TEI BY EXAMPLE

MODULE 0: INTRODUCTION TO TEXT ENCODING AND THE TEI

Ron Van den Branden

Edward Vanhoutte

Melissa Terras

Centre for Scholarly Editing and Document Studies (CTB) , Royal Academy of Dutch Language and Literature, Belgium, Gent, 9 July 2010

Last updated September 2020

TABLE OF CONTENTS

1. Introduction....1
2. LaTeX....3
3. OpenDocument Format....5
4. COCOA....8
5. TEI P3 (SGML)....8
6. TEI P5 (XML)....9

1. Introduction

Contrary to the other examples sections, this examples section of the introductory TBE tutorial will illustrate different types of encoding for one sample text. These markup samples will range from procedural to descriptive markup languages, in a variety of formats (text, SGML, XML). Before starting, have a look at the following document:

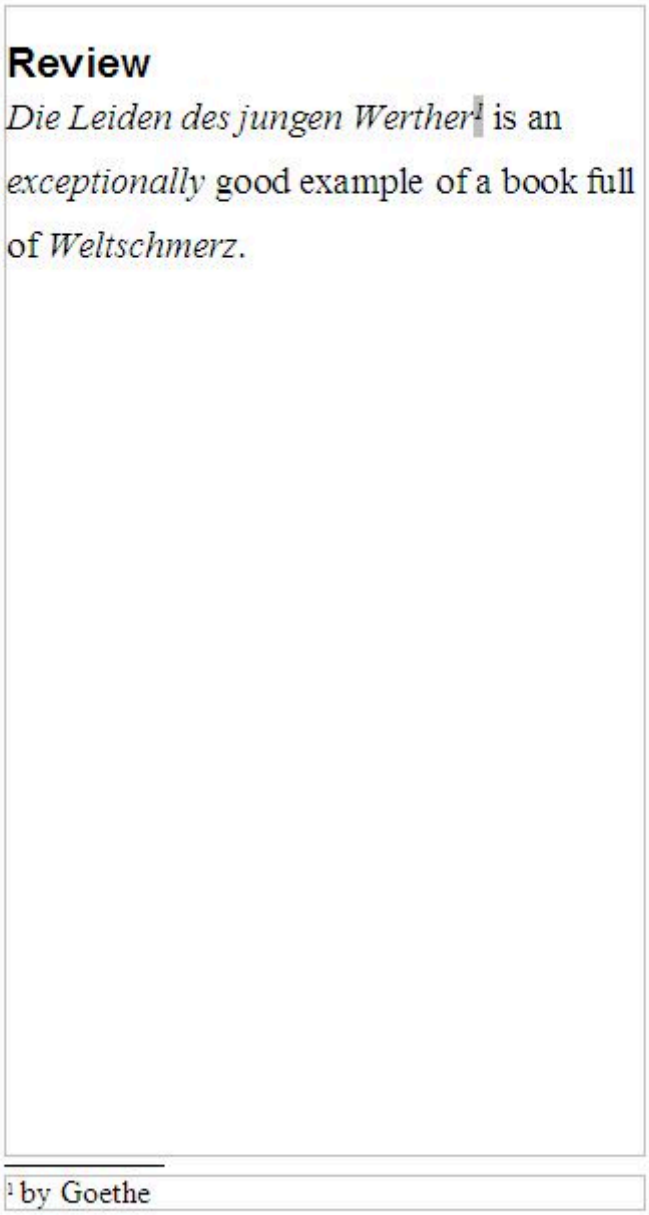

Figure 1. A sample prose document.

In this short piece of prose, following text structures can be distinguished:

- a heading
- a paragraph

- a footnote

Apart from these structures, some textual phenomena can be distinguished:

- a title (“Die Leiden des jungen Werther”)
- emphasised text (“exceptionally”)
- a term (“Weltschmertz”)
- a name (“Goethe”)

Let’s have a look how different encoding flavours treat these phenomena.

2. LaTeX

This example illustrates how the text above could be encoded in LaTeX, an open source typesetting language that can be interpreted by TeX typesetting programs for producing fixed-layout representations such as PDF. LaTeX is not an XML format, and makes use of “procedural” markup, whose meta-information (starting with the `\` character) are instructions for the rendering software on the layout of the text content. As you can see, the actual text contents are preceded by a declaration of several style aspects determining how the text has to be rendered on a page. The text is divided captured as a `{document}`, in which all italicised words are indicated as italicised (`\textit{}`), without difference between the reasons for this typographic emphasis. The footnote is distinguished (`\footnote{}`), but there is no way of telling the computer that “Goethe” is a proper name.

```
 \documentclass[12pt]{article}
\usepackage{makeidx}
\usepackage{multirow}
\usepackage{multicol}
\usepackage[dvipsnames,svgnames,table]{xcolor}
\usepackage[dvips]{graphicx}
\usepackage{ulem}
\usepackage{hyperref}
\author{the TBE crew}
\title{Statement}
\setlength{\paperwidth}{419pt}
\setlength{\paperheight}{595pt}
\setlength{\textheight}{451pt}
\setlength{\textwidth}{239pt}
```

```
\setlength{\voffset}{-72pt}
\setlength{\hoffset}{-72pt}
\setlength{\evensidemargin}{90pt}
\setlength{\oddsidemargin}{90pt}
\setlength{\topmargin}{39pt}
\setlength{\headheight}{13pt}
\setlength{\headsep}{20pt}
\makeatletter
\newenvironment{indentation}[3]%
{\par\setlength{\parindent}{#3}
\setlength{\leftmargin}{#1}       \setlength{\rightmargin}{#2}%
\advance\linewidth -\leftmargin       \advance\linewidth -\rightmargin%
\advance\@totalleftmargin\leftmargin  \@setpar{{\@@par}}%
\parshape 1\@totalleftmargin \linewidth\ignorespaces}{\par}%
\makeatother
% new LaTeX commands
\newcommand{\styleCaption}[1]{\textit{#1}}
\newcommand{\styleEndnoteCharacters}[1]{#1}
\newcommand{\styleFootnoteCharacters}[1]{$^{#1}$}
\newcommand{\styleHeading}[1]{{\large \textsf{#1}}}
\newcommand{\styleIndex}[1]{#1}
\newcommand{\styleStandaardalinealettertypeOne}[1]{#1}
\begin{document}
\section{\textsf{Review}}
{\raggedright
\begin{indentation}{0pt}{0pt}{0pt}
{\large \textit{Die Leiden des jungen Werther\footnote{ by Goethe
}} is an \textit{exceptionally} good example of a book full of
\textit{Weltschmerz}.}
\end{indentation}
}
\end{document}
```

Example 1. A LaTeX example.

3. OpenDocument Format

The same document can be encoded in the OpenDocument Format, an XML encoding scheme for representing electronic documents such as spreadsheets, charts, presentations and word processing documents, that can be interpreted by (desktop) publishing systems such as the Open Office software suite. Notice that, despite ODF being expressed in XML, there are many similarities to the LaTeX approach in the previous example. ODF is a procedural encoding scheme as well, providing an XML vocabulary to describe different *formatting styles*. The text itself is encoded in a `<office:text>` element, in which several structural elements are distinguished: headings, paragraphs, footnotes, each with their own associated rendering instructions in the form of styles. All italicised text is represented in the encoding, with references to different style definitions that are responsible for rendering the text italic in the output. Here, too, there is no way of indicating that the visually unmarked "Goethe" is a proper name.

```
<office:document-content xmlns:oooc="http://openoffice.org/2004/calc"
 xmlns:dom="http://www.w3.org/2001/xml-events" xmlns:xsd="http://www.w3.org/2001/
XMLSchema" xmlns:fo="urn:oasis:names:tc:opendocument:xmlns:xsl-fo-compatible:1.0"
 xmlns:ooo="http://openoffice.org/2004/office" xmlns:xsi="http://www.w3.org/2001/
XMLSchema-instance" xmlns:number="urn:oasis:names:tc:opendocument:xmlns:datastyle:1.0"
 xmlns:of="urn:oasis:names:tc:opendocument:xmlns:of:1.2"
 xmlns:rdfa="http://docs.oasis-open.org/opendocument/meta/rdfa#"
 xmlns:text="urn:oasis:names:tc:opendocument:xmlns:text:1.0"
 xmlns:table="urn:oasis:names:tc:opendocument:xmlns:table:1.0" xmlns:xforms="http://
www.w3.org/2002/xforms" xmlns:svg="urn:oasis:names:tc:opendocument:xmlns:svg-
compatible:1.0" xmlns:draw="urn:oasis:names:tc:opendocument:xmlns:drawing:1.0"
 xmlns:script="urn:oasis:names:tc:opendocument:xmlns:script:1.0"
 xmlns:dr3d="urn:oasis:names:tc:opendocument:xmlns:dr3d:1.0"
 xmlns:form="urn:oasis:names:tc:opendocument:xmlns:form:1.0"
 xmlns:field="urn:openoffice:names:experimental:ooo-ms-interop:xmlns:field:1.0"
 xmlns:meta="urn:oasis:names:tc:opendocument:xmlns:meta:1.0"
 xmlns:rpt="http://openoffice.org/2005/report"
 xmlns:style="urn:oasis:names:tc:opendocument:xmlns:style:1.0" xmlns:math="http://
www.w3.org/1998/Math/MathML" xmlns:ooow="http://openoffice.org/2004/
writer" xmlns:chart="urn:oasis:names:tc:opendocument:xmlns:chart:1.0"
 xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:office="urn:oasis:names:tc:opendocument:xmlns:office:1.0" office:version="1.2">
```

```
<office:scripts/>
<office:font-face-decls>
  <style:font-face style:name="Tahoma1" svg:font-family="Tahoma"/>
  <style:font-face style:name="Times New Roman" svg:font-family="'Times New
   Roman'" style:font-family-generic="roman" style:font-pitch="variable"/>
  <style:font-face style:name="Arial" svg:font-family="Arial" style:font-family-
  generic="swiss" style:font-pitch="variable"/>
  <style:font-face style:name="Lucida Sans Unicode" svg:font-family="'Lucida Sans
   Unicode'" style:font-family-generic="system" style:font-pitch="variable"/>
  <style:font-face style:name="Tahoma" svg:font-family="Tahoma" style:font-family-
  generic="system" style:font-pitch="variable"/>
</office:font-face-decls>
<office:automatic-styles>
  <style:style style:name="P1" style:family="paragraph" style:parent-style-
  name="Heading_20_1" style:master-page-name="Standard">
    <style:paragraph-properties style:page-number="auto"/>
  </style:style>
  <style:style style:name="P2" style:family="paragraph" style:parent-style-
  name="Standard">
    <style:paragraph-properties fo:line-height="150%"/>
  </style:style>
  <style:style style:name="T1" style:family="text">
    <style:text-properties fo:language="nl" fo:country="BE"/>
  </style:style>
  <style:style style:name="T2" style:family="text">
    <style:text-properties fo:font-style="italic" style:font-style-asian="italic"/>
  </style:style>
  <style:style style:name="T3" style:family="text">
    <style:text-properties fo:font-size="14pt" style:font-size-
    asian="14pt" style:font-size-complex="14pt"/>
  </style:style>
  <style:style style:name="T4" style:family="text">
    <style:text-properties fo:font-size="14pt" fo:font-style="italic" style:font-size-
    asian="14pt" style:font-style-asian="italic" style:font-size-complex="14pt"/>
  </style:style>
  <style:style style:name="T5" style:family="text">
```

```
    <style:text-properties fo:font-
    size="12pt" fo:language="nl" fo:country="BE" style:font-size-
    asian="12pt" style:font-size-complex="12pt"/>
  </style:style>
</office:automatic-styles>
<office:body>
  <office:text>
    <office:forms form:automatic-focus="false" form:apply-design-mode="false"/>
    <text:sequence-decls>
      <text:sequence-decl text:display-outline-level="0" text:name="Illustration"/>
      <text:sequence-decl text:display-outline-level="0" text:name="Table"/>
      <text:sequence-decl text:display-outline-level="0" text:name="Text"/>
      <text:sequence-decl text:display-outline-level="0" text:name="Drawing"/>
    </text:sequence-decls>
    <text:h text:style-name="P1" text:outline-level="1">Review</text:h>
    <text:p text:style-name="P2">
      <text:span text:style-name="T4">Die Leiden des jungen Werther</text:span>
      <text:span text:style-name="Footnote_20_Symbol">
        <text:span text:style-name="T4">
          <text:note text:id="ftn1" text:note-class="footnote">
            <text:note-citation>1</text:note-citation>
            <text:note-body>
              <text:p text:style-name="Footnote">
                <text:s/>
                <text:span text:style-name="T5">by Goethe</text:span>
              </text:p>
            </text:note-body>
          </text:note>
        </text:span>
      </text:span>
      <text:span text:style-name="T3"> is an </text:span>
      <text:span text:style-name="T4">exceptionally</text:span>
      <text:span text:style-name="T3"> good example of a book full of </text:span>
      <text:span text:style-name="T4">Weltschmerz</text:span>
      <text:span text:style-name="T3">.</text:span>
    </text:p>
  </office:text>
</office:body>
```

```
</office:document-content>
```

Example 2. An OpenDocument example.

4. COCOA

In the next example, the sample text is encoded in COCOA. This encoding scheme shares with the LaTeX example above its non-XML character, but differs in that COCOA is a "descriptive" markup scheme. It provides a simple means to distinguish user-defined categories in a text, by labeling them unambiguously by means of one-letter tag names. There are two possibilities: either the text is encoded in the tag (e.g., `<H Review>` identifies the text "Review" as belonging to the category "H" (for "heading")), or a tag is numbered (e.g., `<P 1>` indicates that the text following it is part of the first paragraph). This enables the encoder not only to distinguish all text structures (heading ("H"), paragraph ("P"), footnote ("F"); but also to distinguish between the different textual phenomena that occur as italicised text (book title ("B"), emphasis ("E"), term ("T")). Moreover, the typographically unmarked proper name "Goethe" can be tagged as such as well ("N").

```
<H Review>
 <P 1><B Die Leiden des jungen Werther>&lt;F 1>by <N Goethe > is an
 <E exceptionally> good example of a book full of <T Weltschmerz>
```

Example 3. A COCOA example.

5. TEI P3 (SGML)

The sample text could be encoded in TEI P3 as well. Being TEI, this is a descriptive encoding scheme that allows the encoder to explicate the structure and semantics of the textual features s/he wants to analyse. In our sample, we see the typical features of TEI documents (although some of the names have evolved since version P3): a document is encoded in a `<TEI.2>` element, containing both a <teiHeader> section for the meta-information, and a <text> part for the actual text contents. The header must contain a minimal amount of meta-information, while the text content itself is encoded in <body>. Inside the text, the structural elements (heading — <head>, paragraph — <p>, footnote — `<note @place=foot>`), as well as semantic features (title — <title>, emphasis — <emph>, term — <term>) can be fully expressed with comprehensible tag names.

Notice, however, that this is SGML, not XML: some elements can occur without end tags (<title>, <body>, <p>, <head>), and attribute values can occur without surrounding quotes (“type=foot”).

```
<TEI.2>
 <teiHeader>
   <fileDesc>
     <titleStmt>
       <title>Review: an electronic transcription
     </titleStmt>
     <publicationStmt>
       <p>Published as an example for the Introduction module of TBE.
     </publicationStmt>
     <sourceDesc>
       <p>No source: born digital.
     </sourceDesc>
   </fileDesc>
 </teiHeader>
 <text>
   <body>
     <head>Review
     <p><title>Die Leiden des jungen Werther <note place=foot>by <name>Goethe</name> is an
<emph>exceptionally</emph> good example of a book full of <term>Weltschmerz</term>.
 </text>
</TEI.2>
```

Example 4. A TEI P3 SGML example.

6. TEI P5 (XML)

Finally, this example illustrates how a TEI P5 (XML) encoding of the sample text could look. The latest version ot the TEI Guidelines specify a descriptive encoding scheme in XML format. As you’ll see, there are much similarities with the TEI P3 encoding of the previous example: all structural and semantic text features can be indicated and labeled with fairly intuitive element names. Still, some differences stand out:

- in TEI P5, all elements must have end tags
- in TEI P5, all attribute values must be surrounded by quotes
- some basic element names have changed (e.g., the first element of any TEI P5 text is now called <TEI>)

- in TEI P5, many details of the text ontology have been changed, some elements have been revised, improved, deleted, or added

The TBE tutorials will guide you through the most important sections of the TEI Guidelines that should enable you to encode the most common features of different text genres, and derive TEI encoding schemes according to your needs.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Review: an electronic transcription</title>
      </titleStmt>
      <publicationStmt>
        <p>Published as an example for the Introduction module of TBE.</p>
      </publicationStmt>
      <sourceDesc>
        <p>No source: born digital.</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <head>Review</head>
      <p><title>Die Leiden des jungen
       Werther</title><note place="foot">by <name>Goethe</name></note>
       is an <emph>exceptionally</emph> good example of a book full
       of <term>Weltschmerz</term>.</p>
    </body>
  </text>
</TEI>
```

Example 5. A TEI P5 XML example.