Text Encoding Initiative Workshop: Intro to Text Encoding

Michelle Dalmau & John Walsh, Indiana University
Catapult / Scholars' Commons Production

Slides, Exercises & Additional Primers

- http://dcl.slis.indiana.edu/teiworkshop/

Overview: Introduction to TEI

- Introduction to text encoding
  — the what and why
- Introduction to the Text Encoding Initiative / TEI
  — the how

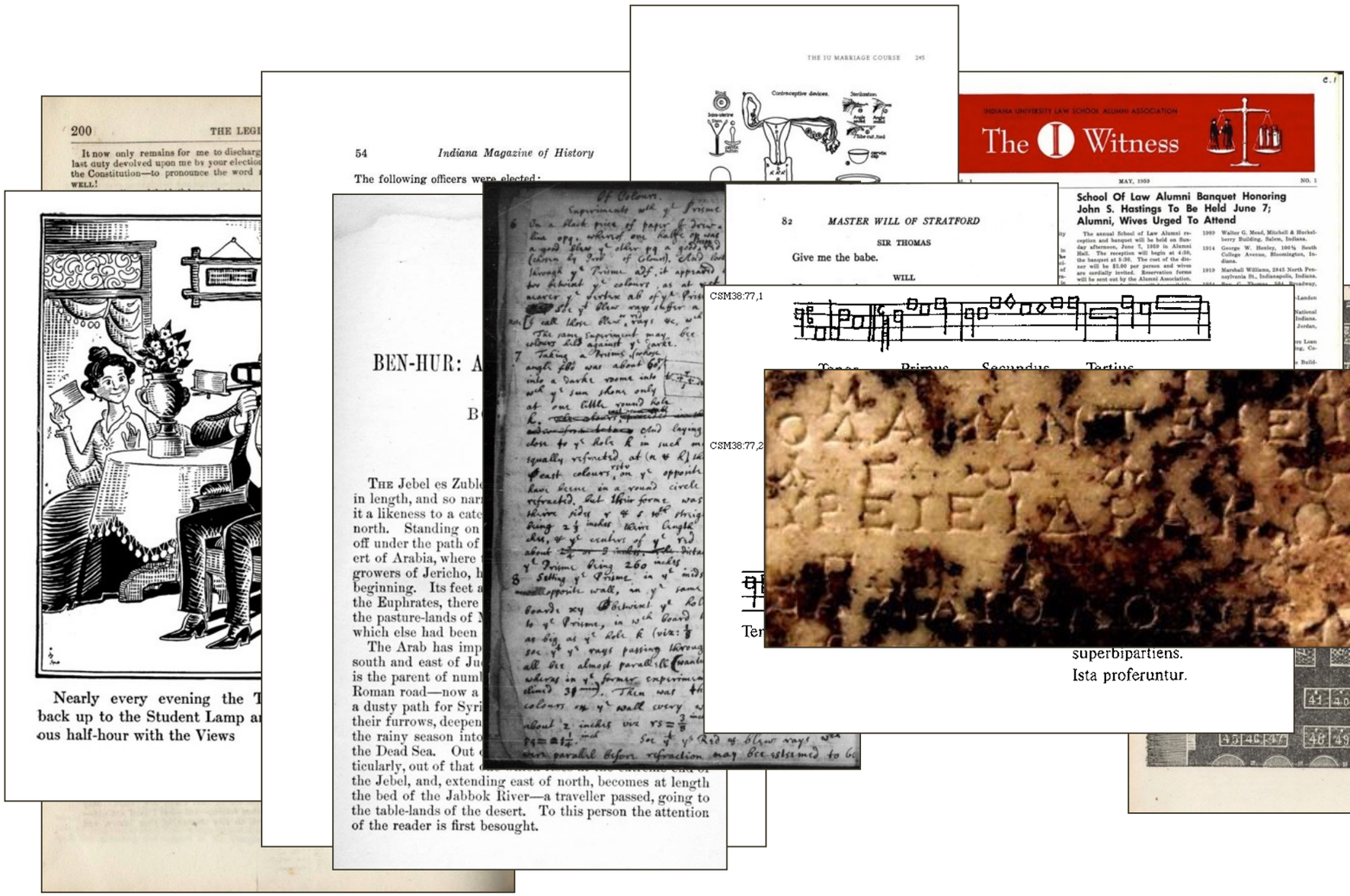
Motivations for Text Encoding

- Access & preservation
- Discovery & dissemination
  - Searching/browsing
  - Interoperability & portability: harvesting/repurposing
- Analysis
  - Linguistic analysis
  - Concordances
  - Topic models
- Visualization
  - Interactive timelines (see VWWP)
  - Map-based interfaces (see Swinburne Project)

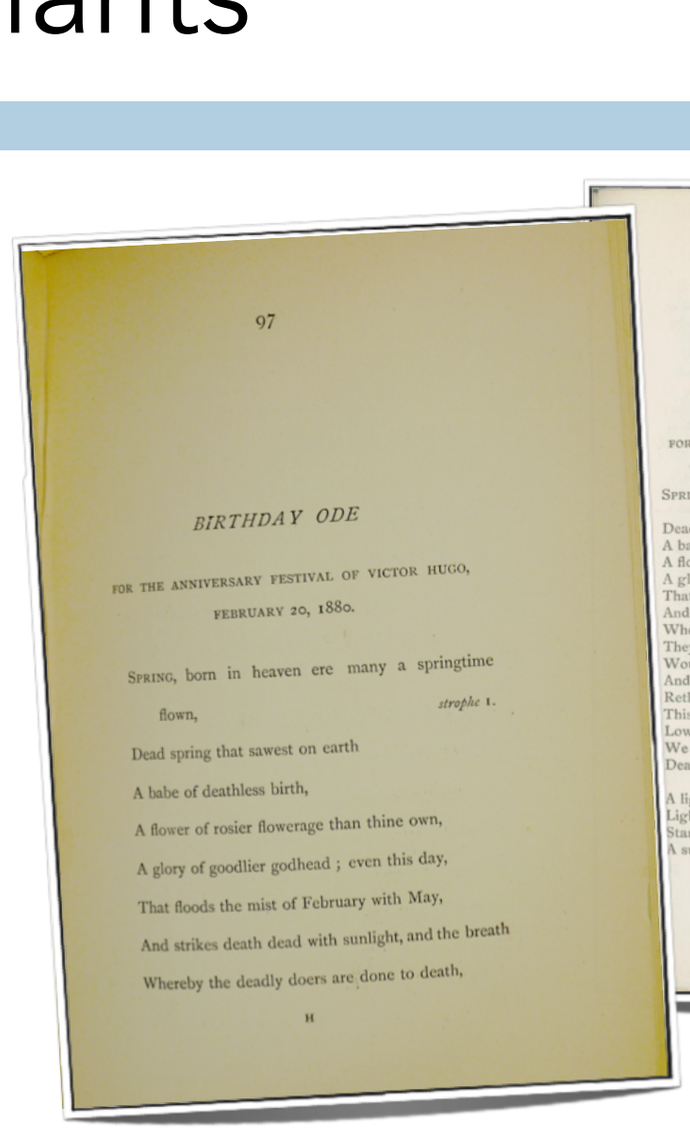
Representing the Text with Encoding

- Structural features
  - Text divisions (chapters, sections, etc.), paragraphs, lists, tables, line groups, lines, etc.
- Content & context
  - Metadata for the electronic and for the source document
  - References to people, places, events, organizations, etc. within the text (phrase-level)
  - Thematic and interpretive annotation
- Formatting & design
  - Bold, italics, small case, indentations, color, dimensions, binding, watermarks, and other features of the material document

What is a Textual Document?



Variants



97

BIRTHDAY ODE

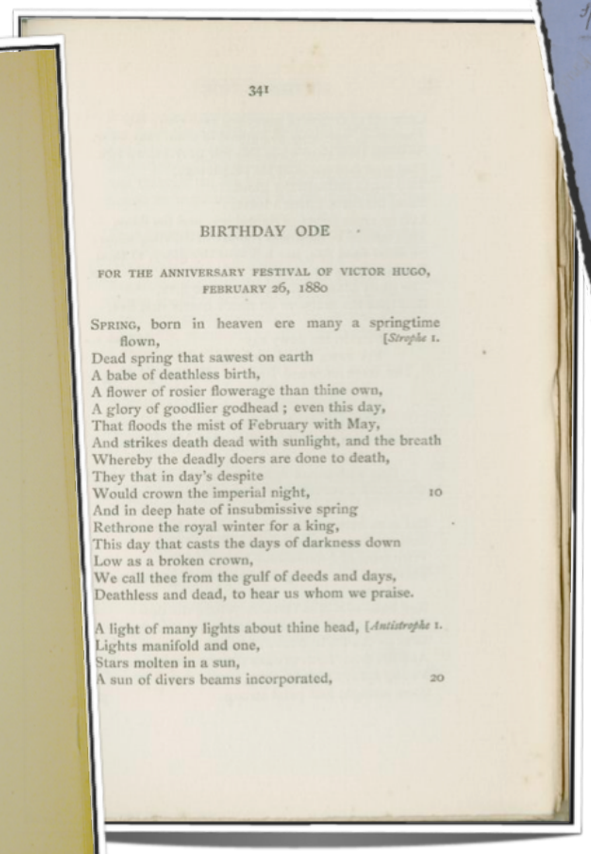
FOR THE ANNIVERSARY FESTIVAL OF VICTOR HUGO,

FEBRUARY 20, 1880.

SPRING, born in heaven ere many a springtime flown, *strophe* 1.
Dead spring that sawest on earth
A babe of deathless birth,
A flower of rosier flowerage than thine own,
A glory of goodlier godhead ; even this day,
That floods the mist of February with May,
And strikes death dead with sunlight, and the breath
Whereby the deadly doers are done to death,

H

Swinburne's *Songs of the Springtides* (1880)



341

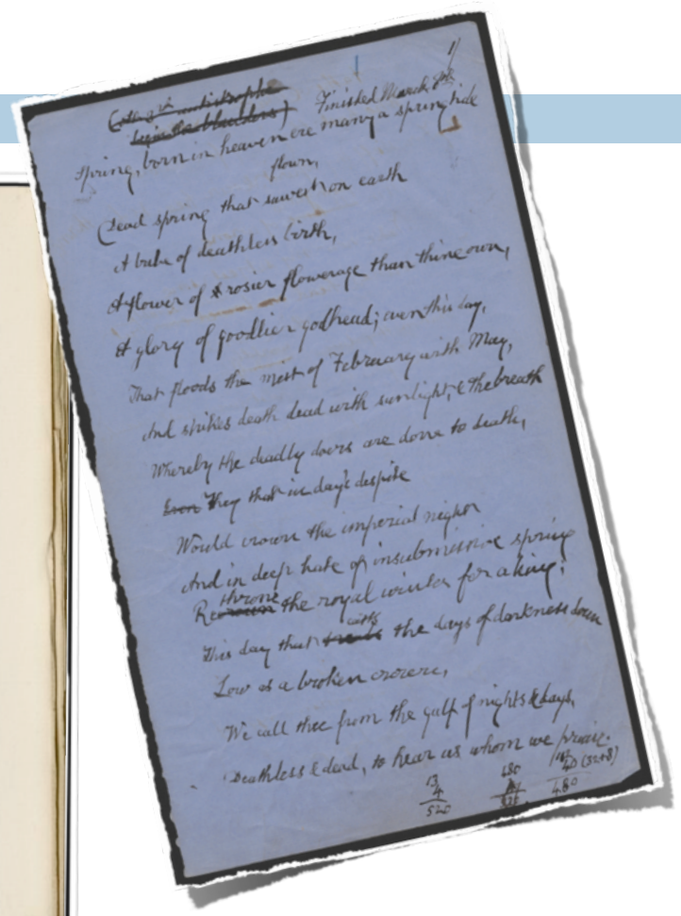
BIRTHDAY ODE

FOR THE ANNIVERSARY FESTIVAL OF VICTOR HUGO,
FEBRUARY 26, 1880

SPRING, born in heaven ere many a springtime flown, [*Strophe* 1.
Dead spring that sawest on earth
A babe of deathless birth,
A flower of rosier flowerage than thine own,
A glory of goodlier godhead ; even this day,
That floods the mist of February with May,
And strikes death dead with sunlight, and the breath
Whereby the deadly doers are done to death,
They that in day's despite
Would crown the imperial night,
And in deep hate of insubmissive spring
Rethrone the royal winter for a king,
This day that casts the days of darkness down
Low as a broken crown,
We call thee from the gulf of deeds and days,
Deathless and dead, to hear us whom we praise.

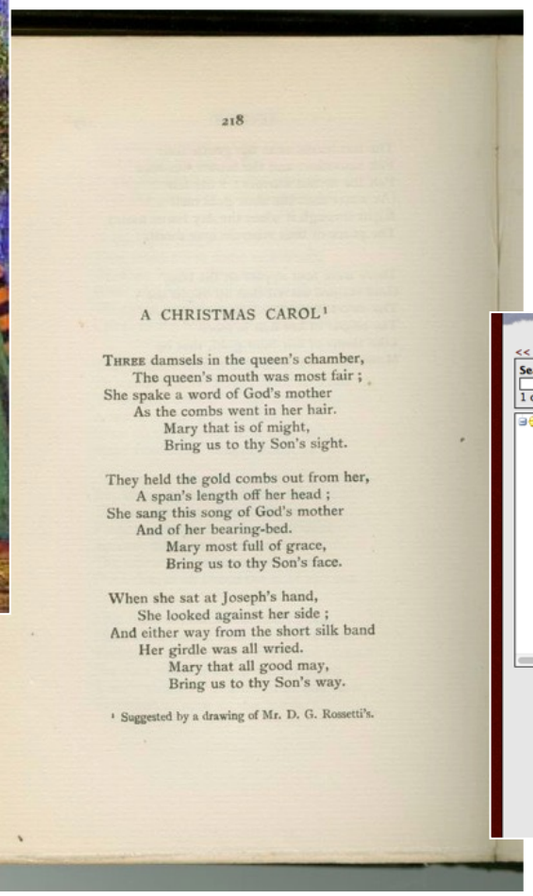
A light of many lights about thine head, [*Antistrophe* 1.
Lights manifold and one,
Stars molten in a sun,
A sun of divers beams incorporated,

Swinburne's *Poems* (1904)



MS. Special Collections Research Center. Syracuse University Library

Intertextual and Contextual Information



218

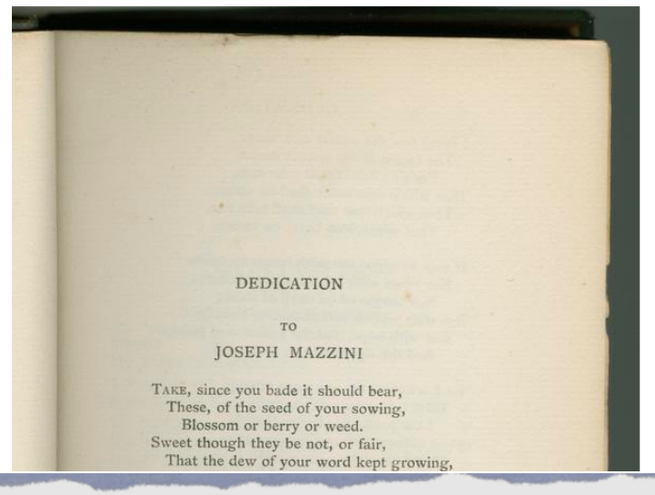
A CHRISTMAS CAROL[1]

THREE damsels in the queen's chamber,
The queen's mouth was most fair;
She spake a word of God's mother
As the combs went in her hair.
Mary that is of might,
Bring us to thy Son's sight.

They held the gold combs out from her,
A span's length off her head;
She sang this song of God's mother
And of her bearing-bed.
Mary most full of grace,
Bring us to thy Son's face.

When she sat at Joseph's hand,
She looked against her side;
And either way from the short silk band
Her girdle was all wried.
Mary that all good may,
Bring us to thy Son's way.

[1] Suggested by a drawing of Mr. D. G. Rossetti's.

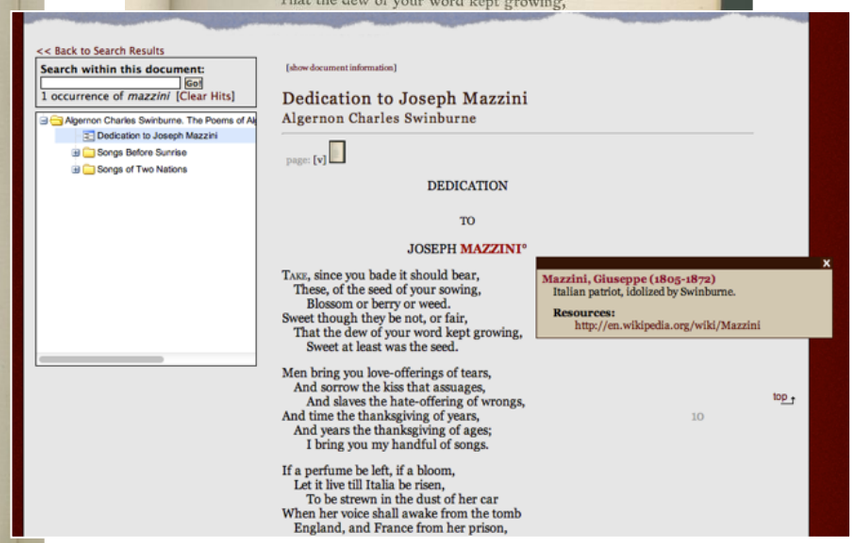


DEDICATION

TO

JOSEPH MAZZINI

TAKE, since you bade it should bear,
These, of the seed of your sowing,
Blossom or berry or weed.
Sweet though they be not, or fair,
That the dew of your word kept growing,



Advantages of Text Encoding

- Re-use and flexibility: build once, use many
- Presentation and output of text controlled by style sheets.
  - Generate different views of the same text and different formats: PDF, HTML, ePub (ebooks), plain text (for text analysis), etc.
- The document *and* the markup can serve as an object of analysis and increased discoverability

Interpretative & Project-specific Encoding

- Text encoding is not necessarily simple data entry/capture; it is not objective but interpretive. Every encoded text is a "reading" of the text.
- There are often many ways to apply a particular markup language to a particular text.
- Individual projects typically require project-specific guidelines and documentation in addition to the general markup language specification or guidelines.

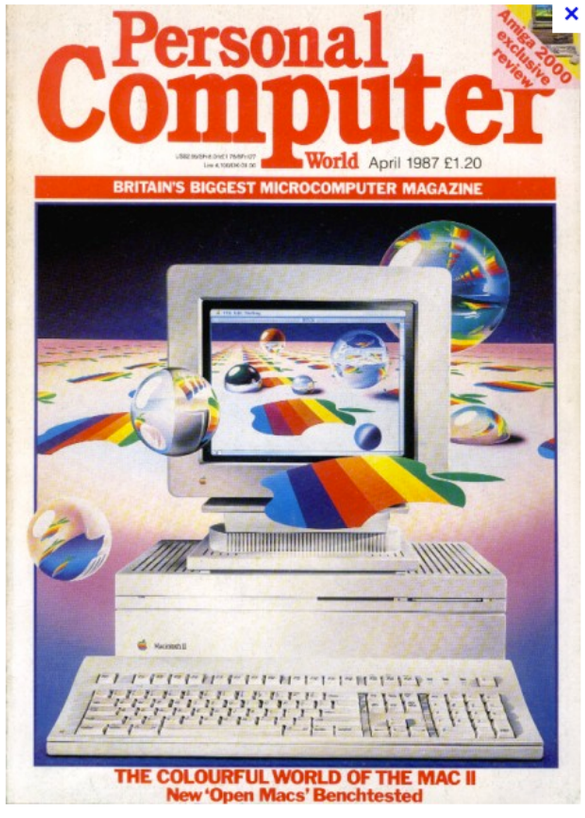
Intro to the Text Encoding Initiative (TEI)

- TEI is:
  - a formally constituted organization, the TEI Consortium;
  - a scholarly community—with an annual conference, open-access journal, and active email discussion list.
  - a text encoding standard produced by that organization, TEI's Guidelines for Electronic Text Encoding and Interchange.
- For our purposes, TEI refers to the technical text encoding standard.

History of TEI

Vanhoutte, Edward. "An Introduction to the TEI and the TEI Consortium." *Literary and Linguistic Computing 19*.1 (2004): 9-16.

It all started in 1987 …



Questions?

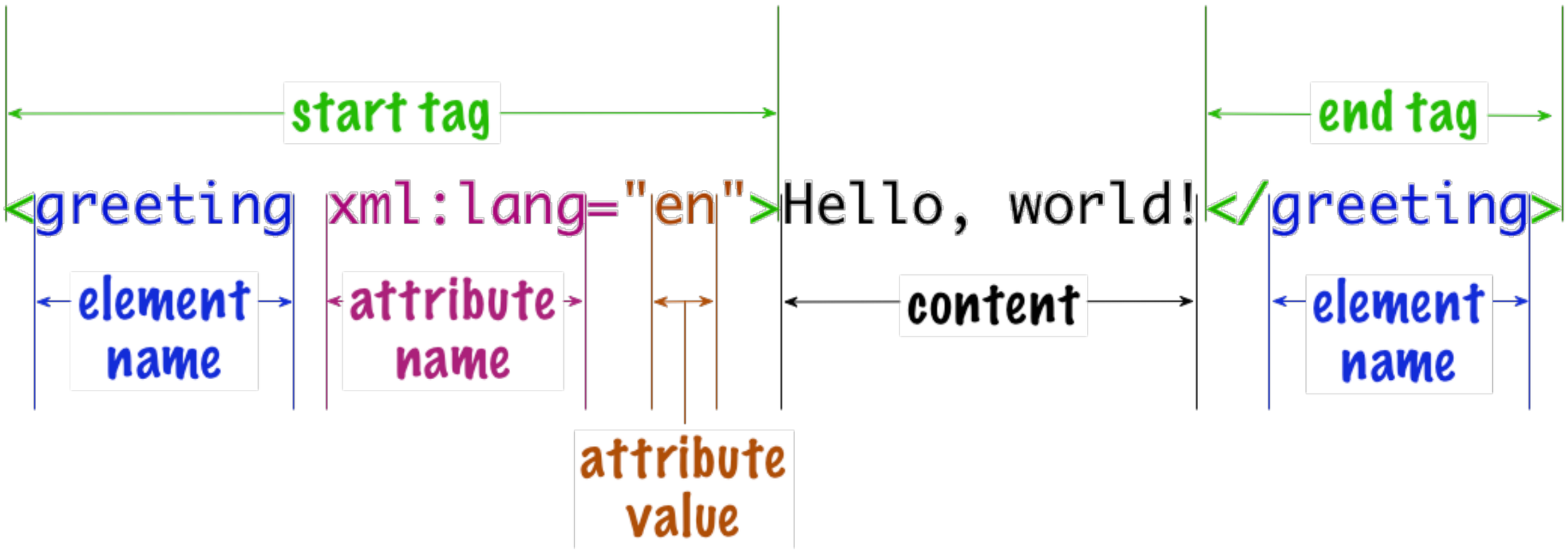
Quick Introduction to XML

- XML, or eXtensible Markup Language, is a **non-proprietary meta language** for creating markup languages suited for different tasks, domains, and disciplines.
- An XML markup language consists of "tags" used to define the structure and other features of a text.
  - HTML:
    - (paragraph of text)
    -
    - Indiana University
  - TEI:
    - <sp who="#rosamond"> (speech) </sp>
    - <lg> (line group, stanza) </lg>
    - <salute>Dear Fred,</salute>

XML Key Terms

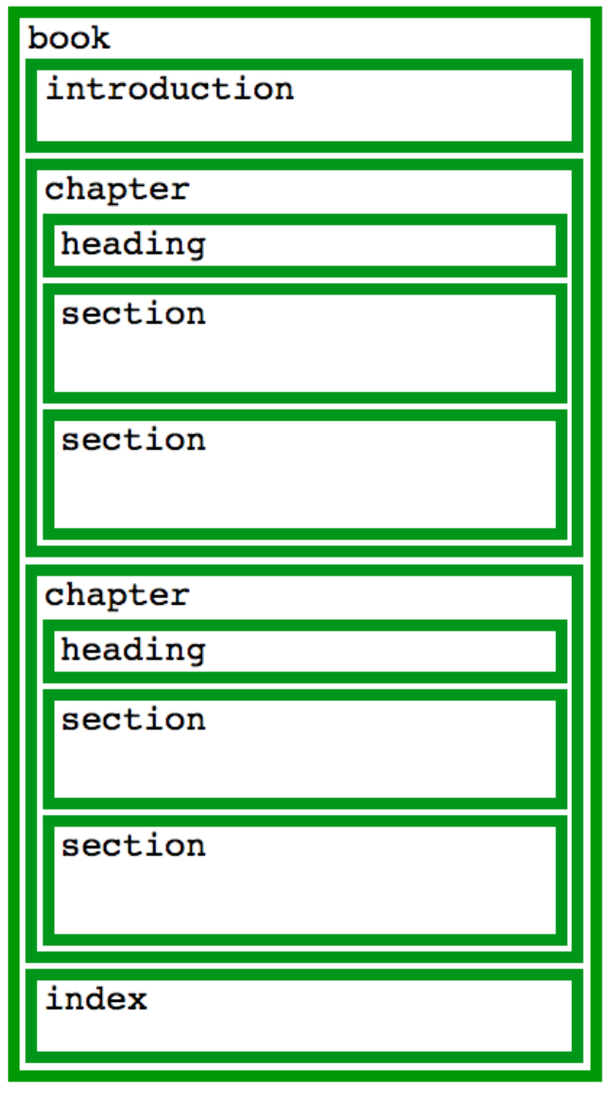
- **Elements** are the basic, named structural units of an XML document (nouns of encoding)
  - <title>The Odyssey</title>
- **Attributes** are name/value pairs (name="value") associated with elements (adjectives of encoding)
  - <creator type="author">Homer</creator>
  - An element may have multiple attributes
- **DTDs** (Document Type Definitions) and Schemas define the rules that govern a particular type of XML document. They declare elements and attributes and the allowable content for those elements and attributes (grammar rules)

XML: Anatomy of an Element

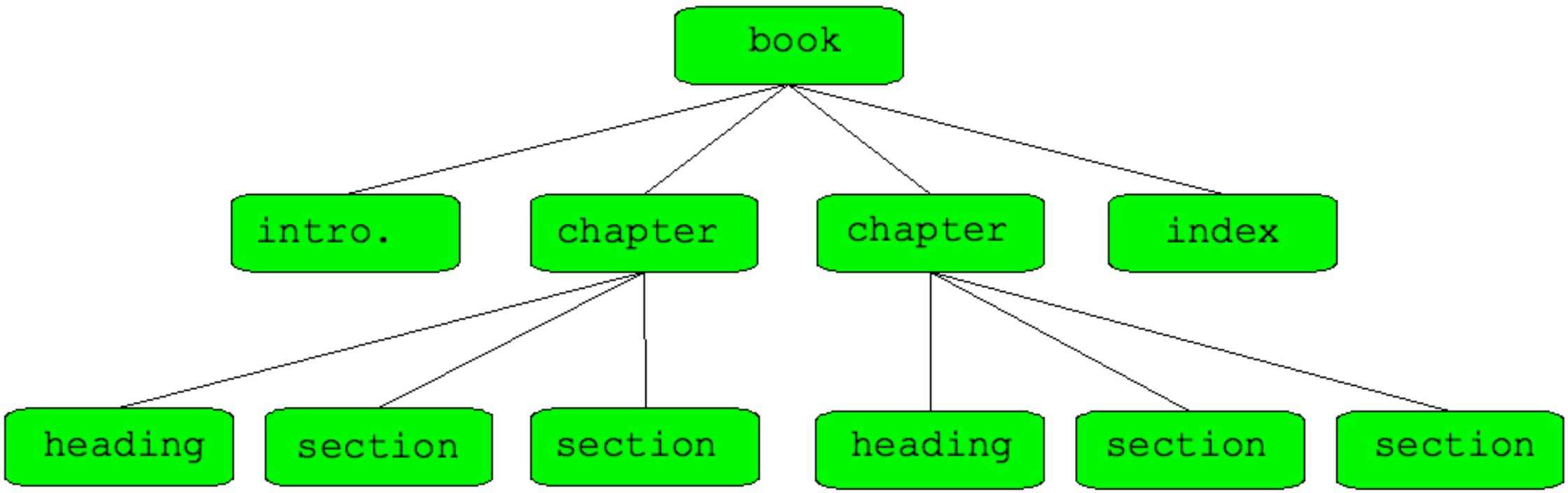


Empty tags or milestone elements: <lb/> = <lb></lb>

XML Representation: Boxes



XML Representation: Tree



XML Representation: Markup

```
<?xml version="1.0" encoding="UTF-8"?>
<book>
   <introduction>Blah blah blah ... </introduction>
   <chapter>
      <heading>Wines</heading>
      <section>White wines ... </section>
      <section>Red wines ... </section>
   </chapter>
   <chapter>
      <heading>Beers</heading>
      <section>Ales ... </section>
      <section>Lagers ... </section>
   </chapter>
   <index> stuff ... </index>
</book>
```

XML: Well-Formed and Valid

- All XML documents need to be well-formed according to some basic rules:
  - Open and close all tags/elements
  - Tags/elements may not overlap
  - Attribute values must be quoted
- XML documents should be valid according to a DTD or Schema:
  - Use the appropriate elements & attributes
  - Adhere to the “grammar rules” (e.g., allowable attributes for elements)
- Software programs help reinforce these principles
  - XML Editors like Oxygen

Questions?

- John Walsh's XML-Primer:
- https://www.youtube.com/watch?v=JhhKyyP0e18

Intro to the TEI Guidelines and Tag Set

- TEI Guidelines: Quick Overview
- TEI P5 Guidelines
- TEI Basic Components
- Basic Markup: Prose
- Basic Markup: Verse
- Basic Markup: Drama
- Basic Markup: Letters

TEI Guidelines: Quick Overview

- Text Encoding Initiative (TEI) / *Guidelines for Electronic Text Encoding and Interchange (TEI)*
- The TEI *Guidelines "are addressed to anyone who works* with any text in electronic form. They provide means of representing those features of a text which need to be identified explicitly in order to facilitate processing of the text by computer programs" (Sperberg-McQueen).
- TEI provides elements, attributes, and other mechanisms for encoding prose, poetry, drama, dictionaries, critical apparatus, linguistic corpora, and other scholarly and non-scholarly texts.

TEI Guidelines: Quick Overview

- The TEI Guidelines:
  - Can be applied lightly or heavily
  - Are designed as a set of modules/mechanisms that can be selected as needed:
    - core: Elements common to all TEI documents
    - figures: Tables, formulae, music notation, and figures
    - linking: Linking, segmentation, and alignment
    - msdescription: Manuscript description
    - namesdates: Names and dates
  - Can adapt to local conditions

TEI P5 Guidelines

- P5 Guidelines: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html
  - Prose documentation with examples
- P5 Tag/Element Set: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html
  - Listing of the tag set with examples and relevant links to prose documentation

TEI P5: Basic Components

- **<TEI>**: The root element of a TEI document
  - **<teiHeader>**: The metadata header for a TEI document. Includes bibliographic, technical, administrative, and other metadata about the digital file and the analog source, if one exists.
  - **<text>**: The text itself, e.g., the title page and chapters of a novel, the acts and scenes of a drama, the books or cantos of a long poem. The <text> element is further subdivided into:
    - **<front>**: Front matter, e.g, the title page(s), table of contents, potentially a preface or dedication
    - **<body>**: The main body of a document, excluding front and back matter
    - **<back>**: Back matter, e.g., indices, appendices

TEI P5: Basic Markup: Prose

- ** **: (division) is used for basic structural divisions of a text, e.g, volumes, chapters, sections, cantos, tables of contents, indices, appendices, etc. The @type attribute may be used to designate the type of the division.
  - …
  - …
  - <div type=”contents">…
  - …
- **<head>**: (heading) contains any type of heading, for example the title of a section, or the heading of a list, figure, table, etc.
- **<p>**: (paragraph)
- **<pb>:** (page break) marks the boundary between one page of a text and the next

TEI P5: Basic Markup: Prose

Chapter 1: The Manor House

Charles hadn't visited the manor house since Easter,1955, and now he remembered why. "Hullo", he called out as he walked up the drive, and then, as if to himself, "To be or not to be?, to walk or not to walk...oh, **hang** it all!" His meditation on Hamlet was interrupted as he collided with a peacock. "Sacré bleu!" he exclaimed with irritation, his sang-froid completely deserting him. It was going to be a long week. His catalog of irritations included:

1. The weather
2. The peacocks
3. His meager grasp of French

TEI P5: Basic Markup: Prose

```
<?xml version="1.0" encoding="UTF-8"?>
<div type="chapter">
    <head>Chapter 1: The Manor House</head>
    <p>Charles hadn't visited the manor house since
        Easter,1955, and now he remembered why.</p>
    <p><said>Hullo</said>, he called out as he walked up the
    drive, and then, as if to himself, <said>To be or
     not to be?, to walk or not to walk...oh,
     <emph rendition="#b">hang</emph> it all!</said>
    His meditation on Hamlet was interrupted as he
    collided with a peacock. <said xml:lang="fr">Sacré
    bleu!</said> he exclaimed with irritation, his
    <foreign xml:lang="fr">sang-froid</foreign> completely deserting him.
    It was going to be a long week. His catalog of irritations included:
        <list type="ordered">
        <item>The weather</item>
        <item>The peacocks</item>
        <item>His meager grasp of French</item>
    </list>
    </p>
</div>
```

TEI P5: Basic Markup: Verse/Poetry

- <lg>: (line group) contains a group of verse lines functioning as a formal unit, e.g. a stanza, refrain, verse paragraph, etc. The @type and @subtype attributes may be used to classify the type of line group
- <l>: (line) contains a line of verse

TEI P5: Basic Markup: Poetry/Verse

THE ROUNDEL

A ROUNDEL is wrought as a ring or a starbright sphere,
With craft of delight and with cunning of sound unsought,
That the heart of the hearer may smile if to pleasure his ear
A roundel is wrought.

Its jewel of music is carven of all or of aught—
Love, laughter, or mourning—remembrance of rapture or fear—
That fancy may fashion to hang in the ear of thought.

As a bird's quick song runs round, and the hearts in us hear
Pause answer to pause, and again the same strain caught,
So moves the device whence, round as a pearl or tear,
A roundel is wrought.

TEI P5: Basic Markup: Poetry/Verse

```xml
<?xml version="1.0" encoding="UTF-8"?>
<div type="poem">
    <head rendition="#center #uppercase">The Roundel</head>
    <lg>
        <l><hi rendition="#small-caps">A roundel</hi> is wrought as a ring or a starbright sphere,</l>
        <l>With craft of delight and with cunning of sound unsought,</l>
        <l>That the heart of the hearer may smile if to pleasure his ear</l>
        <l rendition="#l-indent-03">A roundel is wrought.</l>
    </lg>
    <lg>
        <l>Its jewel of music is carven of all or of aught–</l>
        <l>Love, laughter, or mourning–remembrance of rapture or fear–</l>
        <l>That fancy may fashion to hang in the ear of thought.</l>
    </lg>
    <lg>
        <l>As a bird's quick song runs round, and the hearts in us hear</l>
        <l>Pause answer to pause, and again the same strain caught,</l>
        <l>So moves the device whence, round as a pearl or tear,</l>
        <l rendition="#l-indent-03">A roundel is wrought.</l>
    </lg>
</div>
```

TEI P5: Basic Markup: Drama

- **<sp>**: (speech) contains individual speech in a performance text, or a passage presented as such in a prose or verse text.
- **<speaker>**: contains a specialized form of heading or label, giving the name of one or more speakers in a dramatic text or fragment.
- **<stage>**: (stage direction) contains any kind of stage direction within a dramatic text or fragment.

TEI P5: Basic Markup: Drama

Scene 1

Enter Fay

Fay:

I say, Dinah, has anyone seen my gloves?

Enter Dinah

Dinah:

No, miss, perhaps the parakeet has got them again?

Exit Fay and Dinah

TEI P5: Basic Markup: Drama

```xml
<?xml version="1.0" encoding="UTF-8"?>
<div type="scene">
    <head rendition="#center">Scene 1</head>
    <stage rendition="#i">Enter Fay</stage>
    <sp>
        <speaker>Fay:</speaker>
        <p>I say, Dinah, has anyone seen my gloves?</p>
    </sp>
    <stage rendition="#i">Enter Dinah</stage>
    <sp>
        <speaker>Dinah:</speaker>
        <p>No, miss, perhaps the parakeet has got them again?</p>
    </sp>
    <stage rendition="#i">Exit Fay and Dinah</stage>
</div>
```

TEI P5: Basic Markup: Letters

- **<opener>:** groups together dateline, byline, salutation, and similar phrases appearing as a preliminary group at the start of a division, especially of a letter.
- **<closer>:** groups together dateline, byline, salutation, and similar phrases appearing as a final group at the end of a division, especially of a letter.
  - **<dateline>:** contains a brief description of the place, date, time, etc. of production of a letter, prefixed or suffixed to it as a kind of heading or trailer.
  - **<salute>:** (salutation) contains a salutation or greeting in the closing of a letter, preface, etc.
  - **<signed>:** (signature) contains the closing salutation

TEI P5: Basic Markup: Letters

1906 August the 5th

Cape Cod

My dear Becky

How lovely the oysters are this evening!

Yours very truly

Maria

TEI P5: Basic Markup: Letters

```
<div type="letter">
    <opener>
        <dateline>
            <date when="1906-08-05">1906 August the 5th</date>
            <lb/>
            Cape Cod
        </dateline>
        <salute>
            My Dear Becky
        </salute>
    </opener>
    <p>How lovely the oysters are this evening!</p>
    <closer>
        <salute>Yours very truly</salute>
        <signed>Maria</signed>
    </closer>
</div>
```

Hands-on Exercises: Basic Genres

- http://dcl.slis.indiana.edu/teiworkshop/
- Launch Oxygen
- Complete exercises one at a time: Prose, Verse, Drama and Letters
- Encode a short document of your own choosing