Talk 1: Introduction to Markup, XML, and the TEI Consortium

James Cummings

28 January 2014
Introduction to the Workshop

Aims of Course

1. Introduce the concept of markup and XML encoding
2. Provide hands-on experience in using TEI XML markup
3. Intensively survey the basics of the TEI scheme, its assumptions, and its organization
4. Survey the landscape of the TEI recommendations, including areas like:
   - Markup, XML, and TEI core elements
   - TEI metadata, manuscript: description/transcription/editing
   - Names, people, places, organizations, and events
   - Performance texts, verse, linguistic analysis
   - Customisation of the TEI, and tools for transformation and interrogation
5. Provide routes into more detailed information for exploration at your leisure

@jamecummings – James.Cummings@it.ox.ac.uk – http://tei.it.ox.ac.uk/Talks/2014-01-toronto/
## Workshop Structure

The timings of the course aren’t written in stone: we can go faster or slower depending on you. The intent is to give you a good exposure to the TEI with some practical encoding experience.

<table>
<thead>
<tr>
<th>Time</th>
<th>Tuesday 28 January</th>
<th>Wednesday 29 January</th>
<th>Thursday 30 January</th>
<th>Friday 31 January</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30 – 11:00 (1hr 30mins)</td>
<td>Talk 1: Introduction to Markup, XML, and the TEI Consortium</td>
<td>Talk 5: Document Metadata and the TEI Header</td>
<td>Talk 9: Transcription of Primary Source Documents</td>
<td>Talk 13: Customising and Documenting Your TEI project</td>
</tr>
<tr>
<td>11:00 – 11:45</td>
<td>Exercise 1: Basic Markup &amp; the oXygen XML Editor</td>
<td>Exercise 3: Improving the TEI Header</td>
<td>Exercise 5: Basic Transcription</td>
<td>Exercise 7: Customising the TEI</td>
</tr>
<tr>
<td>11:45 – 13:00 (1hr 15mins)</td>
<td>Talk 2: Basic TEI Structure</td>
<td>Talk 6: Manuscript Description</td>
<td>Talk 10: Names, Dates, People, Places, and Organisations</td>
<td>Talk 14: The TEI Consortium, Guidelines, and Community</td>
</tr>
<tr>
<td>13:00 – 14:00</td>
<td>Lunch</td>
<td>Lunch</td>
<td>Lunch</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00 – 14:45</td>
<td>Exercise 2: Making a valid TEI Document</td>
<td>Exercise 4: Making a Manuscript Description</td>
<td>Exercise 6: Encoding Names and People</td>
<td>Talk to English Students (14:00 – 15:30)</td>
</tr>
<tr>
<td>14:45 – 16:00 (1hr 15mins)</td>
<td>Talk 3: Core Elements of the TEI Guidelines</td>
<td>Talk 7: Critical Apparatus, Facsimiles, and Genetic Editing</td>
<td>Talk 11: Interrogating XML Documents</td>
<td></td>
</tr>
<tr>
<td>16:00 – 16:30 (30 mins)</td>
<td>Talk 4: Performance Texts and Verse</td>
<td>Talk 8: Encoding Linguistic Analysis</td>
<td>Talk 12: Publishing TEI</td>
<td>Discuss REED Markup &amp; Future Work</td>
</tr>
<tr>
<td>16:30 – 17:00 (30 mins)</td>
<td>Discussion / Questions</td>
<td>Discussion / Questions</td>
<td>Discussion / Questions</td>
<td></td>
</tr>
</tbody>
</table>
Workshop Materials

- All course materials including:
  - All slides from lectures
  - All exercises
  - All materials for the exercises

are available at: http://tei.it.ox.ac.uk/Talks/

- All the slides, exercises, and some materials are licensed with a Creative Commons Attribution license, which means they are freely available for re-use (though do let us know!)

- You are not meant to memorise all the elements or topics covered, but be introduced to them so you can read the TEI Guidelines with more confidence later!
After the workshop...

- After the workshop, if you have questions about:
  - The workshop materials or teaching other workshops: tei@it.ox.ac.uk
  - The TEI generally by joining: TEI-L@listserv.brown.edu

If you mail the TEI-L mailing list it is better because:
- we’ll still try to answer as well as we would privately
- you get answers not only from us, but TEI experts around the world
- questions from those of all levels of ability stop the list becoming too technical
- everyone benefits from having the answers be public – and you benefit by reading (and sometimes answering!) others’ problems
Textual Markup

Some of you are already familiar with markup and XML, so we might be able to go faster here! But, in order to talk about texts, markup and encoding of texts, we need to understand what we mean by these basic concepts. When we talk about text encoding, what do we mean by a text? What is in a text and what assumptions do we make in reading them?
What's in a text?

ACT ONE

Scene I. On a ship at sea; a tempestuous noise of thunder and lightning heard.

Enter a Shipmaster and a Boatswain.

Master. Boatswain!

Boats. Here, master; what cheer?

Master. Good! Speak to th’ mariners; fall to ’t yarely, or we run ourselves aground; bestir, bestir. [Exit.

Enter Mariners.

Boats. Heigh, my hearts! cheerly, cheerly, my hearts! yare, yare! Take in the topsail. Tend to th’ master’s whistle. Blow till thou burst thy wind, if room enough.

Enter Alonso, Sebastian, Antonio, Ferdinand, Gonzalo, and Others.

Alon. Good boatswain, have care. Where’s the master? Play the men.

Boats. I pray now, keep below.

Ant. Where is the master, boson?

Boats. Do you not hear him? You mar our labour; keep your cabins; you do

Cheerly, good hearts!—Out of our way, I say. [Exit.

Gon. I have great comfort from this fellow. Methinks he hath no drowning mark upon him; his complexion is perfect gallows. Stand fast, good Fate, to his hanging; make the rope of his destiny our cable, for our own doth little advantage. If he be not born to be hang’d, our case is miserable. [Exeunt.

Re-enter Boatswain.

Boats. Down with the topmast. Yare, lower, lower! Bring her to try wi’ th’ main-course. [A cry within] A plague upon this howling! They are louder than the weather or our office.

Re-enter Sebastian, Antonio, and Gonzalo.

Yet again! What do you here? Shall we give o’er, and drown? Have you a mind to sink?

Seb. A pox o’ your throat, you bawling, blasphemous, incharitable dog!

Boats. Work you, then.

Ant. Hang, cur; hang, you whoreson, insolent wretch! You art afraid to be
What’s in a text?

æt pe garde

na ingen dazum. þæod cynninga
þum se fænon hūda æþelingsæl eða
premodon. oft byld sceaping se æþena
hæææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææææ
What’s in a text?

Hwæt wē Gār-Dena in gēār-dagum
þēod-cyninga þrým gefrūnon,
hū ðā æþelingas ellen fremedon.

Oft Scyld Scēfing sceahēna þrēatum,
5 monegum mēgþum meodo-setla oftēah;
egsode Eorl[e], syððan ærest wearð
fēasceafaft funden; hē þæs frōfre gebād:
wēox under wolcnum, weorð-myndum þāh,
oðbæt him Æghwylc þāra ymb-sittendra
10 ofer hron-rāde hýran scolde,
What’s in a text?

This book is not about heroes. English Poetry is not yet fit to speak of them.
No is it about battle, and glory of battle, or war.

The Poet is not yet fit to speak of battle, or war.

The Poet is not about glory, might, empire, dominion, or power.

The Poet is not about battle; it is not about glory, might, empire, dominion, or power.

The Poet is not about war.

Above all the Poet is not concerned with War.

Rhythm is the subject of War, and the poetry of War.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war.

The Poet is not about war. 
What’s in a text?

PREFACE

This book is not about heroes. English poetry is not yet fit to speak of them.

Nor is it about deeds, or lands, nor anything about glory, honour, might, majesty, dominion, or power, except War.
Above all I am not concerned with Poetry.
My subject is War, and the pity of War.
The Poetry is in the pity.

Yet these elegies are to this generation in no sense consolatory. They may be to the next. All a poet can do today is warn. That is why the true Poets must be truthful.

(If I thought the letter of this book would last, I might have used proper names; but if the spirit of it survives—survives Prussia—my ambition and those names will have achieved fresher fields than Flanders. . . .)

At Ripon, probably in May 1918, WO began this draft Preface for a collection of war poems that he hoped to publish in 1919 (WO, 265–6).

1 Cp. the General Epistle of Jude 25, ‘To the only wise God our Saviour, be glory and majesty, dominion and power, both now and ever.’
2 WO had considered — and decided against — calling his book English Elegies.

CP&F, 535
A MONSEIGNEUR

Le Rererev. Erard Cardinal du Bellay, S.

VEU le Personnage, que tu joues au Spectacle de toute l'Europe, voyez de tout le Monde en ce grand Théatre Roman, veu tant d'affaires et telz, que seul quasi tu soutiens à l'Honneur du Sacré Collège! 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? Epliant donques quelle heure de ce peu de relaiz, que tu prens pour respirer souz le pesant faiz des affaires francoyses (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 occu-

@jamescummings – James.Cummings@it.ox.ac.uk – http://tei.it.ox.ac.uk/Talks/2014-01-toronto/
A MONSEIGNEUR

Le Reuereendissime 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 Joachim du Bellay

Défense et illustration de la langue françoysse (1549)

La Défence, et Illustration de la Langue Francoysse

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

Épitre à Monseigneur le réverendissime cardinal du Bellay S.

Vu le personnage que tu joues au spectacle de toute l'Europe, voye de tout le monde, en ce grand Théâtre Romain, veu tant d'affaires, et tels que seul qui que tu soutiens à l'Honneur du sacré Collège, pècheras pas (comme dit le Pindare Latin) contre le bien public, fi par longues paroles l'empêchey le tens, 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 relaiz, que tu prens pour respirer foubz le pesant faiz des affaires francoysses (chargeurayment digne de si robustes epaules, non moins que le Ciel de celles du grand Hercule) ma Mœe a priz la hardiehe d'entrer au sacré Cabinet de ces saintes, et studieuses occupations: et la entre tant a iij de
A text is not a document

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in 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 encode.
A text is not a document

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in 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 encode.
A text is not a document

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in 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 encode.
A text is not a document

Where is the text?

- in the shape of letters and their layout?
- in the original from which this copy derives?
- in the stories we read into it? or in 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 encode.
Encoding of texts

- A text is more than a sequence of encoded glyphs or lexical tokens
  - It has a *structure* and a *communicative function*
  - It also has multiple possible *readings*

- Encoding, or markup, is a way of making all these things explicit

Only that which is explicit can be reliably processed
What’s the point of markup?

- 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
What’s the point of markup?

- 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
What’s the point of markup?

- 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
What’s the point of markup?

- 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
In the beginning there was *procedural* markup
RED INK ON; print balance; RED INK OFF

which being generalised became *descriptive* markup

<balance type='overdrawn'>some numbers</balance>

also known as *encoding* or *annotation*

descriptive markup allows for easier re-use of data
Some more definitions

- Markup makes explicit the distinctions we want to make when processing a string of bytes.
- Markup is a way of naming and characterizing the parts of a text in a formalized way.
- It’s (usually) more useful to markup what we think things are than what they look like.
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 its re-use more flexible
- 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 its re-use more flexible
- 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 its re-use more flexible
- 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 its re-use more flexible
- 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 its re-use more flexible.
- It also allows easy changes of presentation across a large number of documents.
Introduction to the Workshop

Textual Markup

XML

TEI

Markup as a scholarly activity

- The application of markup to a document can be an intellectual activity
- 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 as a scholarly activity

- The application of markup to a document can be an intellectual activity
- 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 as a scholarly activity

- The application of markup to a document can be an intellectual activity
- 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 as a scholarly activity

- The application of markup to a document can be an intellectual activity
- 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 as a scholarly activity

- The application of markup to a document can be an intellectual activity
- 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 as a scholarly activity

- The application of markup to a document can be an intellectual activity
- 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 as a scholarly activity

- The application of markup to a document can be an intellectual activity

- 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

@jamescummings – James.Cummings@it.ox.ac.uk – http://tei.it.ox.ac.uk/Talks/2014-01-toronto/
Compare the markup

```xml
<hi rend="dropcap">H</hi>
<g ref="#wynn">W</g>ÆT WE GARDE <lb/>na in gear-dagum þeod-cyninga <lb/>þrym gefrunon, hu ða æþelingas <lb/>ellen fremedon. oft scyld scefing sceape
<add>na</add>
<lb/>þreatum, moneg<expan>um</expan> mægbum meodo-setl
<add>a</add>
<lb/>of<damage>
<desc>blot</desc>
</damage>teah ...

<lg>
<l>Hwæt! we Gar-dena in gear-dagum</l>
<l>þeod-cyninga þrym gefrunon,</l>
<l>hu ða æþelingas ellen fremedon,</l>
</lg>
<lg>
<l>Oft Scyld Scefing sceapena þreatum,</l>
<l>monegum mægbum meodo-setla ofteah;</l>
<l>egsode Eorle, syððan ærest wearþ</l>
<l>feasceaf funden...</l>
</lg>
```
A useful mental exercise

Imagine you are going to markup several thousand pages of complex material....

- Which features are you going to markup?
- Why are you choosing to markup this feature?
- How reliably and consistently can you do this?

Now, imagine your budget has been halved. Repeat the exercise!
Extensible Markup Language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). Originally designed to meet the challenges of large-scale electronic publishing, XML also now plays an indispensable role in the exchange of a wide variety of data on the Web, on our phones, and elsewhere.

Its success means that general tools are ubiquitous and how it works is well-understood.
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
- It is one of the best plain text long-term preservation formats for textual data that we have
- You use XML almost every day, never mind the web but in many devices or even derived analogue information sources
Some XML Advantages

XML has four highly distinctive advantages:

1. it places emphasis on descriptive rather than procedural markup;
2. it distinguishes the concepts of syntactic correctness and of validity with respect to a document type definition;
3. it is independent of any one hardware or software system;
4. it has a vast amount of support in tools and infrastructure.
XML is similar to HTML

But:

- XML is extensible: it does not consist of a fixed set of tags;
- XML documents must be well-formed according to a defined syntax;
- an XML document can be formally validated against a schema of some kind;
- XML is more interested in the meaning of data than in its presentation.
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 on any element using `@xml:lang`
- unique identifiers may be specified on any element using `@xml:id`

The '@' before values like `@xml:id` and `@xml:lang` is just a common short-hand way to indicate they are attributes
XML terminology

An XML document may contain:-

- elements, possibly bearing attributes
- processing instructions
- comments
- entity references
- namespaces

An XML document must be well-formed and may be valid
XML terminology Example

```xml
<?xml version="1.0" ?>
<root>
  <element attribute="value"> content </element>
  <!-- comment -->
</root>
```
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" using &lt; or &amp; 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
- Entity references are delimited by & and ;
Parts of an XML document

```xml
<?xml version="1.0"?>
<greetings xmlns="http://www.example.org/greetings">
  <hello type="sarcastic">hello world!</hello>
</greetings>
```

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

```xml
<?xml version="1.0"?>
<greetings xmlns="http://www.example.org/greetings">
  <hello type="sarcastic">hello world!</hello>
</greetings>
```

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

```xml
<?xml version="1.0"?>
<greetings xmlns="http://www.example.org/greetings">
<hello type="sarcastic">hello world!</hello></greetings>
```

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

```xml
<?xml version='1.0'?>
<greetings xmlns='http://www.example.org/greetings'>
<hello type='sarcastic'>hello world!</hello></greetings>
```

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

<?xml version="1.0"?>
<greetings xmlns="http://www.example.org/greetings">
<hello type="sarcastic">hello world!</hello></greetings>

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

```xml
<?xml version="1.0"?>
<greetings xmlns="http://www.example.org/greetings">
<hello type="sarcastic">hello world!</hello></greetings>
```

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

All TEI documents are declared within the TEI namespace:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"> ...
</TEI>
```

XML documents can include elements declared in different name spaces.

- a namespace declaration associates a namespace prefix with an external URI-like identifier
- the default namespace may be declared using a xmlns
- other name spaces must all use a specially declared prefix

```
<TEI xmlns="http://www.tei-c.org/ns/1.0"
     xmlns:math="http://www.mathml.org">
<p>...<math:expr>...</math:expr>...</p>...</TEI>
```

The xml namespace is used by the TEI for global attributes @xml:id and @xml:lang
The Doctype Declaration

You may sometimes find an optional "Document Type" declaration at the start of a document:

```xml
<?xml version="1.0" ?>
<!DOCTYPE greeting SYSTEM "greeting.dtd []">
```

- The DTD is one way of associating the document with its schema (but is not used by W3C or RELAX NG for this purpose)
- The DTD subset is used to provide declarations additional to those in the schema, for example for external files
- The DTD subset may be internal, external, or both

DTDs are now considered old-fashioned – RELAX NG schemas are preferred by most.
The Tempest

<?xml version="1.0" encoding="utf-8"?>
<div n="1">
  <head>SCENE I. On a ship at sea: a tempestuous noise of thunder and lightning heard.</head>
  <stage>Enter a Master and a Boatswain</stage>
  <sp>
    <speaker>Master</speaker>
    <ab>Boatswain!</ab>
  </sp>
  <sp>
    <speaker>Boatswain</speaker>
    <ab>Here, master: what cheer?</ab>
  </sp>
  <sp>
    <speaker>Master</speaker>
    <ab>Good, speak to the mariners: fall to't, yarely,</ab>
    <ab>or we run ourselves aground: bestir, bestir.</ab>
  </sp>
  <stage>Exit</stage> [...]
</div>
Example deconstructed: root node

```xml
<?xml version="1.0" encoding="utf-8" ?> <div n="1">
<!-- .... -->
</div>
```
Example deconstructed: head

<head>SCENE I. On a ship at sea: a tempestuous noise of thunder and lightning heard.</head>
Example deconstructed: stage direction and speech

<stage>Enter a Master and a Boatswain</stage>
<sp>
  <speaker>Master</speaker>
  <ab>Boatswain!</ab>
</sp>
An XML Tree For The Tempest
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 can be combined start-and-end tags for certain elements: `<gap/> <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>

There are of course other mistakes you can make!
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 <seg/>`
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`

There are of course other mistakes you can make!
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>`

There are of course other mistakes you can make!
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>`

There are of course other mistakes you can make!
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>

There are of course other mistakes you can make!
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 <seg/>`  
- `<seg type="text"> some text<gap/> </seg>`
- `<seg type="text">some text</Seg>`

There are of course other mistakes you can make!
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>`

There are of course other mistakes you can make!
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>`

There are of course other mistakes you can make!
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>

There are of course other mistakes you can make!
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>`

There are of course other mistakes you can make!
The Text Encoding Initiative (TEI) is an international membership consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts chiefly in the humanities, social sciences and linguistics.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: ‘The TEI Guidelines’
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: ’The TEI Guidelines’
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.

@jamescummings – James.Cummings@it.ox.ac.uk – http://tei.it.ox.ac.uk/Talks/2014-01-toronto/
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: 'The TEI Guidelines'
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: ’The TEI Guidelines’
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: ’The TEI Guidelines’
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePUB, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: 'The TEI Guidelines'
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: 'The TEI Guidelines'
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community

Whatever you make it... it is a community-developed standard.
The TEI is not (just) a schema!

It is:

- an international consortium supported by libraries and universities with stable large open source community
- a set of definitions, examples and discussion of several hundred useful and mostly textual distinctions
- a set of regularly maintained and updated recommendations: 'The TEI Guidelines'
- a set of customizable tools and stylesheets for transformations to/from many formats (e.g. HTML, Word, PDF, Databases, RDF/Linked Data, Slides, ePub, Schemas, etc.)
- an archivally well-understood, consensus-based way of organizing and structuring textual (and other) resources
- an evolving history of the concerns of the digital humanities community
- whatever you make it... it is a community-developed standard.
Why is it useful?

- Because we need to interchange resources
  - between people
  - (increasingly) between machines
- Because we need to integrate resources
  - of different media types
  - from different technical contexts
- Because we need to preserve resources
  - cryogenics is not the (full) answer!
  - we need to preserve metadata as well as data
XML: One input – many outputs

One of the standard touted benefits of XML is that you need only have a particular piece of data in one location, once, and you can re-use that bit of data in many contexts and in many different forms of output. While this is true, it is an elegant ideal, real life is a bit more complicated:

- Sometimes for convenience you might repeat data
- Sometimes it becomes overly complex to refactor data for different outputs
- Sometimes the data is given to you in a form that is redundant, but deduplication of the data is problematic
- Sometimes the elegant solution implies much more complex processing that a couple copy-and-pastes will simplify

But that said, if you find your repeating your data in more than one location, you are probably doing something wrong.
Common TEI Outputs

TEI P5 XML

- EPub
- HTML
- PDF
- DATABASE
- JSON
- LINKED DATA
- RNG
- WORD

@jamescummings – James.Cummings@it.ox.ac.uk – http://tei.it.ox.ac.uk/Talks/2014-01-toronto/
The virtuous circle of encoding

Resources

encoding

digital resources

analysis

abstract model
The scope of intelligent markup

Even within the original scope of the TEI we have

- basic structural and functional components
- diplomatic transcription, images, annotation
- links, correspondence, alignment
- data-like objects such as dates, times, places, persons, events (named entity recognition)
- meta-textual annotations (correction, deletion, etc)
- linguistic analysis at all levels
- contextual metadata of all kinds
- ... and so on and so forth

Is it possible to delimit encyclopaedically all possible kinds of markup?
Reasons for attempting to define a common framework

- re-usability and repurposing of resources
- modular software development
- lower training costs
- ‘frequently answered questions’ — common technical solutions for different application areas

The TEI was designed to support multiple views of the same resource. The TEI is an evolving model of the concerns of Digital Humanities.
TEI Chapters (1)

In addition to Front Matter and Back Matter, the TEI Guidelines contain chapters on:

1. The TEI Infrastructure
2. The TEI Header
3. Elements Available in All TEI Documents
4. Default Text Structure
5. Non-standard Characters and Glyphs
6. Verse
7. Performance Texts
8. Transcriptions of Speech
9. Dictionaries
10. Manuscript Description
11. Representation of Primary Sources
12. Critical Apparatus

...
TEI Chapters (2)

... 

- 13. Names, Dates, People, and Places
- 14. Tables, Formulæ, and Graphics
- 15. Language Corpora
- 16. Linking, Segmentation, and Alignment
- 17. Simple Analytic Mechanisms
- 18. Feature Structures
- 19. Graphs, Networks, and Trees
- 20. Non-hierarchical Structures
- 21. Certainty and Responsibility
- 22. Documentation Elements
- 23. Using the TEI
TEI Conformance

A document is *TEI Conformant* if and only if it:

- is a well-formed XML document
- can be validated against a *TEI Schema*, that is, a schema derived from the TEI Guidelines
- conforms to the TEI Abstract Model
- uses the *TEI Namespace* (and other namespaces where relevant) correctly
- is documented by means of a TEI Conformant *ODD file* which refers to the TEI Guidelines

or if it can be transformed automatically using some TEI-defined procedures into such a document (it is then considered *TEI-conformable*).

Standardization should not mean ‘Do what I do’, but rather ‘Explain what you do in terms I can understand’
Joining the TEI Consortium

1. Sustaining Partner - $5,000.00 (USD) Bundle (up to 100 members)
2. Patron - $2,500.00 (USD) Bundle (up to 50 members)
3. Friend - $1,500.00 (USD) Bundle (up to 30 members)
4. Contributor - $500.00 (USD) Bundle (up to 10 members)
5. Supporter - $250.00 (USD) Bundle (up to 5 members)
6. Individual - $50.00 (USD)

Individuals are able to cast one vote for TEI Technical Council; 'Institutional' memberships are able to cast a single vote for TEI Board and TEI Technical Council; A wide range of benefits and discounts come with membership.
**oXygen XML Editor**

oXygen XML Editor (http://www.oxygenxml.com/) is:

1. probably the best and most complete XML development IDE available
2. ready to use support for a large number of document types (including TEI including auto-updates)
3. continuous and active development with proactive user community
4. good support: oXygen provides a very responsive support to its users and has a support mailing list; they participate in the communities they support
5. huge academic discounts, additional 20% discount for TEI members: Academic/Non-Commercial named-user licenses are inexpensive (compared to Enterprise license)
We’re going to do the first exercise! I will demonstrate this first if there is time.