XML, TEI, & Enrichment

Editing and reconciliation

XML: The Basics

- XML (eXtensible Markup Language) was designed to store and transport data

- XML contains **headers** that declare versions and encoding standards

- All XML **elements** must have opening <> and closing </> **tags**

- XML requires **validation,** which makes it fragile. On the plus side, it makes sure your data conforms to XML standards.

- XML is flexible and can express different, complex and nested structures

A (very!) simple XML record

example.xml

```xml
<?xml version ="1.0" encoding="UTF-8"?>
<note>
    <to>Sean</to>
    <from>Carolyn</from>
    <heading>Reminder</heading>
    <body>Don't forget about DHSI next week!</body>
</note>
```

TEI: The Basics

- TEI (Text Encoding Initiative) is a standard used to **represent texts** in **digital form.** It specifies encoding methods to make texts machine-readable.
- TEI uses **XML** standards. It is **hierarchical.**
- A TEI record includes a **header** and **body**, which have **elements** and **attributes**
- The latest version of the **TEI guidelines** is available at: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf

The Bierce Letters Dataset

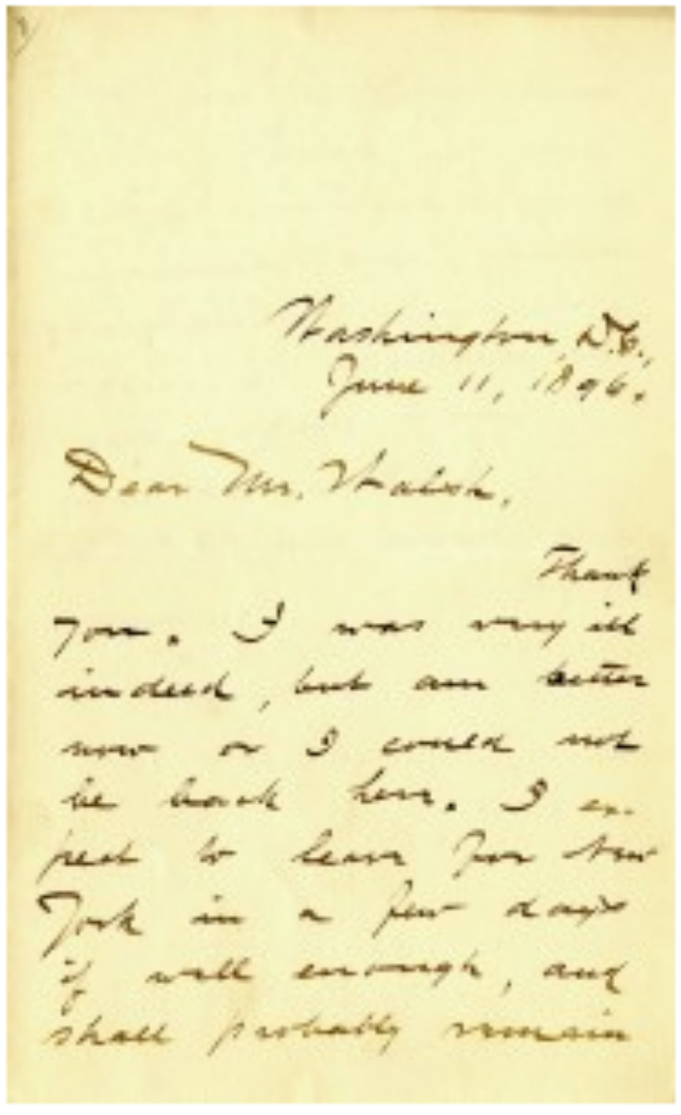
- For our TEI exercises, we'll be working with letters held by the University of Cincinnati's Archives & Rare Books Library

- The letters were written by Bierce to his friend Myles Walsh and Walsh's sister Lily, a young deaf-mute woman

- For more information on the collection, see: http://libapps.libraries.uc.edu/exhibits/bierce/

- Transcribed versions of the letters are available online; Carolyn took 10 of these and created very basic TEI records

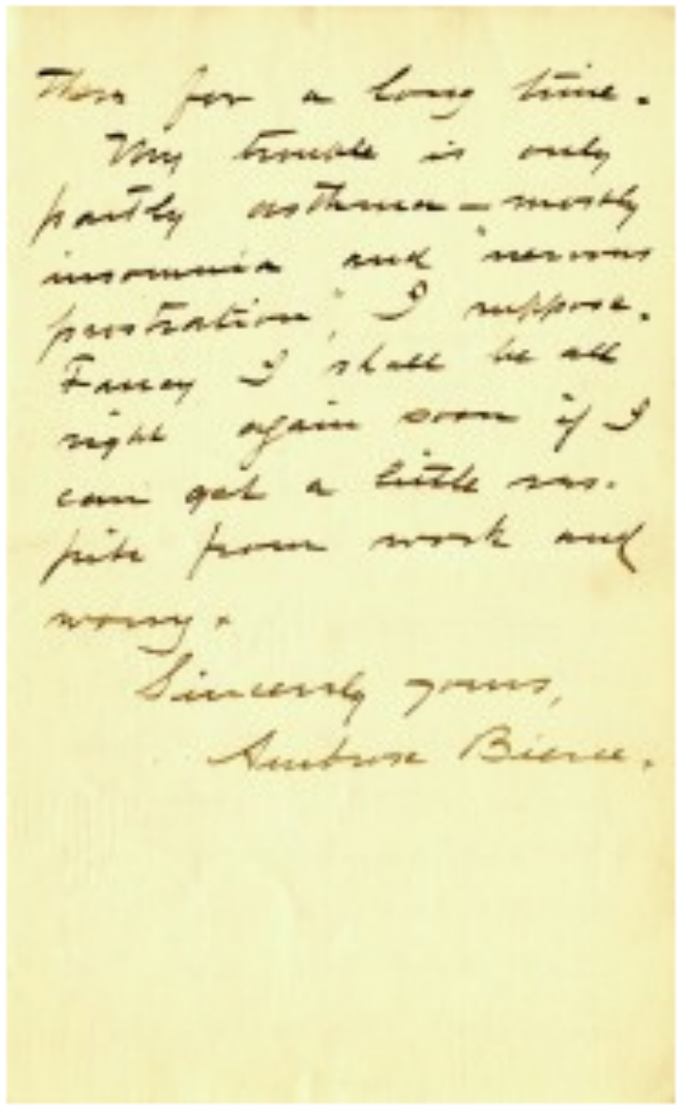
THE LETTERS



Washington D.C.
June 11, 1896.

Dear Mrs. Walsh,

Thank
you. I was very ill
indeed, but am better
now & I count not
be back here. I ex-
pect to leave for New
York in a few days
if well enough, and
shall probably remain



here for a long time.
My trouble is only
partly asthma — mostly
insomnia and nervous
prostration, I suppose.
Fancy I shall be all
right again soon if I
can get a little rest.
pile from work and
worry.

Sincerely yours,
Ambrose Bierce.

Letter 01: Bierce to Elizabeth (Lily) Walsh, undated. (Letter) (Transcription)

Letter 02: Bierce to Elizabeth (Lily) Walsh, October 13, 1895. (Letter) (Transcription)

Working with TEI documents

- Investigate / Visualize
- Transform / Enrich
- Export

https://en.wikipedia.org/wiki/File:Mouse_Trap_Board_and_Boxjpg.jpg

- TEI/XML can be viewed and manipulated with a host of different general or purpose-built tools.

- Oxygen, various text-editors and … OpenRefine!

- We’re going to get our hands dirty with some TEI/XML manipulation

OpenRefine with nested data

- OpenRefine is a great tool for tabular data

- But…

But…

- In most cases, hierarchical data does not easily translate to tabular format without introducing some kind of relational component

- Our TEI documents are pretty deeply hierarchical (most xml is).

- Things may get messy

Exercise 1:

Investigating TEI with OpenRefine

OpenRefine with XML

- OpenRefine **can** import and display hierarchical XML
- OpenRefine **cannot** export hierarchical XML (at least not where you think).

<http://arresteddevelopment.wikia.com/wiki/Teamocil>

OpenRefine with XML

- In our next exercises, we want to enrich our TEI document with some URIs, using OpenRefine for reconciliation

- We can get the TEI into OpenRefine but we can't get it out

OpenRefine with XML

- But we're scrappy, and we don't take no for an answer

- Let's make it work

Exercise 2:

Reconcile and export data

Exercise 3:

Enrich TEI with reconciled data points

Including external metadata

- So we've added some URIs
- We don't need to stop there
- Metadata is everywhere (including the TEIHeader)
- <xenodata> https://github.com/TEIC/TEI/issues/453

Exercise 4:

Migrate TEI header data to RDF

TEI Header Data as Dublin Core

<rdf:Description rdf:about="http://localhost:3333/0">
<rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Description"/>
<dc:title>Letter 01: Bierce to Elizabeth (Lily) Walsh, undated</dc:title>
<dc:publisher>University of Cincinnati</dc:publisher>
<dc:creator>Archives and Rare Book Library</dc:creator>
<dc:date>2015</dc:date>
<dc:description>Letters of Ambrose Bierce to Myles Walsh, 1895-1911</dc:description>
</rdf:Description>