

TEI Internationalization

Sebastian Rahtz

The Text Encoding Initiative

October 2005

TEI I18N and L10N

The Text Encoding Initiative Guidelines have been widely adopted by projects and institutions in many countries in Europe, North America, and Asia, and are used for encoding texts in dozens of languages.

We need to make sure that the TEI and its Guidelines are *internationalized* and *localized* so that they are accessible in all parts of the world.

Building blocks

The P5 revision of the TEI has made substantial changes to support international use:

- Unicode is the only supported character encoding schema
- 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 reference to a non-Unicode character)
- all elements have an attribute `xml:lang`
- there are no places where an attribute is used to hold pure text

Example of declaring a non-Unicode character

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

Declaring non-Unicode

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

Using non-Unicode

With these definitions in place, occurrences of these two special "r"s in the text can be annotated using the element `<g>`:

```
<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>
```

Support for internationalized schemas

The TEI is written in a high-level markup language to describe the schemas, in which:

- there is allowance for translating element name, attribute names, etc, and preserving information to allow canonicalisation
- there technical documentation elements (`<gloss>`, `<desc>`) for TEI elements, attributes etc can be specified multiple times, in different languages
- there is a container (`<equiv>`) to specify relationship of an element, attribute or value to standardised schemes

How does translating names work?

The normal schema:

```
emph =  
  element emph { emph.content, emph.attributes }  
emph.attributes =  
  ...  
  [ a:defaultValue = "emph" ]  
  attribute TEIform { text }?
```

In German:

```
emph =  
  element Betonung { emph.content, emph.attributes }  
emph.attributes =  
  ...  
  [ a:defaultValue = "emph" ]  
  attribute TEIform { text }?
```

Using translated element names

```
</preliminares>
- < cuerpo >
  - <div1 tipo="part">
    - <div2 tipo="act">
      <encabezado tipo="main">Jornada primera</encabezado>
    - <div3 tipo="scene">
      <encabezado tipo="main">Cuadro único</encabezado>
    - <acotacion formato="centered">
      <resaltado formato="bold">(Salen </resaltado>
      REBOLLEDO,
      <resaltado formato="bold">la </resaltado>
      CHISPA
      <resaltado formato="bold"> y soldados </resaltado>
      .
      <resaltado formato="bold">)</resaltado>
    </acotacion>
  - <dialogo>
    <hablante>REBOLLEDO</hablante>
    <verso>¡Cuerpo de Cristo con quien</verso>
    <verso>desta suerte hace marchar</verso>
    <verso>de un lugar a otro lugar</verso>
    <verso parte="I"> sin dar un refresco!</verso>
  </dialogo>
```

What we could do to improve things further

- translate descriptive prose to other languages
- translate technical documentation components (note that this includes gloss for fixed attribute lists)
- translate examples
- localize examples
- add W3C ITS information
- translate processing workflow tool

The components of the TEI Guidelines

- 1 The detailed descriptive prose of the Guidelines chapters and TEI Lite documentation.
- 2 The element, attribute names and suggested attribute values which are put into DTDs and schemas.
- 3 The summary technical descriptions of elements or attributes.
- 4 The examples of usage for each element.
'Internationalization' of these could take the form of simple translation, but in practice *localisation* would be considerably more useful.

Localisation involves choosing examples originating in the target language, which illustrate the element's usage more effectively for a native speaker than a translated example could do.

Examples of translation

- instead of `<addrLine>`, the TEI user might prefer to write `<líneaDirección>`, `<ligneAdresse>`, `<linDireccio>` or `<AdressZeile>`.
- instead of contains a single TEI-conformant document, comprising a TEI header and a text, either in isolation or as part of a `teiCorpus` element., the Spanish-speaking user might find it more helpful to read `contiene un único documento TEI, compuesto de una cabecera TEI (TEI header) y un cuerpo de texto (text), aislado o como parte de un elemento corpusTei (teiCorpus)`

Localisation of examples

What does this

```
<lg>
  <l>Sire Thopas was a doghty swayn;</l>
  <l>White was his face as payndemayn,</l>
  <l>His lippes rede as rose;</l>
  <l>His rode is lyk scarlet in grayn,</l>
  <l>And I yow telle in good certayn,</l>
  <l>He hadde a semely nose.</l>
</lg>
```

mean to a Chinese scholar?

Example of reference documentation

person

<elementSpec>

person	describes a single participant in a language interaction.
<i>Declaration</i>	<pre>element person { tei.global.attributes, ## specifies the role of this participant in the group. attribute role { text }?, ## specifies the sex of the participant. attribute sex { ## (male) "m" ## (female) "f" ## (unknown or inapplicable) "u" }?, ## specifies the age group to which the participant belongs. attribute age { text }?, (p+ tei.demographic*) }</pre>
<i>Attributes</i>	(In addition to global attributes) role specifies the role of this participant in the group. <i>Status:</i> Optional <i>Datatype:</i> datatype.Code <i>Values:</i> a set of keywords to be defined sex specifies the sex of the participant

Example of reference documentation in Japanese

TEI Internationalization

Sebastian Rahtz

person	言語活動の関係者(1件1名)
宣言	<pre>element person { tei.global.attributes, ## 当該関係者の言語活動における役割 attribute role { text }?, ## 関係者の性別 attribute sex { ## (男性) "m" ## (女性) "f" ## (不明または不適切) "u" }?, ## 当該関係者の年齢層 attribute age { text }?, (p+ tei.demographic*) }</pre>
属性	

Example of reference documentation in Spanish

TEI Internationalization

Sebastian Rahtz

person	describe un único participante en una interacción lingüística.
<i>Declaración</i>	<pre>element person { tei.global.attributes, ## especifica el papel de este participante dentro del grupo. attribute role { text }?, ## especifica el sexo del participante. attribute sex { ## (masculino) "m" ## (femenino) "f" ## (desconocido o inaplicable.) "u" }?, ## especifica la edad del grupo al que pertenece el ## participante. attribute age { text }?, (p+ tei.demographic*) }</pre>
<i>Atributos</i>	<p>(Además de los atributos globales)</p> <p>role especifica el papel de este participante dentro del grupo. <i>Estado:</i> Opcional <i>Tipo de datos:</i> datatype.Code <i>Valores:</i> un conjunto de palabras clave a definir.</p> <p>sex especifica el sexo del participante. <i>Estado:</i> Opcional</p>

What are we doing in practice

The TEI Consortium is working with TEI scholars to advance I18N and L10N in various languages:

- Chinese** Marcus Bingenheimer (Chung-hwa Institute of Buddhist Studies, Taipei)
- French** Veronika Lux (University of Nancy)
- German** Werner Wegstein (Würzburg University)
- Hindi** Paul Richards (UGS (The PLM Company))
- Italian** Fabio Ciotti (University of Roma)
- Japanese** Ohya Kazushi (Tsurumi University, Yokohama)
- Polish** Radoslaw Moszczynski (Warsaw University)
- Romanian** Dan Matei (CIMEC - Institutul de Memorie Culturala, Bucharest)
- Slovenian** Tomaž Erjavec, Matija Ogrin (Jozef Stefan Institute, Ljubljana)
- Spanish** Manuel Sánchez (Miguel de Cervantes Digital Library)

The 2006 project

We hope to work on French, Spanish, German, Chinese and Japanese in 2006, and produce

- translated element and attribute names
- translated `<desc>` and `<gloss>` texts
- a mechanism to allow users to easily take advantage of the work

Infrastructure work

We need to change Roma to support the following output schemes:

- English names, descriptions in English
- English names, descriptions in chosen language)
- names designed to make sense to a speaker of the chosen language, descriptions in English
- both names and descriptions in chosen language