing codepoints for encoding and uniquely identifying characters, Character Encoding Standards such as Unicode do provide a number of additional properties for characters[6]. To ensure proper processing, the required properties should also be available for glyphs and characters that have not previously been encoded. I do not want to give the impression that I think a central repository of glyphs is a bad thing per se, but it has to be constructed in a way, that gives also access to the properties necessary for processing. In the case of rare glyphs, an indication of the source is also highly desirable.

Since the issue of glyph registry forms such an integral part of the EGIX proposal, I would like to say a few more words about the current state of the registry, as it is publicly visible from the website at http://www.glocom.ac.jp/iso10036/ (last updated in 2001-05-01). A tally of the glyph listings there reveals the following as of 2003-07-31:

Table 1: Glyphs allocated by ISO 10036 Registration Authority[7]

alpha	7,901
cjk	21,204
hangeul	11,506
(without category)	81,743
total	122,354

While the AFII received only one single glyph registry request in two and a half years, GLOCOM has in about the same span of time registered a whopping 81,743 glyphs, which can be found on the public website mentioned above.

While this certainly is an impressing achievement, there remains a problem: The website does not provide any means to look for the glyphs registered there, except by the registry number. Since the process of assignment of these registry numbers is not explained anywhere, it might as well be random to the unsuspecting user. Now, even if somebody wades through all these pages of glyph images in search for the glyph she wants to use, there is an additional problem: Glyph number 103590 (accessible from http://media.glocom.ac.jp/kmmr/10036/glyph-table.html?cjk&103) and glyph number 10003234 (http://media.glocom.ac.jp/kmmr/10036/glyph_id_03.html) look very similar to me, except maybe that the style of the former is slightly more heavy. Both of them look also quite similar to U+53E5 (句) in the Unicode standard. From the viewpoint of information interchange, this is not desirable: If there exist multiple representations of the same information unit, there is no way to find out about this situation in a system that relies on the information contained in the reference to the glyph. And it seems very possible, given the fact that the Registering Authority for Font-Related Objects apparently is accepting huge batches of glyph sets from third parties[8], without verifying them against the database of already registered glyphs, that such duplicates will exist in large numbers. In fact, the example given above was randomly picked and there are probably several thousands, if not ten-thousands of such duplicated glyphs in the registry.

Conclusion

Given this analysis, the EGIX proposal seems to be on a poor foundation both in terms of the way it employs markup to convey its intended semantics, as well as the way glyphs are referenced. The proposal does allow only the reference to one specific glyph registry, which seems unfortunate. Also the fact that the reference uses Formal Public Identifiers (FPI) rather than URI references seems a bit odd on a technical specification for the World Wide Web. If this were changed to allow for URI references and the specification would employ some markup constructs in accordance with other W3C recommendations (for example

RDF, or even a module for HTML), the usefulness would be much greater. It should also be pointed out, that especially since this proposal addresses glyph variants in the context of East Asian text processing, Unicode already provides so called Ideographic Description Sequences (IDS), which can be used to convey information about how a character is constructed. In the example given in the EGIX proposal, the character would much more usefully be annotated as ⿱土口 . In the context of the CHISE project (see http://www.kanji.zinbun.kyoto-u.ac.jp/projects/chise), such a identifier could be used as a hook into the glyph database and provide ample information about where this glyph is encoded, while work is currently under way to allow on the fly construction of glyphs such referenced.

Disclaimer

The author currently serves as the chair of the TEI workgroup on Character Encoding. He is also a member of the CHISE project.

Endnotes

[1] See http://www.y-adagio.com/public/standards/tr_lsi_xml/lsi_xml.htm. The Japanese title of this document is XML による画像参照交換方式 , a direct translation would be ‘Exchange of image references with XML’, while the English subtitle of the document references is ‘Picture Reference Exchange by XML.’

[2] This document is available online at http://www.glocom.ac.jp/iso10036/docs/main.htm.

[3] Reluctantly, I am quoting here all three examples, although I do not think they are particularly well chosen. The example given here is the name of Yoshida Shigeru, who served as Japanese Prime Minister in the years between 1946 and 1954.

[4] See the document *Final AFII Liaison Statement Concerning ISO/IEC 10036* (SC34N92) of 1999-08-19, available at http://www.y12.doe.gov/sgml/sc34/document/0092.htm.

[5] In the context of markup languages, attaching a prose comment that is clearly not part of the text flow, but rather part of the meta layer and a comment about the digitization of this text would be better served by providing markup constructs for this. Example 2 (quoted above) is introduced with the sentence: ‘Same as Example 1, but includes information for human readers. An human readable comment was inserted. Search processors can ignore the value of span elements. As a result, the comments will not be used for search.’ It is however, difficult to see how the search processor would now about this fact. The purpose of markup certainly would be to contain such a hint. If the example where displayed in any off-the shelf Web browser, this comment would not be ignored, but rather displayed together with the other text.

[6] For more information, please see The Unicode Standard, Chapter 4. The Unicode Character Database is also online at http://www.unicode.org/ucd/

[7] The GLOCOM webpage contains a notice refering to some of these as glyph images inherited from AFII. It is however not clear to which of these this applied, I assume it is meant to apply to the first three categories and the separation into categories was given up after the transition to GLOCOM. Again, there is no explanation of this fact on the public website.

[8] Although the GLOCOM website does not mention this fact, the third of the examples given above does refer to a GIF image on the website of the Mojikyo Font Institute and the fact that there seems to be a systematic relation between the two sets of numbers does suggest this even to the uninformed.