

Paolo Monella

A Saussurean approach to graphemes declaration
in charDecl for manuscripts encoding



Ceci n'est pas un linguiste.

paolo.monella@gmx.net

TEI Conference - Rome, October 3, 2013

Outline

- Saussure

Outline

- Saussure

Outline

- Saussure
- Currently

Outline

- Saussure
- Currently
- Proposed change

Outline

- Saussure
- Currently
- Proposed change
- Interoperability

Outline

- Saussure
- Currently
- Proposed change
- Interoperability
- “Why don't you use...?”

Outline

- Saussure
- Currently
- Proposed change
- Interoperability
- “Why don't you use...?”
- What does it take?

Saussure

Ferdinand De
Saussure

Relational nature
of signs
within a semiotic
system



Ceci n'est pas un linguiste.

Saussure

MS A

a b c d e f g h **i** l m n o p q r s t **u** z . :

MS B

a b c d e f g h **i j** l m n o p q r s t **u v** z . , :

Saussure

MS A

a b c d e f g h **i** l m n o p q r s t **u** z . :

MS B

a b c d e f g h **i** **j** l m n o p q r s t **u** **v** z . , :

Saussure

MS A

a b c d e f g h i l m n o p q r s t u z . :

yes: sir

U+003A

MS B

a b c d e f g h i j l m n o p q r s t u v z . , :

he said: alas



Saussure

MS A

a b c d e f g h i l m n o p q r s t u z . :

Every noun

U+0075

≠

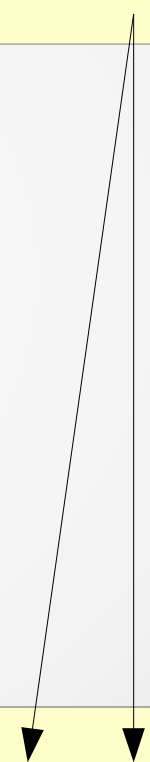
U+0076

U+0075

MS B

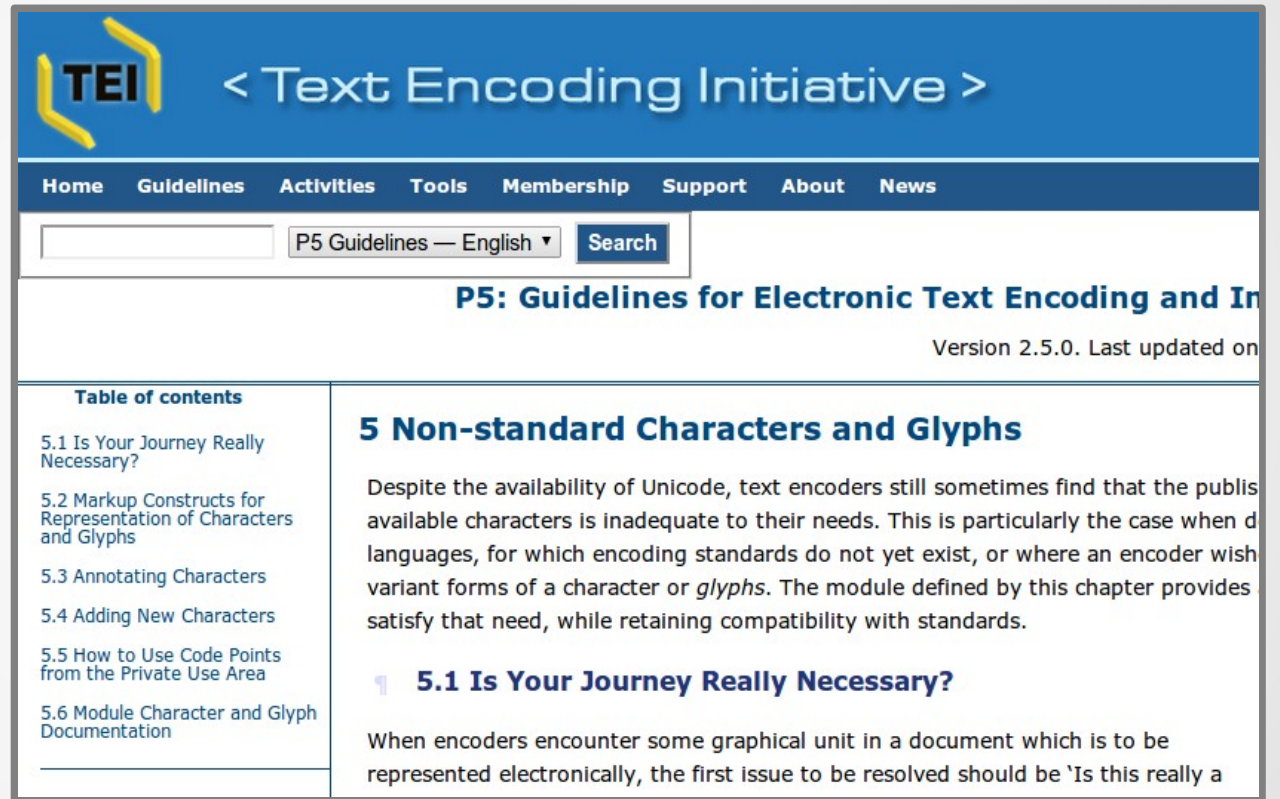
a b c d e f g h i j l m n o p q r s t u v z . , :

Every noun



Currently

- Unicode
(a “u” is a “u”)



The screenshot shows the Text Encoding Initiative (TEI) website. The header features the TEI logo and the title "< Text Encoding Initiative >". Below the header is a navigation bar with links: Home, Guidelines, Activities, Tools, Membership, Support, About, and News. A search bar is present with a dropdown menu set to "P5 Guidelines — English" and a "Search" button. The main content area is titled "P5: Guidelines for Electronic Text Encoding and Interpretation" with a subtitle "Version 2.5.0. Last updated on". A table of contents is on the left, listing sections 5.1 through 5.6. The main content area displays the title "5 Non-standard Characters and Glyphs" and a paragraph explaining that despite the availability of Unicode, text encoders still sometimes find that the published available characters are inadequate to their needs. It also mentions that the module defined by this chapter provides a way to satisfy that need while retaining compatibility with standards. Below this, section "5.1 Is Your Journey Really Necessary?" is introduced, followed by a paragraph stating that when encoders encounter some graphical unit in a document which is to be represented electronically, the first issue to be resolved should be 'Is this really a

TEI < Text Encoding Initiative >

Home Guidelines Activities Tools Membership Support About News

P5 Guidelines — English Search

P5: Guidelines for Electronic Text Encoding and Interpretation
Version 2.5.0. Last updated on

Table of contents

- 5.1 Is Your Journey Really Necessary?
- 5.2 Markup Constructs for Representation of Characters and Glyphs
- 5.3 Annotating Characters
- 5.4 Adding New Characters
- 5.5 How to Use Code Points from the Private Use Area
- 5.6 Module Character and Glyph Documentation

5 Non-standard Characters and Glyphs

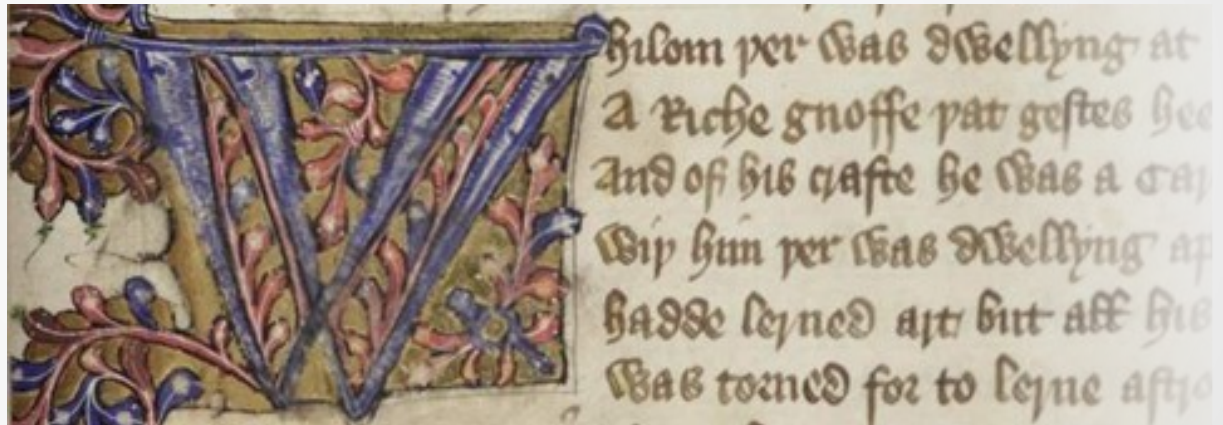
Despite the availability of Unicode, text encoders still sometimes find that the published available characters is inadequate to their needs. This is particularly the case when dealing with languages, for which encoding standards do not yet exist, or where an encoder wishes to represent variant forms of a character or *glyphs*. The module defined by this chapter provides a way to satisfy that need, while retaining compatibility with standards.

5.1 Is Your Journey Really Necessary?

When encoders encounter some graphical unit in a document which is to be represented electronically, the first issue to be resolved should be 'Is this really a

Currently

- Unicode
(a “u” is a “u”)
- Corpus-wide
normalization
(Canterbury
Tales)



Currently

- Unicode
(a “u” is a “u”)
- Corpus-wide
normalization
(Canterbury
Tales)
- Documentation

2. The practice of this transcription

A specially designed computer screen font was used for the transcription.

a b c d e f g h i j k l m n o p q r s t u v w x y
A B C D E F G H I J K L M N O P Q R S T U V W X Y

These were supplemented by the following Middle English characters:

þ ȝ þ ȝ

The following signs of abbreviation were available:

p p P P² 9 9 v f q 7

These characters are available as superscripts:

a e i r t u

Characters usually occurring at word ends are:

7 H h d

Marks of punctuation are:⁹

¶ , ; : . ()

Currently



2. The practice of this transcription

A specially designed computer screen font was used for the transcription.

a b c d e f g h i j k l m n o p q r s t u v w x y
A B C D E F G H I J K L M N O P Q R S T U V W X Y

These were supplemented by the following Middle English characters:

þ ȝ ȝ

The following signs of abbreviation were available:

p p P P 3- 9 + f q ȝ

These characters are available as superscripts:

a e i r t u

Characters usually occurring at word ends are:

ȝ H h d

Marks of punctuation are:⁹

¶ , ; : . ()

a
b
fun-b
c
d
...

Currently

<charDecl>

~~a
b~~

<glyph id="fun-b">



2. The practice of this transcription

A specially designed computer screen font was used for the transcription.

abcdefghijklmnopqrstuvwxyz
ABCDEFGHIJKLMNOPQRSTUVWXYZ

These were supplemented by the following Middle English characters:

١٣٣٣

The following signs of abbreviation were available:

$p, p, p, p, 2-9, 9, v, f, q, j$

These characters are available as superscripts:

acirtu

Characters usually occurring at word ends are:

 $7\text{Hh}\text{d}$

Marks of punctuation are:⁹

9, ; : . ()

- a
- b
- fun-b
- c
- d
- ...

Proposed change: <char>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

a
b
fun-b
c
d
...

Proposed change: <char>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="v">

<charName>Premodern Latin
uncial lowercase v</charName>

<charProp>

<localName>**Expression**

</localName>

<value>U+0076</value>

<localName>**Content**

</localName>

<value>v</value>

<charProp>

<mapping type="Unicode">v

</mapping>

<graphic url="v.jpg"/

</char>

<charDecl>

Proposed change: <char>

<charDecl>

~~a~~

~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

Proposed change: <char>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Proposed change: <char>

<charDecl>

~~a~~

~~b~~

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

Proposed change: <char>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

Proposed change: <char>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

Proposed change: <char>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

Vespa Project

<http://www.unipa.it/paolo.monella/lincei/>

Proposed change: no <g>

<charDecl>

~~a~~
~~b~~

<glyph id="fun-b">

Guide-
lines

<charDecl>

<char xml:id="a">

<char xml:id="b">

<glyph id="fun-b">

<body>

a

b

<g ref="#fun-b">

Techn.

<body>

<g ref="#a">

<g ref="#b">

<g ref="#fun-b">

<body>

a

b

<g ref="#fun-b">

Interoperability

MS A

<charDecl>

</charDecl>

<body>
euery noun
</body>

MS B

<charDecl>

</charDecl>

<body>
e**v**ery noun
</body>



Interoperability

MS A

```
<charDecl>  
  <char xml:id="y">  
    <mapping type="Unicode">  
      y  
    </mapping>  
  </char>  
</charDecl>
```

```
<body>  
  euery noun  
</body>
```

MS B

```
<charDecl>  
  <char xml:id="y">  
    <mapping type="Unicode">  
      y  
    </mapping>  
  </char>  
</charDecl>
```

```
<body>  
  every noun  
</body>
```



Interoperability

MS A

```
<charDecl>
  <char xml:id="u">
    <mapping type="mapto_u">
      u
    </mapping>
    <mapping type="mapto_v">
      v
    </mapping>
  </char>
</charDecl>
```

```
<body>
  e
  <g ref="#u" type="mapto_v"/>
  ery noun
</body>
```

MS B

```
<charDecl>
  <char xml:id="u"> [...]
  <char xml:id="v">
    <mapping type="standard">
      v
    </mapping>
  </char>
</charDecl>
```

```
<body>
  every noun
</body>
```

“Why don't you use...?”

- Medieval Unicode Font Initiative

“Why don't you use...?”

- Medieval Unicode Font Initiative
- Unicode / Private Use Area (PUA)

“Why don't you use...?”

- Medieval Unicode Font Initiative
- Unicode / Private Use Area (PUA)
- Gaiji / <glyph>

“Why don't you use...?”

- Medieval Unicode Font Initiative
- Unicode / Private Use Area (PUA)
- Gaiji / <glyph>
- SGML / Writing System Declaration (WSD)

What does it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

```
<char xml:id="a">
```

```
<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

What does it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

```
<char xml:id="a">
```

```
<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

- More lines of code

What does it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

```
<char xml:id="a">
```

```
↑<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

- More lines of code
- Guidelines: `<char>`

What does it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

```
<char xml:id="a">
```

```
<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

- More lines of code
- Guidelines: `<char>`

- Technical: no `<g>`

What does it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

```
<char xml:id="a">
```

```
<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

- More lines of code
- Guidelines: `<char>`

- Technical: no `<g>`
- Interoperability: `<mapping>`

What does it take?

```
<charDecl>
```

~~a~~
~~b~~

```
<glyph id="fun-b">
```

Guide-
lines

```
<charDecl>
```

```
<char xml:id="a">
```

```
<char xml:id="b">
```

```
<glyph id="fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

Techn.

```
<body>
```

```
<g ref="#a">
```

```
<g ref="#b">
```

```
<g ref="#fun-b">
```

```
<body>
```

a

b

```
<g ref="#fun-b">
```

- More lines of code
- Guidelines: `<char>`

- Technical: no `<g>`
- Interoperability: `<mapping>`