The TEI: private and public concerns

Sebastian Rahtz
Head of Information and Support Group
Oxford University Computing Services

Mapping the Landscape of eResearch, Berlin, February 22nd 2012

Background and expectations

I am assuming that you:

- know what the Text Encoding Initiative is
- have an idea what richly digitized text looks like
- realize that the TEI is an ongoing issue, not a frozen standard
- care about access to resources by computers and not just humans...

I also apologize for my misleading name which contradicts my purely English background.

The problem

The Text Encoding Initiative was designed from the start as a dynamic model which could provide both

- a firmly-anchored model for well-understood structural components of digital texts, *and*
- a framework in which scholars could freely record their work in an open-ended and non-prescriptive way.

Do we end up with interchange, interoperability, or just a private record? Only relatively recently have we started to see enough open availablle TEI texts, and enough tools, for this to really matter.

Many expressions of the same semantics

```
<persName>
  <forename>Edward</forename>
  <forename>George</forename>
  <surname type="linked">Bulwer-Lytton</surname>, <roleName>Baron Lytton of
<placeName>Knebworth</placeName>
  </roleName>
</persName>
```

```
<p>
  <rs type="name">Edward George Bulwer-Lytton, Baron Lytton of
Knebworth</rs>
</p>
```

```
<p>
  <name type="person">Edward George Bulwer-Lytton, Baron Lytton of
Knebworth</name>
</p>
```

```
<p>
  <persName>Edward George Bulwer-Lytton, Baron Lytton of
Knebworth</persName>
</p>
```

continued...

```
<p>
  <persName>
    <forename>Edward</forename>
    <forename>George</forename>
    <surname>Bulwer-Lytton</surname>,
     Baron Lytton of Knebworth </persName>
</p>
```

```
<persName>
  <forename>Edward</forename>
  <forename>George</forename>
  <surname  type="linked">Bulwer-Lytton</surname>, <roleName>Baron Lytton of
<placeName>Knebworth</placeName>
  </roleName>
</persName>
```

(Note the difference in XML whitespace between these two last)

```
<persName  ref="#EBL">
  <forename>Edward</forename>
  <forename>George</forename>
  <surname>Bulwer-Lytton</surname>
  <roleName>Baron Lytton of <placeName>Knebworth</placeName>
  </roleName>
</persName>
```

continued ...

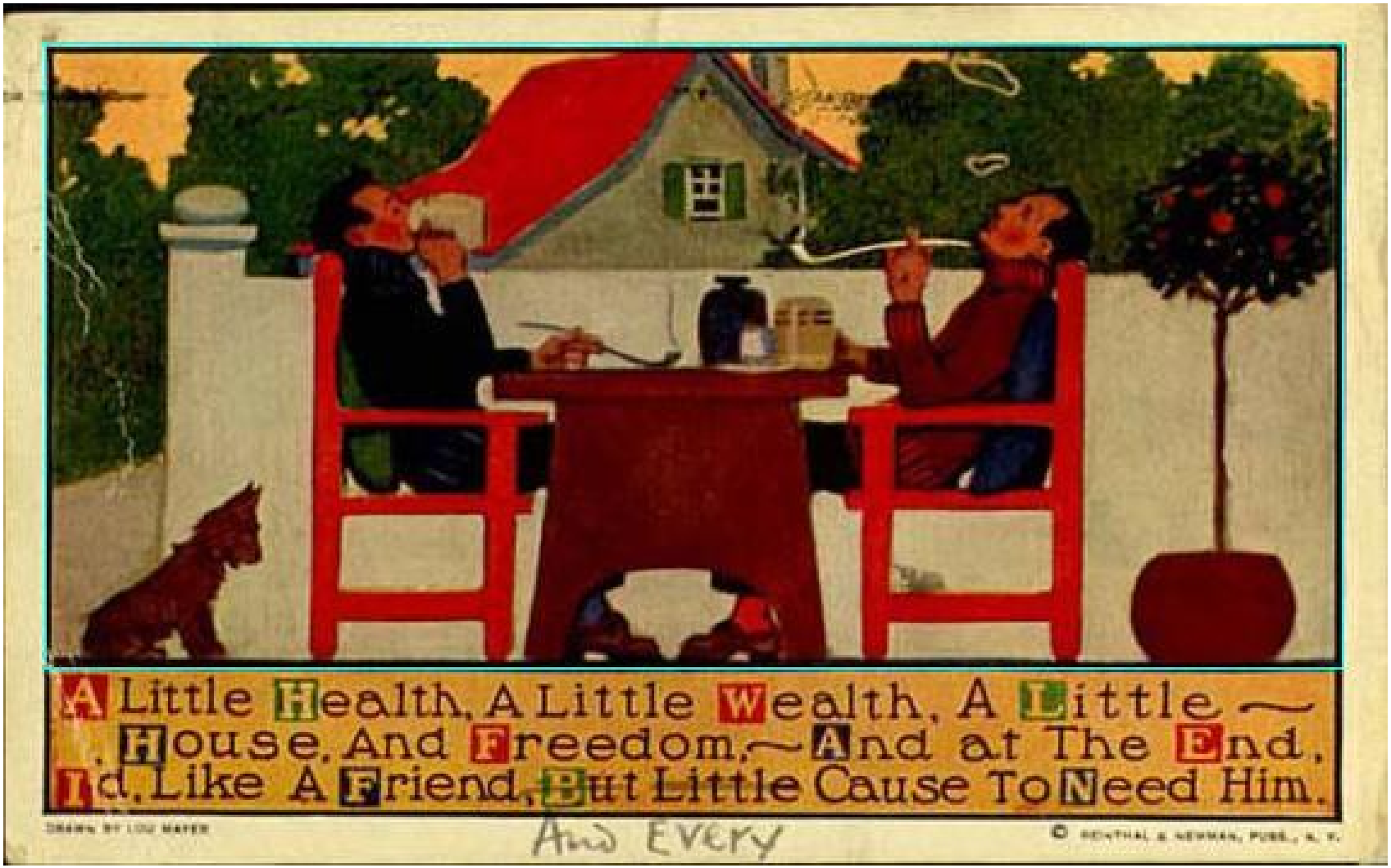
We can't even agree on the difference between

```
<head>CHAPTER I.</head>
<head>CHARLOTTE BROOKS.</head>
```

and

```
<head>CHAPTER I.<lb/>CHARLOTTE BROOKS.</head>
```

A typical artefact



True but useful?

```
<surface
   facs="#postcard_to_Leslie_Gunston_1918_02_17_leaf1_surface2">
  <zone
     facs="#postcard_to_Leslie_Gunston_1918_02_17_leaf1_surface2_zone1"
     subtype="preprinted"
     type="picture">
    <graphic  url="images/1918-02-12-page-2-zone-1.jpg"/>
  </zone>
  <zone
     facs="#postcard_to_Leslie_Gunston_1918_02_17_leaf1_surface2_zone2"
     subtype="preprinted"
     type="verse">
    <line>
      <hi  rend="red_background">A</hi> Little
<hi  rend="green_background">H</hi>ealth, A <hi  rend="green_background">L</hi>ittle
<hi  rend="red_background">W</hi>ealth, A Little  </line>
    <line>
      <hi  rend="blue_background">H</hi>ouse, <hi  rend="blue_background">A</hi>nd
<hi  rend="red_background">F</hi>reedom,  And at The
<hi  rend="red_background">E</hi>nd, </line>
    <line>
      <hi  rend="red_background">I</hi>'d, Like A
<hi  rend="blue_background">F</hi>riend, But little <zone
         facs="#postcard_to_Leslie_Gunston_1918_02_17_leaf1_surface2_zone3"
         type="handwritten"
         subtype="verse">
       <del  rend="stroked"  hand="#Wilfred_Owen">
         <hi  rend="green_background">B</hi>ut Little</del>
       <add  place="below"  hand="#Wilfred_Owen">And Every</add>
```

Many ways to solve a problem — linking text and facsimile

The *Best Practice in Libraries* guidelines say you can choose between:

- *Use the @facs attribute on a <pb> element to point to the corresponding page image using a URI.*
- *Use the <facsimile> element to define a set of images that corresponds to the text, in conjunction with the @facs attribute on a <pb> element to point to the corresponding page image using a URI.*
- *Use the @xml:id attribute on each <pb> element and a METS document to provide correspondence between <pb> elements and one or more facsimile page images (e.g., master, web derivatives, etc.).*

`http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html`

To punctuate or not

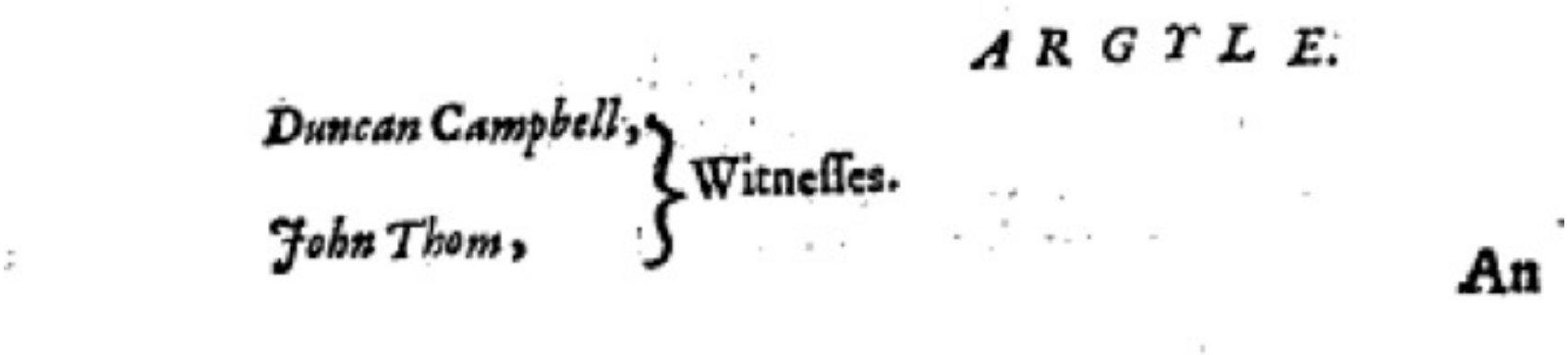
```
<p>She said, <said>Nobody uses the term <soCalled>electronic
text</soCalled>
    anymore</said>!</p>
```

```
<p>She said, <said>“Nobody uses the term <soCalled>‘electronic
text’</soCalled>
    anymore”</said>!</p>
```

```
<p>She said, <said rend="quotes: '“' '”'">Nobody uses the term
<soCalled rend="quotes: '‘' '’'">electronic text</soCalled>
anymore</said>!</p>
```

http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html

A more subtle example, a complex signature


ARGYLE.

Duncan Campbell, } Witneſſes.
John Thom, }

An

(thanks to Martin Mueller for the example

Note that TEI says a <signed> element 'contains the closing salutation, etc., appended to a foreword, dedicatory epistle, or other division of a text. '

One encoding

```
<signed>
  <table>
    <row>
      <cell>Duncan Campbell,</cell>
      <cell  rows="2">Witnesses</cell>
    </row>
    <row>
      <cell>John Thom,</cell>
    </row>
  </table>
</signed>
```

Another encoding

```
<table>
  <row>
    <cell>
      <signed>Duncan Campbell</signed>,</cell>
    <cell  rows="2">Witnesses</cell>
  </row>
  <row>
    <cell>
      <signed>John Thom</signed>,</cell>
  </row>
</table>
```

Is <signed> structural or decorative?

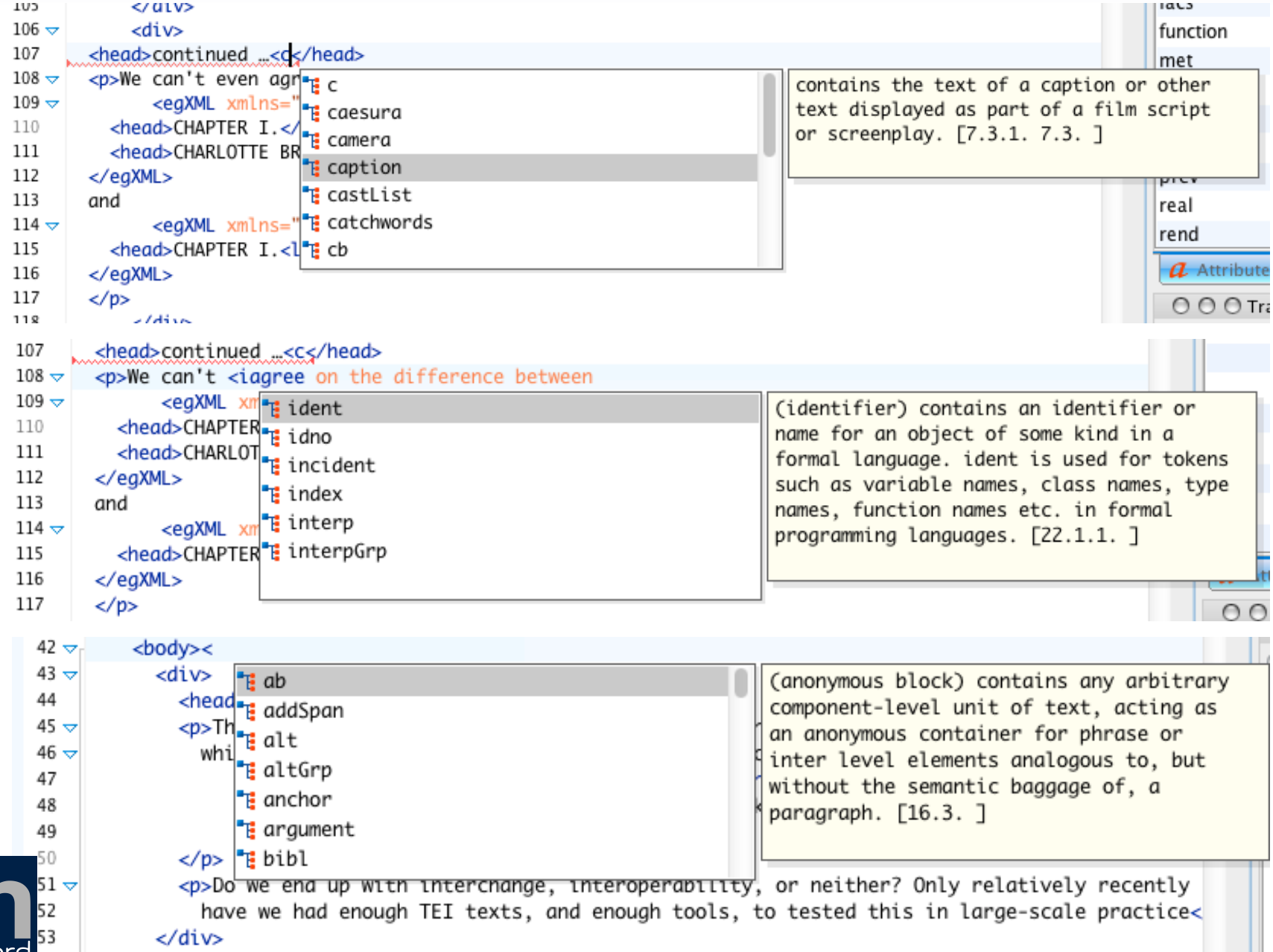
Yet another encoding

```
<ab type="signed">Argyle</ab>
<ab type="signed">
 <table>
  <row>
   <cell>Duncan Campbell,</cell>
   <cell cols="2">Witnesses</cell>
  </row>
  <row>
   <cell>John Thom,</cell>
  </row>
 </table>
</ab>
```

and still more encoding

```
<signed>
  <list type="gloss" rend="braced-right">
    <item>
      <list>
        <item>
          <name>Duncan Campbell</name>,</item>
        <item>
          <name>Iohn Thom</name>,</item>
      </list>
    </item>
    <label>Witnesses.</label>
  </list>
</signed>
```

Choices, absurd choices



What do we find in attribute values?

ECCO values for @rend on <hi> above, below, blackletterType, margDblQuotes, margQuotes, small, sub, sup

ECCO values for @rend on <gap> _____, a, alphabet, blank , book , different, duplicate , from, in, inserted, left, letters , line, lines , line , math, missing , non-Latin, page, pages, page , paragraph , span

ECCO values for @type on <lg> Psalm, address, air, airandduet, airandrecitative, answer, anthem, antistrophe, ballad, canto, catch, chorus, chorusandair, closer, duet, duetwithchorus, elegy, epigram, epigraph, epilogue, epistle, epitaph, epithalamicair, etc

However, there is also the good side

We have

- An interchange format (XML)
- An encoding (UTF-8)
- A very rich descriptive vocabulary
- A lot of agreement on semantics
- A powerful environment for describing variations (ODD)

Taking a step back, some basic questions:

- Why? Why do we care about our texts being interoperable?
- What? Which parts, exactly, of our texts do we want to share?
- How? What methods can we use to carry out our strategy?

and of course a follow-on question, which is

- What next? what do we suggest that the community around the TEI does next? to which the answers may range from 'nothing' to 'abandon the TEI and XML entirely'

Why?

If we don't want to share, why did we do the work in the first place? There does not seem much doubt that people want to combine their work with others to produce a bigger 'answer' than they can find on their own. But what exactly do they want to share and combine? I suggest there are four possible models of sharing and interoperability:

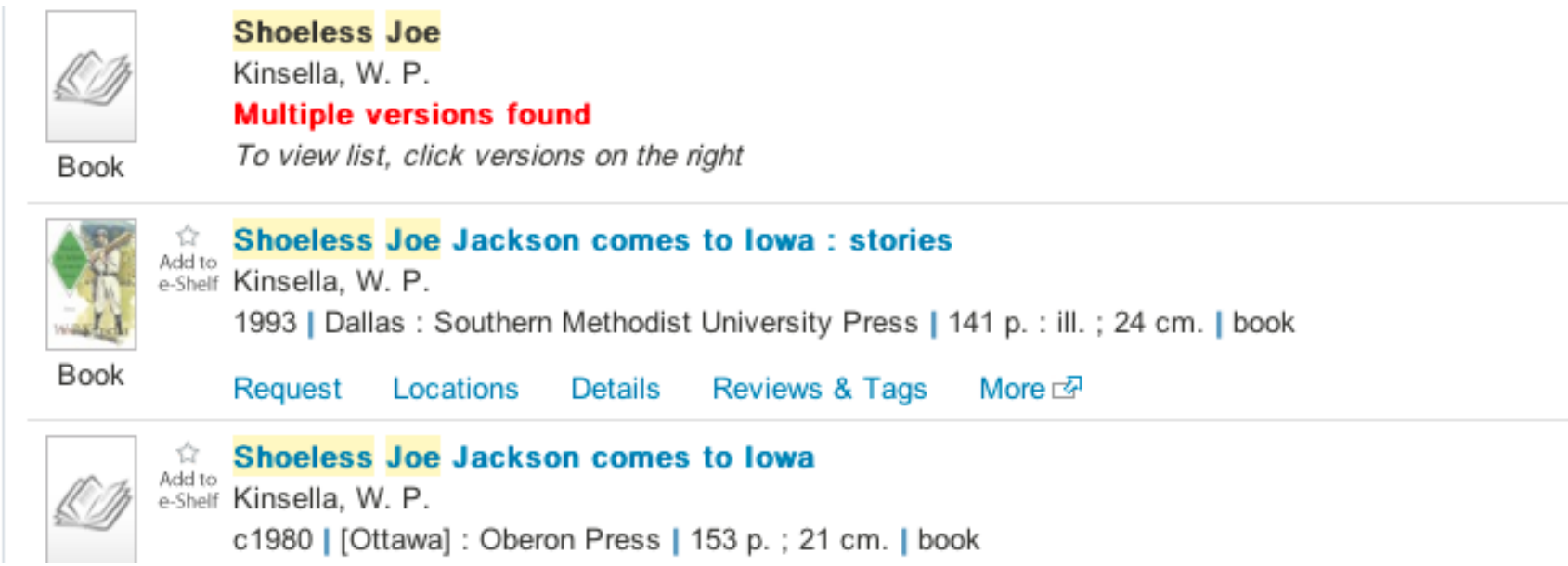
- Bibliography
- Extraction
- Mapping
- The words

Of course, these are not mutually exclusive, so there is also

- Containing

Bibliography

Method	Implications
We release enough information about our text to let others find it in the traditional way, eg author, title, date etc; and then they can simply *read* it. The markup is decoration aimed at describing the original presentation to let us produce facsimiles.	The text markup must be detailed enough to act as typesetting instruction.



Extraction

Method	Implications
We leave clues in our text which lets someone comb through it and produce assertions about the real world. To take an extreme example, we may read the novel *Jane Eyre*, and establish that yes, Jane did marry Rochester. It is a fact about that fictional world. This model treats the text as a database for mining.	We only need the data markup, and the structural markup to provide context.

```
<http://ota.ox.ac.uk/persname/mcturk>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://purl.org/NET/crm-owl#E82_Actor_Appellation> .
<http://ota.ox.ac.uk/persname/mcturk>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#value> "McTurk" .
<http://ota.ox.ac.uk/persname/orrin>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://purl.org/NET/crm-owl#E82_Actor_Appellation> .
<http://ota.ox.ac.uk/persname/orrin>
```

Mapping

Method	Implications
We put markup in our text so that it can all be taken out and directly compared to another text. Every component of the text has a strict meaning. The markup is not describing the presentation but presenting the analysis. A linguistic or critical analysis might be typical of this approach.	Everything must be structural, and have semantics.

2. Christmas Broadcast 1954

It is now two years since my husband and I spent Christmas with our children. And as we do so today we look back upon a Christmas spent last year in Auckland in hot sunshine, thirteen thousand miles away.

Though this was strange for us, we felt at home there, for we were among people who are my own people and whose affectionate greeting I shall remember

The Words

Method	Implications
There is quite a widespread view that in fact interoperability really deals with just the words. Strip out all that dirty markup describing presentation, and you have a stream of Unicode text which you can use anywhere for anything. Interpretation and analysis are dynamic and real time. Effectively, this is what Google Books offer. Scale is what matters.	Markup must be precise enough not be ambigious about eg spaces between words, and line-breaking.

```
the tragedy of hoffman enter hoffman hence clouds of melancholy ile be
no longer subiect to your sismes but thou deare soule whose nerues and
artires in dead resoundings summon vp reuenge and thou shalt hate be
but appeas'd sweete hearse the dead remembrance of my liuing father
strikes ope a cur-taine where ap-peares a body.
```

Containing

Method	Implications
We envisage the text as having two (or more) layers, the presentational layer, and the interpretative layer, and we do not mix the two. There are two distinct documents. Here we enter, of course, the long-standing issue of overlapping markup. A purely technical question, in some ways, but some approaches (stand-off markup, and non-XML solutions) make the layer distinctions clearer than others.	The markup is very concerned to define enough granularity in the text to allow it to be referred to.

What?

Are we

- only wanting the metadata header to make a MARC record
- interpreting the existing markup to render it as a web page
- targetting a specific subset of the markup to answer a question
- ...or do we just not care what the markup is

How?

If we want just words, or just metadata, we're pretty much OK.

But very often we have to interpret markup, and *do* something with it.

Experience in trying to maintain a generic family of TEI stylesheets suggests that this is not at all easy.

What shall we do? Abandon the whole nonsensical farrago of meaningful markup?

- Use a word-processor to replicate format using entirely visual decoration
- Leave it up to the natural language people to extract meaning

We know very well what is happening in an ancient Greek manuscript, without benefit of all the whitespace and typographic techniques we now take for granted.

What shall we do? Separate presentation and interpretation?

- adopt the web model of HTML for making the text look like something, in a reasonable re-useable way
- decorate it with semantic clues which let us perform extraction of embedded assertions

This is what most of the world is up to.

Using HTML5

```
<div class="tei_body">
 <section
    itemscope="itemscope"
    itemtype="http://www.tei-c.org/ns/1.0/"
    class="verse"
    id="index.xml-body.1_div.1">
  <header>
   <h1 itemprop="head">
    <span class="headingNumber">1. </span>
    <span itemprop="head" class="head">
     <span class="capitalize">STRANGE MEETING</span>
    </span>
   </h1>
  </header>
  <div itemprop="lg" class="stanza">
   <div class="l">It seemed
        that out of battle I escaped</div>
   <div class="l">Down some
        profound dull tunnel , long since
        scooped</div>
   <div class="l">Through
        granites which titanic wars had groined .
   </div>
  </div>
 </section>
</div>
```

What shall we do? Go back to the dark ages?

- extract interchange data
- manage it in RDF or a conventional database
- keep a loose bibliographical link back to the original text
- reproduce old texts in eBook format

What shall we do? Keep the faith?

Believe that our complex markup *does* allow for

- repeatable analysis
- better archival recording of decisions
- more objective descriptions

The TEI is better than HTML5 because

- the vocabulary is richer
- the structure is more complex
- the specialist domains are much wider
- the schema is more powerful

Unfortunately it has no default rendering model.

If we keep the faith, how can we resolve the TEI confusion?

- Render down a TEI text into a simpler normalized form, much more tightly bound to a set of TEI elements which are explicitly used to record aspects we want to share.
- Using this so-called 90/10 solution, each encoding project would then have to define a mapping between its own interesting use of TEI elements, and the Ur-elements.
- The mapped form is a generated output, not an archival form.
- The original is interchangeable, the simplified form is interoperable.

Implications of a TEI 90/10 model

We will have to find a way of rephrasing guidance like: 'Alternatively, the <note> element may encode the text of the note at the point it occurs on the page or at another point convenient when converting from a born-digital source document, such as at the end of the containing (or <div1>) or in a special (or <div1>) element within <back>. The point of reference should be encoded using a <ref> or <ptr> element ' ...

http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html

Where do we want to get to?

We want a computer to be able to process a TEI text and *easily* recognize:

1	The distinction between metadata and text	<teiHeader> and <text>
2	Structural components which provide context for both formatting and extraction	<front>, <body>, , <list>, <note> etc
3	Inline strict and loose semantic markup	<hi>, <name>, <foreign>
4	Typologies and links	@type, @ref, etc
5	Editorial interpretation	<interp>, <s>, <lemma> etc
6	Data	<name>, <date>, <placeName> etc

preferably distinguishing those which map to other vocabularies.

Detail: where to store mapping?

It is easy enough to look at the TEI <person> element and say it corresponds to what the CRM calls *E21_Person*

> *This class comprises real persons who live or are assumed to have lived. Legendary figures that may have existed, such as Ulysses and King Arthur, fall into this class if the documentation refers to them as historical figures. In cases where doubt exists as to whether several persons are in fact identical, multiple instances can be created and linked to indicate their relationship. The CRM does not propose a specific form to support reasoning about possible identity. Examples: - Tut-Ankh-Amun - Nelson Mandela*

The <equiv> element allows us to provide a specification in ODD which points from a TEI element to an external identifier, and says how to get there.

Example of <equiv>

```
<elementSpec ident="geo" mode="change">
  <equiv
     filter="crm.xsl"
     mimeType="text/xsl"
     name="E47"
     uri="http://erlangen-crm.org/110404/E47_Place_Spatial_Coordinates"/>
</elementSpec>
```

name	names the underlying concept of which the parent is a representation
uri	references the underlying concept of which the parent is a representation by means of some external identifier
filter	references an external script which contains a method to transform instances
mimeType	gives the MIME media type of filter script

What is in crm.xsl?

Named XSL templates which do creation of RDF XML:

```
<xsl:template name="E47">
  <P87_is_identified_by>
    <E47_Place_Spatial_Coordinates>
      <value>
        <xsl:value-of select="."/>
      </value>
    </E47_Place_Spatial_Coordinates>
  </P87_is_identified_by>
</xsl:template>
<xsl:template name="E69">
  <P100i_died_in>
    <E69_Death>
      <P4_has_time-span>
        <E52_Time-Span>
          <P82_at_some_time_within>
            <E61_Time_Primitive>
              <xsl:call-template name="calc-date-value"/>
            </E61_Time_Primitive>
          </P82_at_some_time_within>
        </E52_Time-Span>
      </P4_has_time-span>
    </E69_Death>
  </P100i_died_in>
</xsl:template>
```

Input

```
<person xml:id="ArnMag01" sex="1" role="scholar">
  <persName xml:lang="is">Árni Magnússon</persName>
  <persName xml:lang="la">Arnas Magnæus</persName>
  <persName xml:lang="da">Arne Magnusson</persName>
  <birth when="1663-11-13">13 November 1663</birth>
  <death when="1730-01-07">7 January 1730</death>
  <residence>
    <date from="1663" to="1680">1663-1680</date>
    <placeName>
      <settlement type="farm">Hvammur</settlement>
      <region type="county">Dalasýsla</region>
      <region type="compass">Western</region>
      <country key="IS">Iceland</country>
    </placeName>
  </residence>
  <residence>
    <date from="1680" to="1683">1680-1683</date>
    <placeName>
      <settlement type="institution">Skálholt</settlement>
      <region type="county">Árnessýsla</region>
      <region type="compass">Southern</region>
      <country key="IS">Iceland</country>
    </placeName>
  </residence>
</person>
```

Result

```
<RDF>
  <E21_Person
      rdf:about="http://www.example.com/idArnMag01">
    <P131_is_identified_by  xml:lang="is">
      <E82_Actor_Appellation
          rdf:about="http://www.example.com/persname/ArnMag01">
        <value>Árni Magnússon</value>
      </E82_Actor_Appellation>
    </P131_is_identified_by>
    <P98i_was_born>
      <E67_Birth>
        <P4_has_time-span>
          <E52_Time-Span>
            <P82_at_some_time_within>
              <E61_Time_Primitive>
                <value>1663-11-13</value>
              </E61_Time_Primitive>
            </P82_at_some_time_within>
          </E52_Time-Span>
        </P4_has_time-span>
      </E67_Birth>
    </P98i_was_born>
  </E21_Person>
</RDF>
```

Result (continued)

```
<RDF>
  <E21_Person
      rdf:about="http://www.example.com/person/ArnMag01">
    <P74_has_current_or_former_residence>
      <E53_Place
          rdf:about="http://www.example.com/place/hvammur">
        <P2_has_type
            rdf:resource="http://www.tei-c.org/type/place/settlement"/>
        <P87_is_identified_by>
          <E48_Place_Name
              rdf:about="http://www.example.com/placename/hvammur">
            <value>Hvammur</value>
          </E48_Place_Name>
        </P87_is_identified_by>
        <P89_falls_within
            rdf:resource="http://www.example.com/place/dalassla"/>
      </E53_Place>
    </P74_has_current_or_former_residence>
  </E21_Person>
  <E53_Place
      rdf:about="http://www.example.com/place/dalassla">
    <P2_has_type
        rdf:resource="http://www.tei-c.org/type/place/region"/>
    <P87_is_identified_by>
      <E48_Place_Name
          rdf:about="http://www.example.com/placename/dalassla">
        <value>Dalasýsla</value>
      </E48_Place_Name>
    </P87_is_identified_by>
  </E53_Place>
</RDF>
```

The ODD is the key

Our TEI metaschema gives us a location to place the rules about:

- whether an element is to be regarded as significant in the 90/10 mapping process
- whether the element is to be transformed for the mapping (eg <name type="person"> to <persName>)
- which data category the element is in
- whether its attributes are significant beyond local typology

Different schemas can be used to maintain several rule sets.

Different markup types

```
<body>
    <div type="letter" n="leaf1">
        <div facs="#letter_to_Leslie_Gunston_1917_07_leaf1_surface2"
            rend="non_line_breaking">
            <head>
                <placeName ref="#Craiglockhart">Craiglockhart</placeName>. <lb/>July 1917.
                <lb/>Wednesday <lb/>Dear <persName ref="#Leslie_Gunston">
                    <choice>
                        <abbr>L.</abbr>
                        <expan>Leslie</expan>
                    </choice>
                </persName>
            </head>
            <p>
                <lb/>Thanks for your <unclear>A</unclear> this morning. I hope <lb/>you <add
                    place="above">have</add> had my card <del rend="stroked">of</del> I
                posted last Monday. <lb/>On <choice>
                    <abbr>Mond.</abbr>
                    <expan>Monday</expan>
                </choice> next I lecture the "<orgName ref="#Field_Club">Field
                    Club</orgName>" - <lb/>a <choice>
                    <abbr>Nat.</abbr>
                    <expan>Natural</expan>
                </choice>
                <choice>
                    <abbr>Hist.</abbr>
                    <expan>History</expan>
                </choice> association in the lines of our <lb/>
                <unclear>old</unclear> Society - Geological, (you + me) - Botanical
                <lb/>(New) Do you remember: <del rend="stroked">my</del> you old <lb/>
                <persName ref="#Leslie_Gunston">Black Molt</persName>? Well, the days have
```

Conclusions

- The TEI does not need defending or justifying. It is a mature, multi-tasking, markup scheme
- The TEI has enough extra power to justify its use as a layer about HTML5 or RDF
- It is also commonly non-interoperable
- TEI's ODD gives you the way to express relationship between public and private concerns
- We need a clean agreement on how to distill interoperable TEI from interchangeable TEI