TEI Internationalization

Sebastian Rahtz

The Text Encoding Initiative

October 2005
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
Building blocks

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

- Unicode is the only supported character encoding scheme
- 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
- 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

```
<charDesc>
  <glyph 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 (either show the PUA character, or the standardized form).
Declaring a variant on a Unicode character

```xml
<charDesc>
  <glyph id="r1">
    <glyphName>LATIN SMALL LETTER R WITH ONE FUNNY STROKE</glyphName>
    <charProp>
      <localName>entity</localName>
      <value>r1</value>
    </charProp>
    <mapping type="standardized">r</mapping>
    <graphic url="r1img.png"/>
  </glyph>
  <glyph id="r2">
    <glyphName>LATIN SMALL LETTER R WITH TWO FUNNY STROKES</glyphName>
    <charProp>
      <localName>entity</localName>
      <value>r2</value>
    </charProp>
    <mapping type="standardized">r</mapping>
    <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>Wo<g ref="#r1">r</g>d in this <g ref="#r2">r</g>ipt are sometimes written in a funny way.</p>
```

What appears could be either the letter "r" in both cases, or the graphic files from the metadata.
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

<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 REBOLLEDO, CHISPA</resaltado> soldados)</acotacion>
        </div3>
      </div2>
    </div1>
  </cuerpo>
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>
<1>Sire Thopas was a doghty swayn;</1>
<1>White was his face as payndemayn,;</1>
<1>His lippes rede as rose;</1>
<1>His rode is lyk scarlet in grayn,;</1>
<1>And I yow telle in good certayn,;</1>
<1>He hadde a semely nose.</1>

</lg>

mean to a Chinese scholar?
1. `<elementSpec> person`

  describes a single participant in a language interaction.

  Declaration

  ```
  <element person
  { attr.global.attributes,
    attribute role { text }?,
    attribute sex { "m" | "f" | "u" }?,
    attribute age { text }?,
    ( p+ | text:demographic* )
  }
  ```

  Attributes: (In addition to global attributes)

  - `role` specifies the role of this participant in the group.
  - `sex` specifies the sex of the participant. Legal values are:
    - `m` male
    - `f` female
    - `u` unknown or inapplicable
  - `age` specifies the age group to which the participant belongs.

  Example

  ```
  <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>
  ```

  May contain a prose description organized as paragraphs, or any sequence of demographic elements in any combination.
**Example of reference documentation in Japanese**

1. `<elementSpec> person`

   宣言

   element person
   {
     att.global.attributes,
     attribute role { text },
     attribute sex { "m" | "f" | "u" },
     attribute age { text },
   }

   属性: (グローバル属性の他)

   role
   sex
   age

   正当な値:
   "m" 男性
   "f" 女性
   "u" 不明または不適切

   例:

   `<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>`
Example of reference documentation in Bulgarian

1. `<elementSpec> person`

Декларация

```
<element person
    att.global.attributes,
    attribute role { data.code }?,
    attribute sex { 'm' | 'f' | 'u' }?,
    attribute age { data.name }?,
    (c+ | model.personPart?)>
```

Атрибути: (Освен глобалните атрибути)

```
role
sex
age
```

Разрешените стойности са:

- Мъж
- Жена
- Неизвестен или несъществуващ

Пример

```
<person sex="f" age="42">
  <p>Female informant, well-educated, born in Shropshire
      UK, 12 Jan 1960, of unknown occupation.
      Speaks French fluently. Socio-Economic status 82.</p>
</person>
```

Могат да се включват произволно организирани текстови описания, или произволна последователност от демографични елементи в произволно съчетание.
1. `<elementSpec> person`

Deklaration

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

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

<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 \texttt{<desc>} and \texttt{<gloss>} texts
- a mechanism to allow users to easily take advantage of the 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