A complete schema definition language for the Text Encoding Initiative

Lou Burnard and Sebastian Rahtz

XML London, June 16th 2013

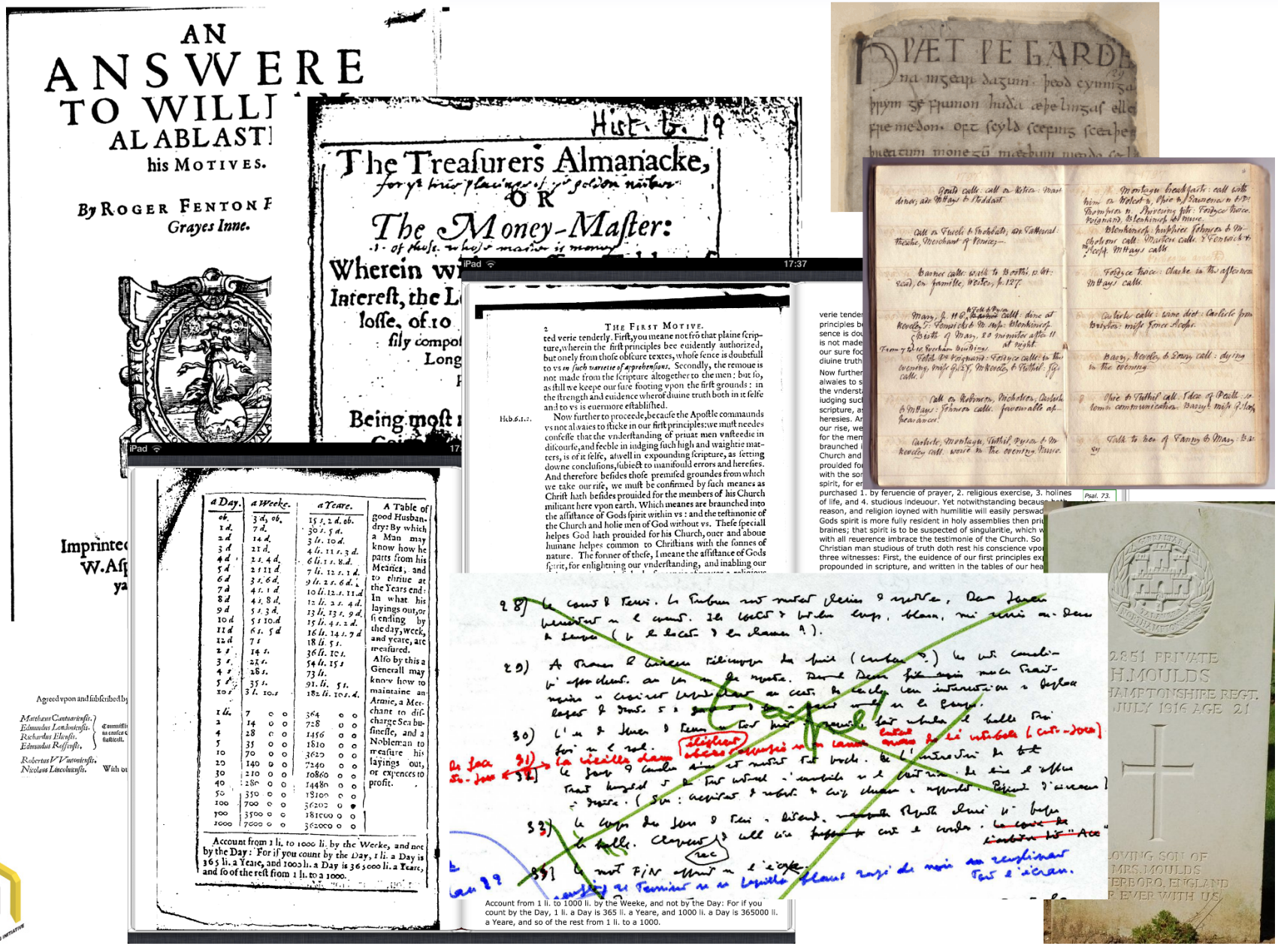
Reminder: what is the TEI?

A 25 year old project to define Guidelines for text encoding:

- mainly targetted at digital editions of existing texts
- covers manuscripts, dictionaries, transcribed text, spoken corpora, and facsimiles, as well as simple books
- governed by an international membership consortium
- defines a very rich language, with about 550 elements managed in 22 modules and an infrastructure of model and attributes classes
- Specialist vocabularies such as XInclude, MathML and SVG are used where appropriate.

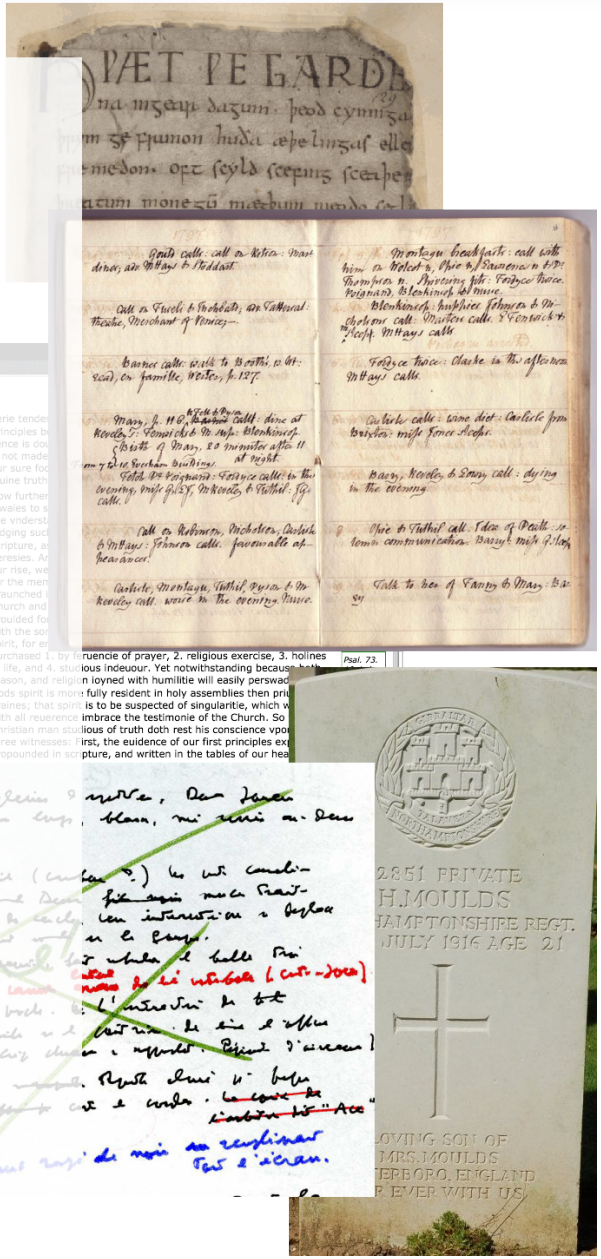
```
http://www.tei-c.org/
```

The domain of the TEI



The domain of the TEI (2)

```
AN
<line>With a thousand <del rend="stroked">fears</del>
    <del rend="stroked">
        <add place="above">
            <unclear>ways</unclear>
        </add>
    </del>
    <add place="above_above">
        <ptr target="#alt4"/>
    </add> that <del rend="stroked">creature</del>
    <add place="above">vision</add>'s face was grained;</line>
<line>Yet no blood <del rend="stroked">sumped</del>
    <add place="above">reached <del rend="stroked">him</del>
    </add>
    <del rend="stroked">here</del>
    <add place="above">there</add> from the upper ground,</line>
<line>And no <del rend="stroked">shell</del>
    <add place="above">guns</add> thumped, or down the flues made
    moan.</line>
<line>
    <del rend="stroked">But all was sleep. And no voice called for
        men.</del>
</line>
<line>“<del rend="stroked">My</del>
    <add place="inspace">Strange</add> friend,” I said, “here is no cause to
    mourn.”</line>
<line>“None”, said that other, “save the undone years,</line>
<line>The hopelessness.<add place="above">
        <del rend="stroked">unachieved.</del>
    </add>
    <add place="below">
        <del rend="stroked">
            <gap reason="unreadable"/>
        </del>
    </add> Whatever hope is yours,</line>
<line>Was my life also; <del>comrade.</del>
    <add place="above">
        <del rend="stroked">for</del>
    </add>
    <add place="below">I went</add>
    <del rend="stroked">I ran</del>
```



The TEI manifesto

1. The Guidelines are descriptive of many different ways and levels of encoding a digital text, not prescriptive
2. The Guidelines should be technology-agnostic. They currently use XML, but are prepared to change
3. The schema is modelled as independently as possible, though it currently uses RELAX NG to describe content models
4. A project is actively encouraged to develop an appropriate subset of the Guidelines, and apply domain-apppropriate constraints

The TEI is built using a literate programming system: ODD (one language does it all)

A set of TEI elements which describe

- elements and attributes
- descriptions (in multiple languages)
- examples
- content models and datatypes
- information about how it can be used
- constraints
- equivalences (eg to formal ontologies like FRBR or CIDOC CRM)

Original tagdoc for <resp> element in TEI P2 (20 years ago)

```
<tagdoc usage=rwa id="resp"><gi>resp</gi>
<name>statement of responsibility</name>
<desc>supplies information about someone other than an author,
sponsor, funder or principal researcher responsible for the
intellectual content of a text, edition, recording, or
series.</desc>
<attlist></attlist>
<exemplum><eg><![ CDATA [
  <resp><role>transcribed from original ms</role>
        <name>Claus Huitfeldt</name>
  </resp>
]]>
</eg></exemplum>
<exemplum><eg><![ CDATA [
  <resp><role>converted to SGML encoding</role>
        <name>Alan Morrison</name>
  </resp>
]]>
</eg></exemplum>
<remarks></remarks>
<part>auxiliary tag set for TEI headers</part>
<classes>
<files names='teihdr2'>
<datadesc></datadesc>
<parents>editionStmt recording seriesStmt titleStmt</parents>
<children>role name</children>
<elemdecl>
<![ CDATA [
<!ELEMENT resp          - o  ((role & name)+)                  >
]]>
</elemdecl>
<attldecl>
<![ CDATA [
<!ATTLIST resp              %a.global;                         >
]]>
</attldecl>
<xref target='hd21'>
</tagdoc>
```

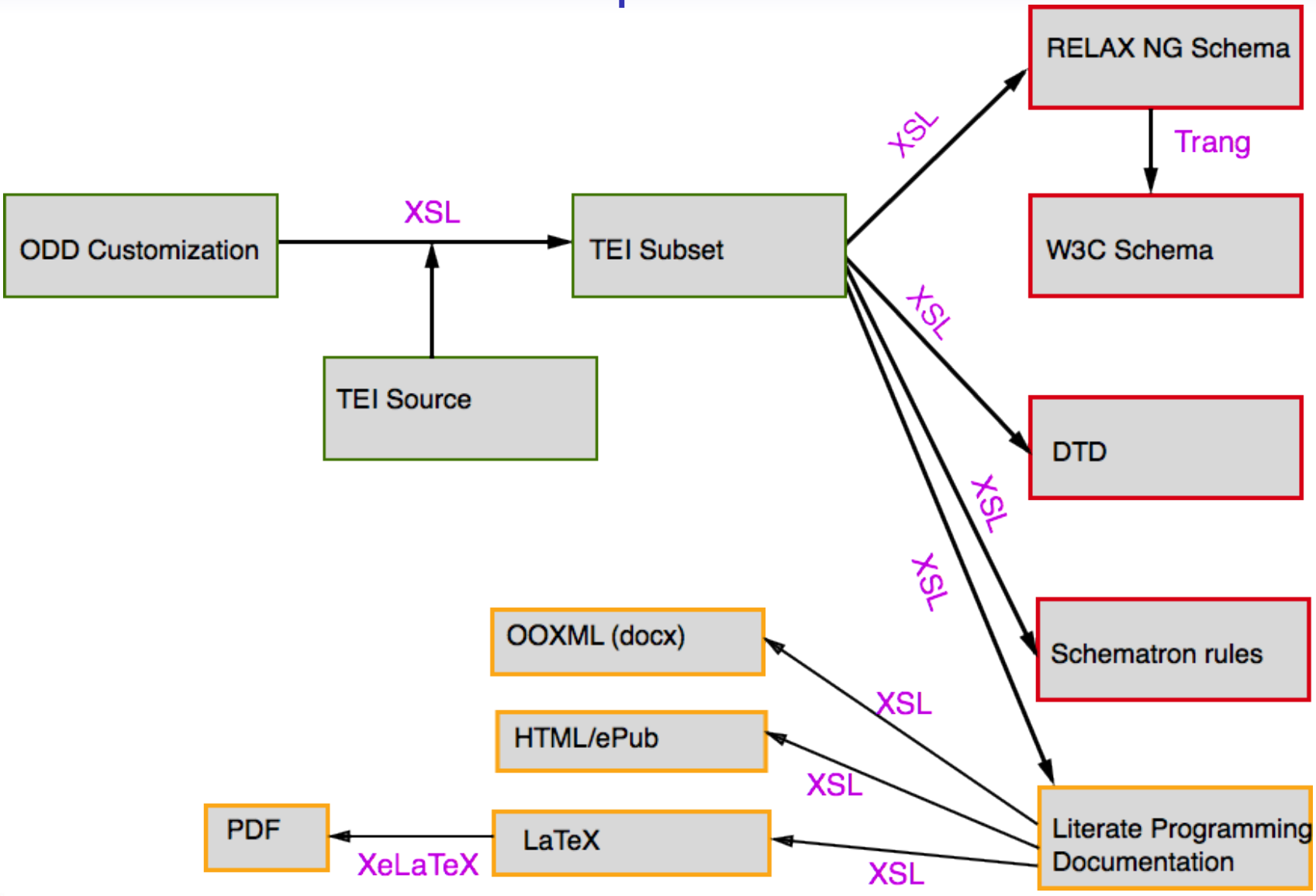
How we do ODD now

```
<elementSpec module="core" ident="respStmt">
  <gloss>statement of responsibility</gloss>
  <desc versionDate="2007-01-21" xml:lang="it">fornisce una dichiarazione di
     responsabilità per qualcuno responsabile del contenuto intelletuale di un
     testo, curatela, registrazione o collana, nel casoin cui gli elementi
     specifici per autore, curatore ecc. non sono sufficienti o non
applicabili.</desc>
  <classes>
    <memberOf key="att.global"/>
    <memberOf key="model.respLike"/>
    <memberOf key="model.recordingPart"/>
  </classes>
  <content>
    <rng:group>
      <rng:oneOrMore>
        <rng:ref name="resp"/>
      </rng:oneOrMore>
      <rng:oneOrMore>
        <rng:ref name="model.nameLike.agent"/>
      </rng:oneOrMore>
    </rng:group>
  </content>
  <exemplum versionDate="2008-04-06" xml:lang="fr">
    <egXML><respStmt>
       <resp>Nouvelle édition originale</resp>
       <persName>Geneviève Hasenohr</persName>
</respStmt>
    </egXML>
  </exemplum>
</elementSpec>
```

We use the same language to define a customization

```
<schemaSpec
   ident="myschema"
   source="http://www.tei-c.org/release/xml/tei/odd/p5subset.xml">
  <moduleRef key="tei"/>
  <moduleRef key="core"/>
  <moduleRef key="header"/>
  <moduleRef key="textstructure"/>
  <moduleRef key="namesdates" include="persName placeName"/>
  <moduleRef key="figures" except="formula"/>
  <elementSpec ident="title" mode="change">
   <attList>
    <attDef ident="type" mode="change">
     <datatype minOccurs="1" maxOccurs="unbounded">
      <rng:text/>
     </datatype>
     <valList mode="replace" type="closed">
      <valItem ident="biography"/>
      <valItem ident="chronology"/>
      <valItem ident="introduction"/>
      <valItem ident="project"/>
     </valList>
    </attDef>
   </attList>
  </elementSpec>
</schemaSpec>
```

The process


RELAX NG Schema
Trang
W3C Schema
XSL
ODD Customization
XSL
TEI Subset
XSL
DTD
XSL
Schematron rules
XSL
OOXML (docx)
XSL
HTML/ePub
XSL
Literate Programming Documentation
XSL
LaTeX
XeLaTeX
PDF
TEI Source

What's the problem?

We're neither one thing nor the other.
Currently in P5:

- Element content models are expressed using a subset of RNG
- Attribute datatypes are expressed using RNG references to W3C datatypes
- Semantic constraints are expressed using ISO Schematron rules

Why don't we just write a huge RELAX NG schema and embed TEI :umentation in it?

Choices

1. Keep on as we are
2. Rewrite everything in pure RELAX NG
3. Define the whole schema language in TEI

1. We have two ways to do things. This is a recipe for confusion
2. We would tie ourselves to one technology
3. We need to show added value from doing so

Looking at element content models

ODD must is intended to support (as far as possible) the intersection of what is possible using the current three different schema languages.

In practice, this reduces our modelling requirements quite significantly.

(It also reduces the scope of what we can model)

Requirements for our content modelling system

1. It must support alternation, repetition, and sequencing of individual elements, element classes, or sub-models (groups of elements)
2. Only one kind of mixed content model — the classic `(#PCDATA | foo | bar)*` — is permitted
3. The SGML ampersand connector — `(a & b)` as a shortcut for `((a,b) | (b,a))` is not permitted
4. A parser or validator is not required to do look ahead and consequently the model must be deterministic, that is, when applying the model to a document instance, there must be only one possible matching label in the model for each point in the document

Change 1: Define new ODD elements to represent syntax of content models

Specifically:

- <sequence> to indicate that its children form a sequence within a content model
- <alternate> to indicate that its children can be alternated within a content model

Change 2: provide new att.repeatable class of attributes

- Attributes *@minOccurs* and *@maxOccurs* are currently defined locally on the <datatype> element
- Instead provide them via a new class, to which existing elements <elementRef>, <classRef> and <macroRef> elements are added
- Default value for both *@minOccurs* and *@maxOccurs* is 1.

Change 3: re-express generic <rng:ref> elements as appropriate XML ODD elements

For example,

```
<rng:ref name="model.pLike"/>
```

becomes

```
<classRef key="model.pLike"/>
```

Example 1 — repeated alternation

`((a, (b|c)*, d+), e?)` is expressed as follows:

```
<sequence>
 <sequence>
  <elementRef key="a"/>
  <alternate minOccurs="0" maxOccurs="unlimited">
   <elementRef key="b"/>
   <elementRef key="c"/>
  </alternate>
  <elementRef key="d" maxOccurs="unlimited"/>
 </sequence>
 <elementRef key="e" minOccurs="0"/>
</sequence>
```

Example 2 — repeated sequence

((a, (b*|c*))+ is expressed as follows:

```
<sequence maxOccurs="unlimited">
 <elementRef key="a"/>
 <alternate>
  <elementRef key="b" minOccurs="0" maxOccurs="unlimited"/>
  <elementRef key="c" minOccurs="0" maxOccurs="unlimited"/>
 </alternate>
</sequence>
```

Example 3 — treatment of class references

Each class reference is understood to mean any one member of the class:

```
<sequence>
 <classRef key="model.a"/>
 <classRef key="model.b" maxOccurs="unlimited"/>
 <alternate minOccurs="0" maxOccurs="unlimited">
  <classRef key="model.c"/>
  <classRef key="model.d"/>
 </alternate>
</sequence>
```

The *@expand* attribute is used to vary this behaviour in the same way as the existing *@generate* on <classSpec>

Examples using *@expand*

Supposing that elements a and b constitute the members of class model.ab:

`<classRef key="model.ab" expand="sequence"/>` is interpreted as `a,b`

`<classRef key="model.ab" expand="sequenceOptional"/>` is interpreted as `a?,b?`

`<classRef key="model.ab" expand="sequenceRepeatable"/>` is interpreted as `a+,b+`

`<classRef key="model.ab" expand="sequenceOptionalRepeatable"/>` is interpreted as `a*,b*`

Example 4 — mixed content

A mixed content model such as `(#PCDATA | a | model.b)*` would be expressed as follows, borrowed the *@mixed* attribute from XSD:

```
<alternate minOccurs="0" maxOc-
curs="unlimited" mixed="true">
 <elementRef key="a"/>
 <classRef key="model.a"/>
</alternate>
```

New and old

```
<alternate>
  <sequence>
    <elementRef key="resp" maxOc-
curs="unbounded"/>
    <classRef key="model.nameLike.agent" ma
Occurs="unbounded"/>
  </sequence>
  <sequence>
    <classRef key="model.nameLike.agent" ma
Occurs="unbounded"/>
    <elementRef key="resp" maxOc-
curs="unbounded"/>
  </sequence>
</alternate>
```

```
<rng:choice>
  <rng:group>
    <rng:oneOrMore>
      <rng:ref name="resp"/>
    </rng:oneOrMore>
    <rng:oneOrMore>
      <rng:ref name="model.nameLike.agent"/>
    </rng:oneOrMore>
  </rng:group>
  <rng:group>
    <rng:oneOrMore>
      <rng:ref name="model.nameLike.agent"/>
    </rng:oneOrMore>
    <rng:oneOrMore>
      <rng:ref name="resp"/>
    </rng:oneOrMore>
  </rng:group>
</rng:choice>
```

The start of added value

The ability to specify repetition at the individual class level gives a further level of control not currently possible.

‘no more than two consecutive sequences of all members of the class model.nameLike’

```
<classRef key="model.nameLike" maxOccurs="2" ex-
pand="sequence"/>
```

Progress so far

1. New and changed elements defined as a TEI customization
2. Processing tools enhanced to cover the new elements
3. Conversion from old content models written and tested

```
https://github.com/TEIC/pureodd
```

A few practical issues to look at

- Investigate how to model embedded MathML and SVG (just use NVDL?)
- Develop native conversion to W3C Schema (remove dependency on trang

Going large (1)

Suppose we forget about supporting only the intersection of current schema language facilities?

- features which are present in one schema language (but not all) are probably there because someone wanted them!
- can we rethink ODD to cater for (potentially) all schema language features, rather than their intersection?
- one possible implementation strategy: use an additional constraint language such as ISO Schematron to mop up the parts that a specific schema language cannot support.
- We would like to recast some constraints which are currently in raw Schematron into pure TEI

Going large (2)

Example 1 we want a content model like `(a&b&c&d)` but only RELAX NG provides `interleave`

Example 2 we want different content models for `teiHeader//p` and for `text//p` but only W3C Schema has concept of `base types`

Going large (3)

Example 1 add an attribute *@preserveOrder* with values `true` or `false` to the <sequence> element

Example 2 add an attribute *@context* with an XPath expression as value to <elementRef> and friends

```
<elementRef key="s" context="ancestor::text" minOc-
curs="1"/>
<macroRef key="macro.limitedContent" con-
text="ancestor::teiHeader"/>
```

In the absence of an exact equivalent in the target schema language, an ODD processor can choose to

- flag the construct as illegal
- overgenerate, by producing code which validates the target construct plus others
- compensate, by over-generating but also producing Schematron code to catch 'false positives'

Autocritique

1. TEI, schmei. Just use HTML5 and stop being obfuscatory
2. you're just re-expressing RELAX in a similar language
3. who cares? validation is so 20th century
4. you're imposing a bottleneck in processing, limited by a single implementation of an under-specified idea

1. is `<span itemProp="unclear">` better than <unclear>?
2. yes, at the start. but now we can extend
3. if you want interoperability, you need a language in which to express 'business rules'
4. a fair cop, sort of. but the old system already the bottleneck, now we are simplifying it

Conclusions

Was this worth it? Yes:

- we expect a lot more human reading and changing of constraints than most schemas
- a single language to express as many constraints as possible helps our users
- we have a coherent platform on which to express more of our semantic rules
- the TEI is positioning itself to be free of XML, if alteratives appear

An extensible independent notation for expressing text encoding Guidelines takes the TEI back to its roots