TEI Simple: state of the nation

James Cummings
(and TEI-Simple Team)

October 2014

Simple Partners

Joint Project between:

- Northwestern University
- University of Nebraska-Lincoln
- University of Oxford
- TEI Consortium
- Mellon Foundation

TEI Simple

- Joint project between TEI Consortium, Mellon Foundation, Northwestern University, University of Nebraska-Lincoln, and the University of Oxford
- TEI Simple aims to ‘define a new highly-constrained and prescriptive subset of the Text Encoding Initiative (TEI) Guidelines suited to the representation of early modern and modern books, and a formally-defined set of processing rules which permit modern web applications to easily present and analyze the encoded texts’
- Initial project runs from September 2014 to July 2015
- All outputs are open source, all working is in the open on github and trello

```
https://github.com/TEIC/TEI-Simple
```

Simple Staff

Project Investigators

- Martin Mueller, Northwestern University
- Brian Pytlik Zillig, University of Nebraska-Lincoln
- Sebastian Rahtz, University of Oxford

Additional Project Members

- Magdalena Turska, DiXiT Project / University of Oxford
- Lou Burnard, Consultant
- James Cummings, DiXiT Project / University of Oxford

Advisory Committee

- Pip Willcox, Bodleian Library, Oxford
- Suzanne Haaf, Deutsches Textarchiv, Berlin
- Matthias Goebel, University of Gottingen
- James Cummings, University of Oxford

Objectives

1. The highly constrained and prescriptive element subset of TEI Simple
2. The processing model (Simple Processing Model: SPM)
3. Formal mapping of the TEI elements used by Simple to the CIDOC CRM
4. TEI-Performance Indicators
5. Integration of TEI Simple into the TEI infrastructure

The CIDOC Conceptual Reference Model (CRM) provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation.

Deliverables

1. A TEI ODD customization with the choices and constraints, SPM notation, and RDF mapping
2. Multi-stage tutorial documentation based on examples using page images and XML
3. A revised set of TEI Stylesheets which implement the SPM

These outputs will be offered to the TEI Technical Council to decide on how to best to incorporate and maintain them

The Simple schema: guidelines

- ‘Simple’ does not have to mean ‘Small’
- The schema is based on analysis of existing usage from corpora:
  - Text Creation Partnership (including Evans, ECCO, EEBO, and unreleased phase 2)
  - Oxford Text Archive: All TEI P5 files
  - Deutsches Textarchiv
  - Documenting the American South
  - CESR
  - OBVIL: corpus critique
- Our biggest enemy is ambiguity for the encoders and developers
- The target is encoding of the <text>; the <teiHeader> and any <sourceDoc> or <facsimile> are much less constrained

‘Simple’ is not necessary ‘Small’ and certainly not ‘Simplistic’.
In this case it is ‘Simple’ as Powerful and Processable.

The customizaton

- We isolate 104 elements which are needed in the body of a text
- The choice is fairly obvious, comparable to DTA, Lite, Tite etc
- We divide them into groups by function, mainly for documentation purposes
- We have started analyzing attribute usage

Texts from all the source corpora representable in Simple

Element groups (1)

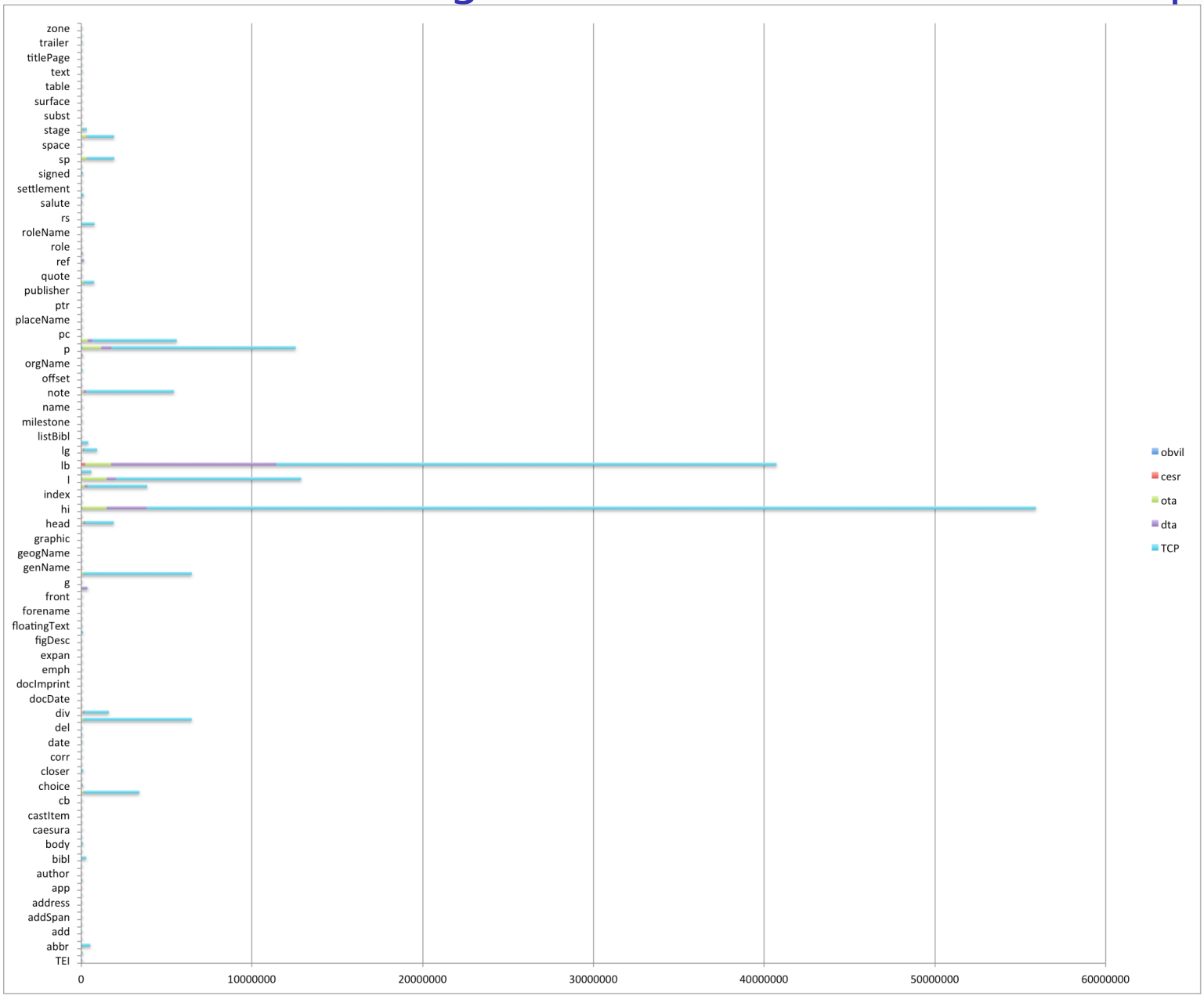
castlist	<actor> <castGroup> <castItem> <castList> <role> <roleDesc>
character	<g>
editorial	<abbr> <add> <addSpan> <am> <choice> <corr> <del> <desc> <ex> <expan> <gap> <handShift> <orig> <reg> <sic> <space> <subst> <supplied> <unclear>
interpretation	<author> <date> <foreign> <hi> <measure> <name> <num> <q> <quote> <ref> <rs> <seg> <time>
linguistic	<c> <pc> <s> <w>
pictures	<figDesc> <figure> <graphic>

Element groups (2)

structure	<bibl> <title> <TEI> <teiCorpus> <ab> <address> <addrLine> <anchor> <back> <body> <cb> <cit> <floatingText> <formula> <front> <fw> <group> <head> <item> <l> <label> <lb> <lg> <list> <listBibl> <milestone> <note> <p> <pb> <sp> <speaker> <spGrp> <stage> <text>
table	<cell> <row> <table>
titlepage	<publisher> <pubPlace> <docAuthor> <docDate> <docEdition> <docImprint> <docTitle> <imprimatur> <titlePage> <titlePart>
wrapper	<argument> <byline> <closer> <dateline> <epigraph> <opener> <postscript> <salute> <signed> <trailer>

The ‘wrapper’ elements remain a major concern

How does usage of those elements stack up?



Element usage (raw count)

	obvil	cesr	ota	dta	TCP	Total
TEI	126	50	2844	848	55820	59688
ab	290	1022	60496	0	59	61867
abbr	0	1482	7015	69	531348	539914
actor	0	0	0	12	0	12
add	0	71	137	0	2124	2332
addName	0	7	0	0	0	7
addSpan	0	0	2	0	4	6
addrLine	0	10	0	0	0	10
address	0	5	0	0	0	5
anchor	233	0	29	0	4	266
app	0	6	0	0	0	6
argument	151	0	1795	2268	66327	70541
author	34	83	1	0	0	118
back	1	29	1007	608	22462	24107
bibl	274	155	16180	4362	279493	300464
biblScope	0	4	0	0	0	4
body	126	50	9246	1155	98314	108891
byline	17	8	3549	1028	11945	16547
caesura	0	0	30299	0	0	30299
castGroup	0	29	0	51	0	80
castItem	0	63	90	744	0	897
castList	0	28	4	74	0	106

Top 20: does it mean anything?

fw	0	20554	15	336239	561	357369
list	621	115	17848	24877	370296	413757
abbr	0	1482	7015	69	531348	539914
label	621	469	19570	0	577272	597932
q	4270	29	110383	0	645279	759961
row	70	1311	30035	12571	738494	782481
lg	316	4481	86396	84157	761815	937165
div	4584	2128	85352	104832	1425044	1621940
head	4961	2526	115775	117506	1670818	1911586
speaker	12	1978	310490	42021	1575243	1929744
sp	12	1997	312812	42050	1583715	1940586
cell	242	3218	138105	34153	3235963	3411681
item	2023	1806	188278	187951	3496780	3876838
note	9451	2874	113910	187675	5122160	5436070
pb	6156	9497	362846	311001	4909429	5598929
desc	186	0	125150	0	6352117	6477453
gap	0	18	125202	5731	6352257	6483208
p	68273	4855	1099529	607881	10786471	12567009
l	21004	39828	1428379	557846	10842122	12889179
lb	3008	237202	1508235	9689271	29293798	40731514
hi	56953	9424	1419289	2367314	52066396	55919376

@rend and @type proposal

- We will *not* usually prescribe *@type*, but instead publish a separate suggested taxonomy for possible interpretative use
- We will preclude use of *@rend* and *@style*
- We will produce a closed list of values for *@rendition* using a private URI of "simple:"

```
<p  rendition="simple:bold">This is quite bold, but this
<hi  rendition="simple:sup">is superscript</hi>
</p>
```

Do we have a single set of values for *@rendition*, or constrained per element?

@type on samples

...case_concerning_exports case_concerning_printing_privileges
case_history case_law case_proper case_reports case_restated
case_studies case_study case_summary cases casestudies casestudy
cast castle castlist casualties casulaties catalogue catalogue_entry
catalogue_of_Calvins_works catalogue_of_Chancellors
catalogue_of_English_martyrs
catalogue_of_English_public_schools_and_European_universities
catalogue_of_Hindu_gods catalogue_of_Lord_Treasurers
catalogue_of_Oxford_University_Press
catalogue_of_apocryphal_biblical_books catalogue_of_archbishops
catalogue_of_army catalogue_of_authors_works catalogue_of_birds
catalogue_of_bishops catalogue_of_books
catalogue_of_books_for_auction catalogue_of_cathedrals
catalogue_of_colleges catalogue_of_compounds
catalogue_of_coronets_colors_and_ensigns ...

A constraint example

```
<classSpec ident="att.pointing"
 mode="change">
 <attList>
  <attDef ident="target" mode="change">
   <constraintSpec ident="validtarget"
    scheme="isoschematron">
    <constraint>
     <s:rule context="tei:*[@target]">
      <s:let name="results"
       value="for $t in tokenize(normalize-space(@target),'\s+') return
starts-with($t,'#') and not(id(substring($t,2)))"/>
      <s:report test="some $x in $results satisfies $x">Error: Every
local pointer in "<s:value-of select="@target"/>" must point to an ID in
this document (<s:value-of select="$results"/>)</s:report></s:rule>
    </constraint>
   </constraintSpec>
  </attDef>
 </attList>
</classSpec>
```


Attribute limiting *@type* on <name>

```
<elementSpec ident="name" mode="change">
  <attList>
    <attDef ident="type" mode="change">
      <valList mode="add" type="closed">
        <valItem ident="person"/>
        <valItem ident="forename"/>
        <valItem ident="surname"/>
        <valItem ident="personGenName"/>
        <valItem ident="personRoleName"/>
        <valItem ident="personAddName"/>
        <valItem ident="nameLink"/>
        <valItem ident="organisation"/>
        <valItem ident="country"/>
        <valItem ident="placeGeog"/>
        <valItem ident="place"/>
      </valList>
    </attDef>
  </attList>
</elementSpec>
```

Starting to limit rendition

```
<classSpec ident="att.global" mode="change">
  <attList>
    <attDef ident="rend" mode="delete"/>
    <attDef ident="style" mode="delete"/>
    <attDef ident="rendition" mode="change">
      <valList mode="add" type="semi">
        <valItem ident="simple:bold"/>
        <valItem ident="simple:allcaps"/>
        <valItem ident="simple:italic"/>
        <valItem ident="simple:normalweight"/>
        <valItem ident="simple:smallcaps"/>
        <valItem ident="simple:doublestrikethrough"/>
        <valItem ident="simple:strikethrough"/>
        <valItem ident="simple:subscript"/>
        <valItem ident="simple:superscript"/>
        <valItem ident="simple:typewriter"/>
        <valItem ident="simple:doubleunderline"/>
        <valItem ident="simple:underline"/>
        <valItem ident="simple:wavyunderlline"/>
<!-- Among others -->
      </valList>
    </attDef>
  </attList>
</classSpec>
```


One for the geeks -- checking *@rendition*

```
<constraintSpec ident="renditionpointer"
 scheme="isoschematron">
 <constraint>
  <s:rule context="tei:*[@rendition]">
   <s:let name="results"
    value="for $val in tokenize(normalize-space(@rendition),'\s+') return
starts-with($val,'simple:') or (starts-with($val,'#') and
//tei:rendition[@xml:id=substring($val,2)])"/>
   <s:assert test="every $x in $results satisfies $x">Error: Each of the
rendition values in "<s:value-of select="@rendition"/>" must point to a
local ID or to a token in the Simple scheme
(<s:value-of select="$results"/>)</s:assert></s:rule>
 </constraint>
</constraintSpec>
```


Limiting elements to the <text>

```
<elementSpec ident="text" mode="change">
  <constraintSpec ident="headeronlyelement"
    scheme="isoschematron">
    <constraint>
      <s:rule context="tei:term|tei:editor|tei:email|tei:att|tei:gi">
        <s:report test="ancestor::tei:text">Error: The element <s:name/> is
not permitted outside the header</s:report></s:rule>
    </constraint>
  </constraintSpec>
</elementSpec>
```


A key tool: lossless transform to Simple

We expect to support a tool (currently an XSL stylesheet called `teitosimple.xsl`) which

- performs a ‘pre-flight’ check on a typical TEI file to see if it can conform
- converts some elements to simpler forms (losslessly), eg <foreName> to `<name type="foreName">`
- maps @rend values to known @rendition where possible

allowing a wider variety of texts to be exposed in Simple mode.

The SPM: Prerequisites

- choices about intended rendering preserved in the ODD file
- project-specific rendering tools generated from the ODD
- support multiple uses of markup
  - appear in rendering
  - contribute to an index or data-table
  - subscribe to a facet
  - appear on a map
- cf <equiv> to map to RDF
- cf <constraint> for providing rules about validity

Implemented as an ODD extension

- We add a <process> instruction for <elementSpec> which will define a way of processing this element
- multiple processing instructions may occur to define expected behaviour in various contexts or output formats

The <process> element attributes

@context	XPath expression defining a context in which this processing instruction is applicable
@name	name of the function from TEI Simple function library to be applied; input for the function supplied as function parameter
@mode	output mode for which this processing instruction is applicable
@class	css class or simple:class name of formatting instruction to be applied to the output
@followRendition	rendition attributes are looked at

Defaults: to be discussed....

- if no <process> for any given mode, emit textual content
- if no @context, means applies to any instance of this element
- <process> rules are additive, not alternates
- @style defaults to element name (equates to HTML @class or Word style)
- @rendition is usually ignored

SPM: first trials

```
<elementSpec ident="choice">
  <process context="not(ancestor::front) and corr and sic"
    name="makeInline(sic)" mode="render"/>
  <process context="corr and sic"
    name="makeMarginalNote(corr)" mode="render"/>
  <process context="ancestor::front and (corr and sic)"
    name="makeLinkedMarginalNote(corr,sic)" mode="render"/>
  <process context="corr and sic"
    name="makeInline(corr)" mode="textextract"/>
</elementSpec>
```

```
<elementSpec ident="speaker">
  <process name="makeInline(.)"
    mode="render"/>
</elementSpec>
```

```
<elementSpec indent="name">
  <process name="makeInline(.)"/>
  <process name="makeMarginalNote(.)"/>
</elementSpec>
```

Is SPM just reinventing the XSLT wheel?

- sort of, in the same way that Schematron does being based on XSLT and XPath
- able to be implemented in simpler languages
- potential to add features we cannot implement yet (cf Pure ODD)
- possibility of validating it in ODD context

TEI Simple: upcoming work

1. Complete customization with closed value lists for attributes
2. (November/December) Complete definition of the SPM
3. (December/January/February) Implementation of the SPM
4. (January/February) Documentation and examples
5. March) TEI Performance indicators
6. (April) Mapping to RDF

TEI Simple FAQ

Are you competing with DTA basisformat or TAPAS? So far as we can, not at all! We expect to share Schematron constraints with DTA, for example

Does Simple apply mostly to relatively modern printed books Yes.

Have you finished that list of elements and attributes? No, we expect to keep on refining that in light of feedback.

How can I tell you what I think and what I want? Raise an issue on Github, or join the Simple maillist (`https://web.maillist.ox.ac.uk/ox/info/teisimple`).

Does TEI Simple work on my phone? Only with the iPhone 6 Plus, sorry.