The Famine Project: Structuring Text

Gary Stringer
Digital Manager, College of Humanities
University of Exeter, UK

Unstructured text

- Plain text, perhaps with headings and basic typography
- Often coupled with images (e.g. Google Books)
- Often quick and relatively inexpensive to produce
- Efficient way of producing a usable archive
- Primarily presentational
  - i.e. encoding the appearance of a text
- Microsoft Word? Google N-grams?

UNIVERSITY OF EXETER

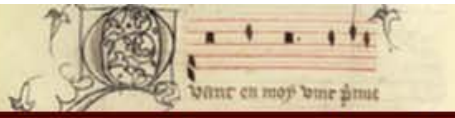
Highly Structured Text

- Semantic structure
  - Document structure (e.g. chapter divisions, marginalia)
  - Semantic annotation (e.g. names, places,
  - Linguistic analysis (root forms, parts of speech, etc.)
  - Apparatus (variant forms, manuscript description, genetics)
- TEI/XML most common in Europe/North America
  - … but other schemes exist

UNIVERSITY OF EXETER

```
    </name>
    <reg>,</reg> si tres
    <app>
       <lem>doucettem
          <rhyme label="a">
             <expan>en</expan>t
          </rhyme>
       </lem>
       <rdg wit="#MSMaC">doucetement</rdg>
       <rdg type="substantive" wit="#MSMaB">doucement</rdg>
    </app>
    <orig>_:</orig>
</l>
<l n="3" met="8">me
    <orig>u</orig>
    <reg>v</reg>ost
    <lb />mo
    <expan>n</expan> cuer enamour
    <rhyme label="b">er</rhyme>
    <orig>_:</orig>
</l>
<l n="4" met="8">que d
    <reg>'</reg>un
    <app>
       <lem>resgart</lem>
       <rdg wit="#MSMaA #MSMaC">regart</rdg>
    </app> me fist pres
    <rhyme label="a">ent</rhyme>
    <orig>_:</orig>
    <reg>,</reg>
</l>
<l n="5" met="8">et tres a
    <lb rend="hyphen" />moureus sentem
    <rhyme label="a">ent</rhyme>
    <orig>_:</orig>
</l>
```

Highly structured markup 'in the raw' – an excerpt of TEI P5 XML – showing features such as modernised and original typography (orig/reg), critical apparatus (app/lem/rdg), and marking of rhyme and metre.



Je chante ung chant

An Archive of Late-Medieval French Lyrics

Home Browse Search About Help

Quant en moy

Music: Motet : Triplum *of* *Quant en moy* / *Amour et biauté* / *Amara valde*

aka: Motet 1: Triplum ()

Guillaume de Machaut

set to music by: Guillaume de Machaut

Edition | Manuscripts | **TEI P5**

Text displayed as TEI P5 conformant XML

```
<body>
   <head>Ci commencent lez Motez</head>
   <lg type="stanza">
      <l n="1" met="8">
         <hi rend="C2">Q</hi>
         <orig>U</orig>
         <reg>u</reg>ant en moy
         <orig>u</orig>
         <reg>v</reg>i
         <expan>n</expan>t premierem
         <rhyme label="a">e
            <expan>n</expan>t
         </rhyme>
         <orig>_:</orig>
      </l>
      <l n="2" met="8">
         <name>
            <orig>a</orig>
            <reg>A</reg>mours
         </name>
         <reg>,</reg> si tres
         <app>
            <lem>doucettem
               <rhyme label="a">
                  <expan>en</expan>t
               </rhyme>
            </lem>
            <rdg wit="#MSMaC">doucetement</rdg>
            <rdg type="substantive" wit="#MSMaB">doucement</rdg>
```

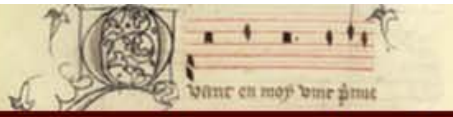
Web view of the raw TEI text, showing extracted metadata including text form, author/composer, associated texts, and available manuscript evidence (witnesses).

Witnesses:

- MaVg f. 260v-261r
- MaA f. 414v-415r
- MaB f. 258v-259r
- MaC f. 206v-207r
- MaE f. 131v-132r (from)
- MaG f. 102v-103r (from)
- MaW f. 74v (from)

Remarks:

None available.



Je chante ung chant

An Archive of Late-Medieval French Lyrics

Home Browse Search About Help

Quant en moy

Music: Motet : Triplum *of* *Quant en moy / Amour et biauté / Amara valde*
aka: Motet 1: Triplum ()

Guillaume de Machaut
set to music by: Guillaume de Machaut

Edition | **Manuscripts** | TEI P5

Text according to MaVg

View: MaVg

Ci commencent lez Motez

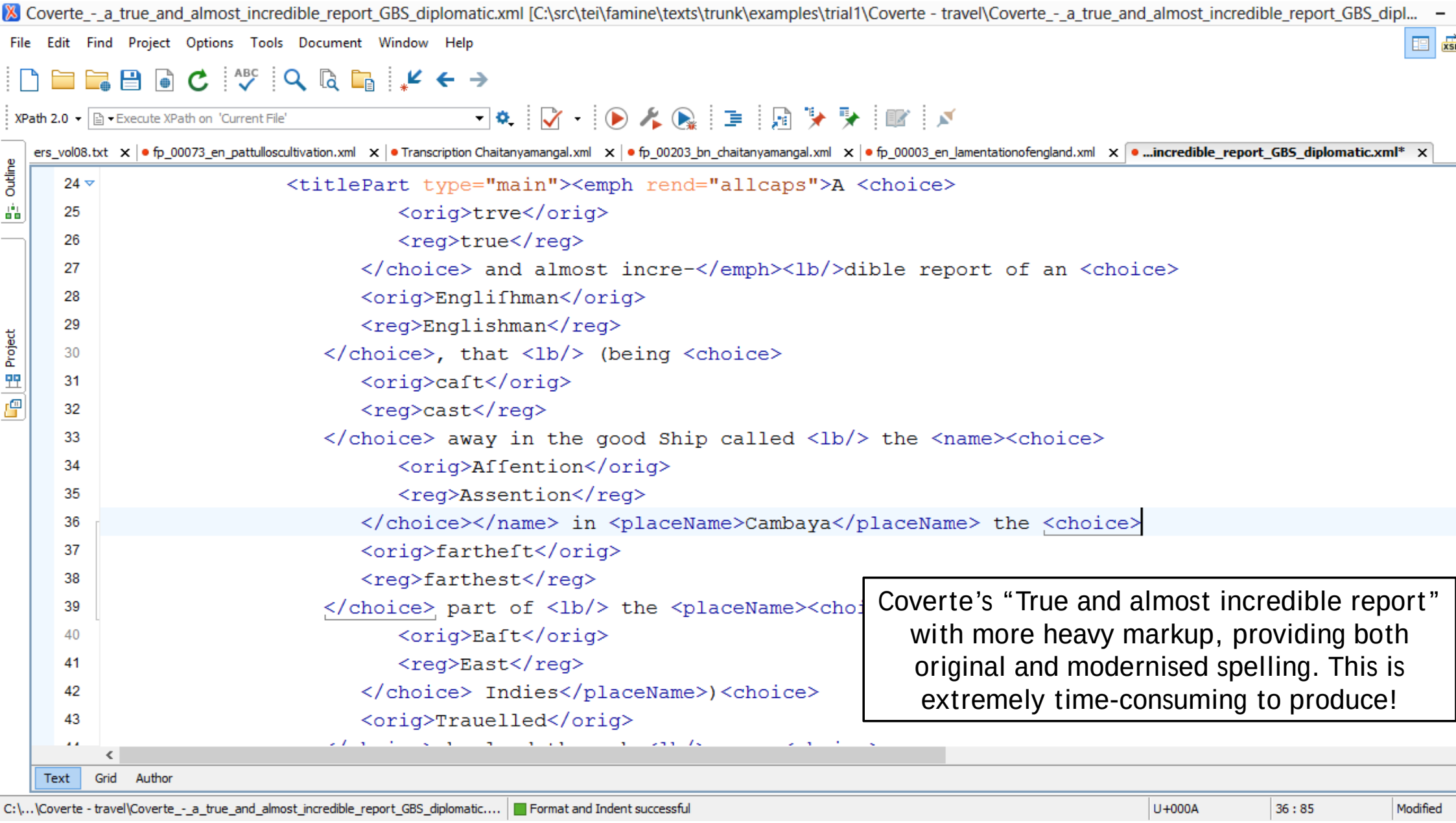
Quant en moy vint premierement
Amours, si tres doucettement — MaC doucetement MaB doucement
me vost mon cuer enamourer
que d'un resgart me fist present, — MaA MaC regart
et tres amoureus sentement
me donna aveuc doulz penser, — MaA MaC avec MaA MaC dous
espoir d'avoirmerci sanz refuser. — MaB MaC mercy MaA MaC sans
Mais onquez en tout mon vivant — MaA MaC onques
hardement ne me vost donner; — MaA volt

et si me fait en desirant
penser si amoureusement
que, par force de desirer,
ma joie convient en tourment
muer, se je n'ay hardement.
Las! et je n'en puis recouvrer,
qu' Amours secours ne me vuet nul prester — MaB vieult

Witnesses:

- MaVg **f. 260v-261r**
- MaA f. 414v-415r
- MaB f. 258v-259r
- MaC f. 206v-207r
- MaE f. 131v-132r (from Earp)
- MaG f. 102v-103r (from Earp)
- MaW f. 74v (from Earp)

View of text generated from structured markup (TEI P5), to form a 'dynamic critical edition', including critical apparatus from variant readings, and normalised spelling. Changing the 'view' will show the text from any of the encoded mss, or a diplomatic edition representative of the original.



Coverte's "True and almost incredible report" with more heavy markup, providing both original and modernised spelling. This is extremely time-consuming to produce!

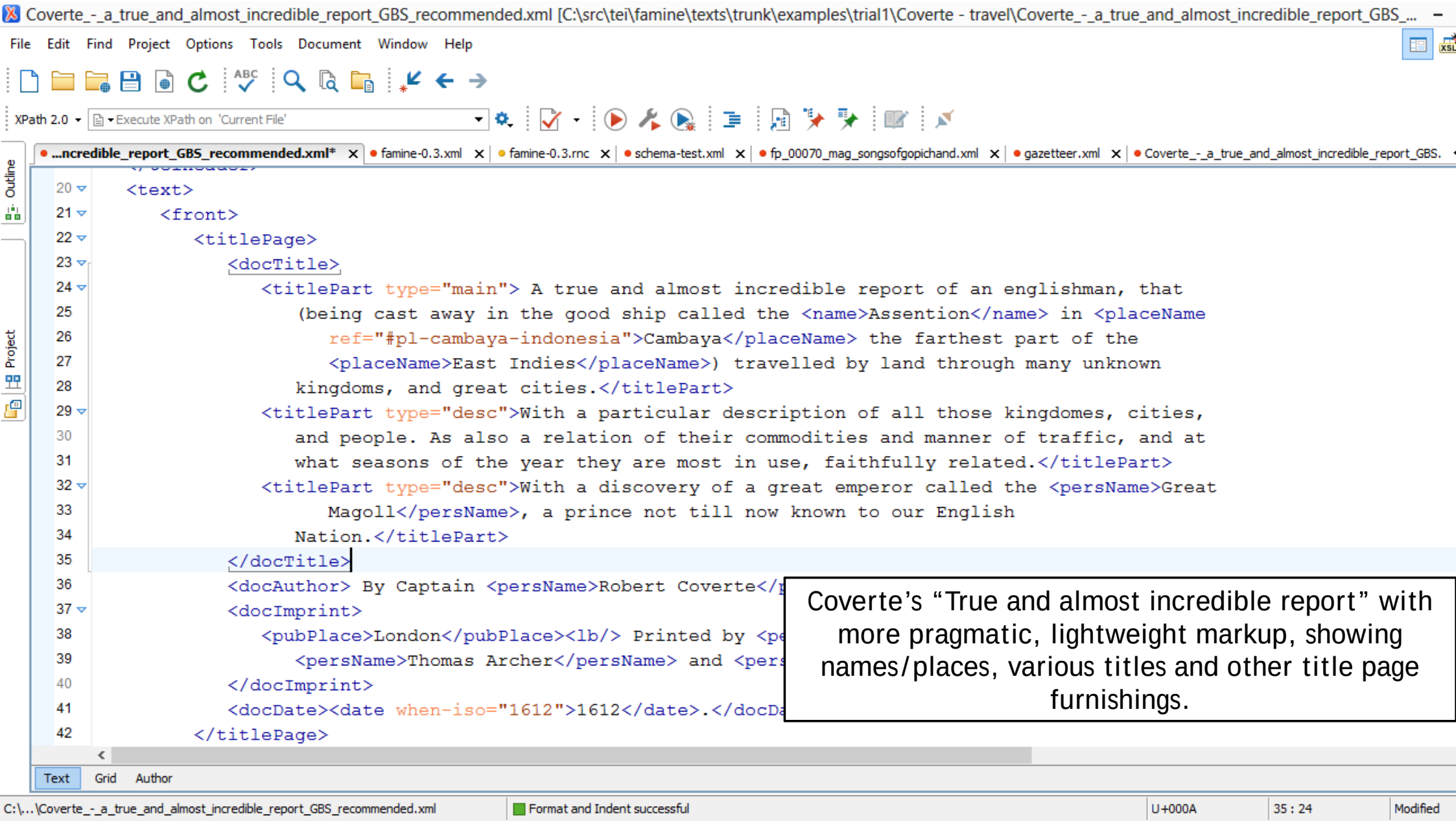
Lightweight standardised structure (Famine Project)

- A rational and pragmatic balance
- Features to be encoded are decided upon based around research questions
- Classification of texts using keywords
- Thematic markup is important (see tag list)
- Names and places are also vital
- Encoding multilingual documents – xml:lang and Unicode

UNIVERSITY OF EXETER

```
3850            encreasing,<lb n="1769"/>Hourely ioyes, be still vpon yo
3851          <pb n="B2"/>
3852          <milestone unit="compo" n="C"/>
3853          <lb n="1770"/>
3854          <hi rend="italic">Iuno sings her blessings on you.
3855            <lb n="1771"/>Earths increase, foyzon plentie,
3856            <lb n="1772"/>Barnes, and Garners, neuer empty.
3857            <lb n="1773"/>Vines, with clustring bunches growing,
3858            <lb n="1774"/>Plants, with goodly burthen bowing:
3859            <lb n="1775"/>Spring come to you at the farthest,
3860            <lb n="1776"/>In the very end of Haruest.
3861            <lb n="1777"/>Scarcity and want shall shun you,</hi>
3862            <lb n="1778"/>Ceres <hi rend="italic">blessing so is on yo
3863          <lb n="1779"/>
3864        </ab>
3865      </sp>
3866      <sp>
3867        <speaker rend="italic">Fer.</speaker>
3868        <ab>This is a most maiesticke vision, and<lb n="1780"/>Harmoni
```

Excerpt from the First Folio edition of The Tempest, very lightly marked up with speakers, line breaks/numbers, and some typographical markers. Taken from Oxford Text Archive 3014, searchable from University of Chicago's PhiloLogic interface



Coverte's "True and almost incredible report" with more pragmatic, lightweight markup, showing names/places, various titles and other title page furnishings.

Decision factors

- Time and effort required/available
- Purpose of text archive and research questions
- Preference of funding bodies
- Breadth of genres and types to be encoded
- What can be automated?
  - E.g. marking up names can automate lists and indexes
- Balancing effort required with usefulness of outcome

UNIVERSITY OF EXETER

Practicalities - TEI and XML

- TEI is de facto standard in Western Humanities
  - E.g. Oxford Text Archive, EEBO, and many others
- Defining and maintaining standards
  - Comprehensiveness of TEI means restricting scope
  - Documentation and training
- Developing support structures, e.g. Gazetteer
  - Allows consistent referencing
  - Saves repetition of markup

UNIVERSITY OF EXETER

Practicalities:
Workflow, standardisation and versioning

- Phased and well-defined workflow
- Choosing the right tools (e.g. <oXygen/>)
- Producing *workable* standards
- Selecting file names and file structures
- Keeping track of everything (Subversion)
- Standardisation of terminology
  - Define terms, create ontology

UNIVERSITY OF EXETER

Problems and benefits

- Hierarchical structure particularly problematic for thematic markup
- Very dense/verbose markup can make proofing difficult
+ Texts can be shared and reused (interoperable)
+ Texts can be ‘mined’ semantically (more later…)
+ Texts are self-documenting (metadata)

UNIVERSITY OF EXETER

Faceted browsing and searching and lists

- Search within specific structures or semantic values
- E.g. search within:
  - Specific genre (e.g. only travel narratives)
  - Names, placenames, themes,
  - Text marked as addressing (e.g.) socio-economic groups
- Easy to list all terms used for themes, names, etc.
- All in addition to standard full-text search

UNIVERSITY OF EXETER

Basic Search

Search Options
Search wildcards

*
a sequence of letters

?
an optional single letter

Search for [ ] Submit

Search for [ ] within [4] words of each other Submit

Search for [douce dame] within [the full text ▾] Submit
Experimental - under development

- the full text
- any stanza
- the refrain
- any single line

Visualisation

- Graphical representation of encoded data
- Can provide automated analysis and hypothesis testing
- Can interpret the texts in innovative and engaging ways
- Can lead to new research questions

- Examples of dynamic visualisation:
  - Visualisations of dataset on Russian poets

UNIVERSITY OF EXETER

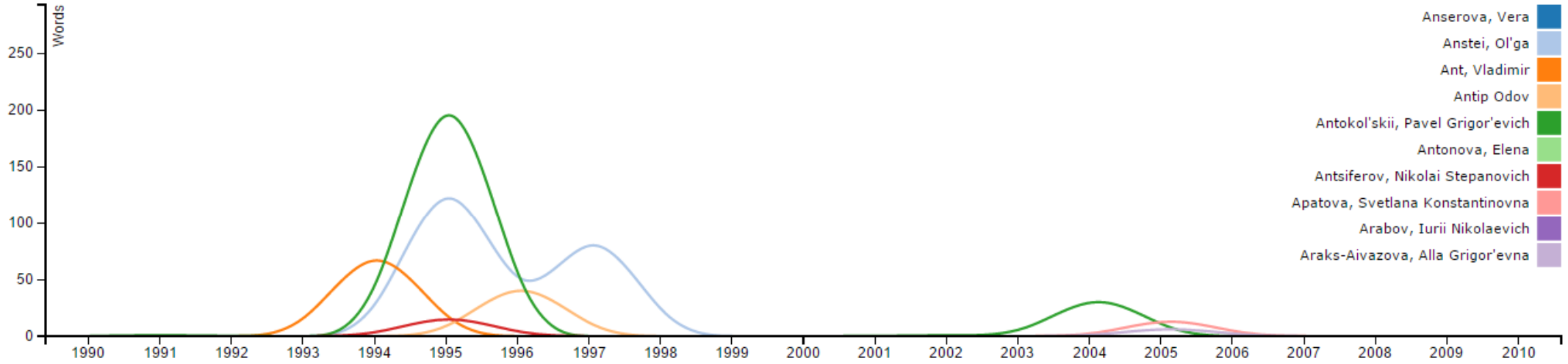
Russian Poetry Project

RECONFIGURING THE CANON OF TWENTIETH-CENTURY RUSSIAN POETRY 1991-2008

Home | Poets | Poems | Publications | Graphs

Poet mentions

10 poets selected | Create report

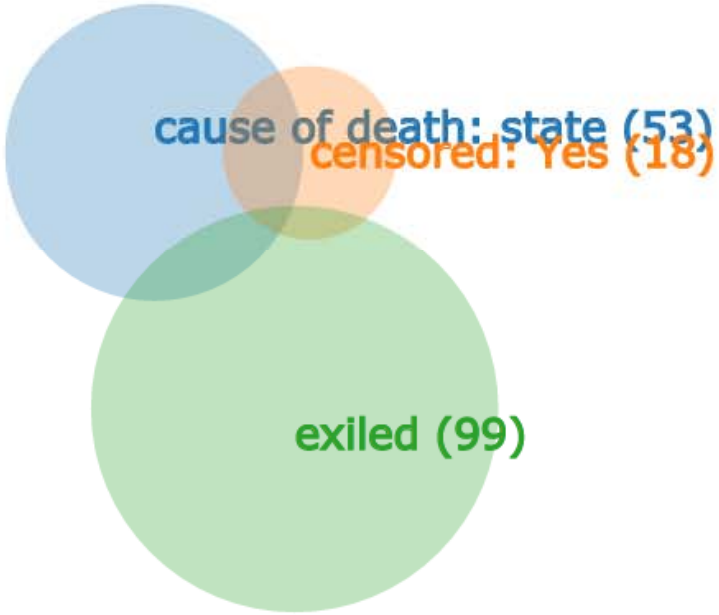


Automatically generated graph for 10 selected poets showing number of 'mentions' (citations, poem publications, reviews, etc.) per year

Russian Poetry Project

RECONFIGURING THE CANON OF TWENTIETH-CENTURY RUSSIAN POETRY 1991-2008

Three-set venn diagram



Select three sets exactly:

Cause of death:	○ Natural causes	◉ State violence	○ n/a
Censored:	◉ Yes	○ n/a	
Party member:	○ Yes	◉ n/a	
Exiled:	◉ Yes	○ n/a	
Exiled (internal):	○ Yes	◉ n/a	
Emigre:	○ Yes	◉ n/a	
Imprisonned:	○ Yes	◉ n/a	

Create report

Mapping and linking

- Geodata added can be linked to geographical databases
- Texts can be mapped onto Google Maps
- Waypoints and journeys extracted from geodata markup
- Possible overlays of contemporary maps
- Exploring possibilities of ‘emotional mapping’
  - Adding subjective responses to physical data

UNIVERSITY OF EXETER

T.R. Uthco (Diane Hall, Doug Hall, Jody Proctor), *Walking Mission Street* (Spring 1975)

0:50

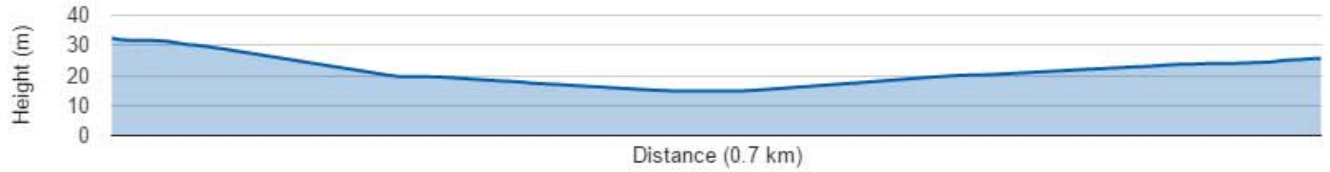
Example of mapping of journeys (from siteworks.ex.ac.uk) showing overlay of waymarkers and locations on modern (Google) mapping

Example of mapping of journeys (from siteworks.ex.ac.uk) showing route directions and nearby locations, including gradients on the walking route

Route to site | Windows: No. 1 to 2 | Walking ▾

Reset map

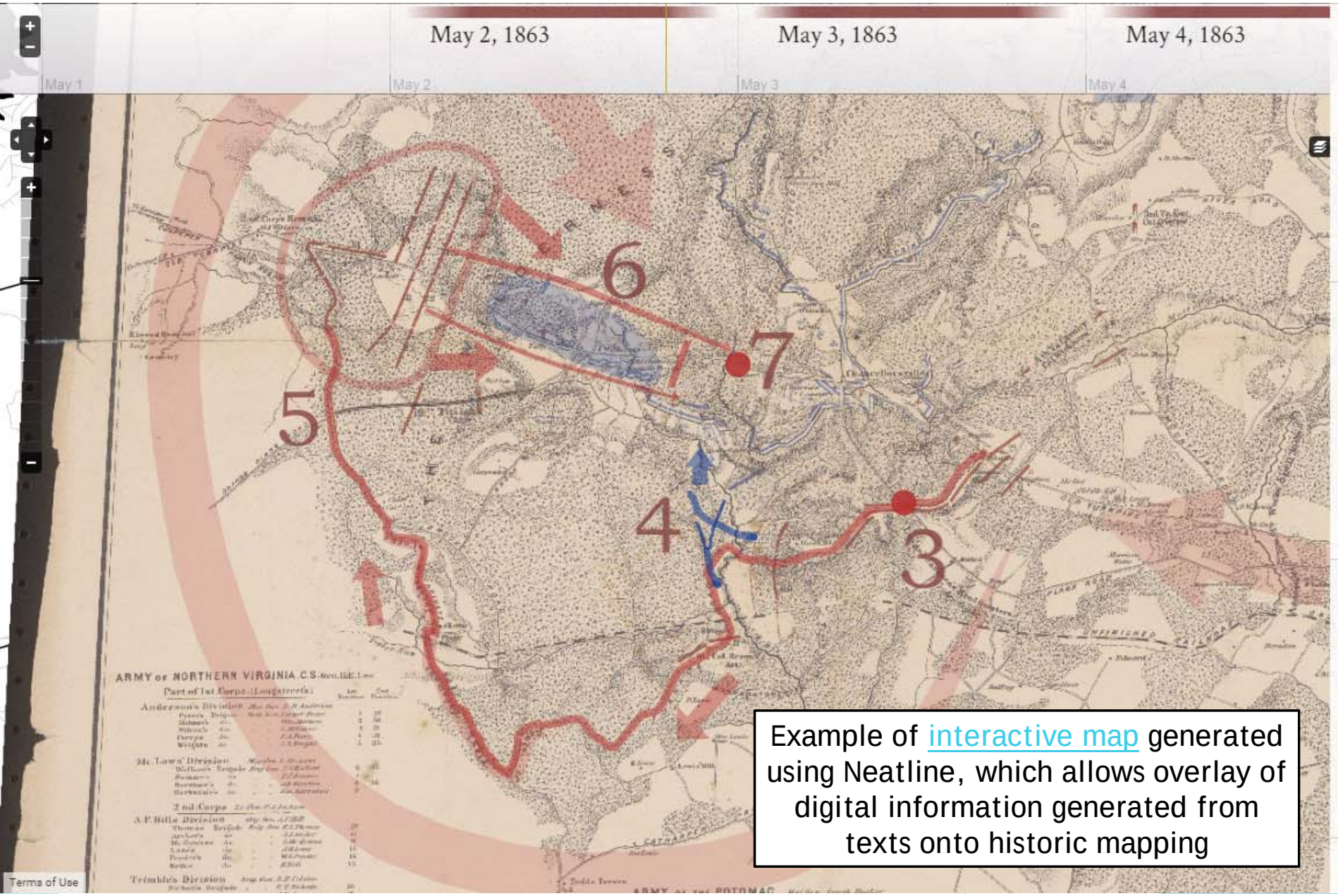
Elevation



Directions

Walking directions are in beta. Use caution – This route may be missing sidewalks or pedestrian paths.

2161-2199 Jones Street, San Francisco, CA 94133, USA



Example of interactive map generated using Neatline, which allows overlay of digital information generated from texts onto historic mapping

April 27 - 30

Hooker splits his army and tries to "double envelop" Lee with two forces, one approaching from the north by way of Chancellorsville and the other from the east at Fredericksburg.

April 30

Lee splits his army, sending Jubal Early to hold off Sedgwick at Fredericksburg and moving the rest of his force west to engage Hooker at Chancellorsville.

May 2, 7:30 a.m.

Lee splits his army again, sending Jackson's Second Corps on a 12-mile, 10-hour flanking march around the Union positions around Chancellorsville and the Wilderness

May 2, 11:00

Preservation, sustainability and archiving

- Standardised texts are long-lived
- TEI standard has been relatively unchanged since 1990
  - Tools can convert what has changed
  - Very long-lived in computer terms!
- Designed to be robust and easy to reuse
- TEI Header contains catalogue metadata for archiving

UNIVERSITY OF EXETER

Further reading...

- TEI Consortium (2007-2015) "TEI: P5 Guidelines". Available from http://www.tei-c.org/Guidelines/P5/.
- Burnard, L., et al. (2006) "Electronic Textual Editing", New York: MLA.
- Scholar's Lab, "Neatline". Available from http://neatline.org/.

- Famine project Wiki: http://humrestest.ex.ac.uk/faminetrac/

- All projects cited are linked in the relevant section

UNIVERSITY OF EXETER

Gary Stringer

Assistant College Manager (Infrastructure & Technical Services)

College of Humanities, University of Exeter, UK

G.B.Stringer@exeter.ac.uk

Web profile