[DOCX 2 TEI](http://nl.ijs.si/tei/convert/)

Example document

**Tomaž Erjavec**

2013-06-01

Contents

[Introduction 2](#_Toc357845414)

[Web interface 2](#_Toc357845415)

[Bugs and suggestions 2](#_Toc357845416)

[Figures and Tables 2](#_Toc357845417)

[Other features 4](#_Toc357845418)

[Standard character level styles 4](#_Toc357845419)

[Notes 4](#_Toc357845420)

[Indexes 4](#_Toc357845421)

[Bibliography 4](#_Toc357845422)

[Page breaks 4](#_Toc357845423)

[TEI element styles 6](#_Toc357845424)

[Paragraph level styles and elements 6](#_Toc357845425)

[Character level styles / phrase level elements 7](#_Toc357845426)

[Janus elements 7](#_Toc357845427)

[Auto-generated sections 8](#_Toc357845428)

[Index 8](#_Toc357845429)

[List of Figures 8](#_Toc357845430)

[List of Tables 8](#_Toc357845431)

# Introduction

This document is meant as an exemplar and test .docx file for a docx2tei profile of the [TEI Stylesheets](http://www.tei-c.org/Tools/Stylesheets/). Its other function is as the source for a .dotx Word template than can serve for authoring new or editing existing Word files (primarily books) with the intention of converting them to TEI with, say, [OxGarage](http://www.tei-c.org/oxgarage/).

This profile imports Stylesheets/profiles/default/docx/from.xsl and extends it a bit, but also requires a certain style while editing the Word document to be converted:

* The more is done with Word styles, the better; and formatting and esp. layout is not well preserved.
* The division before the first heading, with Title, Subtitle, Author and Date goes to the element tei:front, with the rest of the text in this division ignored; the first heading and the rest is tei:body.
* Styles starting with “tei:” are substituted by their TEI element, c.f. Section “TEI elements”;
* Some styles also have shorthand notation, e.g. a series of Bibliography-style paragraphs go into tei:listBilb/tei:bibl+ and tei:orgName goes to tei:name[@type=”org”], and some styles get extra processing,

## Web interface

A test interface is at <http://nl.ijs.si/tei/convert/>

### Bugs and suggestions

1. Styles don’t work well in notes, as e.g. “tei:bibl” is there “teibibl”, c.f. footnote 3.
2. Bad handling of Table captions, are not in the tei:table.
3. Problems with cross references to tables, footnotes, bookmarks, …
4. For included Excel OxGarage produces embeddings/\*.xsl, but Stylesheets don’t.
5. Bad (generated) Index and “Table of Figures”; maybe: identify, remove & insert divGen?
6. Broken elements should probably be joined (e.g. <name>P<name><name>odravje</name> → <name>Podravje</name>), c.f. also example with Add
7. Problems with phrase level elements containing initial or final space, which should probably go outside the element (e.g. “foo<term> bar </term>baz” → “foo <term>bar</term> baz”)
8. Word leaves many dead <anchor>s , these could be cleaned up
9. Embedded XSL could also be converted and inserted into TEI?
10. pb’s at start / end of elements could be promoted upwards
11. Facsimiles! Ideally maybe submit zip with docx+pdf, and make pointers into pdf if poss., or split into page images and link.

# Figures and Tables

Here we reference Figure 1 as well as Table 1: Overview and Table 2.



Figure 1: An included image

Figure 2. An embedded .xslx graph, doesn’t work yet.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Lendava | Murska Sobota | Beltinci | okoliške vasi | skupaj |
| 1778 | 14 | 0 | 0 | 0 | 14 |
| 1793 | 19 | 14 | 21 | 6 | 60 |
| 1812 | 23 | 13 | 40 | 0 | 76 |

Table 1 A Simple table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SLOVENSKE DEŽELE** | | | | |
|  | 144 | 192 | 182 |
|  | 96 | 89 | 145 |
| *Ljubljana* | *(74)* | *(76)* | *(95)* |
| **Sodni okraji z ozemlja današnje slovenske Primorske** | | | | |

Table 2 A bit more Complex table

# Other features

## Standard character level styles

Things like **bold**, *italic*, underline, and ~~strikethrough~~ and red are stored as hi/@rend, as well as ***~~italic bold underline strikethrough~~***, but the exact details, such as the colour of underline and more **fancy text effects** are not preserved. Information on fonts is not preserved.

We can also have dates in paragraphs, e.g. 1984-04-04.

## Notes

Margin Note Left

Margin Note Right

Footnotes are probably ok,[[1]](#footnote-1) more problematic are marginal notes. However, one DOCX2TEI profile has a template.docx file that defines 4 MarginNote styles, used here. But I couldn’t find any from.xsl that would actually do anything with them. So, now they get @place=”margin\_xx” where xx ∈ {left, right, inner, outer}.[[2]](#footnote-2)

Margin Note Inner

Margin Note Outer

## Indexes

You need to reveal formatting to see that we are here indexing the Web, Web services, and Web apps, but also bugs and errors. Note that the index terms can be in Word also formatted, which is lost in the TEI. We can have ranges though, like this.

## Bibliography

This is maybe a too strict style? Also, currently no integration with Words bibl handling?, but maybe that is impossible to do anyway.

1. AUMÜLLER, Jutta: Assimilation: Kontroversen um ein migrationspolitisches Konzept. Bielefeld: Transcript Verlag, 2009.
2. BAGARIĆ, Marija: Obitelj Kapetanović. V: Pravednici među narodima [online], <http://www.geoskola.hr/hr/projekti/pravednici/kapetanovici.htm> (Pridobljeno 27. 9. 2007).
3. BARKAI, Avraham: Jüdisches Leben unter der Verfolgung. V: Michael A Meyer in Michael Brenner (ur.): Deutsch-jüdische Geschichte in der Neuzeit, 4. del: Aufbruch und Zerstörung, München: Verlag C. H. Beck, 2000, str. 225-248.
4. BELLER, Steven: Wien und die Juden 1867-1938. Wien, Köln in Weimar: Böhlau, 1993.

# Page breaks

Page breaks are problematic, esp. the soft ones, e.g. if they appear in tables or two-column layout. Also, they can appear inside any otherwise empty element, like p, head, div.  
Here was a hard line break, and here is a page break, first soft, then hard:

# TEI element styles

These styles start with “tei:” and the string following is - typically - the name of the TEI element. It some cases, to make them easier to use, the styles are mapped to more complicated structures; to see which and how it is easiest to compare this file with the derived TEI. The styles presented here use lots of eye watering effects, which need some work.

## Paragraph level styles and elements

These styles are paragraph level, i.e. they mark complete paragraphs.

First, a quote, which is always outside paragraphs:

The doer alone learneth.

A citation with bibliographic item:

A spectre is haunting Europe; the spectre of Communism. The Communist Manifesto (1848), by Karl Marx and Friedrich Engels

Next, a bit of poetry, using tei:lg:

There once was a man from Nantucket

Who kept all his cash in a bucket.

But his daughter, named Nan,

Ran away with a man

And as for the bucket, Nantucket.

We can also have individual lines of poetry, without line group:

There once was a man from Nantucket

Who kept all his cash in a bucket.

A drama speech, with the convention that the first paragraph goes to speaker and the rest are lines:

Polonius

Though this be madness, yet there is method in ’t.

Will you walk out of the air, my lord?

Hamlet

Into my grave.

## Character level styles / phrase level elements[[3]](#footnote-3)

These styles appear inside paragraphs:

Name text Organisation text Person text Place text.

Bibl text Foreign text “Quoted” text.

Damage text (reason for) Gap text Supplied text Surplus text Unclear text.

### Janus elements

The following pairs of elements get, if they are adjacent, a parent element: tei:subst for tei:del + tei:add and tei:choice for the others:

Add text Delete Add text Delete.

Abbreviation text Abbreviation Expansion text Expansion.

text Orig text Orig Regularised text Regularised text.

text Sic text Sic Corr text Corr text.

Catch-word (typically printers mark)

# Auto-generated sections

## Index

bugs

horrible bugs. *See* bugs

error, 4

**formatting**, 4

Range

Subrange, 7

Web, 4

Web app, 4

Web service, 4

## List of Figures

Can you do this at all? And what exactly is “Index of authorities”?

## List of Tables

[Table 1 A Simple table 3](#_Toc357777361)

[Table 2 A bit more Complex table 3](#_Toc357777362)

1. Although there used to be problems with them in HTML, but we've yet to get there.. [↑](#footnote-ref-1)
2. Again, worries how to set this in HTML… [↑](#footnote-ref-2)
3. It turns out to be problematic to have phrase-level styles in notes, c.f. AUMÜLLER, Jutta: Assimilation: Kontroversen um ein migrationspolitisches Konzept. Bielefeld: Transcript Verlag, 2009. [↑](#footnote-ref-3)