Digital Humanities Observatory
Ireland's window on humanities e-scholarship

A project of the RIA ROYAL IRISH ACADEMY ACADAMH RÍOGA NA HÉIREANN

Brief Introduction to TEI

Kevin S. Hawkins
k.hawkins@dho.ie

15.01.2010 • University College Cork, Ireland

~ FUNDED UNDER PRTLI . CYCLE 4 2008 - 2011 ~

Our objectives: to …

1. Learn about digital text
2. Learn about the Text Encoding Initiative (TEI)
3. Create a TEI Lite document of a W.B. Yeats poem
4. Consider more complicated texts
5. See TEI encoding in action
6. Find out how to learn more

raster images vs. vector images

proprietary vs. non-proprietary formats

closed vs. open standards

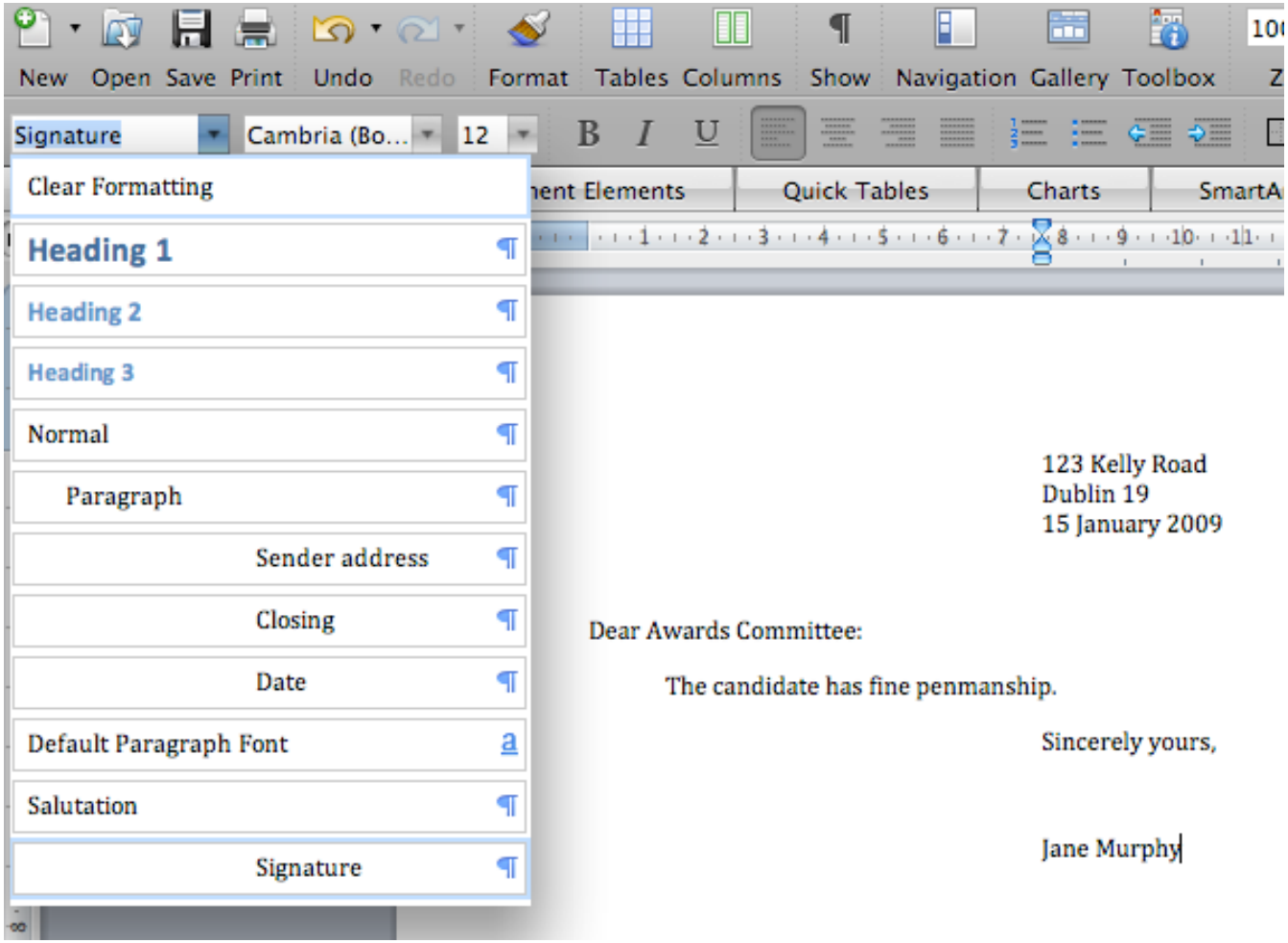
Plain text isn’t good enough

```
123 Kelly Road
Dublin 19
15 January 2009

Dear Awards Committee:

The candidate has fine penmanship.

Sincerely yours,

Jane Murphy
```

What if you want to …

- Publish a collection of letters and decide after beginning that you want to have the sender's address and closing always right-aligned?
- Search your collection of letters to extract a list of all senders and another list of all recipients?

Word processor styles: create your own!



Extensible Markup Language (XML): word processor styles on steroids

- Can have styles for spans of text, not just whole blocks.
- You can give properties to these styles, e.g.,
  - This salutation is formal.
  - This sentence contains sarcasm.
  - This word is misspelled.

XML in brief (1)

- Open, non-proprietary standard
- Stored in plain text but usually thought of as contrasting with it (as above)
- Marks beginning and ends of spans of text using tags:

<sentence>This is a sentence.</sentence>

XML in brief (2)

- Spans of text must nest properly:

Wrong:
<sentence>Overlap is <emphasis>not allowed!</sentence></emphasis>

Right:
<sentence>Overlap is <emphasis>not allowed!</emphasis></sentence>

Elements (tags), attributes, values, content

<sentence type="declarative">This is a sentence.</sentence>

<sentence type="interrogative">Is this is a sentence?</sentence>

Wait, this all looks a lot like HTML!

HTML is a specific implementation of XML (well, actually, its predecessor SGML) that has pre-defined elements and attributes. You can't create your own elements, so its usefulness is limited.

Schemas (DTDs and others)

A syntax for your XML documents, specifying:

- Which elements may nest inside of others
- In what order these elements must occur
- How many times they may repeat
- What attributes they may have
- What values those attributes may have

Why would you want to constrain your document structure like this?

- Prevent errors in creating the XML
- Make it easier to search the text

  Remember we were going to extract names of senders and recipients? You know where to expect to find them within your XML documents.

Exercise 1: Create a schema for a business letter. Start working individually, but we'll finish as a group.

Exercise 2: Encode this letter according to our schema using <oXygen/>.

```
123 Kelly Road
Dublin 19
15 January 2009

Dear Awards Committee:

The candidate has fine penmanship.

Sincerely yours,

Jane Murphy
```

Questions so far?

Brilliant. We can agree on a syntax for our business letters. But how do we keep from reinventing the wheel, and how do we make sure we use the same vocabulary of element names as our colleagues so that we can use each other's texts?

Use an existing schema!

What is the TEI?

The Text Encoding Initiative (TEI) was formed in 1987 by humanities scholars interested in developing guidelines for the interchange of electronic text.

The Guidelines "apply to texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content. They treat both continuous materials ('running text') and discontinuous materials such as dictionaries and linguistic corpora."*

The Guidelines are maintained by the TEI Consortium.

* From chapter vi, About These Guidelines (p. xxiii)

What sorts of textual features can you encode?

- Structural elements

  headings, paragraphs, lines of verse, dialogue, tables, …
- Content features

  foreign words, names, rhyme, …
- Editorial interventions

  insertions, deletions, …
- Linguistic features

  part of speech, syntactic phrase, …
- Physical appearance of a source document

  page breaks, water stains, change in hand of a manuscript, …

Why encode (mark up) texts?

- Making explicit (to the computer) what is implicit to a human reader to aid in searching.

  Extracting names for an index

- Scholars making explicit their interpretation(s) of a text.

  Producing a scholarly edition

- Producing texts that can be interchanged with other projects and reused in the future.

What sorts of things is TEI *not* designed to encode?

- Instructions on how to display the text

  "Make headings bold in 14-point font."

This information is stored separately in a stylesheet.

Must I really label the part of speech of each word in my text?

No! You only bother adding tags that are relevant to your needs.

Besides, you'll quickly discover that markup decisions involve interpretation of the structure of the text. Not everyone agrees on the same way to encode a text.

What else can TEI encoding do for me?

- It has a rich vocabulary of elements for providing metadata about an encoded text and its source (stored in 'the header').
- It can work with other XML vocabularies (for describing non-textual content), so you could even encode figures and equations within your TEI document!

Exercise 3: What would you encode?

You don’t know the names of elements or attributes in TEI yet. But say you were going to use TEI to encode a literary text. What sorts of things would you want to mark?

Choose from the sample documents and discuss in groups.

What is TEI Lite?

"TEI Lite is a specific customization of the TEI tagset, designed to meet '90% of the needs of 90% of the TEI user community'. Due to its simplicity and the fact that it can be learned with relative ease, TEI Lite has been widely adopted, particularly by beginners and by big institutional projects that rely on large teams of encoders to markup their documents."

* From http://www.tei-c.org/Guidelines/Customization/Lite/

Exercise 4

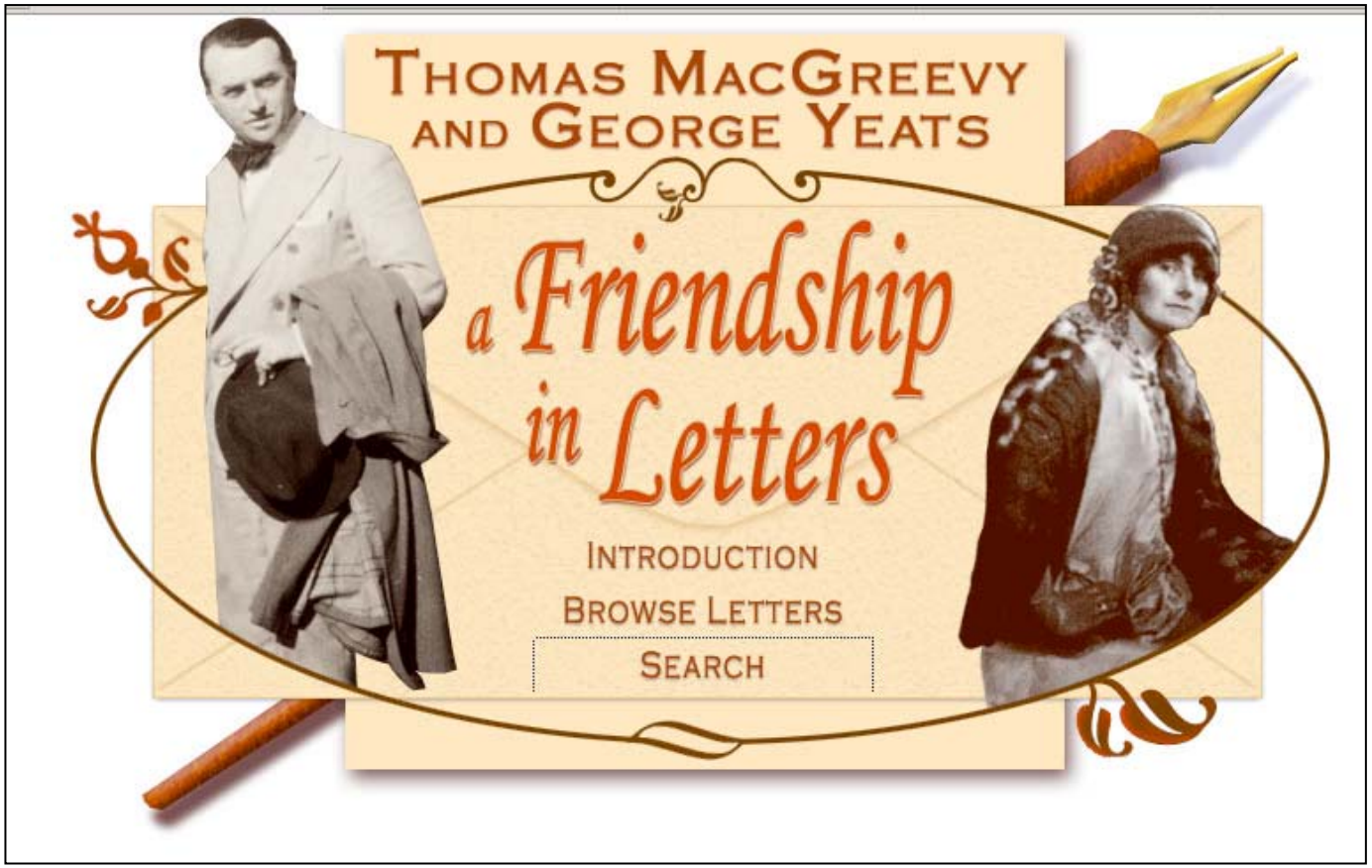
Handout: Encode a Yeats poem using TEI Lite, render it in a web browser and create a PDF of it.

If you finish early:

1. Do the challenge exercise at the end of the handout.
2. Look at the TEI Guidelines to investigate how you would encode the document you looked at in Exercise 3.

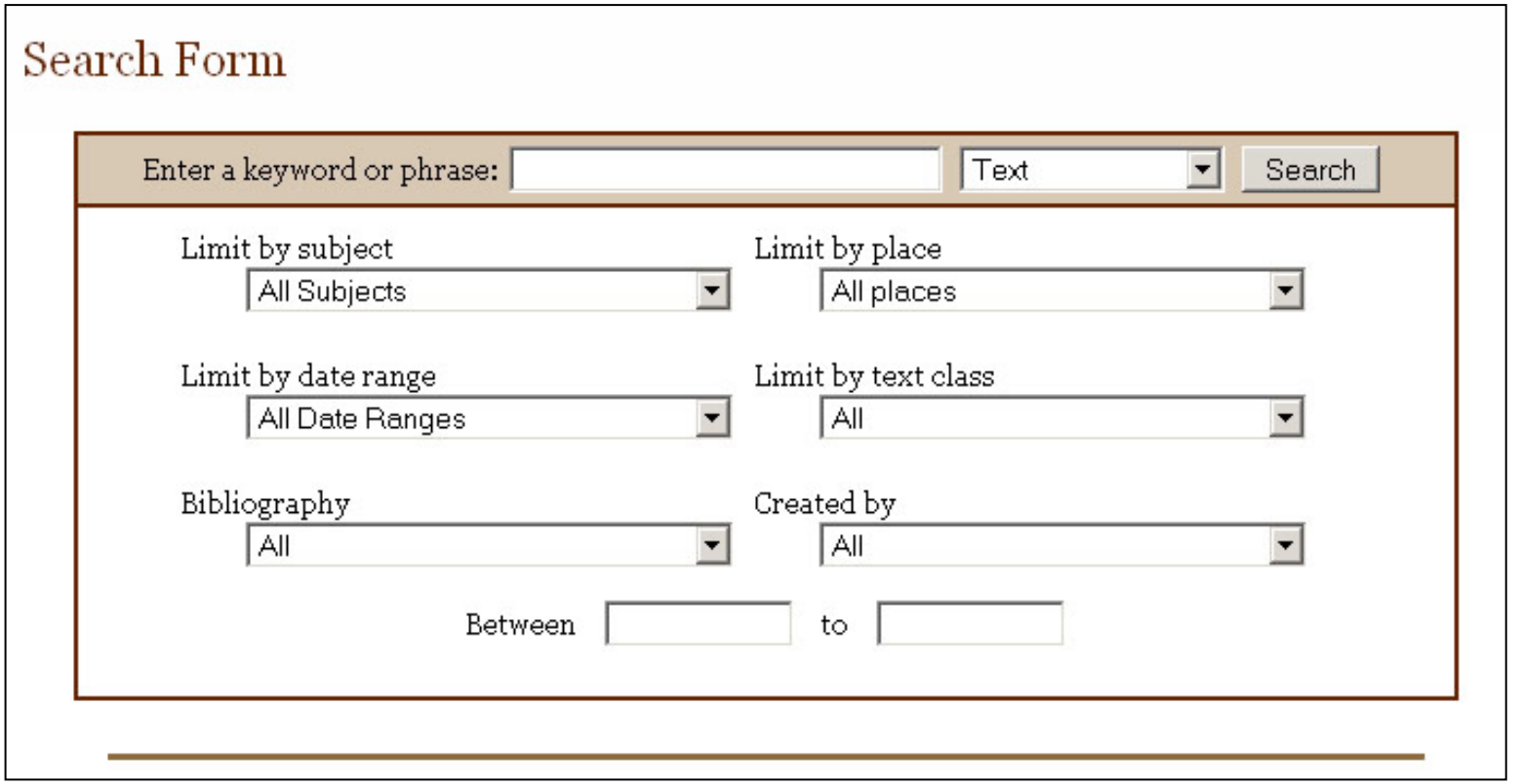
Some examples of TEI in action …

Thomas MacGreevy and George Yeats: a Friendship in Letters

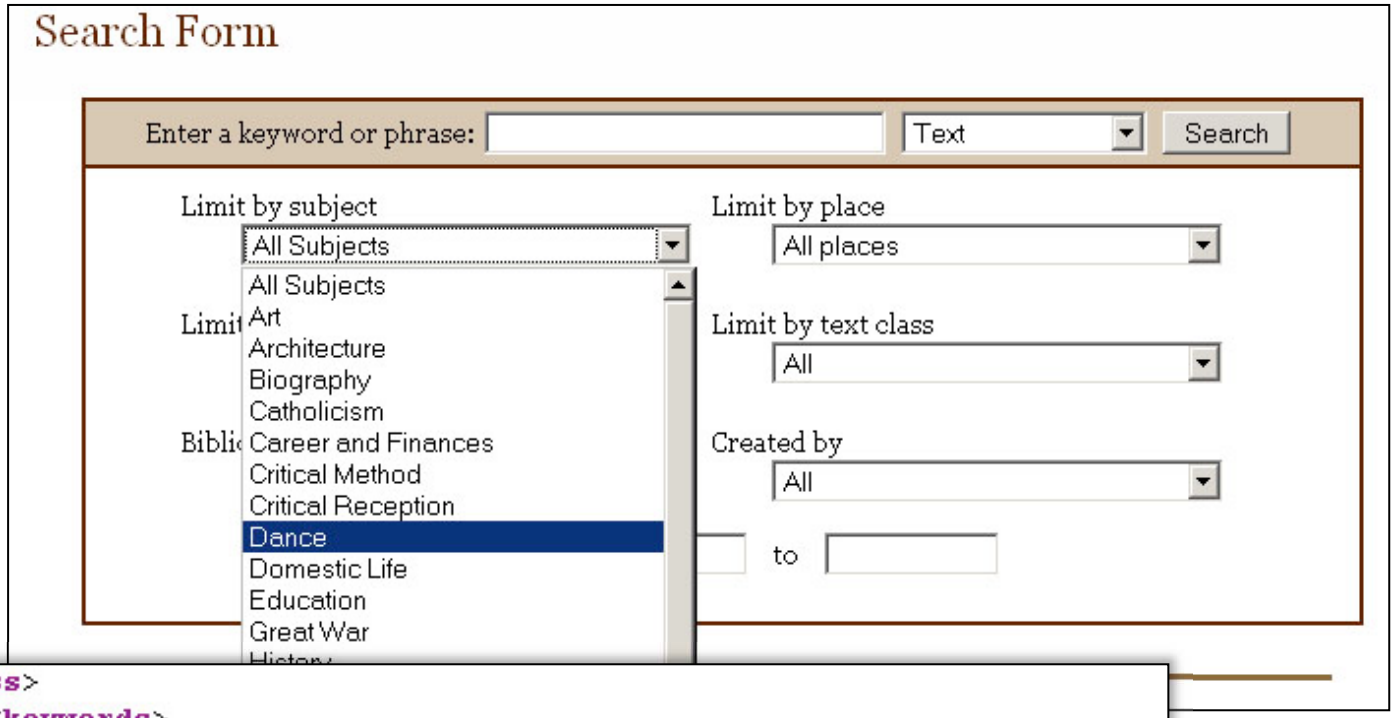


http://www.macgreevy.org/collections/gyeats

Thomas MacGreevy and George Yeats: a Friendship in Letters

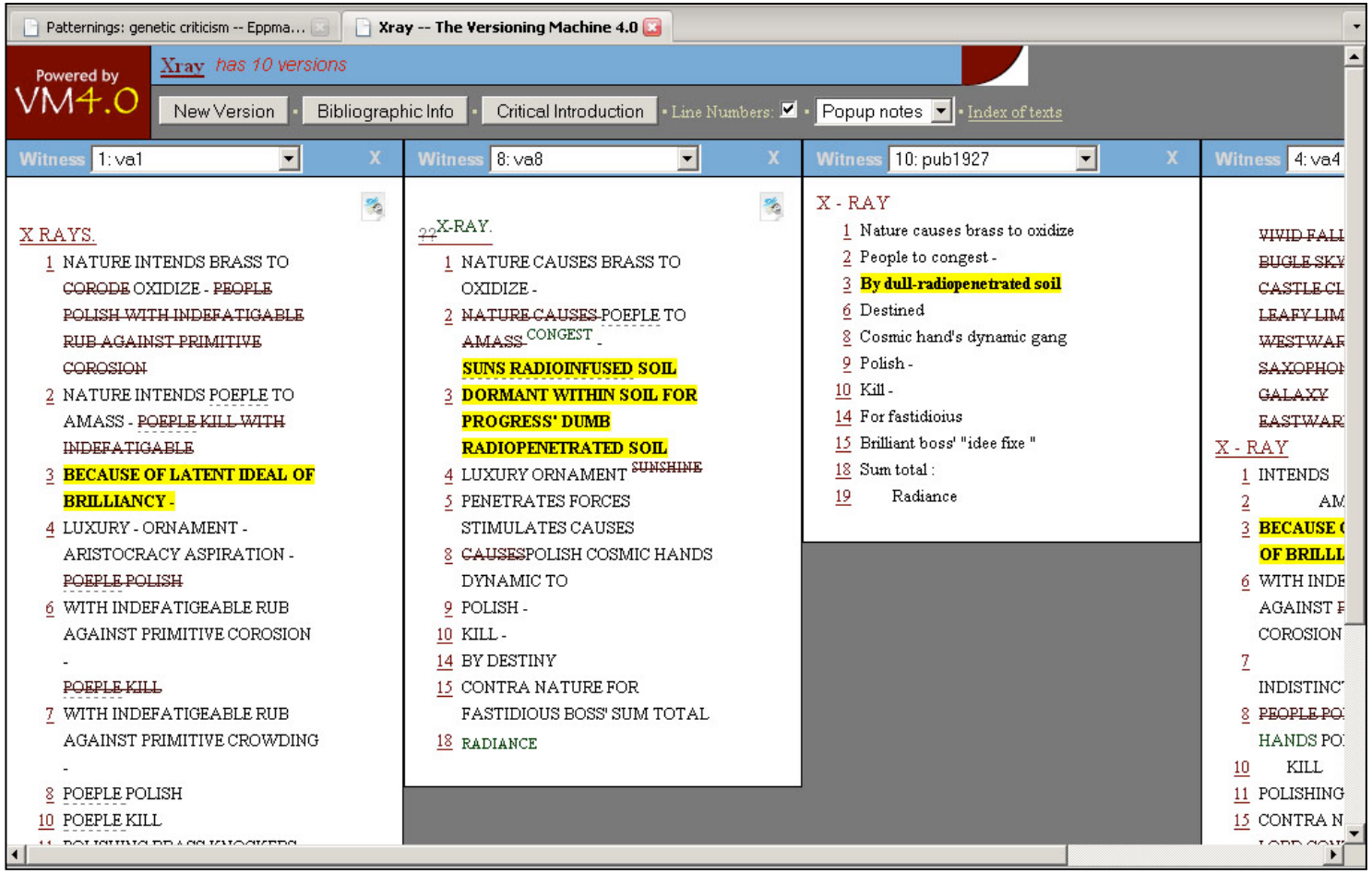


http://www.macgreevy.org/collections/gyeats




```
<textClass>
        <keywords>
                <list type="keyword">
                        <item type="subject">Irish Culture</item>

                        <item type="subject">Catholicism</item>
                        <item type="subject">Travel</item>
                        <item type="subject">Career and Finances</item>
                        <item type="subject">Opera</item>
                        <item type="subject">Dance</item>
                        <item type="date">1900-1999</item>

                        <item type="nationality">Irish</item>
                        </list>
        </keywords>
        <keywords>
```

In Transition: Selected Poems by the Baroness Elsa von Freytag-Loringhoven



http://www.lib.umd.edu/digital/transition/

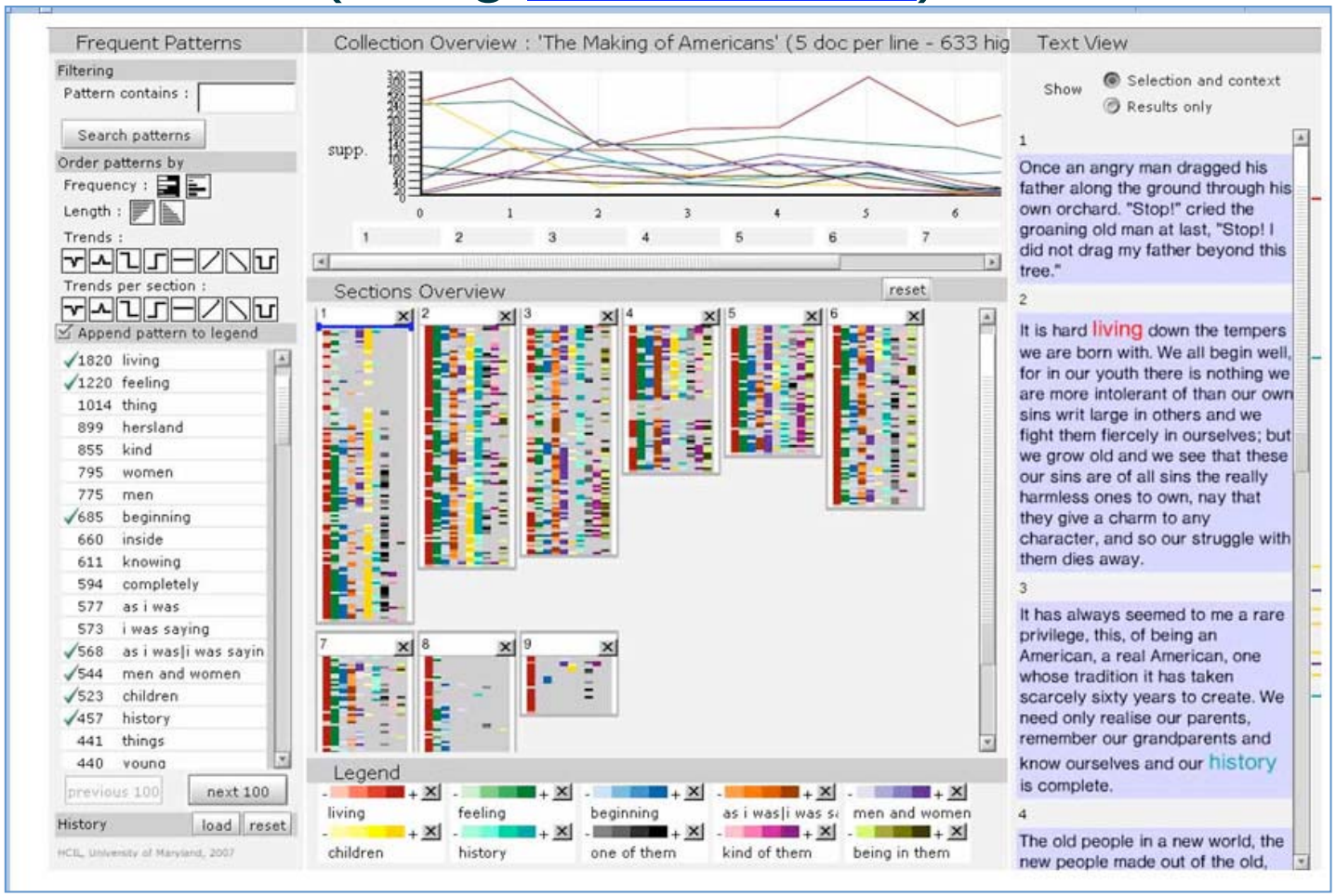
```
<l n="3">
  <app type="line" xml:id="a6" loc="a6">
    <rdg wit="#va1 #va2 #va3 #va4 #va5 #va6">BECAUSE OF </rdg>
    <rdg wit="#va6">RADIO </rdg>
    <rdg wit="#va7"><del rend="overstrik
    <rdg wit="#va1">LATENT</rdg>
    <rdg wit="#va2">
      <del rend="overstrike">LATENT</del>
      <add place="above">DORMANT</add>
    </rdg>
    <rdg wit="#va3 #va4 #va5">DORMANT </rdg>
    <rdg wit="#va1 #va2 #va3 #va4 #va5">IDEAL OF </rdg>
    <rdg wit="#va1 #va3 #va4">BRILLIANCY -</rdg>
    <rdg wit="#va5"><del rend=
    <rdg wit="#va7 #va8 #va6"
    <rdg wit="#va8">DUMB </
    <rdg wit="#va9">BY DULL
    <rdg wit="#va8 #va9">RAD
    <rdg wit="#pub1927">By d
  </app>
</l>
```

<app> element (for an entry in a critical apparatus)

```
<l>
<app>
<rdg wit="#v1"></rdg>
<rdg wit="#v2 #v3></rdg>
</app>
</l>
```

Some tools for exploring your TEI documents …

Repetition in Gertrude Stein's *The Making of Americans* (using FeatureLens)



Contact Tanya Clement (tclement@umd.edu) for more information.

Text Analysis Portal for Research (TAPoR)

This site is useful as you are exploring text encoding. You can:

- Manage TEI documents.
- Experiment with text tools online.
- Learn about digital textuality and text analysis.

Want to learn more?

- The Electronic Publishing Unit (EPU) at UCC is available to help!
- The TEI-C website is a bit overwhelming, but that's where everything is.
- Help is available on TEI-L, the mailing list for the TEI. Try not to be put off by the high level of discussion and the curtness of some replies. The community really does welcome newcomers!

Slides from this two-day workshop are available at http://dho.ie/node/667

Kevin S. Hawkins
k.hawkins@dho.ie