Talk 13: TEI, Unicode, and Non-Standard Characters

James Cummings

July 2009

Non-Standard Characters

In the majority of cases, Unicode already copes for most characters that scholars need. However, there are some particular medieval abbreviations, and other uncommon glyphs which have yet to make it into Unicode. Moreover, sometimes you may wish to record palaeographic variants in a single character in order to facillitate scribal identification. (Double-compartment versus single-comparment 'a' for example.)

Since this is uncommon in the majority of text encoding, but a necessary component for those who need it, the TEI deals with this by having another optional module 'Gaiji'.

Characters come in all different fonts

Donec eris felix, multos numerabis amicos

Μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος

Þa ȳðan ȝetacniað þine deopan cræft, and . . .

phonetician /fəʊnəˈtɪʃən/ dog /dɒg/ bird /bɜːd/

И речѐ бг҃ъ: да бꙋ́детъ свѣ́тъ. И бысть свѣ́тъ.

א בְּרֵאשִׁית בָּרָא אֱלֹהִים אֵת הַשָּׁמַיִם וְאֵת הָאָרֶץ׃

ΑΒΓΔΕΥΖΗΨΙΚΛΜΝϹΠΠΥΚΣΤΥϜΧΟϤ

अथ कलेन महता स मत्स्यः सुमहानभूत् ।

Characters in TEI: Unicode

- Unicode is the only supported character encoding schema. This means that entities for characters are deprecated, and the recommended daily use is for UTF-8 encoded text, as in

`<persName xml:lang="el-grc">Φλ. Θάλλος</persName>`

- There is a clean mechanism to use non-Unicode characters
- all appropriate text content models are set to allow a mixture of CDATA and <g> (where <g> is a reference to a non-Unicode character)
- all elements have an attribute *@xml:lang* to record the language used
- there are no places where an attribute is used to hold pure text

Non-Unicode characters

If you wish to encode characters, or specific glyphs, which do not appear in Unicode:

1. define them in a series of `<charDesc>` elements, inside `<encodingDesc>` in the TEI header
2. refer to them using using the `<g>` element in the body of your text

Inside `<charDesc>`:

- the `<char>` element defines a character which is not available in the current document character set
- the `<glyph>` element annotates an existing character (usually providing a glyph that shows how a character appeared in the original document)

TEI

'Gaiji' Module Elements

- The TEI Gaiji module includes:
    - `<charDecl>` (declaration about nonstandard characters and glyphs)
    - `<char>` (descriptive information about a character)
    - `<charName>` (name of the character)
    - `<charProp>` (some property of the character)
    - `<g>` (represents a non-standard character or glyph)
    - `<glyph>` (descriptive information about a glyph)
    - `<glyphName>` (name of the glyph)
    - `<localName>` (locally defined name for a property)
    - `<unicodeName>` (name of a registered Unicode normative property)
    - `<value>` (single value for some property or attribute)
    - `<mapping>` (one or more characters related to the current character or glyph)

    You can also use `<graphic>` to provide a picture of your character or glyph.

TEI

`<charDecl>` Example

```xml
<charDecl>
  <glyph xml:id="z103">
    <glyphName>LATIN LETTER Z WITH TWO STROKES</glyphName>
    <mapping type="standardized">Z</mapping>
    <mapping type="PUA">U+E304</mapping>
  </glyph>
</charDecl>
```

Another `<charDecl>` Example

```
<charDecl>
  <glyph xml:id="r1">
    <glyphName>LATIN SMALL LETTER R WITH ONE FUNNY STROKE</glyphName>
    <charProp>
      <localName>entity</localName>
      <value>r1</value>
    </charProp>
    <graphic url="r1img.png"/>
  </glyph>
  <glyph xml:id="r2">
    <glyphName>LATIN SMALL LETTER R WITH TWO FUNNY STROKES</glyphName>
    <charProp>
      <localName>entity</localName>
      <value>r2</value>
    </charProp>
    <graphic url="r2img.png"/>
  </glyph>
</charDecl>
```

Referencing Character Declarations

One references the character declarations and glyphs by pointing to their *@xml:id* attribute with a *@ref* attribute on the <g> element:

```
<p>Wo<g ref="#r1">r</g>ds in this
manusc<g ref="#r2">r</g>ipt are sometimes
written in a funny way. </p>
```

TEI

Defining a character

A new character can assigned to a position in the Unicode Private Use Area (PUA), and also described in terms of Unicode combining characters:

```
<charDecl>
  <char  xml:id="ydotacute">
    <charName>LATIN SMALL LETTER Y WITH DOT ABOVE AND ACUTE</charName>
    <charProp>
      <localName>entity</localName>
      <value>ydotacute</value>
    </charProp>
    <mapping  type="composed">#x0079;#x0307;#x0301;</mapping>
    <mapping  type="PUA">U+E0A4</mapping>
  </char>
</charDecl>
```

Defining a local glyph

A new glyph variant can also be assigned to a position in the Unicode Private Use Area (PUA) and provide standardized form as a fallback:

```
<charDecl>
  <glyph xml:id="z103">
    <glyphName>LATIN LETTER Z WITH TWO STROKES</glyphName>
    <mapping type="standardized">Z</mapping>
    <mapping type="PUA">U+E304</mapping>
  </glyph>
</charDecl>
```

This can now be referred to using the <g> element, as in

```
<g ref="#z103"/>
```

Examples

```
<p> ... <g ref="#Filig">Fi</g>lthy riches... </p>
<!-- in the charDecl -->
<glyph xml:id="Filig">
  <glyphName>LATIN UPPER F AND LATIN LOWER I LIGATURE</glyphName>
  <graphic url="Filig.png"/>
</glyph>
```

More use of <g>

It is also possible to override what appears in the text by using markup like this

```
<g ref="#z103">z</g>
```

where the content of the <g> element can be used immediately without any lookup.

Note the likelihood that few pieces of TEI-processing software will have implemented support for the <charDecl> / <g> markup yet.

TEI

PUA: Private Use Area

The developers of the Unicode Standard have set aside an area of the codespace for the private use of software vendors, user groups, or individuals. In Unicode 5.0 there were around 137,000 code points available in this area, which should be enough for most needs. No code point assignments will be made to this area by standard bodies and only some very basic default properties have been assigned (which may be overwritten where necessary by the mechanism outlined in this chapter). Therefore, unlike all other code points defined by the Unicode Standard, PUA code points should not be used directly in documents intended for blind interchange.

TEI

The ENRICH gBank

As part of the ENRICH project we are developing a web service based *Gaiji Bank* that contains non-Unicode Characters that others can reference.

```xml
<char xml:id="f161">
  <desc>PUNCTUS ELEVATUS</desc>
  <charProp>
    <localName>entity</localName>
    <value>punctelev</value>
  </charProp>
  <mapping type="standardized">; </mapping>
  <mapping type="PUA" subtype="MUFI">U+F161</mapping>
  <graphic url="images/f161.png"/>
</char>
```

These are taken from the Medieval Unicode Font Initiative and the corresponding images automatically generated with their version of the Andron Scriptor Web font.