SCIENTIFIC PAPER

UDC 004.439:004.55XML
UDC 519.76

TEXT ENCODING INITIATIVE GUIDELINES AND THEIR LOCALISATION

Tomaž Erjavec*
Jožef Stefan Institute, Department of Knowledge Technologies
Ljubljana, Slovenia

Abstract: The paper introduces the Text Encoding Initiative Guidelines, a formal specification and accompanying documentation defining a vocabulary of XML elements, meant for the annotation of texts for scholarly purposes. TEI is widely used for the encoding of e.g. complex texts in digital libraries, for a wide variety of text types. The paper describes the history and organisation of the Text Encoding Initiative, the structure of the Guidelines and gives several use cases. It then moves on to the question of internationalisation: while TEI elements can describe text in any language, the Guidelines themselves are nevertheless in English. However, TEI offers several possibilities of how to translate (parts of) the Guidelines. We discuss them, and offer a new, resource light alternative.

Keywords: XML, TEI, Text encoding standardisation, internationalisation.

*tomaz.erjavec@ijs.si

1. Introduction

Digitally stored text has been, for a long time, viewed only as a digital representation of its printed image, and was closely bound to the software that created it. Such text was thus not suitable for task-based machine processing and could become quickly obsolescent. To address these concerns the ISO standard SGML was created, which allowed platform and output independent means of representing text, by allowing users to define their own tagsets, suitable for encoding arbitrary text types. The most widely known application of SGML is without doubt HTML, the Hypertext markup language, used to encode documents for display in Web browsers. But neither SGML nor HTML were suited as a broad basis for encoding arbitrary texts for arbitrary purposes: SGML was too complex, while HTML defined only a small vocabulary of elements, mostly visually oriented.

These were the reasons why a new standard was created by the World Wide Web Consortium, the Extensible Markup Language, XML.[1] XML is a subset of SGML, and as such, unlike HTML, specifies a means to define arbitrary tagsets, while being much simpler than SGML, so that it became much easier to write software to process it. Since its inception, XML has become a true success story as the interchange format for digital data, with a host of software and related standards to support its authoring, validation, and transformations.

Of these related standards we should mention XSLT,[2] the transformation language for XML documents. It is possible to write scripts in XSLT that transform XML documents into other, differently structured documents. As XML is primarily used as a storage and interchange format, XSLT enables its transformation into more directly usable formats, e.g., HTML for Web browsing.

XML, as was SGML, is a meta-language: it does not define a particular tagset, but only provides a means to define tagsets, and how a tagset is defined depends on the type of data under consideration and the envisaged usage scenarios. While developers are free to define their own tagsets, or, more precisely, XML schemas that specify the tags (XML elements and attributes) and the valid interrelations between tags, it is in general better to make use of pre-defined standard schemas, as they have been carefully designed, are well documented and maintained, ensure interoperability and can be accompanied by dedicated software.

By now there exist a large number of standard XML schemas, for areas as diverse as mobile phone applications, representing mathematical formulas, music notation and technical documentation.

But what about the field of humanities texts, i.e. texts that are meant to be the subject of humanities research, regardless of what text type they are, what language they are written in, what period they come from, or what methods are used in their analysis? To date, there is only one standard (or, better, recommendation) that attempts to cover this vast area, namely the Text Encoding Initiative.

2. The Text Encoding Initiative

The Text Encoding Initiative (TEI) developed a community standard for representing digital texts in a way that is both powerful and responsive to the research needs of humanities scholars. TEI can be characterised as:[3]

– a freely available set of guidelines for encoding humanities texts using XML,

– an international consortium that exists to support the maintenance and development of the guidelines, and

[1] http://www.w3.org/XML/

[2] http://en.wikipedia.org/wiki/XSL_Transformations

[3] This section closely follows the excellent introduction to the TEI given in http://www.wwp.brown.edu/encoding/seminars/tei.html

– a community of projects and individuals who use the TEI Guidelines.

Like many standardization efforts, the TEI faces the challenges inherent in such a project. How can the many disciplines and communities within the humanities domain find common ground in a single encoding language? How do we agree on the level of detail that is necessary or appropriate in describing our textual materials? How do we reconcile the advantages to be gained by consistency and agreement with the need for individual specialization? How do we handle the truly idiosyncratic and unexpected?

Unlike many standardization efforts, the TEI addresses these and similar questions by explicitly accommodating variation and debate within its technical framework. The TEI Guidelines are designed to be both modular and customizable, so that specific projects can choose the relevant portions of the TEI and ignore the rest, and can also if necessary create extensions of the TEI language to describe facets of the text which the TEI does not yet address. Because the TEI itself is complex, the customization process is not entirely trivial, but it is designed to be as straightforward as possible.

2.1. What is TEI used for?

The predominant use of the TEI Guidelines is in creating large digital library collections, which provide access to large quantities of textual material, often focusing on rare or fragile materials which would otherwise be inaccessible. The text encoding in collections like these emphasizes features that will be of immediate, general use for searching and retrieval: information such as bibliographic data, basic text structure, and metadata such as subject keywords to help users find materials of interest.

The TEI Guidelines are also used by more specialized research projects to represent smaller, more thematically focused collections of texts, often organized around a single genre, author, period, or country (or a combination of these. Projects of this type often use more detailed markup to represent particular features of the text that are relevant to the specific collection or important for the specific scholarly audience being served. For instance, a collection focusing on an author whose writings are an important source of information about famous contemporaries might usefully encode all references to names and works, perhaps including links to more detailed indexes of biographical or critical information. Similarly, an electronic edition of a particular author or work might well include a representation of variant readings, authorial revisions, editorial emendations, and similar editorial information. Some collections of this sort are intended to serve very specific research goals, such as linguistic analysis, or as the basis for a dictionary, and in these cases the markup may be very highly specialized.

The uses described above are all some species of publication: the goal is to create a digital collection that can be used online by some public audience of greater or lesser extent—perhaps limited to a small community or to paying subscribers, perhaps open to the general public. More rarely, the TEI is also used by individuals to create digital representations of textual materials to support their own personal research, in forms which may or may not be published. Where the scope and purpose of the larger collections may be determined by audience and funding, in the case of an individual's work the constraints are more personal and professional: the encoded material might serve as a private research tool, or might develop into the equivalent of a digital monograph that represents the author's analysis of a set of texts. The use of the TEI Guidelines in these cases may be as detailed as the author finds useful–limited only by time, energy, imagination, and the constraints of usefulness.

2.2. Learning TEI

The TEI web site[4] is a good source of general information about the TEI, and is also the place to find the TEI Guidelines. However, the TEI web site is only one of many sources of useful information about how to use the TEI and how to understand its significance. Although there is no single source that can give a complete picture, there is now a growing literature on the role of text encoding in scholarly work. The bibliography[5] of the Brown University gives some useful sources, but probably the best way to get an initial grounding in the TEI is to attend a workshop, a number of which are given each year, mostly in the U.S.A. and U.K., and are advertised on the TEI web site.

For those who need a more detailed understanding of the actual encoding process, workshops are a good start, but they are not sufficient. Learning the TEI is like learning a language – it has a fairly extensive vocabulary and a complex range of usage. An introductory workshop gives a good understanding of what the language is for and of its basic terms, but it needs to be followed with both practice and a detailed exploration of how the language is really used. Having a project to work on – a set of documents that are of interest – is a great help. To gain a detailed knowledge of the TEI Guidelines as you encode, there are several things one can do in addition to reading the Guidelines themselves. The TEI maintains the TEI-L mailing list, on which questions, even if asked by beginners, will almost invariably receive helpful answers. The list is also archived, and the answers to a number of common (and less common) questions can be found in the TEI-L archives; the range of opinions and approaches can also give a valuable sense of how different kinds of projects use the TEI. A good way of learning TEI is also to look at the work of actual text encoding projects. Many projects have documentation and some have exceptionally good documentation that explains the rationale behind their encoding decisions and the criteria by which they recognize and encode specific textual features. Many are also happy to share sample encoded texts, which can be a very useful way of getting a more detailed view of the encoding landscape.

3. History of TEI[6]

The TEI began in 1987 at a meeting at Vassar College, which brought together a diverse group of scholars from many different disciplines and representing leading professional societies, libraries, archives, and projects in a number of countries in Europe, North America, and Asia. The initial phase of their work resulted in the release of the first draft (known as ‘P1’) of the Guidelines in 1990. A second phase immediately began and released its results throughout 1990–1993. Then, after another round of revisions, extensions, and supplements, the first official version of the Guidelines (‘P3’) was released in 1994. As more scholars became acquainted with the Guidelines, comments, corrections, and requests for extensions arrived from around the world. In the end there were nearly 200 scholars from many disciplines, professions, and countries in the core group that was developing the TEI Guidelines.

In 2000 the TEI Consortium was established. It is an international membership organization, which now maintains, continues developing, and promotes the TEI. The goal of establishing the TEI Consortium was to maintain a permanent home for the TEI as a democratically constituted, academically and economically independent, self-sustaining, non-profit organization.

[4] http://www.tei-c.org/

[5] http://www.wwp.brown.edu/encoding/seminars/readings.html

[6] For a more detailed account of the TEI history see http://www.tei-c.org/About/history.xml on which this section is also based.

Following the establishment of the TEI Consortium, a critical priority was the release of an XML version of the TEI Guidelines, updating P3, itself based on SGML, to enable users to work with the emerging XML toolset. The P4 version of the Guidelines was published in 2002. It was essentially an XML version of P3, making no substantive changes to the constraints expressed in the schemas apart from those necessitated by the shift from SGML to XML, and changing only corrigible errors identified in the prose of the P3 Guidelines. However, given that P3 had by this time been in steady use since 1994, it was clear that a substantial revision of its content was necessary, and work began immediately on the P5 version of the Guidelines. This was planned as a thorough overhaul, involving a public call for features and new development in a set of crucial areas including character encoding, graphics, manuscript description, and the language in which the TEI Guidelines themselves are written. The P5 version of the Guidelines was released at the end of 2007.

The impact of the TEI on digital scholarship has been enormous. Today, the TEI is internationally recognized as a critically important tool, both for the long-term preservation of electronic data, and as a means of supporting effective usage of such data in many subject areas. It is the encoding scheme of choice for the production of critical and scholarly editions of literary texts, for scholarly reference works and large linguistic corpora, and for the management and production of detailed metadata associated with electronic text and cultural heritage collections of many types.

The TEI's recommendations have been endorsed by many organizations, including the US National Endowment for the Humanities, the UK's Arts and Humanities Research Board, the Modern Language Association, the European Union's Expert Advisory Group for Language Engineering Standards, and many other agencies around the world that fund or promote digital library and electronic text projects. Recognizing its importance in the emerging digital library community, the Library of Congress has produced guidelines for best practice in applying the TEI metadata recommendations for interoperability with other standards.

The success of the TEI has also gone a long way to ensuring that our cultural heritage will be brought forward into the emerging new networked world, and made broadly available to the students, scholars, and the wider public.

4. The TEI Guidelines

The TEI *Guidelines for Electronic Text Encoding and Interchange* define and document a markup language for representing the structural, renditional, and conceptual features of texts. They focus (though not exclusively) on the encoding of documents in the humanities and social sciences, and in particular on the representation of primary source materials for research and analysis. These guidelines are expressed as a modular, extensible XML schema, accompanied by detailed documentation, and are published under an open-source license.

The TEI Guidelines are available in several formats: the printed and bound copies can be purchased, while the online HTML and PDF versions are available for free from the TEI home page. It is also possible to download the schemas, the XML source files of the Guidelines, documentation etc. as zip packages from the TEI SourceForge site; more detailed instructions for accessing and using these materials are available from the TEI homepage.

The TEI Guidelines are themselves written in XML and follow Donald Knuth's literate programming paradigm, where the same document contains the formal specifications of XML schemas and the accompanying documentation (prose text).

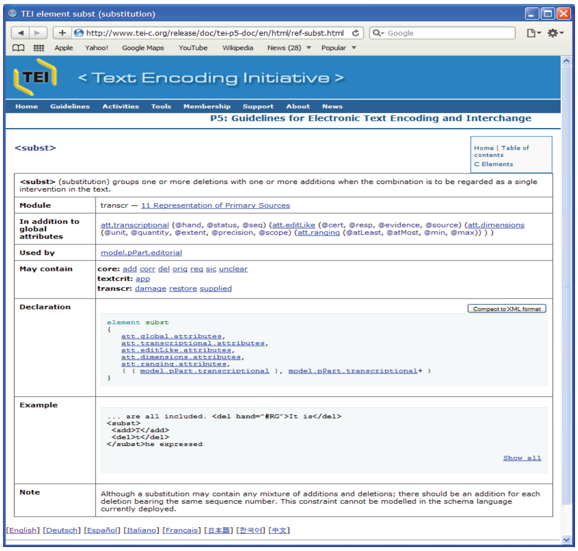
The TEI Guidelines define several hundred elements and attributes for marking up documents of any kind. Each definition has the following components:

– a prose description,

– a formal declaration, expressed using a special-purpose XML vocabulary defined in the Guidelines in combination with elements taken from the ISO schema language RELAX NG[7], and

– usage examples.

For illustration, Figure 1 gives the HTML view of the definition of the TEI <subst> element. It specifies that the element belongs to the module for Transcription of Primary Sources, and gives a hyperlink to the full text of the chapter explaining this module (c.f. next section). The prose description states that the element “groups one or more deletions with one or more additions when the combination is to be regarded as a single intervention in the text.” The definition further documents which attributes the element uses, which TEI model the element is used by, which elements it can contain, the formal schema definition, an example of usage, and notes.



P5: Guidelines for Electronic Text Encoding and Interchange

<subst>

<subst> (substitution) groups one or more deletions with one or more additions when the combination is to be regarded as a single intervention in the text.	
Module	transcr — 11 Representation of Primary Sources
In addition to global attributes	att.transcriptional (@hand, @status, @seq) (att.editLike (@cert, @resp, @evidence, @source) (att.dimensions (@unit, @quantity, @extent, @precision, @scope) (att.ranging (@atLeast, @atMost, @min, @max))))
Used by	model.pPart.editorial
May contain	**core:** add corr del orig reg sic unclear **textcrit:** app **transcr:** damage restore supplied
Declaration	`element subst { att.global.attributes, att.transcriptional.attributes, att.editLike.attributes, att.dimensions.attributes, att.ranging.attributes, ( ( model.pPart.transcriptional ), model.pPart.transcriptional+ ) }`
Example	`... are all included. <del hand="#RG">It is</del> <subst> <add>T</add> <del>t</del> </subst>he expressed`
Note	Although a substitution may contain any mixture of additions and deletions; there should be an addition for each deletion bearing the same sequence number. This constraint cannot be modelled in the schema language currently deployed.

Declaration:

```
element subst
{
   att.global.attributes,
   att.transcriptional.attributes,
   att.editLike.attributes,
   att.dimensions.attributes,
   att.ranging.attributes,
   ( ( model.pPart.transcriptional ), model.pPart.transcriptional+ )
}
```

Example:

```
... are all included. <del hand="#RG">It is</del>
<subst>
 <add>T</add>
 <del>t</del>
</subst>he expressed
```

[English] [Deutsch] [Español] [Italiano] [Français] [日本語] [한국어] [中文]

Figure 1. Definition of the TEI element <subst>

4.1. TEI Modules

Each chapter of the Guidelines presents a group of related elements, and also defines a corresponding set of declarations, called a module. All the definitions are collected together in the reference sections provided as an appendix. Formal declarations for a given chapter are collected together within the corresponding module. For convenience, each element is assigned to a single module, typically for use in some specific application area, or to support a particular kind of usage. A module is thus simply a convenient way of grouping together a number of associated element declarations. In the simple case, a TEI schema is made by combining together a small number of modules.

Table 1 lists all the modules defined by the TEI Guidelines P5.

Name of the module	*Formal public identifier*	*Chapter of the Guidelines where the module is defined*
analysis	Analysis and Interpretation	17 Simple Analytic Mechanisms
certainty	Certainty and Uncertainty	21 Certainty, Precision, and Responsibility
core	Common Core	3 Elements Available in All TEI Documents
corpus	Metadata for Language Corpora	15 Language Corpora
dictionaries	Print Dictionaries	9 Dictionaries
drama	Performance Texts	7 Performance Texts
figures	Tables, Formulae, Figures	14 Tables, Formulæ, and Graphics
gaiji	Character and Glyph Documentation	5 Representation of Non-standard Characters and Glyphs
header	Common Metadata	2 The TEI Header
iso-fs	Feature Structures	18 Feature Structures
linking	Linking, Segmentation, and Alignment	16 Linking, Segmentation, and Alignment
msdescription	Manuscript Description	10 Manuscript Description
namesdates	Names, Dates, People, and Places	13 Names, Dates, People, and Places

[7] http://www.relaxng.org/

nets	Graphs, Networks, and Trees	19 Graphs, Networks, and Trees
spoken	Transcribed Speech	8 Transcriptions of Speech
tagdocs	Documentation Elements	22 Documentation Elements
tei	TEI Infrastructure	1 The TEI Infrastructure
textcrit	Text Criticism	12 Critical Apparatus
textstructure	Default Text Structure	4 Default Text Structure
transcr	Transcription of Primary Sources	11 Representation of Primary Sources
verse	Verse	6 Verse

Table 1. TEI Modules

4.2. Constructing a TEI XML schema

To determine that an XML document is valid (as opposed to merely well-formed), its structure must be checked against a schema. For a valid TEI document, this schema must be a conformant TEI schema.

The specification for a conformant TEI schema is itself a TEI document, using elements from the module for Documentation Elements. Such a document is informally referred to as an ODD document, from the design goal originally formulated for the system: ‘One Document Does it all’. Stylesheets for maintaining and processing ODD documents are maintained by the TEI, and the Guidelines are also written as such a document.

A TEI-conformant schema gives a specific combination of TEI modules, possibly also including additional declarations that modify the element and attribute declarations contained by each module, for example to suppress or rename some elements. The same system may also be used to specify a schema which extends the TEI by adding new elements explicitly, or by reference to other XML vocabularies.

An ODD document can be processed by an ODD processor, which will generate an appropriate XML schema from this set of declarations, expressed using any of the standard schema languages:

− the XML DTD language (part of XML itself),
− the ISO RELAX NG language,
− the W3C Schema language,[8]
− or in principle any other adequately powerful schema language.

These output schemas can then be used by any XML processor such as a validator or editor to validate or otherwise process documents.

The TEI site provides an ODD processor called Roma, which gives a Web interface that helps in the process of creating a customized TEI schema.

5. Three Examples of TEI texts

While the previous sections have introduced TEI in broad strokes, this section gives some concrete examples from our own work on Slovenian texts.

All the texts discussed below are written as TEI compliant XML documents: some older ones in TEI P4, while those recently completed in TEI P5. The delivery mechanism is different for different cases, depending on the particular usage scenario, but in all cases rests on XSLT stylesheets, which take the source TEI and transform it into a use-oriented format, usually HTML suitable for viewing in a Web browser.

For the three examples discussed below, we first briefly introduce the texts, or, rather, projects in the scope of which they were compiled, then illustrate the encoding on a small example, and finally present how the XML source of the texts was converted for actual use of the resource.

5.3. The eZISS library

The eZISS digital library[9], being developed in cooperation between the Scientific Research Centre of the Slovenian Academy of Sciences and Arts and the Jožef Stefan Institute, offers

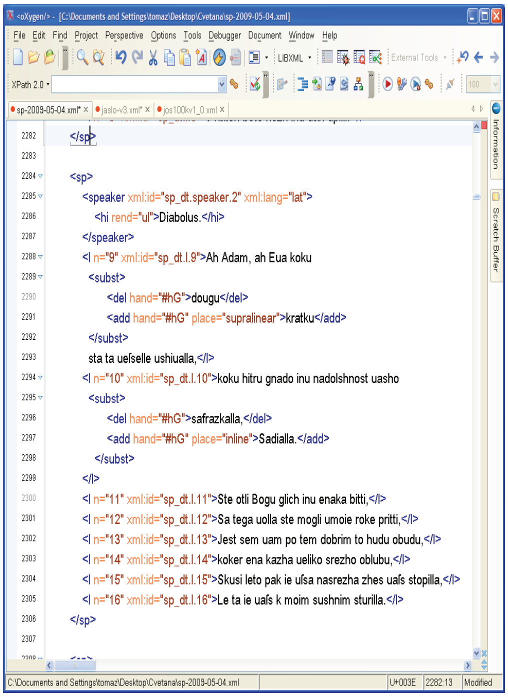
[8] http://www.w3.org/XML/Schema
[9] http://nl.ijs.si/e-zrc/

selected, typically older Slovenian texts with integrated facsimiles, transcriptions and scholarly commentary, in some cases including audiovisual recordings. The editions are intended for lasting public use, and are freely accessible on the world-wide Web for browsing, as well as downloading in source TEI or derived HTML. Abiding by the Creative Commons[10] licensing provisions, the editions can be also distributed to others.

The eZISS editions use TEI schemas which include modules for Manuscript Descriptions, Transcription of Primary Sources and Textual Criticism, as well as other modules, e.g. for Performance Texts.

The markup in the various editions varies considerably, as the texts are very different in structure and thus need different TEI elements for their description. To briefly illustrate the kind of markup used in our editions, Figure 2 contains a short excerpt from the eZISS edition of the Љkofja Loka Passion Play, the oldest performance text in Slovenian, written at the beginning of the 18th century. The excerpt shows one speech (TEI element <sp>) from the play, spoken (<speaker>) by the Devil. Note that the attribute xml:id gives a unique identifier to this speaker and that xml:lang marks the language in which the speaker is referred to as Latin, “lat” being the three letter ISO 639 code for this language. Note also that “Diabolus.” is marked as highlighted, with the required rendering as underlined. The speech then contains eight lines (<l>), each one given the running number of the line in the play (attribute n) and a unique identifier.

The first two lines also contain a substitution (<subst>, see also Figure 1 for the definition), where the scribe (the full description of the scribe is given in the TEI header, and the value of the attribute hand refers to this description) crossed out (<del>) a passage of the text and added (<add>) some other text. The place attribute specifies where the addition was made: in the first line inline, in the next above the line.


<oXygen/> - [C:\Documents and Settings\tomaz\Desktop\Cvetana\sp-2009-05-04.xml]
File Edit Find Project Perspective Options Tools Debugger Document Window Help
LIBXML
External Tools
XPath 2.0
100
sp-2009-05-04.xml* ×
jaslo-v3.xml* ×
jos100kv1_0.xml ×
Information
Scratch Buffer
U+003E
2282:13
Modified

```
    </sp>

    <sp>
      <speaker xml:id="sp_dt.speaker.2" xml:lang="lat">
        <hi rend="ul">Diabolus.</hi>
      </speaker>
      <l n="9" xml:id="sp_dt.l.9">Ah Adam, ah Eua koku
        <subst>
            <del hand="#hG">dougu</del>
            <add hand="#hG" place="supralinear">kratku</add>
        </subst>
        sta ta uefselle ushiualla,</l>
      <l n="10" xml:id="sp_dt.l.10">koku hitru gnado inu nadolshnost uasho
        <subst>
            <del hand="#hG">safrazkalla,</del>
            <add hand="#hG" place="inline">Sadialla.</add>
         </subst>
      </l>
      <l n="11" xml:id="sp_dt.l.11">Ste otli Bogu glich inu enaka bitti,</l>
      <l n="12" xml:id="sp_dt.l.12">Sa tega uolla ste mogli umoie roke pritti,</l>
      <l n="13" xml:id="sp_dt.l.13">Jest sem uam po tem dobrim to hudu obudu,</l>
      <l n="14" xml:id="sp_dt.l.14">koker ena kazha ueliko srezho oblubu,</l>
      <l n="15" xml:id="sp_dt.l.15">Skusi leto pak ie ufsa nasrezha zhes uafs stopilla,</l>
      <l n="16" xml:id="sp_dt.l.16">Le ta ie uafs k moim sushnim sturilla.</l>
    </sp>
```

C:\Documents and Settings\tomaz\Desktop\Cvetana\sp-2009-05-04.xml

Figure 2. A speech from the eZISS edition of the Škofja Loka Passion Play.

For viewing, the editions are transformed into HTML, using XSLT stylesheets. The older (TEI P4) editions each had a dedicated stylesheet to transform it into HTML, while editions written in TEI P5 have a much simpler dedicated stylesheet that only converts the (complex) source XML into a simplified, although still TEI compliant XML. This XML is then transformed to HTML by the standard TEI XSLT stylesheets, available form the TEI site.

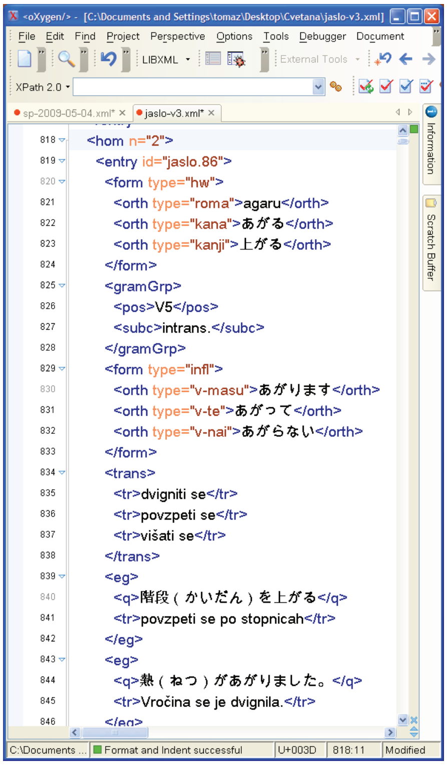
5.4. The jaSlo dictionary

The on-line Japanese-Slovene Dictionary jaSlo,[11] being developed in cooperation between the Arts Faculty of the University of Ljubljana and the Jožef Stefan Institute is a learner’s dictionary meant for Slovenian students of the Japanese language and currently contains about 10,000 entries.

[10] http://creativecommons.org

[11] http://nl.ijs.si/jaslo/

The schema for jaSlo dictionary uses the TEI P4 module for Dictionaries, and the start of an example entry is given in Figure 3. As can be seen the entry first contains the Japanese headword (<form type=”hw”>), which is further subdivided into the word written three orthographies: in Romaji (transliteration into Latin script), in Hiragana (phonetic Japanese script) and in Kanji (Chinese ideograms). The grammar group (<gramGrp>) gives the part-of-speech of the word (Verb, class 5) and its subcategorisation (intransitive), followed by three inflected forms of the verb, which determine its inflectional behaviour. The translation gives three translation equivalents in Slovene. Follows an example, in Japanese and translated into Slovene. Lexical entries have further information as well: the difficulty level of the word, a reference to the year and chapter of the textbook where the word is first introduced definitions (mostly for proper nouns), etymology, cross references, etc.


<oXygen/> - [C:\Documents and Settings\tomaz\Desktop\Cvetana\jaslo-v3.xml]
File Edit Find Project Perspective Options Tools Debugger Document
LIBXML
External Tools
XPath 2.0
sp-2009-05-04.xml*
jaslo-v3.xml*
Information
Scratch Buffer
C:\Documents ...
Format and Indent successful
U+003D
818:11
Modified

```
<hom n="2">
  <entry id="jaslo.86">
    <form type="hw">
      <orth type="roma">agaru</orth>
      <orth type="kana">あがる</orth>
      <orth type="kanji">上がる</orth>
    </form>
    <gramGrp>
      <pos>V5</pos>
      <subc>intrans.</subc>
    </gramGrp>
    <form type="infl">
      <orth type="v-masu">あがります</orth>
      <orth type="v-te">あがって</orth>
      <orth type="v-nai">あがらない</orth>
    </form>
    <trans>
      <tr>dvigniti se</tr>
      <tr>povzpeti se</tr>
      <tr>višati se</tr>
    </trans>
    <eg>
      <q>階段（かいだん）を上がる</q>
      <tr>povzpeti se po stopnicah</tr>
    </eg>
    <eg>
      <q>熱（ねつ）があがりました。</q>
      <tr>Vročina se je dvignila.</tr>
    </eg>
```

Figure 3. Start of a lexical entry in the jaSlo dictionary

The dictionary is offered on-line via a Web search interface. The interface is implemented as a Perl CGI service, which, according to the search criteria searches the TEI encoded dictionary, returns the relevant entries, and transforms the XML into HTML for display.

5.5. The JOS corpora

The JOS project[12] is developing Slovene linguistically annotated corpora and associated resources meant to facilitate developments in Human Language Technologies for the Slovene language. Current results include the JOS morphosyntactic specifications, two word-level manually annotated corpora, and two Web services. The developed resources are available under the Creative Commons licences.

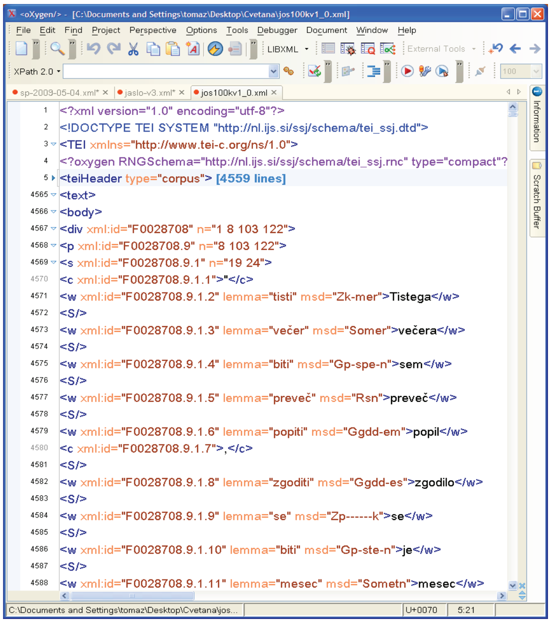
The JOS corpora contain sampled paragraphs from the Slovene reference corpus FidaPLUS,[13] annotated with context-disambiguated morphosyntactic descriptions and lemmas. The jos100k corpus contains 100,000 words and has been extensively manually validated, while jos1M contains 1 million words with partially hand validated annotations. The corpora can serve as training sets for trainable part-of-speech taggers and lemmatisers for Slovene.

The corpora are encoded in TEI P5, with the schema using the modules for Metadata for Language Corpora, Linguistic Analysis, and Feature Structures with some JOS extensions. Figure 4 shows the first part of an annotated sentence (<s>) from the jos100k corpus. The sentence contains words (<w>), punctuation marks (<c>), and whitespace (<S>). Each word is annotated for its morphosyntactic description (MSD), and the word lemma. The MSDs are compact strings, which have a direct decomposition into features structures, defined in the

[12] http://nl.ijs.si/jos/

[13] http://www.fidaplus.net/

TEI header. So, for example, the MSD “Zk-mer” decomposes into “zaimek vrsta=kazalni spol=moљki љtevilo=ednina sklon=rodilnik”, or, in English “Pronoun Type=demonstrative Gender=masculine Number=singular Case= genitive”.


<oXygen/> - [C:\Documents and Settings\tomaz\Desktop\Cvetana\jos100kv1_0.xml]
File Edit Find Project Perspective Options Tools Debugger Document Window Help
LIBXML
External Tools
XPath 2.0
100
sp-2009-05-04.xml*
jaslo-v3.xml*
jos100kv1_0.xml
Information
Scratch Buffer
C:\Documents and Settings\tomaz\Desktop\Cvetana\jos...
U+0070
5:21

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE TEI SYSTEM "http://nl.ijs.si/ssj/schema/tei_ssj.dtd">
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<?oxygen RNGSchema="http://nl.ijs.si/ssj/schema/tei_ssj.rnc" type="compact"?
<teiHeader type="corpus"> [4559 lines]
<text>
<body>
<div xml:id="F0028708" n="1 8 103 122">
<p xml:id="F0028708.9" n="8 103 122">
<s xml:id="F0028708.9.1" n="19 24">
<c xml:id="F0028708.9.1.1">"</c>
<w xml:id="F0028708.9.1.2" lemma="tisti" msd="Zk-mer">Tistega</w>
<S/>
<w xml:id="F0028708.9.1.3" lemma="večer" msd="Somer">večera</w>
<S/>
<w xml:id="F0028708.9.1.4" lemma="biti" msd="Gp-spe-n">sem</w>
<S/>
<w xml:id="F0028708.9.1.5" lemma="preveč" msd="Rsn">preveč</w>
<S/>
<w xml:id="F0028708.9.1.6" lemma="popiti" msd="Ggdd-em">popil</w>
<c xml:id="F0028708.9.1.7">,</c>
<S/>
<w xml:id="F0028708.9.1.8" lemma="zgoditi" msd="Ggdd-es">zgodilo</w>
<S/>
<w xml:id="F0028708.9.1.9" lemma="se" msd="Zp------k">se</w>
<S/>
<w xml:id="F0028708.9.1.10" lemma="biti" msd="Gp-ste-n">je</w>
<S/>
<w xml:id="F0028708.9.1.11" lemma="mesec" msd="Sometn">mesec</w>
```

Figure 4. The start of the annotated sentence “Tistega večera sem preveč popil, zgodilo se je…” from the jos100k corpus.

The corpora are available in source XML and as tabular files, converted with an XSLT script, which are more suitable for tagger and lemmatiser training. As mentioned, the JOS project also offers a Web service to MSD tag and lemmatise uploaded Slovene texts, with the two tools having been trained on the JOS corpora.

6. Internationalisation

The TEI 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. The TEI community is broadly international and multilingual, and its geographical reach increases every year. However, the complex encoding of texts at which the TEI excels requires a close understanding of the available elements (numbering over 500), and non-English speakers are at a considerable disadvantage in learning and using the Guidelines.

The TEI is therefore working to produce a working architecture for

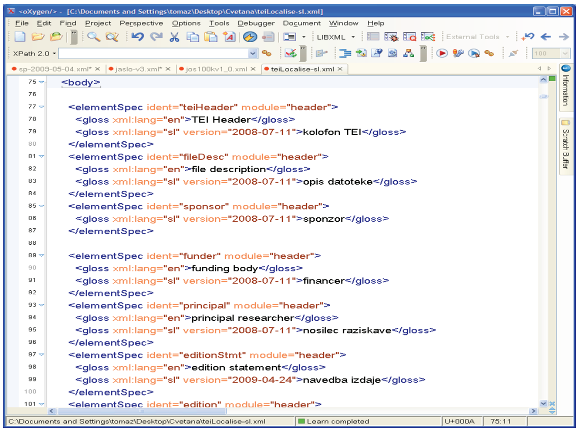
1. internationalisation of the TEI source;
2. internationalised stylesheets for delivery of the reference section of TEI guidelines;
3. translations of the TEI reference documentation.

This work is well advanced for six “large” languages, supported by a grant from the ALLC. However, it is unlikely that the Guidelines will be translated any time soon into languages with a small number of speakers, such as Slovene. The complete printed Guidelines have over 1300 pages, with the reference section, giving the precise descriptions and definitions of all the elements, etc. having almost 500 pages.

The precise XML encoding of texts is also a specialist area, so the potential readership of the translated Guidelines into Slovene would most likely remain small. In spite of it being unreasonable to expect full or reference translation of the Guidelines into every language, there are, however, some areas where a small effort produces already useful results.

For Slovene point 2 above has been partially completed, so that the TEI XSLT stylesheets that produce HTML (also performing tasks such as producing the table of contents, and splitting large TEI documents into multiple HTML files) have their strings translated, e.g. “Table of Contents” or “Next” in the HTML become “Kazalo” and “Naslednji”.

We also translated the names for all TEI elements into Slovene; this does not mean that the element names in XML are translated, but that the element definitions are given glosses in Slovene, as shown in Figure 5. As can be seen, the structure of an element specification is here quite simple: the element "teiHeader" belongs to the module "header", and has an English gloss "TEI Header" with the Slovene translation "kolofon TEI". It should be noted that it is in fact the minority of TEI elements that do have an English gloss, as most elements, like "sponsor", have a name that is itself a gloss.



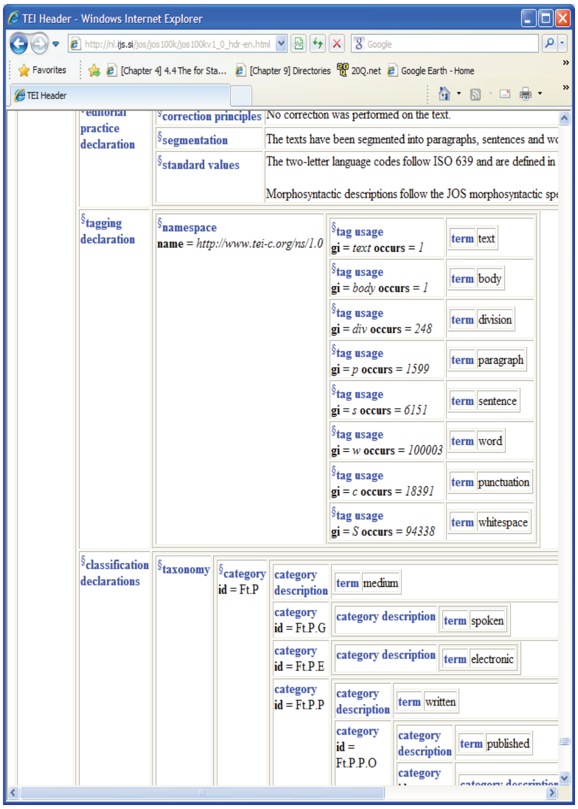
```
<body>

  <elementSpec ident="teiHeader" module="header">
    <gloss xml:lang="en">TEI Header</gloss>
    <gloss xml:lang="sl" version="2008-07-11">kolofon TEI</gloss>
  </elementSpec>
  <elementSpec ident="fileDesc" module="header">
    <gloss xml:lang="en">file description</gloss>
    <gloss xml:lang="sl" version="2008-07-11">opis datoteke</gloss>
  </elementSpec>
  <elementSpec ident="sponsor" module="header">
    <gloss xml:lang="sl" version="2008-07-11">sponzor</gloss>
  </elementSpec>

  <elementSpec ident="funder" module="header">
    <gloss xml:lang="en">funding body</gloss>
    <gloss xml:lang="sl" version="2008-07-11">financer</gloss>
  </elementSpec>
  <elementSpec ident="principal" module="header">
    <gloss xml:lang="en">principal researcher</gloss>
    <gloss xml:lang="sl" version="2008-07-11">nosilec raziskave</gloss>
  </elementSpec>
  <elementSpec ident="editionStmt" module="header">
    <gloss xml:lang="en">edition statement</gloss>
    <gloss xml:lang="sl" version="2009-04-24">navedba izdaje</gloss>
  </elementSpec>
  <elementSpec ident="edition" module="header">
```

Figure 5. Using Documentation Elements for localisation of TEI element glosses into Slovene.

The file containing such abbreviated specification for all TEI elements together with Slovene translations is available on the Web[14] in the hope that other languages will be added in the future.

While such glosses could be somewhat interesting as an addition to the actual Guidelines, we use them in a different way in our editions. We wrote an XSLT stylesheet that converts the TEI header into simple HTML, but substitutes the names of the elements by their glosses, in English or Slovene, or any other language that would be added to the specification file. This enables the translation of the document metadata; as a typical TEI header will also contain information about tag usage in the document, it also gives the translations of all the element names that are used in the document.

Figure 6 gives a part of the TEI header for the jos100k corpus, which also contains the tag usage part of the header. The jos100k header is written both in English and in Slovene (distinguishing the languages with the xml:lang attribute), so that the XSLT stylesheet inserts into HTML the appropriate language of the header content, as well as the appropriate language of the element glosses. Such HTML files are then used to present each edition either in English or in Slovene.

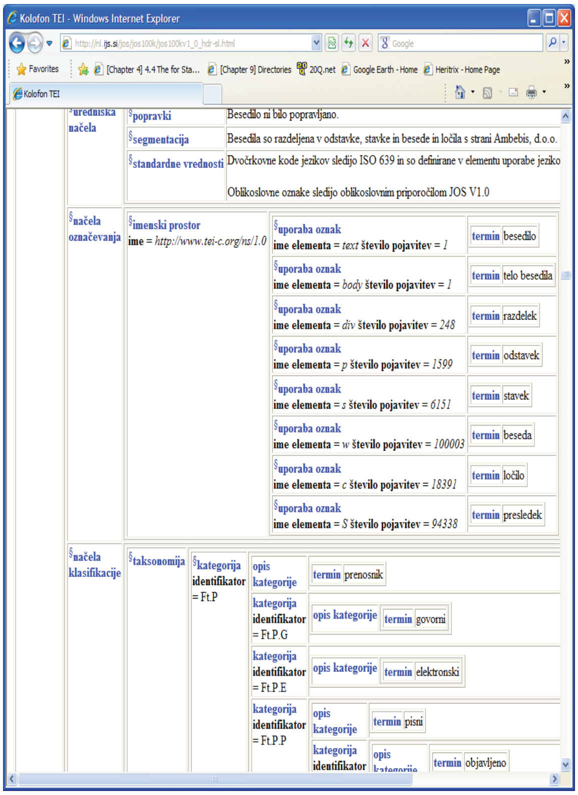

TEI Header - Windows Internet Explorer
http://nl.ijs.si/jos/jos100k/jos100kv1_0_hdr-en.html
Google
Favorites
[Chapter 4] 4.4 The for Sta...
[Chapter 9] Directories
20Q.net
Google Earth - Home
TEI Header

§editorial practice declaration
§correction principles No correction was performed on the text.
§segmentation The texts have been segmented into paragraphs, sentences and wo
§standard values The two-letter language codes follow ISO 639 and are defined in
Morphosyntactic descriptions follow the JOS morphosyntactic spe

§tagging declaration
§namespace name = *http://www.tei-c.org/ns/1.0*
§tag usage gi = *text* occurs = *1* term text
§tag usage gi = *body* occurs = *1* term body
§tag usage gi = *div* occurs = *248* term division
§tag usage gi = *p* occurs = *1599* term paragraph
§tag usage gi = *s* occurs = *6151* term sentence
§tag usage gi = *w* occurs = *100003* term word
§tag usage gi = *c* occurs = *18391* term punctuation
§tag usage gi = *S* occurs = *94338* term whitespace

§classification declarations
§taxonomy
§category id = Ft.P
category description term medium
category id = Ft.P.G category description term spoken
category id = Ft.P.E category description term electronic
category id = Ft.P.P category description term written
category id = Ft.P.P.O category description term published
category
category description

[14] http://nl.ijs.si/tei/localise/teiLocalise-sl.xml


Kolofon TEI - Windows Internet Explorer
http://nl.ijs.si/jos/jos100k/jos100kv1_0_hdr-sl.html
Google
Favorites
[Chapter 4] 4.4 The for Sta...
[Chapter 9] Directories
20Q.net
Google Earth - Home
Heritrix - Home Page
Kolofon TEI

§uredniška načela
§popravki Besedilo ni bilo popravljano.
§segmentacija Besedila so razdeljena v odstavke, stavke in besede in ločila s strani Ambebis, d.o.o.
§standardne vrednosti Dvočrkovne kode jezikov sledijo ISO 639 in so definirane v elementu uporabe jeziko

Oblikoslovne oznake sledijo oblikoslovnim priporočilom JOS V1.0

§načela označevanja
§imenski prostor
ime = *http://www.tei-c.org/ns/1.0*

§uporaba oznak **ime elementa** = *text* **število pojavitev** = *1* — termin besedilo
§uporaba oznak **ime elementa** = *body* **število pojavitev** = *1* — termin telo besedila
§uporaba oznak **ime elementa** = *div* **število pojavitev** = *248* — termin razdelek
§uporaba oznak **ime elementa** = *p* **število pojavitev** = *1599* — termin odstavek
§uporaba oznak **ime elementa** = *s* **število pojavitev** = *6151* — termin stavek
§uporaba oznak **ime elementa** = *w* **število pojavitev** = *100003* — termin beseda
§uporaba oznak **ime elementa** = *c* **število pojavitev** = *18391* — termin ločilo
§uporaba oznak **ime elementa** = *S* **število pojavitev** = *94338* — termin presledek

§načela klasifikacije
§taksonomija
§kategorija **identifikator** = Ft.P
opis kategorije — termin prenosnik
kategorija **identifikator** = Ft.P.G — opis kategorije — termin govorni
kategorija **identifikator** = Ft.P.E — opis kategorije — termin elektronski
kategorija **identifikator** = Ft.P.P — opis kategorije — termin pisni
kategorija **identifikator** — opis kategorije — termin objavljeno

Figure 6. The jos100k TEI header transformed into HTML with English or Slovene element name glosses and header content.

7. Conclusions

The paper introduced the Text Encoding Initiative, an international effort to develop a community standard for representing digital texts in a way that is both powerful and responsive to the research needs of humanities scholars. We discussed the TEI organisation, its history, Guidelines, and schema construction, and gave three examples from our own work, were the TEI encoding scheme is used. The paper concluded with a discussion of the internationalisation of the TEI Guidelines.

We presented a "light" approach to localisation, appropriate for smaller languages, where the glosses of the TEI elements are translated, and can be then used e.g. for a HTML presentation of TEI headers; the XML file with all the elements, their English glosses and their translations into Slovene is freely available on the Web, and can be used for adding the translations into other languages.

As can be seen from the paper, the TEI Guidelines are a large and rather complex document, so a legitimate question is whether it is worth the time and effort it takes to learn and use this recommendation – it is not better for particular projects to develop their own encoding, perfectly suited to their needs? The answer, based on our own experience is not straightforward. For simple texts, which will be used by a particular individual, or inside a particular institution, they answer might well be "no". However, if a substantial investment of time is required to annotate the texts, making them valuable and worthy of long-term preservation, and possibly usable in more ways than one, and especially if the texts are to be made generally available in their source form, then TEI most likely is the answer: the markup of the texts can be formally validated, it is by definition well-documented, and the projects can make use of the substantial (and increasing) software that is associated with TEI.

Web References

Below we give the more important URLs already mentioned in the text; we do not list classical bibliographic references, which, however, can be found by following the links given below. All URLs have been accessed on 2009-09-10.

XML – Extensible Markup Language: http://www.w3.org/XML/

ISO Relax NG – Schema language for validating XML documents: http://www.relaxng.org/

XSLT – XSL Transformations: http://en.wikipedia.org/wiki/XSL_Transformations

TEI Consortium Web Site: http://www.tei-c.org/

eZISS – a digital library of Slovenian Literary texts: http://nl.ijs.si/e-zrc/

jaSlo - an on-line Japanese-Slovene Learner's Dictionary: http://nl.ijs.si/jaslo/

JOS – Slovene Language Corpora for HLT Research: http://nl.ijs.si/jos/

TEI element name translations into Slovene: http://nl.ijs.s^{s}i/tei/localise/teiLocalise-sl.xml