

Towards an internationalized and localized TEI

Sebastian Rahtz, Oxford University

Abstract

The Text Encoding Initiative Guidelines have been widely adopted by projects and institutions in many countries in Europe, the Americas, and Asia, and are used for encoding texts in dozens of languages. However, the Guidelines are written in English, the examples are largely drawn from English literature, and even the names of the elements are abbreviated English words. 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.

The paper describes how the TEI project can develop internationally, including

- A review of why localisation and internationalisation matter
- A discussion of how the TEI architecture can be leveraged to support internationalised versions
- The application of the W3C ITS guidelines to the TEI work
- Practical results from a pilot project, and future translation plans
- The tools needed to make use of an internationalised TEI
- The steps towards ontologies in the TEI

1. TEI, internationalisation, and localisation

The Text Encoding Initiative Guidelines [TEI] 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. For example, the projects listed at <http://www.tei-c.org/Applications/> have examples of work involving Chinese, Danish, Dutch, Finnish, French, German, Greek, Hungarian, Italian, Japanese, Latin, Norwegian, Serbian, Spanish, Welsh, and some African languages; but given that the Guidelines are c. 1400 pages of fairly dense technical English, it is possible that only the more dedicated scholars get involved.

It may be useful to distinguish between what we might call ‘traditional’ or documentary approaches to translation, which focus on translating the descriptive prose of the Guidelines *as a document*, and ‘formal’ approaches which focus instead on translating the individual components (examples, element and attribute names, technical descriptions) in a way that enables these components to be used within the formal structures of the TEI as a technical standard. While the first approach may be very useful, the results are more difficult to maintain over the long term and are also more difficult to produce, since they cannot be accomplished in discrete chunks. The latter approach is the one we propose here, since it is more easily maintainable (only the affected elements need to be updated when changes are

made to the Guidelines) and can be more easily undertaken in a distributed fashion by collaborative groups.

Some translation work has already been undertaken:

- There have already been six ‘traditional’ translations of the TEI Lite (<http://www.tei-c.org/Lite/>) documentation into other languages. These have not covered translation of the element names or technical reference documentation. They are in wide use, however, and have created a need for more extensive translations of the Guidelines themselves.
- The French *Groupe d'experts n° 8* within CN 357 (*Commission de normalisation «Modélisation, production et accès aux documents»*) of the CG 46 (*Commission générale «Information et documentation»*) at AFNOR has an interest in TEI translation. Amongst other goals, this group intends to translate the definitions of the TEI elements and attributes into French. So far, they have worked in a ‘traditional way’ on some chapters of the P4 and P5 versions. Dissemination of the resulting French version of these chapters is very limited.
- Some ‘formal’ work has also been undertaken on translating element and attribute names; Alejandro Bia (for his background work see eg [BIA]) and Arno Mittelbach have prepared translation sets for Catalan, Spanish, and German. This work is integrated into the Roma (<http://www.tei-c.org.uk/Roma/>) application, allowing users to create tailored schemas in one of the supported languages.

Translation of documentation is only part of the issue. 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. The W3C define these processes as follows:

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]

Localization primarily concerns examples in the TEI context. There are over 1100 formal examples (that is to say, syntactically complete and valid) scattered through the text of TEI Guidelines, and another 775 in the formal definitions of elements; nearly all are in English. This is usually acceptable for examples like this:

```
Lexicography has shown little sign of being affected by the work of
followers of J.R. Firth, probably best summarized in his slogan,
<cit>
  <quote>You shall know a word by the company it keeps.</quote>
  <ref>(Firth, 1957)</ref>
</cit>
```

which is in the field of discourse of many scholars, but many others require considerably greater familiarity with Anglo-Saxon culture. Even Shakespeare:

```
<sp>
<speaker>First Servant</speaker>
  <ab>O, I am slain! My lord, you have one eye left</ab>
```

```

    <ab>To see some mischief on him. O!</ab>
  </sp>
<stage>Dies</stage>
<sp>
  <speaker>CORNWALL</speaker>
  <ab>Lest it see more, prevent it. Out, vile jelly!</ab>
  <ab>Where is thy lustre now?</ab>
</sp>
<sp>
<speaker>GLOUCESTER</speaker>
  <ab>All dark and comfortless. Where's my son Edmund?</ab>
  <ab>Edmund, enkindle all the sparks of nature,</ab>
  <ab>To quit this horrid act.</ab>
</sp>

```

is not easy, while older English is even harder:

```

<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>

```

It will be countered that the words of these examples do not matter much, since it all that is required is to appreciate the markup constructs being used (most people will recall that Shakespeare wrote plays, and this is all that matters). However, sometimes the point of the markup is not obvious, as in this example:

Next morning a boy in that dormitory confided to his bosom friend, a **<distinct type="psSlang">fag</distinct>** of Macrea's, that there was trouble in their midst which King **<distinct type="archaic">would fain</distinct>** keep secret.

Here there is the English word ‘psSlang’ (expandable to ‘public school slang’) for the type attribute of **<distinct>** to consider, where the value of ‘fag’ gives little help.

When the general context itself is clear, and the English text perhaps easy to translate, the names of the elements may stand in the way of easy comprehension. Thus:

```

<persName key="EGBR1">
  <roleName type="office">Governor</roleName>
  <forename sort="2">Edmund</forename>
  <forename full="init" sort="3">G.</forename>
  <addName type="nick">Jerry</addName>
  <addName type="epithet">Moonbeam</addName>
  <surname sort="1">Brown</surname>
  <genName full="abb">Jr</genName>.
</persName>

```

Can only really be take advantage of by someone who

1. appreciates the cultural context of ‘forename’ and ‘surname’

2. can mentally expand ‘nick’ to ‘nickname’ (and knows what a nickname is)
3. can appreciate whether a ‘Governor Edmund G. Jerry Moonbeam Brown Jr.’ is a politician, a kind of food, or a new dance

The user of the Guidelines may accordingly prefer to:

1. 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)’ 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.’ in the documentation
2. use element names of <líneaDirección>, <ligneAdresse>, <linDireccio> or <AdressZeile> instead of <addrLine>
3. see examples from daily life, as in:

```
<Adresse>
  <AdressZeile>Herrn Jürgen Jemandem</AdressZeile>
  <AdressZeile>Computer+Software GmbH </AdressZeile>
  <AdressZeile>Albrecht-Thär-Straße 22</AdressZeile>
  <AdressZeile>48147 Münster</AdressZeile>
  <AdressZeile>GERMANY</AdressZeile>
</Adresse>
```

(thanks to <http://www.columbia.edu/kermit/postal.html#germany> for the example).

We will consider later how these translated element names can be reconciled with the English names.

It should be noted that element name translation by itself is quick and useful, but necessarily the most effective way to proceed. For example, many of the element names are in an abbreviated form of English (eg <respStmt>) which are not easy to translate sensibly. Furthermore, unless the reference descriptions are also translated, the element names by themselves do not give a clear idea of what the element is for. Using <infoResp> instead of <respStmt> is not as helpful as translating the description ‘supplies a statement of responsibility for someone responsible for the intellectual content of a text, edition, recording, or series, where the specialized elements for authors, editors, etc. do not suffice or do not apply.’

2. TEI architecture support for the I18N and L10N process

2.1. Unicode

The first priority in internationalizing the TEI is to ensure clean support for character sets throughout the system. With this in mind, the P5 revision of the TEI made substantial changes in its dealings with characters. As the W3C (<http://www.w3.org/International/>) recommend, in the TEI scheme:

- 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">Φ λ . Θ á λ λ ο ς</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

A non-Unicode character can be defined using the <glyph> element in the TEI header. In the following example, we define a new character and assign it to a position in the Unicode Private Use Area (PUA); we also provide a standardized form as a fallback:

```
<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>
```

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

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

At this point we will expect the processing application to work out what to do (either show the PUA character, if it can, or the standardized form). Other facilities in the <charDesc> element allow the user to provide an image file which has a picture of the character. 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.

Where a character is simply a relatively unimportant variant on a Unicode character, the user does not need to define a point in PUA, but can simply use <charDesc> to describe the variation.

2.2. TEI literate programming

The TEI is written in a high-level markup language for specifying XML schemas and their documentation. This language is an XML vocabulary known as ODD (*One Document Does it all*), and is one of the TEI modules. This provides a literate programming language for production and documentation of any XML schema, with three important characteristics:

1. The element and attribute sets making up the schema are formally specified using a special XML vocabulary
2. The specification language also includes support for macros (like DTD entities, or schema patterns), a hierarchical class system for attributes and elements, and the creation of pre-defined groups of elements known as modules.
3. Content models for elements and attributes are written using an embedded RELAXNG XML notation, but tools are available to generate schemas in any of RELAXNG, DTD language, or W3C schema.

4. Documentation describing the supported elements, attributes, value lists etc is managed along with their specification, together with use cases, examples, and other supporting material.

The expectation is that many people wish to use only a subset of the TEI, so the TEI's 22 modules (containing 500 elements) can be combined together and customized as desired using the ODD language, to produce a schema suitable for use by a project. Customization may include tightening the constraints on existing elements, removing unused elements, and even adding new elements or attributes (though this will make the text not portable).

The ODD language has allowance for translating element name, attribute names, and descriptions, and for preserving information to allow canonicalisation. The technical documentation elements (<gloss> and <desc>) for TEI elements and attributes etc can be specified multiple times, in different languages, distinguished by the standard xml:lang attribute. There is also a container (<equiv>) to specify the relationship of an element, attribute or value to standardised schemes.

Each definition of a new primary object (element or attribute) has associated description and examples. A complete example of a definition is as follows:

```
<elementSpec module="header" ident="taxonomy">
  <gloss>taxonomy</gloss>
  <desc>defines a typology used to classify texts either
implicitly, by means of a bibliographic citation, or explicitly
by a structured taxonomy.</desc>
  <content>
    <rng:choice>
      <rng:oneOrMore>
        <rng:ref name="category"/>
      </rng:oneOrMore>
      <rng:group>
        <rng:group>
          <rng:ref name="model.biblLike"/>
        </rng:group>
        <rng:zeroOrMore>
          <rng:ref name="category"/>
        </rng:zeroOrMore>
      </rng:group>
    </rng:choice>
  </content>
  <exemplum>
    <egXML>
      <taxonomy xml:id="tax.b">
        <bibl>Brown Corpus</bibl>
        <category xml:id="tax.b.a">
          <catDesc>Press Reportage</catDesc>
          <category xml:id="tax.b.a1">
            <catDesc>Daily</catDesc>
          </category>
          <category xml:id="tax.b.a2">
            <catDesc>Sunday</catDesc>
          </category>
        </category>
      </taxonomy>
    </egXML>
  </exemplum>
</elementSpec>
```

```

</category>
<category xml:id="tax.b.a3">
  <catDesc>National</catDesc>
</category>
<category xml:id="tax.b.a4">
  <catDesc>Provincial</catDesc>
</category>
<category xml:id="tax.b.a5">
  <catDesc>Political</catDesc>
</category>
<category xml:id="tax.b.a6">
  <catDesc>Sports</catDesc>
</category>
</category>
<category xml:id="tax.b.d">
  <catDesc>Religion</catDesc>
  <category xml:id="tax.b.d1">
    <catDesc>Books</catDesc>
  </category>
  <category xml:id="tax.b.d2">
    <catDesc>Periodicals and tracts</catDesc>
  </category>
</category>
</taxonomy>
</egXML>
</exemplum>
</elementSpec>

```

The important things to note here are that the content model for the element is expressed in RELAXNG, which references other elements only by the names of classes to which they belong; and that the worked example is well-formed XML embedded in its own namespace. This specification may be processed to produce a DTD, a RELAXNG schema, an XSD schema, or documentation in various forms.

The objects identified by the `ident` attribute in the TEI can be given an alternate name by use of the `<altIdent>` element; so the example above could be rewritten as

```

<elementSpec module="header" ident="taxonomy">
  <altIdent xml:lang="fr">taxinomie</altIdent>
  ....
</elementSpec>

```

Providing a French name for the element. How does this work in the schema, where other elements might refer to ‘taxonomy’? The normal schema, using RELAXNG compact syntax, has the definition

```

taxonomy =
  ## (taxonomy) defines a typology used to classify texts either
  ## implicitly, by means of a bibliographic citation,
  ## or explicitly by a structured taxonomy.
  element taxonomy { taxonomy.content, taxonomy.attributes }
taxonomy.content = category+ | (model.biblLike, category*)
taxonomy.attributes = att.global.attributes, empty

```

in which the *element* <taxonomy> is defined by the containing pattern ‘taxonomy’; it is the *pattern name* which other elements use, not the element name. If the schema were translated into Greek, it would look like this:

```
taxonomy =
  element τ α ξ ι ν ο μ ι α { taxonomy.content,
taxonomy.attributes }
...
```

where the *pattern name* remains the same. This type of schema markup is generated by the TEI tools, picking up the information from <altIdent>. The descriptions work in the same way. We can expand the TEI source to add French translations alongside the English originals, and the appropriate text can be passed to the generated schemas or documentation:

```
<elementSpec module="header" ident="taxonomy">
  <altIdent xml:lang="fr">taxinomie</altIdent>
  <gloss>taxonomy</gloss>
  <gloss xml:lang="fr">Taxinomie</gloss>
  <desc>defines a typology used to classify texts either
implicitly, by means of a bibliographic citation, or explicitly
by a structured taxonomy.</desc>
  <desc xml:lang="fr">L'élément Taxinomie <gi>taxonomy</gi>
définit une typologie employée pour classer des textes soit
implicitement au moyen d'une citation bibliographique, soit
explicitement au moyen d'une taxinomie structurée.</desc>
....
</elementSpec>
```

[We thank Pierre Yves Duchemin for these translations.]

What does a translated schema look like in practice? If we take a Spanish play, and translate the element names to Spanish (thank to Alejandro Bia for this work), a text like this will be much more familiar-looking to encoders in Spanish-speaking countries:

```
<cuervo>
  <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">
soldados</resaltado>.<resaltado formato="bold">)</resaltado>
        </acotacion>
        <dialogo>
          </dialogo>
        </div3>
      </div2>
    </div1>
  </cuervo>
```

This file will not work with normal TEI publishing tools, or be suitable for archiving, but it is straightforward to write a transformation (eg in XSL) which reads the TEI source with the element names and `<altIdent>` information, and puts the text back to canonical form.

2.3. TEI applications

TEI applications, as well as the texts themselves, need to have developed internationalised interfaces. For example, an application which turns TEI XML into HTML for web display, and provides a heading such as ‘Contents’ when it meets `<divGen type="toc"/>`, will have to provide appropriate translations. The TEI XSL family maintained by Sebastian Rahtz, for example (<http://www.tei-c.org/Stylesheets/teic/>), can operate in many languages:

ISO Language code Text

en	Contents
de	Inhalt
ro	Cuprins
fr	Contenu
pt	Índice geral
es	Contenidos
slv	Vsebina
sv	Innehåll
sr	Sadržaj
pl	Spis treści
hi	Mula Shabda
nl	Inhoud
tr	İçerik
el	Π ε ρ ι ε χ ό μ ε ν α

[Bulgarian, Chinese, Japanese, Russian, and Thai also available; they are omitted here to avoid printing problems.]

2.4. TEI schema-making tools

The ODD language files need to be processed to produce schemas in the chosen language. This is done by a set of XSLT scripts, which can either be run on a command-line, or as a web service called Roma (<http://www.tei-c.org.uk/Roma/>). This currently has support for varying the languages of its interface, but must also allow for supporting the following output schemes:

- canonical: English names, descriptions in English
- local descriptions: English names, descriptions in chosen language
- local names: names designed to make sense to a speaker of the chosen language, descriptions in English

- fully localized: both names and descriptions in chosen language

This work is in progress; while the underlying XSLT supports the generation of documentation in different languages, the web interface has still to be implemented.

2.5. *The application of the W3C ITS guidelines to TEI work*

An *Internationalisation Tag Set* working group (under the chairmanship of Yves Savourel, Enlaso) is writing a Recommendation (if it is accepted) for the World Wide Web Consortium about markup which encodes information for translators and localisers. The current state can be found at <http://www.w3.org/International/its> (this document is itself written using the TEI ODD language). The ITS consists of a set of elements and attributes for annotating a text with information for further processing, covering **Internationalization**:

- Markup for bidirectional text
- Ruby annotation
- Language identification

and **Localization**

- Translatability of content
- The localization process in general
- Terminology markup

It is intended that the ITS annotation elements be added at several stages. The simplest is at the content authoring stage, by technical writers, developers of authoring systems, localizers or translators. In addition, specialist terminologists might annotation a text with terminological information, or localization engineers and translators may add information.

The primary ITS notion is that information about elements and attributes can be supplied

- in a document schema
- in an external rules file
- in a rule section in an instance file
- attached to instance elements

where the information consists of a set of *data categories*. On an instance element, for example, the following attributes may be attached

translate

should this object be translated?

locInfo

Is there some localisation hint?

locInfoType

What type of hint is it?

term

Does this object describe a technical term?

termRef

Where is the term defined?

dir

What is the text direction?

rubyText

Is there some Ruby annotation?

A complete example of a TEI text marked up with a combination of ITS rules and ITS local markup looks like this:

```
<TEI xmlns:tei="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <its:rules>
      <its:translateRule translate="no" selector="//tei:body/tei:p"/>
    </its:rules>
  </teiHeader>
  <text>
    <body>
      <p>Hello <hi>world</hi>
    </p>
    <p
      its:translate="yes">translate me</p>
    </body>
  </text>
</TEI>
```

where the ITS rules say that <p> elements should not normally be translated, but the second <p> has an explicit override.

If we take a TEI ODD document, we can express the relationship between the structural elements and the documentation elements with the following ITS rules, which says that the default is to **not** translate anything, but gives a set of elements which **are** to be translated:

```
<its:rules xmlns:tei="http://www.tei-c.org/ns/1.0">
  <its:translateRule translate="no" selector="//tei:*/>
  <its:translateRule translate="yes" selector="//tei:desc"/>
  <its:translateRule translate="yes" selector="//tei:gloss"/>
  <its:translateRule translate="yes" selector="//tei:valDesc"/>
  <its:translateRule translate="yes" selector="//tei:p[@rend='dataDesc']"/>
  <its:translateRule translate="yes" selector="//tei:remarks"/>
</its:rules>
```

Using this information, we can show graphically in [Figure 1, Example of ITS implementation](#), using an ITS tool, which elements need a translated equivalent (those in green).

```

<elementSpec module="corpus" id="PERSON" usage="opt" ident="person">
  <equiv/>
  <gloss/>
  <desc>describes a single participant in a language interaction. </desc>
  <content>
    <rng:choice>
      <rng:oneOrMore>
        <rng:ref name="model.pLike"/>
      </rng:oneOrMore>
      <rng:zeroOrMore>
        <rng:ref name="model.personPart"/>
      </rng:zeroOrMore>
    </rng:choice>
  </content>
  <attList>
    <attDef ident="role" usage="opt">
      <equiv/>
      <desc>specifies the role of this participant in the group.</desc>
      <datatype>
        <rng:ref name="data.code"/>
      </datatype>
      <valDesc>a set of keywords to be defined</valDesc>
    </attDef>
    <attDef ident="sex" usage="opt">
      <equiv/>
      <desc>specifies the sex of the participant.</desc>
      <datatype>

```

Done

Figure 1. Example of ITS implementation

For the purposes of the formal translation procedure advocated by this paper, the ITS procedure provides a good framework.

3. Results so far

We present here some examples showing work completed so far:

```

<elementSpec module="corpus" xml:id="PERSON" usage="opt"
  ident="person">
  <equiv/>
  <gloss/>
  <desc>言語活動の関係者(1件1名)</desc>
  <content>
    <rng:choice>
      <rng:oneOrMore>
        <rng:ref name="p"/>
      </rng:oneOrMore>
      <rng:zeroOrMore>
        <rng:ref name="tei.demographic"/>
      </rng:zeroOrMore>
    </rng:choice>
  </content>
  <attList>
    <attDef ident="role" usage="opt">
      <equiv/>
      <desc>当該関係者の言語活動における役割</desc>

```

```

    <datatype>
      <rng:ref name="datatype.Code"/>
    </datatype>
    <valDesc>定義済みキーワード</valDesc>
  </attDef>
  <attDef ident="sex" usage="opt">
    <equiv/>
    <desc>関係者の性別</desc>
    <datatype>
      <rng:ref name="datatype.Sex"/>
    </datatype>
    <valList type="closed">
      <valItem ident="m">
        <equiv/>
        <gloss>男性</gloss>
      </valItem>
      <valItem ident="f">
        <equiv/>
        <gloss>女性</gloss>
      </valItem>
      <valItem ident="u">
        <equiv/>
        <gloss>不明または不適切</gloss>
      </valItem>
    </valList>
  </attDef>
  <attDef ident="age" usage="opt">
    <equiv/>
    <desc>当該関係者の年齢層</desc>
    <datatype>
      <rng:ref name="datatype.Code"/>
    </datatype>
    <valDesc>推定年齢</valDesc>
  </attDef>
</attList>
<exemplum xml:lang="en">
  <egXML>
    <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>
  </egXML>
</exemplum>
<exemplum xml:lang="ja">
  <egXML>
    <person sex="f" age="42">

```

<p>女性、教養あり、1950 年 1 月 12 日英国シュロプシア生まれ、 不明、フランス語を流暢に話す、社会経済状態：中	職業
---	----

```

</p>
</person>
</egXML>
</exemplum>
<remarks>
  <p rend="dataDesc">
    段落単位の記述、または人口統計学のデータが混在して含まれる
  </p>
</remarks>
<listRef>
  <ptr target="#CCAHPA"/>
</listRef>
</elementSpec>

```

Figure 2. Example of translated ODD

1. <elementSpec> person

describes a single participant in a language interaction.

Declaration

```

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

Figure 3. Example of reference documentation

1. <elementSpec> person

宣言

```

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

```

属性: (グローバル)属性の他

role

sex

正当な値:

m 男性

f 女性

u 不明または不適切

age

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>
</person>
<person sex="f" age="42" lang="ja">
  <p>女性、教育あり、1950年1月12日英国シェロプシア生まれ、
    職業不明、フランス語を流暢に話す、社会経済状態：中
  </p>
</person>

```

段落単位の記述、または人口統計学のデータが混在して含まれる

Figure 4. Example of reference documentation in Japanese

1. <elementSpec> person

Декларация

```

element person
{
  att.global.attributes,
  attribute role { data.code }?,
  attribute sex { "m" | "f" | "u" }?,
  attribute age { data.name }?,
  ( p+ | model.personPart* )
}

```

Атрибути: (Освен глобалните атрибути)

role

sex

Разрешените стойности са:

m

мъж

f

жена

u

неизвестен или несъществуващ

age

Пример

```

<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 U2.</p>
</person>

```

May contain a prose description organized as paragraphs, or any sequence of demographic elements in any combination.

Модул: corpus

Figure 5. Example of reference documentation in Bulgarian



Figure 6. Interface translation in Bulgarian

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>
<person sex="f" age="42" lang="ja">
  <p>女性、教養あり、1950年1月12日英国シェロプシア生まれ、
    職業不明、フランス語を流暢に話す、社会経済状態：中
  </p>
```

Figure 7. Reference documentation in Japanese, with German annotation

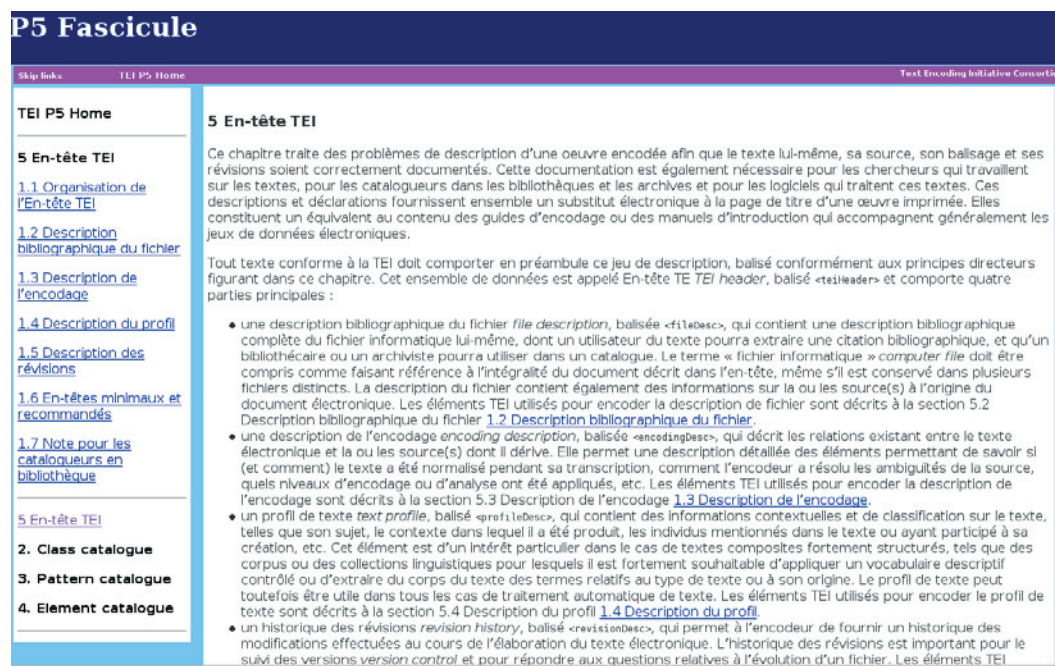


Figure 8. TEI Guidelines in French

4. Future directions

The TEI Consortium is working with TEI scholars to advance I18N and L10N in various languages (listed in). 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, and a mechanism to allow users to easily take advantage of the work. The scale of work involved is not impossible to contemplate. The TEI contains

- 494 elements
- 489 attributes
- 1203 <desc> elements, 106666 characters
- 1177 <gloss> elements, 32385 characters

The work needed for each language is to

- 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 infrastructure challenges are not inconsiderable. We need, at least:

- An infrastructure to allow translators to submit material, and get prompt feedback
- Integrating the translations into the P5 source
- Ensuring that translations are flagged as decayed when the English original changes, and that translators are notified
- Managing multi-language examples

By the end of 2006, we expect to be well on the way to meeting these goals.

Appendix A Acknowledgements

The first steps in formalized internationalization of the TEI (as opposed to the translations of the Lite document) were made by Alejandro Bia, to whom many thanks are due. Translation examples in this paper come from Pierre Yves Duchemin (French), Marcus Bingenheimer (Chinese), Arno Mittelbach (German) and Alejandro Bia (Spanish). Veronika Lux and Julia Flanders co-wrote some of the explanations of TEI I18N.

Appendix B References

1. Manuel Sánchez, Alejandro Bia, Régis Déau, *Multilingual Markup of Digital Library Texts Using XML, TEI and XSLT*. Presented in XML Europe 2003
2. *The CIDOC Conceptual Reference Model*. Draft International Standard ISO/DIS 21127.
3. Christian Lieske and Felix Sasaki (eds). *Internationalization Tag Set (ITS) Version 1.0*. <http://www.w3.org/International/its/itstagset/>. World Wide Web Consortium, 2006.
4. Lou Burnard and Syd Bauman (eds). *Text Encoding Initiative Guidelines development version (P5)*. TEI Consortium, Charlottesville, Virginia, USA, Text Encoding Initiative.

Appendix C TEI internationalisation partners

The following peoples and bodies have agreed to coordinate their respective languages:

Chinese	Marcus Bingenheimer	Chung-hwa Institute of Buddhist Studies, Taipei
Dutch	Bert Van Elsacker	-
French	Laurent Romary	Nancy
French	Veronika Lux	Nancy
German	Christian Wittern	Institute for Research in Humanities, Kyoto University
German	Werner Wegstein	Wuerzburg University
Hindi	Paul Richards	UGS (The PLM Company), http://www.ugs.com/
Hungarian	Király Péter	-
Italian	Fabio Ciotti	University of Roma
Japanese	OHYA Kazushi	Tsurumi University, Yokohama

Norwegian	Øyvind Eide	-
Polish	Radoslaw Moszczynski	Warsaw University
Portuguese	Leonor Barroca	Open University
Romanian	Dan Matei	CIMEC - Institutul de Memorie Culturala, România
Serbian	dr Cvetana Krstev	-
Slovenian	Tomaž Erjavec, Matija Ogrin	Dept. of Knowledge Technologies, Jozef Stefan Institute, Slovenia
Spanish	Manuel Sánchez	Miguel de Cervantes Digital Library
Swedish	Matt Zimmerman	NYU
Tibetan	Linda Patrik, Tensin Namdak	www.nitartha.org