An introduction to XML and the TEI

Lou Burnard Consulting
Goals of this session

1. What is a text? what is a document?
2. What is "markup"?
3. What is XML markup? Why is it so important?
The digital turn

The humanities are all about text

- (non-digital) books, manuscripts, archival papers...
- ... as well as other -- increasingly digital -- cultural manifestations such as sounds, images, blogs, tweets ...

The digital humanities are all about digital technologies and techniques for manipulating such manifestations in an integrated way

Markup (aka encoding or tagging) is one of the key technologies behind such integration.
Texts and digital texts

Texts are four dimensional:

- a document has a physical presence with visual aspects which may be transferred more or less automatically from one physical instance to another
- a text has linguistic and structural properties which may be transcribed, translated, and transmitted with some human intervention
- a text conveys information about the real world, which may be understood (or not), annotated, and even used to generate new texts
- A text usually has associated metadata, documenting what it is, where it came from, etc.

Good markup thus has to operate in all of these dimensions
A digital text may be ...

a ‘substitute’ (surrogate) simply representing the appearance of a document existant

Diary of Robert Graves 1935-39 and ancillary material

Copyright St John’s College Robert Graves Trust

July 27 Saturday

Get up at 11.30. Rose came to work at inserts of Richardson’s first chapter. donna had a talk with Carl about department.
She slept until 5 (I worked on kitchen). I went to Fabrics. I am going to order my new custom, about 10 pounds in two window (shutting shutters) fun out lights, take out away pictures.
Then worked at sonria’s life, after L. went over it. Carl brought motion, & we had coffee &c. haven’t smoked today.
I went to fabrica or worked here.
... or it may be

a representation of its linguistic content and structure, with additional annotations about its meaning and context.
What do we do when we digitize?

Resources

encoding

digital resources

analysis

abstract model
Maybe the TEI can help ...

The TEI (of which more anon) provides a well established conceptual model of text able to support:

- conversion of existing data
- creation of new datasets
- integration of existing data sets derived from a variety of sources

It is an open and non proprietary format, based on open technologies

It also expresses an explicit theory about textual ontology
EV le Personnaje, que tu joues au Speckle de toute l'Europe, voye de tout le Monde en ce grand Théatre Romain, seu tant d'affaires, et telz, que seul quasi tu soutiens: à l'Honneur du sacré Cola lége! pecheroy'-je pas (comme dit le Pindare Latin) contre le bien publicq', si par longues paroles j'empeschoy le tens, que tu donnes au service de ton Prince, au profit de la patrie, et à l'accroissement de ton immortelle renommée? Epiant donques quelque heurte de ce peu de relaiz, que tu prins pour respirer soubz le pesant faiz des affaires Françoyses (charge vrayement digne de si robustes epaules, non moins que le Ciel de celles du grand Hercule), ma Muse a pris la hardiesse d'entrer au sacré cabinet de tes Saintes et studieuses oc-

A MONSEI-
GNEVR LE REVE-
rendissime Cardinal
du Bellay.
S.

A MONSEIGNEUR

Le Peverrendissime Cardinal du Bellay, S.

Veu le personnage que tu joues au spectacle de toute l'Europe, voye de tout le monde, en ce grand theatre romain; veu tant d'affaires et telz, que seul quasi tu soutiens: à l'honneur du sacré College! pecheroy'-je pas (comme dit le Pindare latin) contre le bien publicq', si par longues paroles j'empeschoy le tens que tu donnes au service de ton Prince, au profit de la patrie, et à l'accroissement de ton immortelle renommée? Epian donques quelque heurte de ce peu de relaiz, que tu prins pour respirer soubz le pesant faiz des affaires Françoyses (charge vrayement digne de si robustes epaules, non moins que le Ciel de celle du grand Hercule), ma Muse a pris la hardiesse d'entrer au sacré cabinet de tes saintes et studieuses oc-
How about these?

A MONSEIGNEUR LE REV. ET ÉRENDISSIME CARDINAL DU BELLAY.

S.

EV le Personnage que tu joues au Spectacle de toute l'Europe, voyre de tout le Monde en ce Grand Théâtre Romain, veu tant d'affaires, et telz que seul quasiz tu soutiens à l'Honneur du Sacré Calice, pecheroy' je pas (comme dit le Pindare Latin) contre le bien public, si par longues paroles l'empeschoy le sens que tu donnes au service de ton Prince, au profit de la Patrie, et à l'accroissement de ton immortelle renommée? Épient doncques quelque heure de ce peu de repos, que tu prens pour respirer, souuz le pesant faiz des affaires François (charge vraiment digne de si robustes epaules, non moins que le Ciel de celles du grand Hercule) ma Muse a pris la hardiesse d'écrire au sacré Cabinet de ses sainctes, et studieuses occupations, et entre tant a fuy de

Joachim du Bellay

Défense et illustration de la langue françoys (1549)

La Défense, et Illustration de la Langue Fransoyse

L'auteur prie les lecteurs differer leur jugement jusques à la fin du livre, et ne le condamner sans avoir premièremen bien vu, et examiné ses raisons.

Épitre à Monseigneur le rév. et érendissime cardinal du Bellay S.

Vu le personnage que tu joues au spectacle de toute l'Europe, voyre de tout le monde, en ce grand Théâtre Romain, vu tant d' affaires, et tels que seul quasiz tu soutiens à l'honneur du sacré Collège, pécheray je pas (comme dit le Pindare Latin) contre le bien public, si par longues paroles l'empeschay le temps que tu donnes au service de ton prince, au profit de la patrie, et à l'accroissement de ton immortelle renommée? Épient doncques quelques heures de ce peu de repos que tu prends pour respirer sous le pesant faiz des affaires françaises (charge vraiament digne de si robustes epaules, non moins que le ciel de celles du grand Hercule), ma Muse a pris la hardiesse d'écrire au sacré cabinet de ses saintes et studieuses occupations, et entre tant de riches et excellents voeux de jour en jour dédiés à l'image de ta grandeur, prendre le sien humble et petit, mais toujours bien heureux s'il rencontre quelque faveur devant les yeux de ta bonté, semblable à celles des Dieux immortels, qui n'ont moins agréables les pauvres présents d'un bien riche vouloir que les superbes et ambitionnes espresses.
A text is not a document

What is the essential part of a text?

- the shape of letters and their layout?
- the original from which this copy derives?
- the stories we read into it? its author's intentions?

A "document" is something that exists in the world, which we can digitize.

A "text" is an abstraction, created by or for a community of readers, which we can markup.
A text is not a document

What is the essential part of a text?

- the shape of letters and their layout?
- the original from which this copy derives?
- the stories we read into it? its author's intentions?

A "document" is something that exists in the world, which we can digitize.

A "text" is an abstraction, created by or for a community of readers, which we can markup.
A text is not a document

What is the essential part of a text?

- the shape of letters and their layout?
- the original from which this copy derives?
- the stories we read into it? its author's intentions?

A "document" is something that exists in the world, which we can digitize.

A "text" is an abstraction, created by or for a community of readers, which we can markup.
A text is not a document

What is the essential part of a text?

- the shape of letters and their layout?
- the original from which this copy derives?
- the stories we read into it? its author's intentions?

A "document" is something that exists in the world, which we can digitize.

A "text" is an abstraction, created by or for a community of readers, which we can markup.
A text is more than a sequence of encoded glyphs or characters!
A text is more than a sequence of linguistic forms!
- It has a structure and a communicative function
- It also has multiple possible readings
- Its meaning can be enriched by annotation

Markup makes these things explicit

Only that which is explicit can be reliably found and displayed, counted, analysed...
The Babel effect

Naturally, there are many possible readings of most texts ...

I

Loomings

Call me Ishmael. Some years ago – never mind how long precisely – having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the ... and (alas) many different ways of expressing each reading!
Encodings

- Good news: there is software capable of converting amongst hundreds of different formats
- Bad news: we still need it
Good news: there is software capable of converting amongst hundreds of different formats

Bad news: we still need it
Encodings

- Good news: there is software capable of converting amongst hundreds of different formats
- Bad news: we still need it
Information exchange (1)

without a standard interchange format: 20 translations are needed for 5 formats (n*n-n)
Information exchange (2)

with a standard interchange format: 10 translations are needed for $n^2$ formats ($n^*2$)
What's the markup for?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

we don't have to be limited to the view of one editor or consumer
What's the markup for?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

we don't have to be limited to the view of one editor or consumer
What's the markup for?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

we don't have to be limited to the view of one editor or consumer
What's the markup for?

- To make explicit (to a machine) what is implicit (to a person)
- To add value by supplying multiple annotations
- To facilitate re-use of the same material
  - in different formats
  - in different contexts
  - by different users

we don't have to be limited to the view of one editor or consumer
Styles of markup

- In the beginning there was procedural markup
  RED INK ON; print balance; RED INK OFF
- which naturally led to thinking of markup as primarily presentational
  <redInk>some numbers</redInk>
- but maybe descriptive markup is preferable?
  <balance type='overdrawn'>some numbers</balance>
- also known as encoding or annotation

Isn't it more useful to markup what we think things are than what they look like?
Markup is a scholarly activity

- The application of markup to a document is not an automatic process.
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor.
- There is (almost) no such thing as neutral markup -- all of it involves interpretation.
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself.
- Good textual encoding is never as easy or quick as people would believe.
- Detailed document analysis is needed before encoding for the resulting markup to be useful.
Markup is a scholarly activity

- The application of markup to a document is not an automatic process
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor
- There is (almost) no such thing as neutral markup -- all of it involves interpretation
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself
- Good textual encoding is never as easy or quick as people would believe
- Detailed document analysis is needed before encoding for the resulting markup to be useful
Markup is a scholarly activity

- The application of markup to a document is not an automatic process.
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor.
- There is (almost) no such thing as neutral markup -- all of it involves interpretation.
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself.
- Good textual encoding is never as easy or quick as people would believe.
- Detailed document analysis is needed before encoding for the resulting markup to be useful.
Markup is a scholarly activity

- The application of markup to a document is not an automatic process
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor
- There is (almost) no such thing as neutral markup -- all of it involves interpretation
  - Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself
  - Good textual encoding is never as easy or quick as people would believe
  - Detailed document analysis is needed before encoding for the resulting markup to be useful
Markup is a scholarly activity

- The application of markup to a document is not an automatic process.
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor.
- There is (almost) no such thing as neutral markup -- all of it involves interpretation.
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself.
- Good textual encoding is never as easy or quick as people would believe.
- Detailed document analysis is needed before encoding for the resulting markup to be useful.
Markup is a scholarly activity

- The application of markup to a document is not an automatic process.
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor.
- There is (almost) no such thing as neutral markup -- all of it involves interpretation.
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself.
- Good textual encoding is never as easy or quick as people would believe.
- Detailed document analysis is needed before encoding for the resulting markup to be useful.
Markup is a scholarly activity

- The application of markup to a document is not an automatic process
- In deciding what markup to apply, and how this represents the original, one is undertaking the task of an editor
- There is (almost) no such thing as neutral markup -- all of it involves interpretation
- Markup can assist in answering research questions, and the deciding what markup is needed to enable such questions to be answered can be a research activity in itself
- Good textual encoding is never as easy or quick as people would believe
- Detailed document analysis is needed before encoding for the resulting markup to be useful
Separation of form and content

- Presentational markup cares more about fonts and layout than meaning.
- Descriptive markup says what things are, and leaves the rendition of them for a separate step.
- Separating the form of something from its content makes it much easier to re-use.
- It also allows easy changes of presentation across a large number of documents.
Separation of form and content

- Presentational markup cares more about fonts and layout than meaning
- Descriptive markup says what things are, and leaves the rendition of them for a separate step
- Separating the form of something from its content makes it much easier to re-use
- It also allows easy changes of presentation across a large number of documents
Separation of form and content

- Presentational markup cares more about fonts and layout than meaning.
- Descriptive markup says what things are, and leaves the rendition of them for a separate step.
- Separating the form of something from its content makes it much easier to re-use.
- It also allows easy changes of presentation across a large number of documents.
Separation of form and content

- Presentational markup cares more about fonts and layout than meaning.
- Descriptive markup says what things are, and leaves the rendition of them for a separate step.
- Separating the form of something from its content makes it much easier to re-use.
- It also allows easy changes of presentation across a large number of documents.
Separation of form and content

- Presentational markup cares more about fonts and layout than meaning.
- Descriptive markup says what things are, and leaves the rendition of them for a separate step.
- Separating the form of something from its content makes it much easier to re-use.
- It also allows easy changes of presentation across a large number of documents.
What then shall we markup?

Compare ...

```xml
<pb n="4"/>A MONSEI- <lb/>GNEUR LE REVE-
<lb/>rendissime Cardinal <lb/>du Bellay. <lb/>S
<lb/>
<c rend="lettrine">V</c>EU le
Personnaige, <lb/>que tu joues au Spec- <lb/>tacle de toute
l'Europe...
```

```xml
<div type="dedicace">
  <head>A MONSEIGNEUR LE REVERENDISSIME CARDINAL DU
  BELLAY</head>
  <salute>S<ex>ire</ex></salute>
  </salute>
  <p>
    <c rend="lettrine">V</c>EU le Personnaige, que tu joues
    au Spectacle de toute
    l'Europe... </p>...
</div>
```
... and

```xml
<pb n="4"/>
<s>
  <w pos="PPJ" lemma="voir">VEU</w>
  <w pos="ART" lemma="le">le</w>
  <w pos="SBC" lemma="personnage">Personnage</w>
  <pc>,</pc>
  <w pos="COO" lemma="que">que</w> ...
</s>
```

or even

```xml
<s>
  <choice>
    <reg>Vu</reg>
    <orig>Veu</orig>
  </choice> le
  <choice>
    <reg>Personnage</reg>
    <orig>Personnage</orig>
  </choice>, que tu joues au Spectacle...
</s>
```
Hwæt! we Gar-dena in gear-dagum þeod-cyninga þrym gefrunon, 
þeða æþelingas ellen fremedon. oft scyld scefing sceæpe

Oft Scyld Scefing sceæpæna þreatum, 
monegum mægbam meodo-setla ofteah;
egsode Eorle, syððan ærest wearþ
feasceæft funden...
The markup language we use must be able to:

- specify all the characters found
- make explicit the structures perceived
- represent that structure in a linear processable form
- additionally supply a variety of metadata or annotations

XML is a good fit... mostly
Some alphabet soup

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGML</td>
<td>Standard Generalized Markup Language</td>
</tr>
<tr>
<td>HTML</td>
<td>Hypertext Markup Language</td>
</tr>
<tr>
<td>W3C</td>
<td>World Wide Web Consortium</td>
</tr>
<tr>
<td>XML</td>
<td>eXtensible Markup Language</td>
</tr>
<tr>
<td>DTD</td>
<td>Document Type Definition (or Declaration)</td>
</tr>
<tr>
<td>CSS</td>
<td>Cascading Style Sheet</td>
</tr>
<tr>
<td>Xpath</td>
<td>XML Path Language</td>
</tr>
<tr>
<td>XSLT</td>
<td>eXtensible Stylesheet Language - Transformations</td>
</tr>
<tr>
<td>XQuery</td>
<td>XML Querying</td>
</tr>
<tr>
<td>RELAXNG</td>
<td>Regular Expression Language for XML (New Generation)</td>
</tr>
</tbody>
</table>

Oh, and then there's also **TEI**, the Text Encoding Initiative
XML: what it is and why you should care

- XML is **structured data** represented as strings of text
- XML looks like HTML, except that:-
  - XML is **extensible**
  - XML must be **well-formed**
  - XML can be **validated**
- XML is application-, platform-, and vendor- independent
- XML empowers the **content provider** and facilitates data integration
Nearly all there is to know about XML, on one slide

- An XML document contains at least one element
- An element is represented by a start-tag, some optional content, and an end-tag
- The content of an element can be a string of Unicode characters, or one or more other elements
- An element may also have attributes, each consisting of a name and a value
- An XML document may also contain processing instructions, comments or entity references

An XML document must be well-formed and may be valid
For example

```xml
<?xml version="1.0" ?>
<root>
  <element attribute="value"> content 
  </element>
<!-- comment -->
</root>
```
Parts of an XML document

- The XML declaration
- Namespace declarations
- The root element of the document itself
- Other elements and content
- Attribute and value

```xml
<?xml version="1.0"?>
<greetings xmlns="http://www.example.org/greetings">
  <hello type="sarcastic">hello world!</hello>
</greetings>
```
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

• An XML document represents a (kind of) tree
• It has a single root and many nodes
  • Each node can be
    • a subtree
    • a single element (possibly bearing some attributes)
    • a string of character data
• Each element has a name or generic identifier
• XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
    - a single element (possibly bearing some attributes)
    - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
The rules of the XML Game

- An XML document represents a (kind of) tree
- It has a single root and many nodes
- Each node can be
  - a subtree
  - a single element (possibly bearing some attributes)
  - a string of character data
- Each element has a name or generic identifier
- XML elements and attributes are case sensitive
Representing an XML tree

- An XML document is encoded as a linear string of characters.
- It begins with a special processing instruction.
- Element occurrences are marked by start- and end-tags.
- The characters `<` and `&` are Magic and must always be "escaped" if you want to use them as themselves.
- **Comments** are delimited by `<!--` and `-->`.
- Attribute name=value pairs are supplied on the start-tag and may be given in any order, separated by spaces.
- Entity references (eg `&lt;` for `<` or `&#199;` for Ç) are delimited by `&` and `;`.
- **CDATA sections** are delimited by `<![CDATA[ and ]]>` (but you probably don't need these).
The XML declaration

An XML document must begin with an XML declaration which does three things:

- specifies that this is an XML document
- specifies which version of the XML standard it follows
- specifies which character encoding the document uses

```xml
<?xml version="1.0" ?>
<?xml version="1.0" encoding="iso-8859-1" ?>
```

The default, and recommended, encoding is ‘UTF-8’ (Unicode)
Namespaces

- A namespace is a label, declared with URL syntax, used to identify a group of XML element names and distinguish them from others.
- An XML document can include elements from many different namespaces. All TEI documents show that by default their component elements belong to the TEI namespace by beginning with the following declaration:
  `<TEI xmlns="http://www.tei-c.org/ns/1.0"> ... </TEI>`
- The `xml namespace` is also used by the TEI for global attributes `@xml:id` and `@xml:lang`
- Other namespaces can appear in a TEI document: for example, MathML:

  ```xml
  <TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:math="http://www.mathml.org">
    <p>...<math:expr>...</math:expr>...</p>...
  </TEI>
  ```
XML syntax: the small print

What does it mean to be well-formed?

1. There is a single root node containing the whole of an XML document
2. Each subtree is properly nested within the root node
3. Element/attribute/etc. names are always case sensitive
4. Start-tags and end-tags are always mandatory (except there is a combined start-and-end tag `<pb/>`)
5. Attribute values are always quoted

Note: You can be valid in addition to being well-formed. This means you obey the rules of a specified schema, such as the TEI.
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

- Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

- Which are correct?
  - `<seg>some text</seg>`
  - `<seg> <foo>some</foo> <bar>text</bar> </seg>`
  - `<seg> <foo>some <bar></foo> text</bar> </seg>`
  - `<seg type="text">some text</seg>`
  - `<seg type='text'>some text</seg>`
  - `<seg type=text>some text</seg>`
  - `<seg type=text>some text</seg>`
  - `<seg type="text"> some text <seg/>`
  - `<seg type="text"> some text<gap/> </seg>`
  - `<seg type="text">some text</Seg>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text"> some text <seg/>`
- `<seg type="text"> some text<gap/>`
- `<seg type="text">some text</Seg>`
Test your XML knowledge

Which are correct?

- `<seg>some text</seg>`
- `<seg> <foo>some</foo> <bar>text</bar> </seg>`
- `<seg> <foo>some <bar></foo> text</bar> </seg>`
- `<seg type="text">some text</seg>`
- `<seg type='text'>some text</seg>`
- `<seg type=text>some text</seg>`
- `<seg type="text">some text <seg/`>`
- `<seg type="text">some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`
XML is an international standard

- XML requires use of ISO 10646 (also known as Unicode)
  - a 31 bit character repertoire including most human writing systems
  - encoded as UTF8 or UTF16

- other encodings may be specified at the document level

- language may be specified at the element level using @xml:lang to supply values taken from another ISO standard (ISO 639)

Characters or glyphs not defined in Unicode must therefore be represented using markup
XML validation

A valid XML document is well formed, and also conforms to some additional structural rules, which make up what we call a schema. A schema allows you to specify:

- which elements can appear as the root element of a document
- which elements and attributes can appear where
- names, datatypes, and default values for all attributes

A schema thus allows you to check that ‘every chapter must begin with a heading’ or ‘recipes must include an ingredient list’ or ‘values for the attribute @when are all valid dates’

A namespace by contrast just allows you to label elements as being somehow related. A schema provides a formal specification and is written in a formal language.
Schema languages

A schema can be expressed using:

- DTD: an older ISO standard
- RNG: ISO standard RELAXNG
- WSD: the W3C Schema language

All have different tool kits, different syntaxes, and different methods of doing things, particularly for content validation.

The TEI uses RELAXNG