WHAT THE COMPUTER DOESN'T KNOW...

Representing Primary Source Documents in TEI

Download these slides from http://bit.ly/cromptonworkshops

Constance Crompton
Assistant Professor of Digital Humanities, UBCO

with thanks to Julia Flanders, Syd Bauman, and Lee Zickel

INTRODUCTIONS

- Where are you from? What material or texts do you work with?
- What project do you have underway or in mind?
- What experience do you have with the TEI, if any?

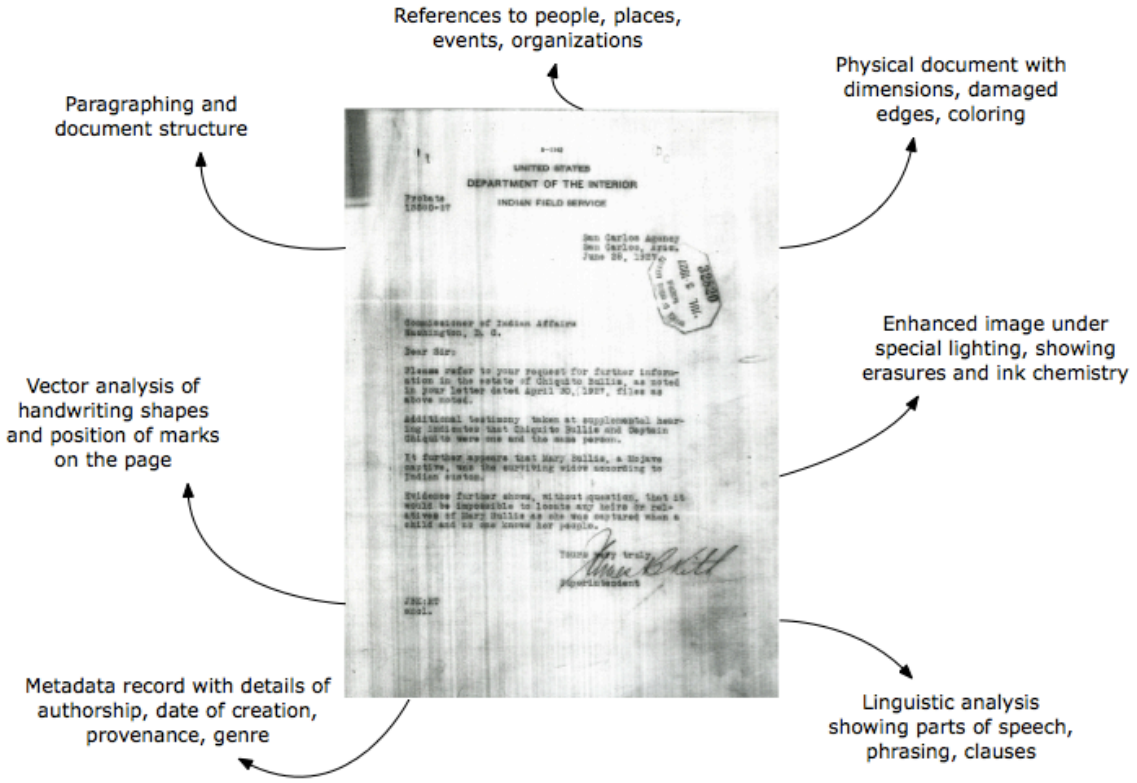
WHAT IS THE TEXT ENCODING INITATIVE?

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form" www.tei-c.org

- Developed by an international consortium founded in 1987 to meet community need for ways to represent text that are
  - Hardware and software independent
  - Computationally tractable
  - Formal
  - Reusable

- Now has modules for representing objects other than text
- Both a community standard and a community research effort

SOME THINGS WE MIGHT WANT TO REPRESENT



XML MARKUP IS EVERYWHERE

eXensible Mark-up Language makes the world go round

- KML (maps)
- TEI (editions)
- XHTML and parts of HTML5 (for browser display)
- WordprocessingML (Microsoft Word)
- CMBL (Comic Book Markup Language)
- DocBook (hardware and software documentation)
- ePub (electronic publishing, iBook)
- MathML (math)
- RSS (web syndication or web feeds)
- RDF (enables advanced inference-based web search)
- MARCXML (library records)
- … and dozens more

XML ANATOMY

I've "drawn a box" around the text UBC Okanagan. The opening tag and closing tag mark the sides of the box.

On Wednesday I called Arlene. She will be visiting <location>UBC Okanagan</location>before the end of the summer.

NB: I've added colour for demonstration purposes. Code doesn't have to be in colour to work.

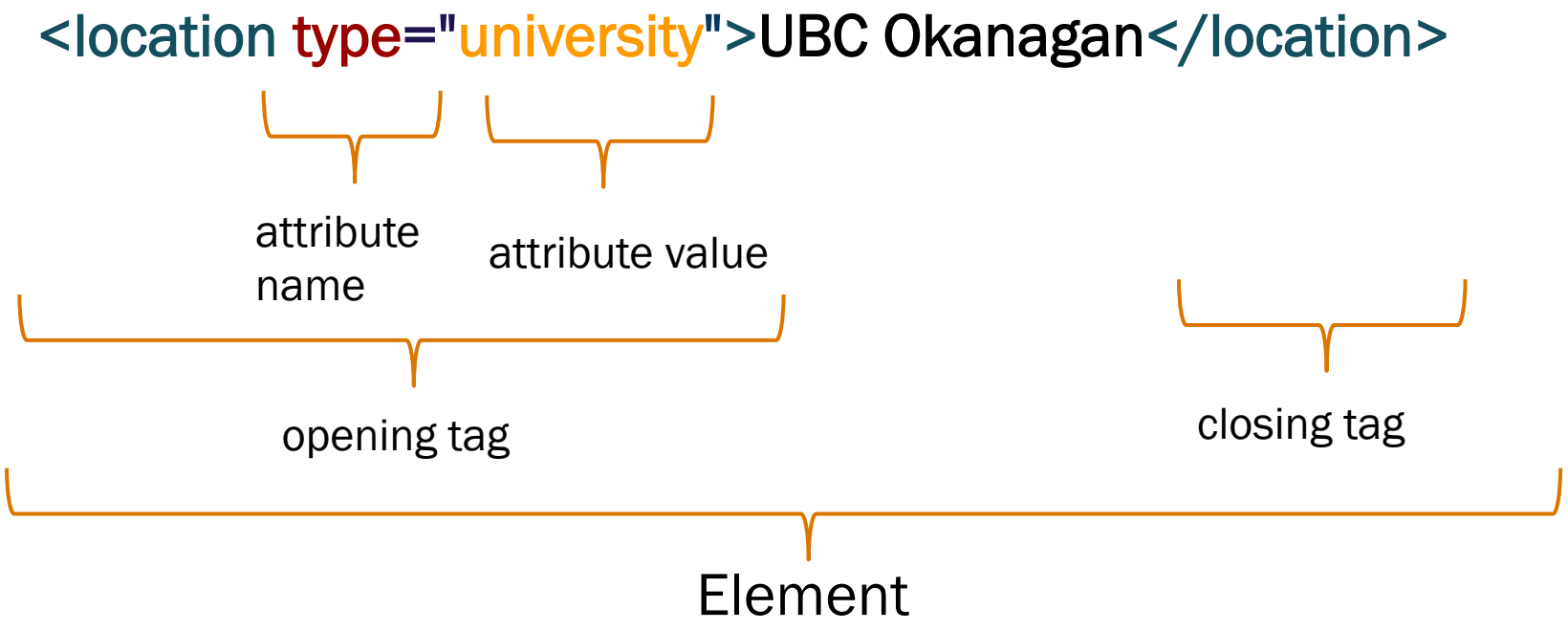
XML ANATOMY

I've "drawn a box" around the text UBC Okanagan. The opening tag and closing tag mark the sides of the box, and have added more information with an attribute name and attribute value (which go inside the opening tag). Wherever you have an attribute name, you must have an attribute value, set off from them name with an = symbol and double quotation marks

On Wednesday I called Arlene. She will be visiting <location type="university">UBC Okanagan</location> before the end of the summer.

Empty elements are those without content. Instead of writing an opening and closing tag with no content in between them, we write a singe tag with the forward slash to the right of the element name (e.g. <lb></lb> as <lb/> to represent a line break).

XML ANATOMY


<location type="university">UBC Okanagan</location>
attribute name
attribute value
opening tag
closing tag
Element

XML RULES

- XML documents must be “well formed”
- XML must be “valid”
- XML elements must *nest*, and never *overlap*
- XML documents must have a single *root element* and can be expressed as a hierarchy, or tree.

Let’s talk through what these mean.

VALID XML

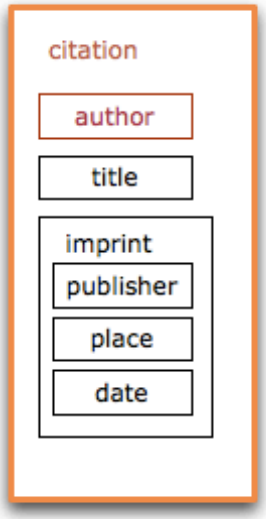
- TEI must be validated against a schema

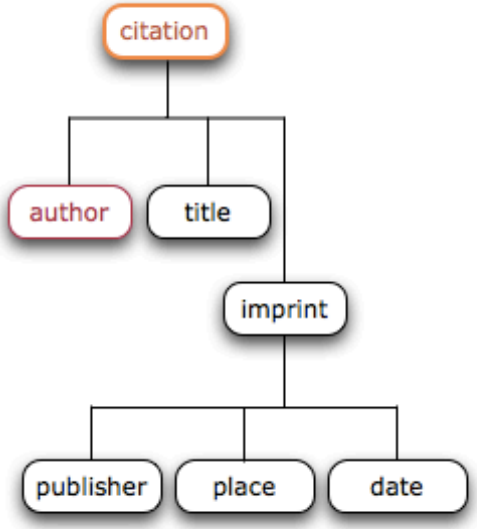
- A valid XML document uses correct vocabulary – only includes elements and attributes specified by the markup language (e. g. TEI, XHTML, WordProcessingML).

- A valid XML document uses correct grammar – the elements are in the right place, in the right order

ELEMENTS, NESTING, ONE ROOT

- XML documents all have elements, attributes, and values.
- All XML-based languages' elements *nest*, and never *overlap*
- XML documents have a single *root element* and can be expressed as a hierarchy, or tree.






```
<?xml version="1.0"
encoding="UTF-8"?>
<citation>
  <author>Katherine Hayles</author>
  <title>Writing Machines</title>
  <imprint>
    <publisher>MIT Press</publisher>
    <place>Cambridge, MA</place>
    <date>2002</date>
  </imprint>
</citation>
```

TEI-XML

The TEI help us model our research materials in a way that is

- Sustainable
- Sharable
- Platform Agnostic
- Explicit
- Formal

TEI-XML

XML

Syntax

```
<element>
    <element attribute="value">
        content
    </element>
</element>
```

TEI

Language:
vocabulary and grammar

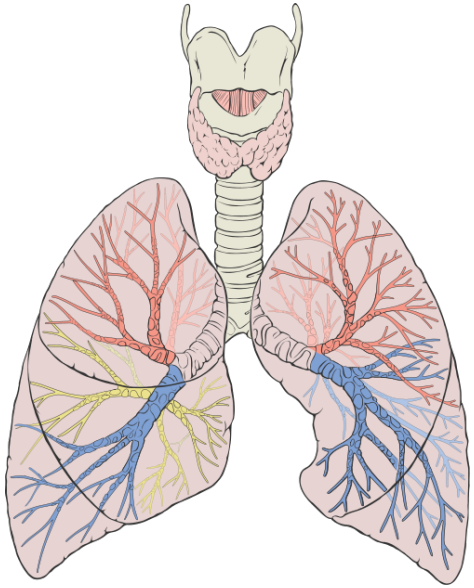
```
<p>

<note type="foot">

<head>
```

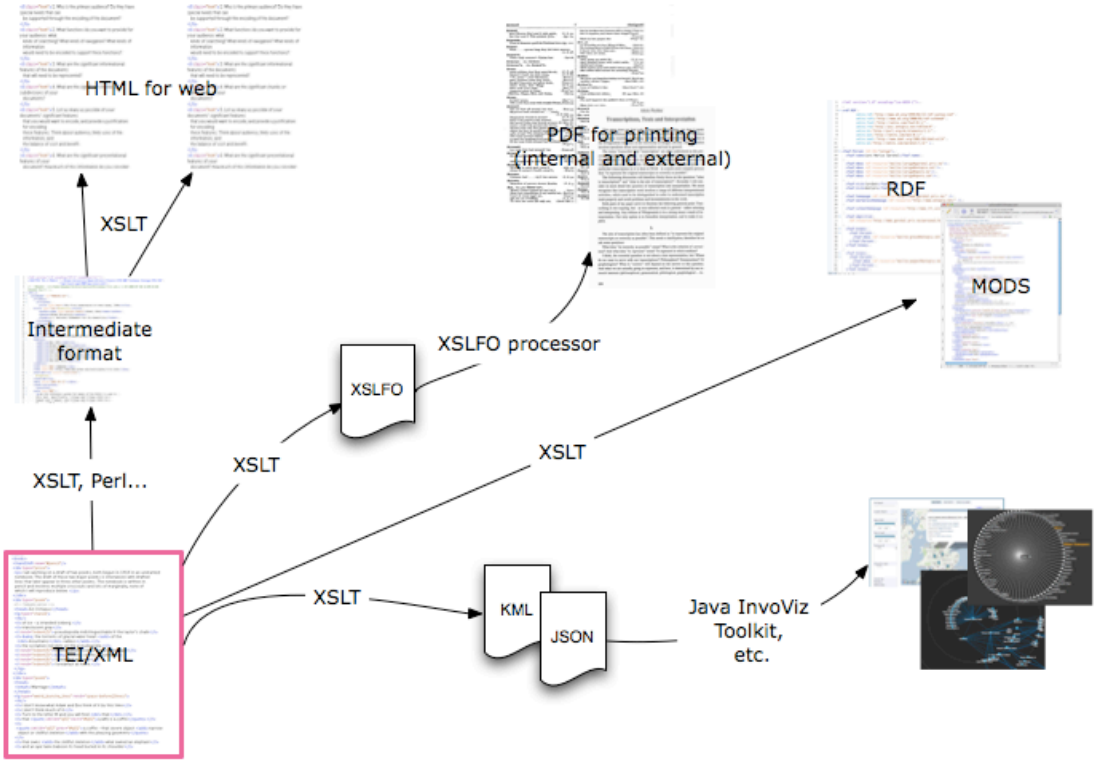
Concepts

WHEN IS TEI KNOWLEDGE USEFUL?
MODELING



A model is a tool for thinking – a representation or surrogate for a real-world object or phenomenon, that helps us better understand a certain aspect of that object or phenomenon. We can model using pen and paper, wax, or in our case, code.

WHEN IS TEI KNOWLEDGE USEFUL? SINGLE ENCODING, MULTIPLE OUTPUTS



WHEN IS TEI KNOWLEDGE USEFUL? CLOSE READING AND PATTERNS AT SCALE

Close reading (via markup) and distant reading (via visualization) are both valuable

- Jockers: “The truth, however, seems to be that you’re interested in the same thing I am, which is to say, how “macro” and “micro” approaches are interconnected and interdependent.”
- Flanders: “the detailed markup we’re doing isn’t aimed at the individual text: I would call it “detailed data at scale”—we are encoding a collection of texts in a consistent way that captures a set of repeating features: e.g. named entities, rhetorical structures, intertextual references, etc. These features operate meaningfully both within the ecology of a single text, and also within the ecology of the collection as a whole: so the “micro” view represented by the markup is very much in the service of the “macro” view represented by the collection and the collection-level tools/interface.”

Flanders, Julia and Matt Jockers "A Matter of Scale”
http://digitalcommons.unl.edu/englishfacpubs/106/

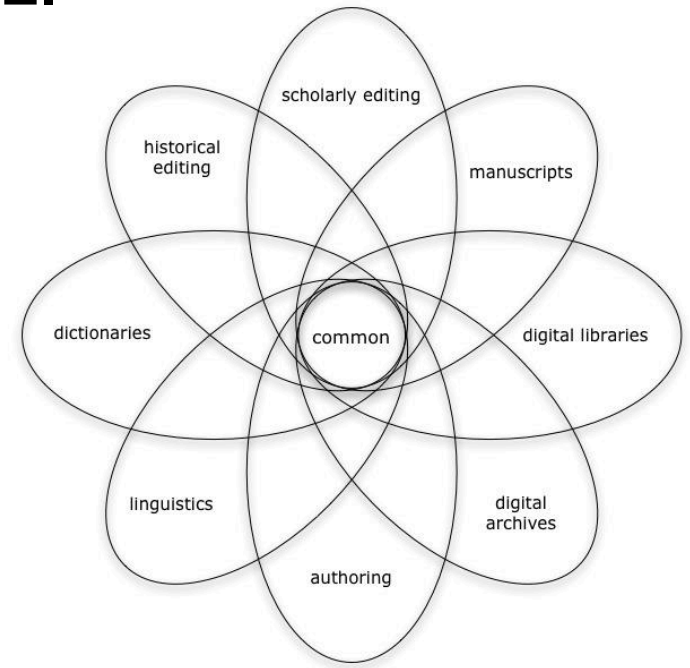
WHEN IS TEI KNOWLEDGE USEFUL? CLOSE READING AND PATTERNS AT SCALE

- Jockers: “When you drive through the fine state of Nebraska, you see the individual rows of corn, the silos, the barns, and so on. These are the details you see from up close. You don’t see how each farm is connected to another farm in a beautiful and organized patchwork of 640-acre sections. You don’t see how the landscape of one farm is the result of and dependent upon the landscape of those surrounding it. You need a plane or a satellite to reveal those particulars.”

- Flanders: “So, close or distant reading, whether done with the assistance of the computer or with the naked eye, is a method dependent upon different levels of focus and attention; at one moment we focus our critical lens in a way that is meant to call certain aspects of that poem or passage into our field of view and in that moment we necessarily ignore other facets that might be seen using a different focus.”

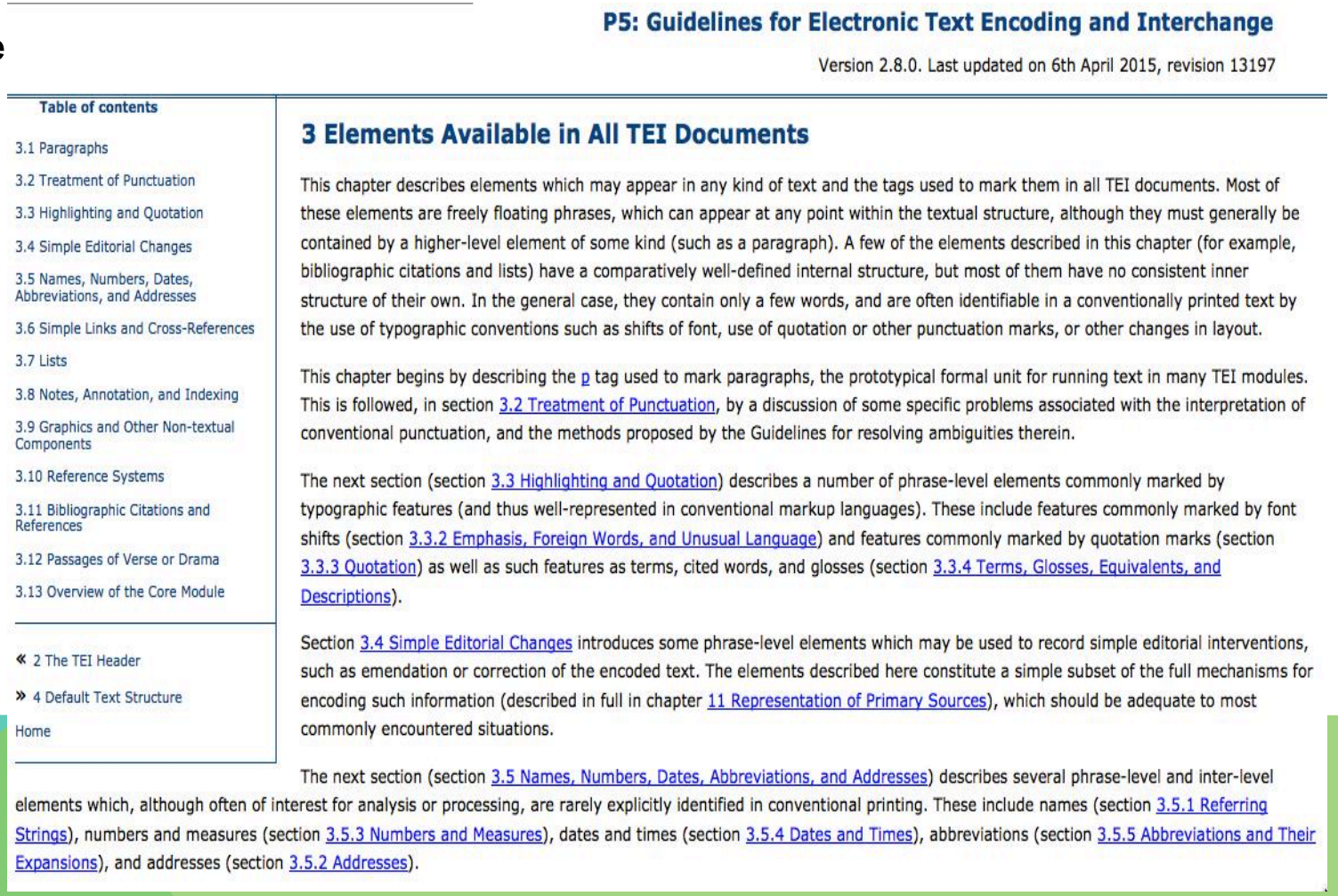
SCHEMAS: CUSTOMIZING THE TEI

- The TEI is capacious.
- No project uses all 500+ elements
- Schema customization is a formal modeling activity – representing textual sources in a computational tractable way
- Schema customization is the way to keep data from chaos.



THE TEI GUIDELINES: CHAPTERS

The chapter prose explains how the elements work together to describe a document.



P5: Guidelines for Electronic Text Encoding and Interchange

Version 2.8.0. Last updated on 6th April 2015, revision 13197

Table of contents

« 2 The TEI Header
» 4 Default Text Structure
Home

3 Elements Available in All TEI Documents

This chapter describes elements which may appear in any kind of text and the tags used to mark them in all TEI documents. Most of these elements are freely floating phrases, which can appear at any point within the textual structure, although they must generally be contained by a higher-level element of some kind (such as a paragraph). A few of the elements described in this chapter (for example, bibliographic citations and lists) have a comparatively well-defined internal structure, but most of them have no consistent inner structure of their own. In the general case, they contain only a few words, and are often identifiable in a conventionally printed text by the use of typographic conventions such as shifts of font, use of quotation or other punctuation marks, or other changes in layout.

This chapter begins by describing the p tag used to mark paragraphs, the prototypical formal unit for running text in many TEI modules. This is followed, in section 3.2 Treatment of Punctuation, by a discussion of some specific problems associated with the interpretation of conventional punctuation, and the methods proposed by the Guidelines for resolving ambiguities therein.

The next section (section 3.3 Highlighting and Quotation) describes a number of phrase-level elements commonly marked by typographic features (and thus well-represented in conventional markup languages). These include features commonly marked by font shifts (section 3.3.2 Emphasis, Foreign Words, and Unusual Language) and features commonly marked by quotation marks (section 3.3.3 Quotation) as well as such features as terms, cited words, and glosses (section 3.3.4 Terms, Glosses, Equivalents, and Descriptions).

Section 3.4 Simple Editorial Changes introduces some phrase-level elements which may be used to record simple editorial interventions, such as emendation or correction of the encoded text. The elements described here constitute a simple subset of the full mechanisms for encoding such information (described in full in chapter 11 Representation of Primary Sources), which should be adequate to most commonly encountered situations.

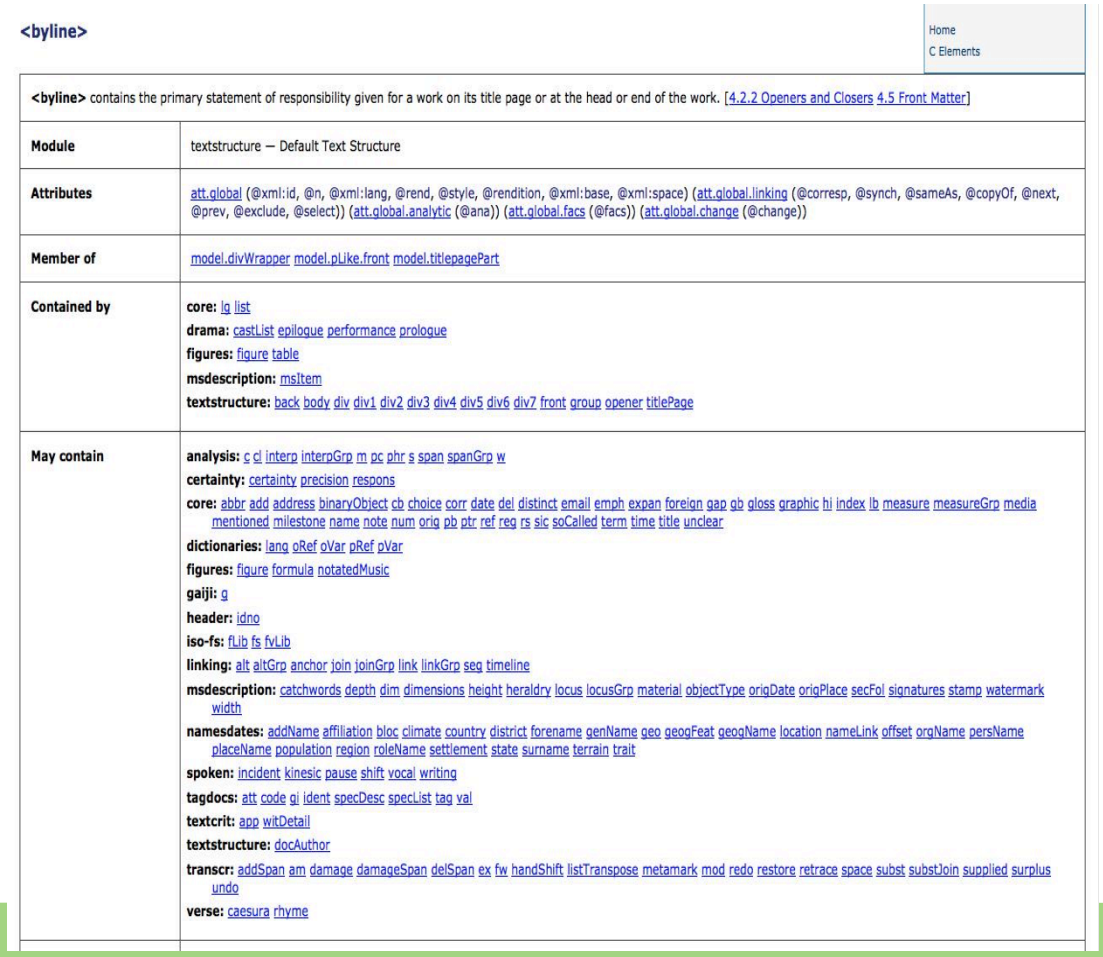
The next section (section 3.5 Names, Numbers, Dates, Abbreviations, and Addresses) describes several phrase-level and inter-level elements which, although often of interest for analysis or processing, are rarely explicitly identified in conventional printing. These include names (section 3.5.1 Referring Strings), numbers and measures (section 3.5.3 Numbers and Measures), dates and times (section 3.5.4 Dates and Times), abbreviations (section 3.5.5 Abbreviations and Their Expansions), and addresses (section 3.5.2 Addresses).

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index-toc.html

THE TEI GUIDELINES: APPENDICES

The appendices are handy quick reference. The element appendix comprises

- **Definition**
- **Chapter in the Guidelines**
- **The attributes this element can have**
- **Elements that can contain this element**
- **Elements this element can contain**



<byline>

Home
C Elements

<byline> contains the primary statement of responsibility given for a work on its title page or at the head or end of the work. [4.2.2 Openers and Closers 4.5 Front Matter]

Module	textstructure — Default Text Structure
Attributes	att.global (@xml:id, @n, @xml:lang, @rend, @style, @rendition, @xml:base, @xml:space) (att.global.linking (@corresp, @synch, @sameAs, @copyOf, @next, @prev, @exclude, @select)) (att.global.analytic (@ana)) (att.global.facs (@facs)) (att.global.change (@change))
Member of	model.divWrapper model.pLike.front model.titlepagePart
Contained by	**core:** lg list **drama:** castList epilogue performance prologue **figures:** figure table **msdescription:** msItem **textstructure:** back body div div1 div2 div3 div4 div5 div6 div7 front group opener titlePage
May contain	**analysis:** c cl interp interpGrp m pc phr s span spanGrp w **certainty:** certainty precision respons **core:** abbr add address binaryObject cb choice corr date del distinct email emph expan foreign gap gb gloss graphic hi index lb measure measureGrp media mentioned milestone name note num orig pb ptr ref reg rs sic soCalled term time title unclear **dictionaries:** lang oRef oVar pRef pVar **figures:** figure formula notatedMusic **gaiji:** g **header:** idno **iso-fs:** fLib fs fvLib **linking:** alt altGrp anchor join joinGrp link linkGrp seg timeline **msdescription:** catchwords depth dim dimensions height heraldry locus locusGrp material objectType origDate origPlace secFol signatures stamp watermark width **namesdates:** addName affiliation bloc climate country district forename genName geo geogFeat geogName location nameLink offset orgName persName placeName population region roleName settlement state surname terrain trait **spoken:** incident kinesic pause shift vocal writing **tagdocs:** att code gi ident specDesc specList tag val **textcrit:** app witDetail **textstructure:** docAuthor **transcr:** addSpan am damage damageSpan delSpan ex fw handShift listTranspose metamark mod redo restore retrace space subst substJoin supplied surplus undo **verse:** caesura rhyme

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html

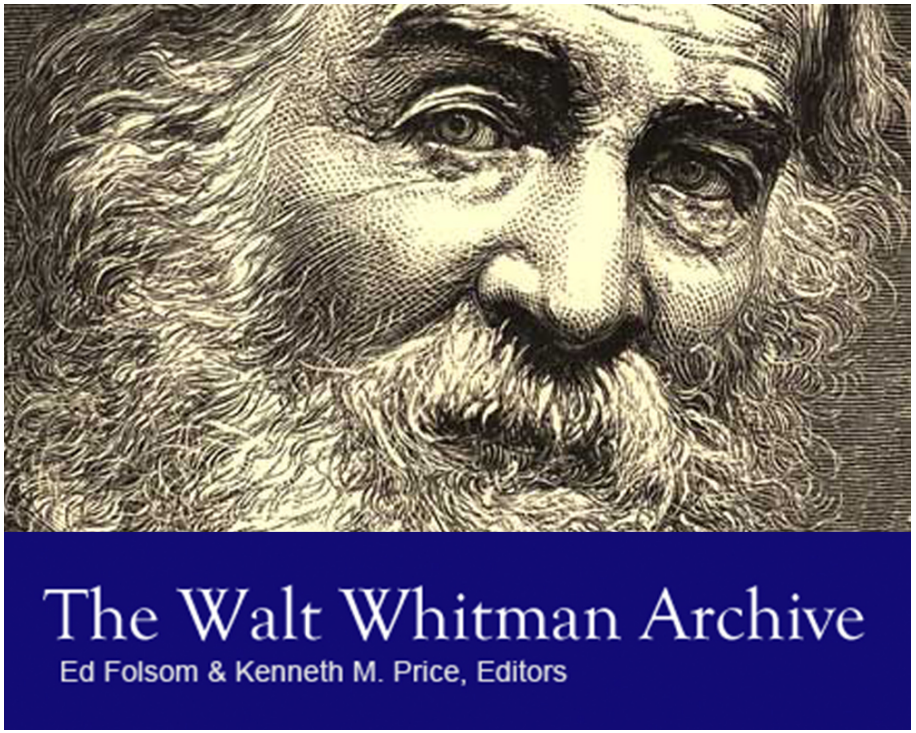
WHEN IS TEI KNOWLEDGE USEFUL? AGGREGATION, PEER REVIEW, GRANTS

- Aggregation and sharing
- Peer Review (by organizations such as NINES)
- Grants (agencies like to see that projects will use disciplinary standards)

TEI-BASED PROJECTS: THE MAP OF EARLY MODERN LONDON

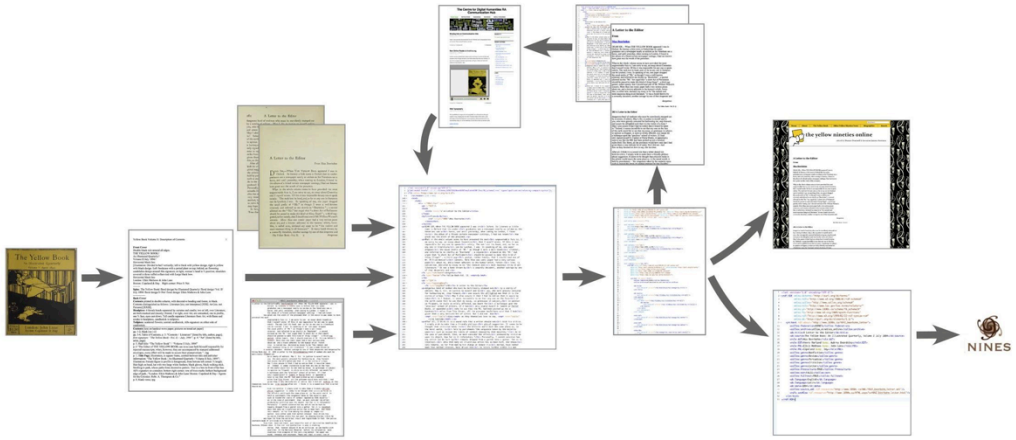
MoEML: http://mapoflondon.uvic.ca/index.htm
Let's look at their code:
http://mapoflondon.uvic.ca/dev/codesharing.htm

TEI-BASED PROJECTS: WALT WHITMAN ARCHIVE



http://www.whitmanarchive.org/

TEI-BASED PROJECTS: THE YELLOW NINETIES ONLINE



The Yellow Book

Bibliographic audit
a detailed description of the contents and physical features

PDF
a reference for italics, line breaks, page breaks, and catchwords

OCR
a plaintext version of the volume

Communication Hub
tracks documents' movement through the workflow; includes training material, revision lists, and weekly updates

TEI
converted to HTML to aid revision

Extensible Stylesheet Language documents
small custom programs that convert TEI into other formats

HTML
displays The Yellow Nineties content online

RDF
documents that make the The Yellow Nineties searchable through the Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES)

http://www.1890s.ca/

FURTHER RESOURCES: SUBJECT-BASED TEI COMMUNITY

Get in touch with TEI-encoding groups who share your research subject area

NINES - 19thc Scholarship
18thConnect - 18thC scholarship
MESA - Medieval scholarship
Modernist Journals - Modernism
CWRC - Canadian Lit studies
ModNets - Modernist Networks

Looking to host an MVP version as you learn/ move from institution to institution while on the job market/ generally get things underway? Try TAPAS, the TEI Archiving, Publishing, and Access Service.

FURTHER RESOURCES

Online Resources

- WWP web site: Encoding Guide and seminar materials
- TEI Guidelines and web site
- TEI-L mailing list and its archives
- WWP-ENCODING mailing list and its archives
- TEI by Example
- Digital Humanities Questions and Answers
- TAPAS, the TEI Archiving, Publishing, and Access Service

Events

- WWP Workshops in Digital Humanities, Northeastern University
- Digital Humanities conference (this year late June to early July in Sydney, Australia)
- Digital Humanities Summer Institute, University of Victoria
- Humanities Intensive Learning and Teaching, University of Maryland
- Nebraska Digital Workshop, University of Nebraska
- Rare Book School, University of Virginia
- *Balisage*: The Markup Conference, annually in early August (NB a student support award)
- TEI conference (this year 26–31 Oct in Lyon, France)
- THATCamps

THANKS!

Keep in touch!

constance.crompton@ubc.ca |@clkcrompton

With special thanks to Michelle Schwartz, Julia Flanders, Syd Bauman, Lee Zickel, Travis White, Raymon Sandhu, Seamus Riordan-Short, Arlene Johnson, WestGrid, Compute Canada, and the Social Science and Humanities Research Council of Canada