

Towards P5

Lou Burnard
Sebastian Rahtz
Syd Bauman
November 2003



Towards P5: overview

The next release of the TEI Guidelines has three aims:

Interoperability taking advantage of the work done by others

Expansion addressing areas as yet untamed

Internal audit cleaning up the accretions of a decade

Warning!

1. P5 will not necessarily be backward compatible with P4
2. P4 is now frozen (apart from bug fixes)



Interoperability

A lot of other people have been working in this area since 1987!

TEI P5 must fit into a joined-up digital world, along with

- ➡ W3C standards (XLink, schema, etc)
- ➡ Unicode character encoding
- ➡ Specialized markup vocabularies (MathML, SVG, DocBook, etc)
- ➡ Other metadata schemas (METS, EAD, etc)
- ➡ ISO standards for terminology and data registration



Expansion and revision

The P5 Guidelines will contain (at least) some new materials

- ☞ Manuscript description
- ☞ Multimedia and graphics
- ☞ Authoring and tag documentation

And some that is substantially revised:

- ☞ feature structures
- ☞ manuscript transcription
- ☞ character encoding
- ☞ terminology, header, linking and stand off...

And some materials may be ruthlessly excised...



Here be Dragons!

Your old files will not work with P5:

1. `<TEI.2>` is now `<TEI>`; `<teiCorpus.2>` is now `<teiCorpus>`
2. TEI elements are in the <http://www.tei-c.org/P5/> namespace
3. attribute values may have changed... `Y|N` is now `true|false`
4. ... or disappeared (Unicode issues; `lang` may become `xml:lang`; ID/IDREF may disappear)
5. Your TEI extension files will require almost certainly require re-writing—but there will be substantial payoffs



Here be Treasures!

Some new technical advantages:

1. Unicode
2. Multiple schema validation facilities
3. Classes can be implemented directly in RelaxNG
4. Better modularization tools
5. Proper integration with W3C standards (eg linking)



Internal audit

A lot of water has passed under the bridge since LB and CMSMcQ defined ODD on a Norwegian table cloth to provide a literate programming environment for the TEI.

- ➡ Remove dependencies on SGML DTD technology
- ➡ Simplify the ODD language
- ➡ Document the intended behaviour of ODD processors
- ➡ Generate DTDs and schemas
- ➡ Modularize documentation production
- ➡ Use class system more consistently
- ➡ Build new generation of ODD-aware tools



Progress report

- ➡ Content models for elements replaced by RelaxNG patterns
- ➡ Attribute values normalized where possible, linked to W3C schema datatypes where relevant
- ➡ ODD language revised, simplified and turned into a tagset
- ➡ SGML remnants (largely) removed
- ➡ Examples now part of markup, in own namespace

We can directly convert all current P4 sources to new P5 format, and generate HTML, schema and DTD for the whole of the current source: using perl and xslt.



What does a new ODD look like?

```
<tagDoc id="XREF" usage="opt">
  <Name>xref</Name>
  <equiv/>
  <gloss>extended reference</gloss>
  <classes names="CLLOC XPOINTER TERMINCL"/>
  <elementContent>
    <rng:ref name="paraContent"/>
  </elementContent>
  <desc>defines a reference to another location in the
current document, or an external document,
using an extended pointer notation,
possibly modified by additional text or
comment.</desc>
</tagDoc>
```



Attributes in an ODD

```
<attList>
  <attDef usage="req">
    <Name>sigil</Name>
    <equiv/>
    <datatype target="datatype.Text"/>
    <valDesc>the identifier to be used for this witness
or witness group in the <att>wit</att> attribute
of readings in the apparatus.</valDesc>
    <desc>indicates the sigil for one witness or for
one group of witnesses to which readings are assigned
in a critical apparatus.</desc>
    <eg/>
    <remarks>
      <p>In local encoding schemes, the value of the
<att>id</att> attribute can be used as the sigil, and
the declared value of the <att>wit</att> attribute may
be changed to IDREF, so as to ensure that only witnesses
referred to in a <gi>witness</gi> element contained
within a <gi>witList</gi> may occur in the value of
any <att>wit</att> attribute on a reading element within
an apparatus.</p>
    </remarks>
  </attDef>
```



Examples in an ODD

```
<exemplum>
  <xmleg xmlns="http://www.tei-c.org/P5/Examples/">
    <persName><foreName>Edward</foreName>
      <foreName>George</foreName><surname
type="linked">Bulwer-Lytton</surname>,
      <roleName>Baron Lytton of
      <placeName>Knebworth</placeName></roleName></persName>
    </xmlScreen>
  </exemplum>
```



A fragment of schema

```
<define name="xref">
  <element name="xref">
    <ref name="content.xref"/>
  </element>
</define>
<define name="content.xref">
  <ref name="attributes.xref"/>
  <ref name="paraContent"/>
</define>
<define name="class.loc" combine="choice">
  <ref name="xref"/>
</define>
<define name="class.terminologyInclusions"
  combine="choice">
  <ref name="xref"/>
</define>
<define name="attributes.xref"
  combine="interleave">
  <ref name="a.global"/>
  <ref name="a.xPointer"/>
  <optional>
    <attribute name="TEIform" a:defaultValue="xref">
      <text/>
    </attribute>
  </optional>
</define>
```



A simple user schema

```
<grammar ns="http://www.tei-c.org/P5/"  
  xmlns="http://relaxng.org/ns/structure/1.0"  
  datatypeLibrary=  
    "http://www.w3.org/2001/XMLSchema-datatypes">  
<include href="../../../Schema/tei.rng"/>  
<include href="../../../Schema/verse.rng"/>  
<include href="../../../Schema/figures.rng"/>  
<include href="../../../Schema/analysis.rng"/>  
<include href="../../../Schema/linking.rng"/>  
</grammar>
```



A more complex user schema

```
...  
<include href="../../../Schema/linking.rng">  
  <define name="ab"> <notAllowed/></define>  
  <define name="when"> <notAllowed/></define>  
  <define name="attributes.xref" combine="interleave">  
    <ref name="a.global"/>  
    <ref name="a.xPointer"/>  
    <optional>  
      <attribute name="url"><text/></attribute>  
    </optional>  
    <optional>  
      <attribute name="TEIform" a:defaultValue="xref">  
        <text/>  
      </attribute></optional>  
    </define>  
</include>
```



...and in compact notation

```
...  
include "../Schema/linking.rnc" {  
  ab = notAllowed  
  when = notAllowed  
  attributes.xref &=  
    a.global,  
    a.xPointer,  
    attribute url { text }?,  
    ([a:defaultValue="xref"] attribute TEIform {text}?)  
}
```

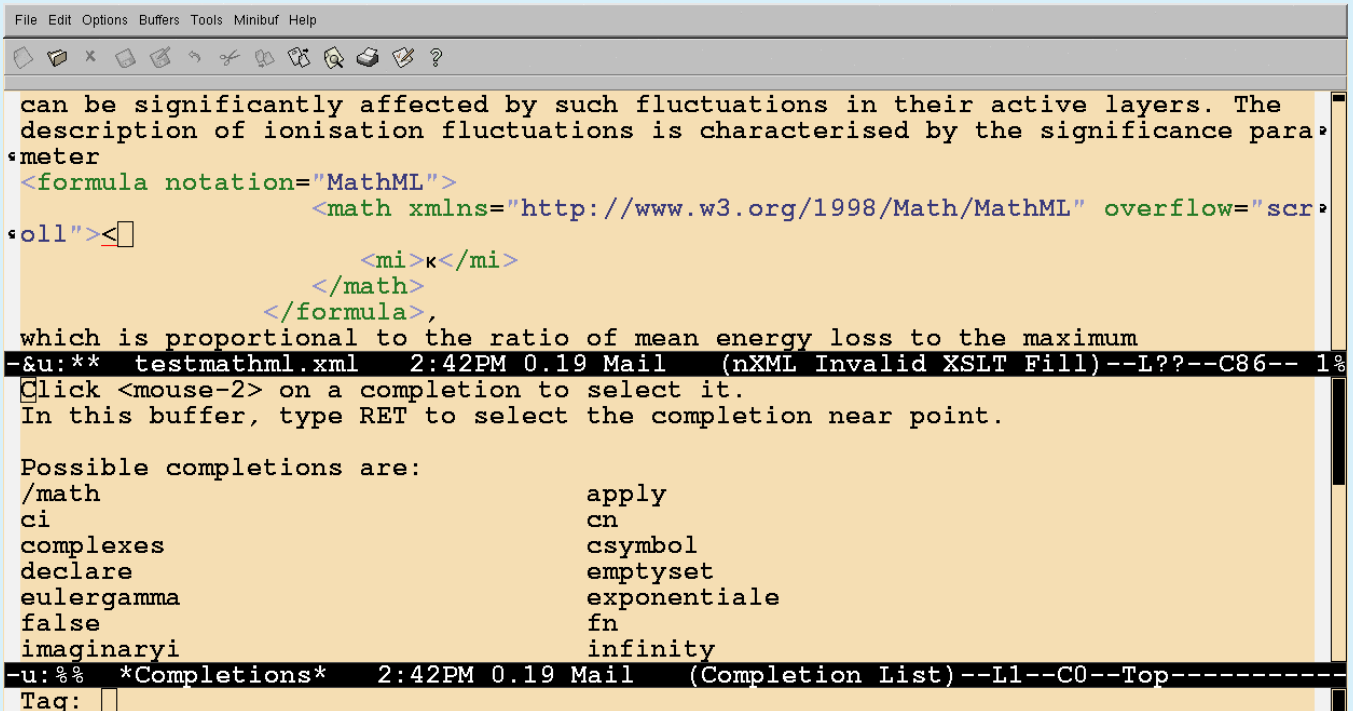


Including math?

```
...  
<include href="../../../Schema/mathml2-main.rng"/>  
  
<include href="../../../Schema/tei.rng">  
    <define name="datatype.Formula">  
        <ref name="mathml.math"/>  
    </define>  
</include>  
  
<include href="../../../Schema/figures.rng">
```



Mixed namespace editing, in Emacs



The screenshot shows the Emacs editor window with a menu bar (File, Edit, Options, Buffers, Tools, Minibuf, Help) and a toolbar. The main text area contains XML code for a formula. A completion list is displayed at the bottom, listing possible completions for the tag. The status bar at the bottom shows the current buffer, time, and other information.

```
File Edit Options Buffers Tools Minibuf Help

can be significantly affected by such fluctuations in their active layers. The
description of ionisation fluctuations is characterised by the significance para-
<meter
<formula notation="MathML">
  <math xmlns="http://www.w3.org/1998/Math/MathML" overflow="scr-
<oll"><math>
  <mi>κ</mi>
</math>
</formula>,
which is proportional to the ratio of mean energy loss to the maximum
-&u:** testmathml.xml 2:42PM 0.19 Mail (nXML Invalid XSLT Fill)--L??--C86-- 1%
Click <mouse-2> on a completion to select it.
In this buffer, type RET to select the completion near point.

Possible completions are:
/math apply
ci cn
complexes csymbol
declare emptyset
eulergamma exponentiale
false fn
imaginaryi infinity
-u:%% *Completions* 2:42PM 0.19 Mail (Completion List)--L1--C0--Top-----
Tag: <math>
```



Where from here? (1) external

1. Start editing in P5 ODD format
2. Get stuck into new manuscript chapter
3. New chapter on character encoding
4. Rewrite chapter on linking
5. Rewrite text of Guidelines to cover schema world



And where from here? (2) internal

1. Complete revision of ODDs
2. Rewrite HTML and PDF display
3. Generate W3C schemas
4. Pizza-chef replacement



Web display of guidelines, showing compact Relax syntax

File Edit View Go Bookmarks Tools Help

file:///home/rahtz/TEI/P5/Guidelines/ref-ABBR.html


Mozilla Firebird Help Mozilla Firebird Disc... Plug-in FAQ



Text Encoding Initiative

<abbr>

<abbr>	(abbreviation) contains an abbreviation of any sort.
Module	TEI core (file teicore2)
Class	class data
May contain	#PCDATA abbr add addSpan address alt altGrp anchor app c caesura cb certainty cl corr damage date dateRange dateStruct del delSpan distinct emph expan fLib foreign formula fs fsLib fvLib fw gap geogName gloss handShift hi index interp interpGrp join joinGrp lang lb link linkGrp m measure mentioned milestone name num oRef oVar orgName orig pRef pVar pb persName phr placeName ptr ref reg respons restore rs s seg sic soCalled space span spanGrp supplied term time timeRange timeStruct timeline title unclear w xptr xref
May occur within	ab abbr activity actor add addName addLine admin affiliation author authority bibl biblScope birth bloc byline camera caption case castItem catDesc cell channel cl classCode closer colloc constitution corr country creation damage date dateRange def del derivation descrip dictScrap distance distinct distributor docAuthor docDate docEdition docImprint domain edition editor education emph entryFree etym expan extent factuality figDesc firstLang foreName foreign form funder fw gen genName gloss gram gramGrp head headItem headLabel hi hyph imprimatur interaction item itype l label lang langKnown language lbl lem locale measure meeting mentioned mood name nameLink note num number occasion occupation opener orgDivn orgName orgTitle orgType orig orth otherForm p per persName phr placeName pos preparedness principal pron pubPlace publisher purpose q quote rdg re ref reg region rendition residence resp restore role roleDesc roleName rs s salute seg sense settlement sic signed soCalled soccStatus sound speaker sponsor stage street stress subc supplied surname syll symbol tagUsage tech term time timeRange title titlePart tns tr trailer trans u unclear usg view wit witDetail witness writing xr xref
Declaration	<pre>element abbr { phrase.seq, a.global, attribute expan { datatype.Text }?, attribute resp { datatype.IDref }?, attribute cert { datatype.Text }?, attribute type { datatype.Text }? }</pre>
Attributes	(In addition to global attributes and those inherited from class data) expan gives an expansion of the abbreviation. Datatype: datatype.Text resp signifies the editor or transcriber responsible for supplying the expansion of the abbreviation held as the value of the expan attribute.



Pizza to Sushi

The Pizza Chef is a front end to a DTD compiler. The user has to

- ➡ Download a pair of DTD extension files and edit them by hand
- ➡ Create extensions or changes in DTD language
- ➡ Create her own documentation

Can we do better with `roma`? This

- ➡ Works with Relax NG instead of DTD
- ➡ Can output Relax, W3C schema, and DTD
- ➡ Automates a number of common extensions
- ➡ Can generate reference documentation of selected elements



Roma stage 1

First choose which base tagsets and extra modules, and what sort of output is required:

- ☞ RelaxNG schema
- ☞ compiled RelaxNG schema
- ☞ compact RelaxNG schema
- ☞ W3C schema
- ☞ compiled DTD

You will also say if you want to

- ☞ Leave elements as they are
- ☞ Configure elements, including them by default
- ☞ Configure elements, excluding them by default




Roma stage 1, verbose interface

Roma: generating validators for the TEI - Mozilla

File Edit View Go Bookmarks Tools Window Help

http://localhost/roma.xsp

Search



Text Encoding Initiative

Roma: generating validators for the TEI

These pages will help you design your own TEI-conformant validator, as DTD, Relax NG or W3C Schema.

Base tagset

- [Prose](#) This tagset is suitable for most documents most of the time
- [Verse](#) This tagset adds specialist tagging for metrical analysis, rhyme-scheme etc to the basic verse markup already included in the core
- [Drama](#) This tagset adds specialist tagging for cast lists, records of first performance, etc. to the basic drama markup already included in the core
- [Speech](#) This tagset replaces the basic structure by one suitable for linguistic analysis of speech acts, etc.
- [Dictionaries](#) This tagset replaces the basic structure with one containing detailed lexicographic features
- [Terminology](#) This tagset replaces the basic structure with one specific to terminological databases
- [General base](#) This tagset allows you to combine tags from different base tagsets, with the proviso that any single text division can contain tags from only one of the base tagsets you choose.
- [Mixed base](#) This tagset allows you to combine tags from different base tagsets, with no restriction at all as to where tags from different base tagsets can appear.

Additional tagsets


Done



Roma stage 1, expert interface

File Edit View Go Bookmarks Tools Window Help

← → ↶ × http://localhost/roma.xsp?expert=true Search

 Text Encoding Initiative

Roma: generating validators for the TEI

Base tagset

* ☒ [Prose](#) * ☒ [Verse](#) * ☐ [Drama](#) * ☐ [Speech](#) * ☐ [Dictionaries](#) * ☐ [Terminology](#) * ☐ [General base](#) * ☐ [Mixed base](#)

Additional tagsets

* ☒ [Linking](#) * ☒ [Figures](#) * ☐ [Analysis](#) * ☐ [FS](#) * ☐ [Certainty](#) * ☐ [Transcription](#) * ☐ [Textcrit](#) * ☐ [Names & Dates](#) * ☐ [Nets](#) * ☐ [Corpora](#)

Configuring the next stage

Do you want to do selection of elements within the additional tagsets?

▼

Do you want to add new elements? ☒

Which sort of output do you want? ▼

| [TEI Home](#) | [Search](#) | [Feedback](#)


Done

Towards P5

Roma stage 2, choosing entity sets

File Edit View Go Bookmarks Tools Window Help

http://localhost/roma2.xsp?base=prose&tagsets=figures,&elements=include&output=.dtd&expert=true Search



Text Encoding Initiative

Roma: customizing elements and output

Configuring elements

Figures

	Include	Exclude	Tag Name
cell	<input type="radio"/>	<input checked="" type="radio"/>	cell
figDesc	<input checked="" type="radio"/>	<input type="radio"/>	caption
figure	<input checked="" type="radio"/>	<input type="radio"/>	graphic
formula	<input checked="" type="radio"/>	<input type="radio"/>	formula
row	<input checked="" type="radio"/>	<input type="radio"/>	row
table	<input checked="" type="radio"/>	<input type="radio"/>	table

Entity Sets

☒ [ISOlat1](#): * ☐ [ISOlat2](#): * ☐ [ISOgrk1](#): * ☐ [ISOgrk2](#): * ☐ [ISOcyr1](#): * ☐ [ISOcyr2](#): * ☐ [ISOnum](#): * ☐ [ISOdia](#): * ☐ [ISOpub](#): * ☐ [ISObox](#): * ☐ [ISotech](#): * ☐ [ISOgrk3](#): * ☐ [ISOgrk4](#): * ☐ [ISOamso](#): * ☐ [ISOamsb](#): * ☐ [ISOamsr](#): * ☐ [ISOamsn](#): * ☐ [ISOamsa](#): * ☐ [ISOamsc](#):

Additional features

☒ url attribute for <figure>, <xref> and <xptr>

☒ TEI Lite standard extensions

Done



Roma stage 2, expert mode

File Edit View Go Bookmarks Tools Window Help

http://localhost/roma2.xsp?base=&tagsets=analysis,linking,figures,&elements=include&output=.rng Search

or an external document, using an extended pointer notation, possibly modified by additional text or comment.

figures

	Include	Exclude	Tag Name	Description
cell	<input type="radio"/>	<input checked="" type="radio"/>	cell	contains one cell of a table.
figDesc	<input checked="" type="radio"/>	<input type="radio"/>	figDesc	contains a brief prose description of the appearance or content of a graphic figure, for use when documenting an image without displaying it.
figure	<input checked="" type="radio"/>	<input type="radio"/>	figure	indicates the location of a graphic, illustration, or figure.
formula	<input type="radio"/>	<input checked="" type="radio"/>	formula	contains a mathematical or other formula.
row	<input type="radio"/>	<input checked="" type="radio"/>	row	contains one row of a table.
table	<input type="radio"/>	<input checked="" type="radio"/>	table	contains text displayed in tabular form, in rows and columns.

Additional features

Enforce validation of date elements ☒

url attribute for <figure>, <xref> and <xptr> ☒

TEI Lite standard extensions ☒

MathML as content of <formula> ☐

SVG as content of <figure> ☐

Submit

[TEI Home](#) | [Search](#) | [Feedback](#)


March 2003 (revised 27/02/2003) Sebastian Rahtz (revised rahtz).
Copyright TEI Consortium 2003

Type ahead find stopped.

Roma stage 2, renaming elements

File Edit View Go Bookmarks Tools Window Help

http://localhost/roma2.xsp?base=prose&tagsets=figures,&elements=include&output=.rng&expert=true Search



Text Encoding Initiative

Roma: customizing elements and output

Configuring elements

Figures

	Include	Exclude	Tag Name
cell	<input type="radio"/>	<input type="radio"/>	cell
figDesc	<input type="radio"/>	<input type="radio"/>	caption
figure	<input type="radio"/>	<input type="radio"/>	graphic
formula	<input type="radio"/>	<input type="radio"/>	formula
row	<input type="radio"/>	<input type="radio"/>	row
table	<input type="radio"/>	<input type="radio"/>	table

Additional features

Enforce validation of date elements ☒

url attribute for <figure>, <xref> and <xptr> ☒

TEI Lite standard extensions ☐

MathML as content of <formula> ☐

SVG as content of <figure> ☐

Submit

Towards P5

General options which can be turned on and off

1. date elements to be validated against an ISO date format (Schema only)
2. `<xptr>`, `<xref>` and `<figure>` elements to support a `url`
3. TEI Lite features to be activated
4. `<formula>` element should force content to be MathML (Schema only)
5. `<figure>` element should allow SVG (Scaleable Vector Graphics) elements (Schema only)



Simple result

```
<grammar ...>
<include href="tei.rng"/>
<include href="figures.rng">
  <define name="formula"><notAllowed/></define>
  <define name="table"><notAllowed/></define>
  <define name="figDesc">
    <element name="caption">
      <ref name="c.figDesc"/>
    </element>
  </define>
  <define name="figure">
    <element name="graphic">
      <ref name="c.figure"/>
    </element>
  </define>
</include>
</grammar>
```



Element renaming

Input:

```
<define name="figure">
  <element name="figure">
    <ref name="c.figure"/>
  </element>
```

Redefinition:

```
<define name="figure">
  <element name="graphic">
    <ref name="c.figure"/>
  </element>
</define>
```

ie define an element called `<graphic>`, which has the *same content model* as the old `<figure>`, and is inside the pattern called `figure`. An attribute `TEIform` identifies the original name.



Generation of alternate outputs

1. DTD compilation performed by `carthago`
2. RelaxNG schema flattening performed by an XSLT transform
3. compact RelaxNG generated by James Clark's `trang`
4. W3C Schema generated by James Clark's `trang`


MathML and SVG inclusion are managed by simply `<include>`ing the relevant RelaxNG grammars, each in their own namespace.



Creating new elements

File Edit View Go Bookmarks Tools Window Help

http://localhost/roma3.xsp?base=prose&tagsets=linking,figures,&elements=include&output=.rng&exp Search



Text Encoding Initiative

Roma: creating new elements

New element:

- Name:
- Class: (or) Clone of:
- Description:

Add more elements? ☐

[TEI Home](#) | [Search](#) | [Feedback](#)

March 2003 (revised 27/02/2003) Sebastian Rahtz (revised rahtz).
Copyright TEI Consortium 2003

Done

Towards P5

What is missing?

- ➡ An interface to add elements with arbitrary content models
- ➡ An interface for adding and removing attributes from classes
- ➡ A method for adding new classes
- ➡ A way of generating module-specific documentation



Where have we got to (Nov 2003)? (1)

- ➡ Character encoding nearly complete
 - ➡ to go to TEI Council 1Q04
- ➡ Tag documentation (ODDs for ODD) nearly complete
 - ➡ A P4-derived P5 Schema now exists
 - ➡ Prototype now under test in Oxford
 - ➡ Available by end of 2003
- ➡ Multimedia progressing
 - ➡ Many relevant recommendations from SO WG
 - ➡ Much testing still to be done



Where have we got to (Nov 2003)? (2)

- ➡ Manuscripts started
 - ➡ WG has agreed way of unifying existing proposals
 - ➡ Area to be scoped out by SIG
 - ➡ New ODD to be drafted 1Q04
- ➡ Collaborative work with ISO TC37/SC4
 - ➡ FS Part 1 due to move to DIS in Feb 2004
 - ➡ FS Part 2 work to commence 1Q04
 - ➡ Metadata and Terminology waiting to start
- ➡ Authoring
 - ➡ Area to be scoped out by SIG
 - ➡ New ODD to be drafted 1Q04



When will it be published?

err... Council has to review... Board has to decide...

- 👉 when it's ready!
- 👉 first draft for comment and testing by end March 2004

