TEI Internationalization

Sebastian Rahtz

The Text Encoding Initiative

October 2005
TEI internationalisation and localisation

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.
Definitions

**Internationalization (I18N)** Internationalization is the process of generalizing a product so that it can handle multiple languages and cultural conventions without the need for redesign. Internationalization takes place at the level of program design and document development.

**Localization (L10N)** Localization is the process of taking a product and making it linguistically and culturally appropriate to a given target locale (country/region and language) where it will be used.

http://www.w3.org/TR/itsreq/#intro_definitions
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 glyph in the private use area

```xml
<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>
```
Referencing the glyphs

We may now refer to

<g ref="#z103"/>

and expect the processing application to work out what to do
Declaring a variant on a Unicode character

```
<charDesc>
  <glyph xml:id="r1">
    <glyphName>LATIN 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 R WITH TWO FUNNY STROKES</glyphName>
    <charProp>
      <localName>entity</localName>
      <value>r2</value>
    </charProp>
    <graphic url="r2img.png"/>
  </glyph>
</charDesc>
```
Referencing declared glyphs

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

```xml
<p>Words in this manuscript are sometimes written in a funny way.</p>
```

What appears could be either the letters "r1" and "r2" or the graphic files.
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 are technical documentation elements (\texttt{<gloss>}, \texttt{<desc>}) for TEI elements, attributes etc can be specified multiple times, in different languages
- there is a container (\texttt{<equiv>}) to specify relationship of an element, attribute or value to standardised schemes
How does translating names work?

The normal schema:

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

In German:

```xml
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">(Sale)</resaltado>
      REBOLLEDO,
      <resaltado formato="bold">la</resaltado>
      CHISPA
      <resaltado formato="bold">y soldados</resaltado>
    - <acotacion formato="bold">) (Sale)</acotacion>
  </dialogo>
  - <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 part="1"> 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

```xml
<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

- **Declaration**

  ```xml
  <elementSpec>
  <element name="person">
    <!-- TEI global attributes, specify the role of this participant in the group. -->
    <attribute name="role" type="text"/>
    <!-- specifies the sex of the participant. -->
    <attribute name="sex">
      <!-- (male) -->
      "m"
      <!-- (female) -->
      "f"
      <!-- (unknown or inapplicable) -->
      "u"
    </attribute>
    <!-- specifies the age group to which the participant belongs. -->
    <attribute name="age" type="text">
      ( p+ | tei:demographic+ )
    </attribute>
  </element>
  </elementSpec>
  ```

- **Attributes**

  - **role** specifies the role of this participant in the group.
    - **Status:** Optional
    - **Datatype:** `datatype.code`
    - **Values:** a set of keywords to be defined

---

*Note: The above code snippet is a simplified example of how to define an XML element in the TEI (Text Encoding Initiative) standard.*
Example of reference documentation in Japanese

<table>
<thead>
<tr>
<th>person</th>
<th>言語活動の関係者(1件1名)</th>
</tr>
</thead>
<tbody>
<tr>
<td>宣言</td>
<td>element person</td>
</tr>
<tr>
<td></td>
<td>{</td>
</tr>
<tr>
<td></td>
<td>tei.global.attributes,</td>
</tr>
<tr>
<td></td>
<td>## 当該関係者の言語活動における役割</td>
</tr>
<tr>
<td></td>
<td>attribute role { text }?,</td>
</tr>
<tr>
<td></td>
<td>## 関係者の性別</td>
</tr>
<tr>
<td></td>
<td>attribute sex</td>
</tr>
<tr>
<td></td>
<td>{</td>
</tr>
<tr>
<td></td>
<td>## (男性)</td>
</tr>
<tr>
<td></td>
<td>&quot;m&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;f&quot;</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;u&quot;</td>
</tr>
<tr>
<td></td>
<td>}?,</td>
</tr>
<tr>
<td></td>
<td>## 当該関係者の年齢層</td>
</tr>
<tr>
<td></td>
<td>attribute age { text }?,</td>
</tr>
<tr>
<td></td>
<td>( p+</td>
</tr>
<tr>
<td>属性</td>
<td>(プロードバーメプランの他)</td>
</tr>
<tr>
<td>person</td>
<td>describe un único participante en una interacción lingüística.</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------</td>
</tr>
<tr>
<td>Declaración</td>
<td>element person</td>
</tr>
<tr>
<td></td>
<td>tei.global.attributes,</td>
</tr>
<tr>
<td></td>
<td>## especifica el papel de este participante dentro del grupo.</td>
</tr>
<tr>
<td></td>
<td>attribute role { text }?,</td>
</tr>
<tr>
<td></td>
<td>## especifica el sexo del participante.</td>
</tr>
<tr>
<td></td>
<td>attribute sex</td>
</tr>
<tr>
<td></td>
<td>{</td>
</tr>
<tr>
<td></td>
<td>## (masculino)</td>
</tr>
<tr>
<td></td>
<td>'m'</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>'f'</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>'u'</td>
</tr>
<tr>
<td></td>
<td>},</td>
</tr>
<tr>
<td></td>
<td>## especifica la edad del grupo al que pertenece el</td>
</tr>
<tr>
<td></td>
<td>## participante.</td>
</tr>
<tr>
<td></td>
<td>attribute age { text }?,</td>
</tr>
<tr>
<td></td>
<td>( p+</td>
</tr>
<tr>
<td>Atributos</td>
<td>(Además de los atributos globales)</td>
</tr>
<tr>
<td>role</td>
<td>especifica el papel de este participante dentro del grupo.</td>
</tr>
<tr>
<td></td>
<td><em>Estado</em>: Opcional</td>
</tr>
<tr>
<td></td>
<td><em>Tipo de datos</em>: <code>datatype.Code</code></td>
</tr>
<tr>
<td></td>
<td><em>Valores</em>: un conjunto de palabras clave a definir.</td>
</tr>
<tr>
<td>sex</td>
<td>especifica el sexo del participante.</td>
</tr>
<tr>
<td></td>
<td><em>Estado</em>: Opcional</td>
</tr>
</tbody>
</table>
Example of reference documentation in Japanese, with German annotation

1. `<elementSpec> person`

Deklaration

```xml
element person {
    att.global.attributes,
    attribute role { text }?,
    attribute sex { "m" | "f" | "u" }?,
    attribute age { text }?,
    ( p+ | tei démographic* )
}
```

**Attribute:** (Neben global gültigen Attributen)

- **role**
- **sex**

  Gültige Werte:
  - m  
    男性
  - f  
    女性
  - u  
    不明または不適切

- **age**

**Beispiel**

```xml
<person sex="f" age="42">
    <p>Female informant, well-educated, born in Shropshire
    UK, 12 Jan 1950, of unknown occupation.
    Speaks French fluently. Socio-Economic status B2.</p>
</person>
```
What are we doing in practice

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

- Chinese
- Dutch
- French
- German
- Hindi
- Italian
- Japanese
- Polish
- Portuguese
- Romanian
- Serbian
- Slovenian
- Spanish
- Swedish
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