

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?

Highly Structured Text

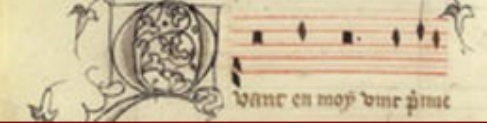
- ▶ Semantic structure
 - ▶ Document structure (e.g. chapter divisions, marginalia)
 - ▶ Semantic annotation (e.g. names, places,
 - ▶ Linguistic analysis (root forms, parts of speech, etc.)
 - ▶ Apparatus (variant forms, manuscript description, genetics)
- ▶ TEI/XML most common in Europe/North America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

```

</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>
</1>
<1 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>
</1>
<1 n="4" met="8">que d
  <reg>'</reg>un
  <app>
    <lem>regart</lem>
    <rdg wit="#MSMaA #MSMaC">regart</rdg>
  </app> me fist pres
  <rhyme label="a">ent</rhyme>
  <orig>_:</orig>
  <reg>,</reg>
</1>
<1 n="5" met="8">et tres a
  <lb rend="hyphen" />moureus sentem
  <rhyme label="a">ent</rhyme>
  <orig>_:</orig>
</1>

```

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.



Quant en moy

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

aka: Motet 1: Triplum 0

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>
      </app>
    </l>
  </lg>
</body>
```

[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.

Quant en moy

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

aka: Motet 1: Triplum 0

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 doucetement

me vost mon cuer enamourer

[MaC](#) doucetement [MaB](#) doucement4 que d'un resgart me fist present,[MaA](#) [MaC](#) regart

et tres amoureux sentement

me donna avec doulz penser,[MaA](#) [MaC](#) avec [MaA](#) [MaC](#) dousespoir d'avoirmerci sanz refuser.[MaB](#) [MaC](#) mercy [MaA](#) [MaC](#) sanz8 Mais onquez en tout mon vivant[MaA](#) [MaC](#) onqueshardement ne me vost donner;[MaA](#) volt

et si me fait en desirant

penser si amoureuxment

12 que, par force de desirer,

ma joie convient en tourment

muer, se je n'ay hardement.

Las! et je n'en puis recouvrer,

16 qu' Amours secours ne me vust nul prester

[MaB](#) rmeult

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.



Outline

Project



```

24 <titlePart type="main"><emph rend="allcaps">A <choice>
25 <orig>trve</orig>
26 <reg>>true</reg>
27 </choice> and almost incre-</emph><lb/>dible report of an <choice>
28 <orig>Englifthman</orig>
29 <reg>Englishman</reg>
30 </choice>, that <lb/> (being <choice>
31 <orig>caft</orig>
32 <reg>cast</reg>
33 </choice> away in the good Ship called <lb/> the <name><choice>
34 <orig>Affention</orig>
35 <reg>Assention</reg>
36 </choice></name> in <placeName>Cambaya</placeName> the <choice>
37 <orig>fartheft</orig>
38 <reg>farthest</reg>
39 </choice> part of <lb/> the <placeName><cho
40 <orig>Eaft</orig>
41 <reg>East</reg>
42 </choice> Indies</placeName>)<choice>
43 <orig>Trauelled</orig>

```

Covert'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

3850 encreasing,<lb n="1769"/>Hourly 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](#)

```

20 <text>
21   <front>
22     <titlePage>
23       <docTitle>
24         <titlePart type="main"> A true and almost incredible report of an englishman, that
25           (being cast away in the good ship called the <name>Assention</name> in <placeName
26             ref="#pl-cambaya-indonesia">Cambaya</placeName> the farthest part of the
27             <placeName>East Indies</placeName>) travelled by land through many unknown
28             kingdoms, and great cities.</titlePart>
29         <titlePart type="desc">With a particular description of all those kingdomes, cities,
30           and people. As also a relation of their commodities and manner of traffic, and at
31           what seasons of the year they are most in use, faithfully related.</titlePart>
32         <titlePart type="desc">With a discovery of a great emperor called the <persName>Great
33           Magoll</persName>, a prince not till now known to our English
34           Nation.</titlePart>
35       </docTitle>
36       <docAuthor> By Captain <persName>Robert Covert</persName>
37     <docImprint>
38       <pubPlace>London</pubPlace><lb/> Printed by <persName>Thomas Archer</persName> and <persName>
39       <persName>Thomas Archer</persName> and <persName>
40     </docImprint>
41     <docDate><date when-iso="1612">1612</date>.</docDate>
42   </titlePage>
  
```

Covert'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
- ▶ 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

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

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](#))

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)

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.
- ▶ In addition to standard full-text search

Basic Search

Search Options
 Search wildcards

*
 a sequence of letters

?
 an optional single letter

Search for

Search for within words of each other

Search for within

Experimental - under development

- the full text ▼
- 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

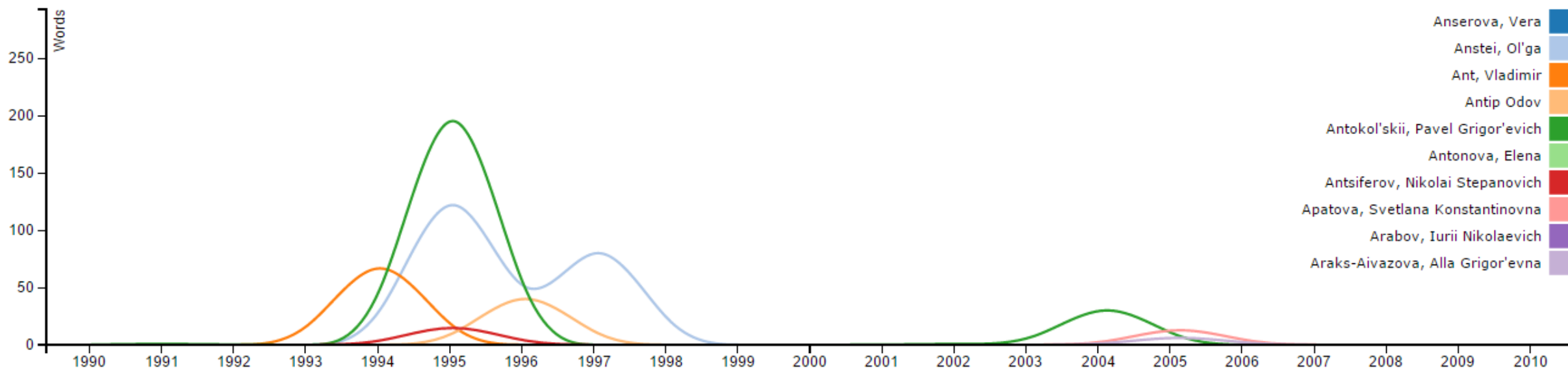
- ▶ Examples of dynamic visualisation:
 - ▶ [Visualisations of dataset on Russian poets](#)

[Home](#) [Poets](#) [Poems](#) [Publications](#) [Graphs](#)

Poet mentions

10 poets selected

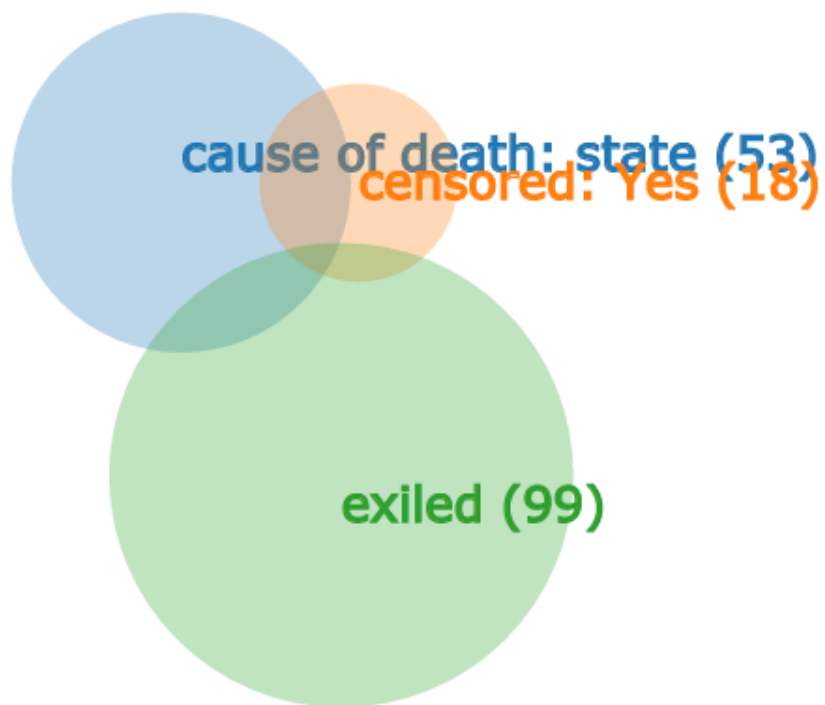
Create report



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

[Home](#) [Poets](#) [Poems](#) [Publications](#) [Graphs](#)

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

Imprisoned: Yes n/a

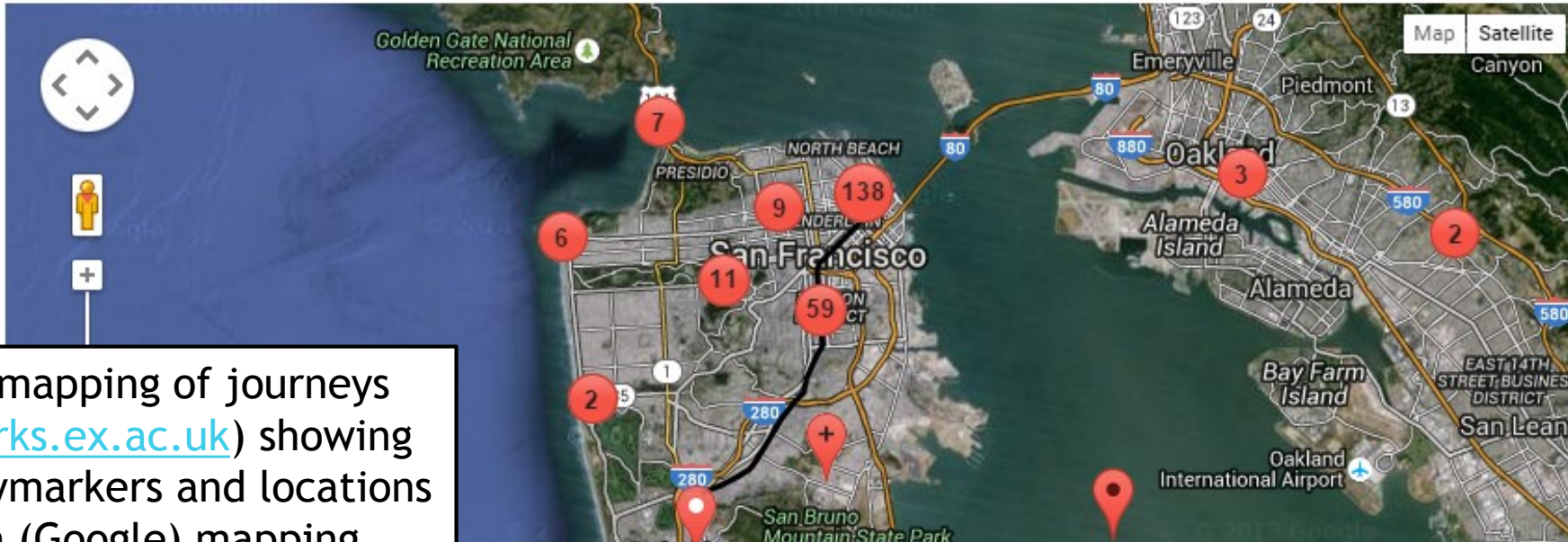
[Create report](#)

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



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



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

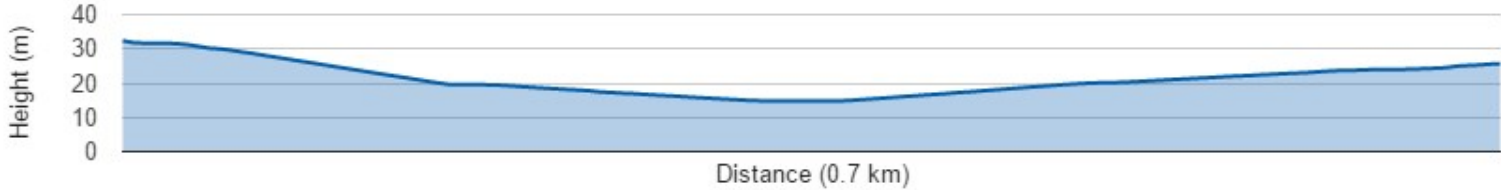
Example of mapping of journeys (from siteworks.ex.ac.uk) showing route directions and nearby locations, including gradients on the walking route



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

May 2, 1863

May 3, 1863

May 4, 1863

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.

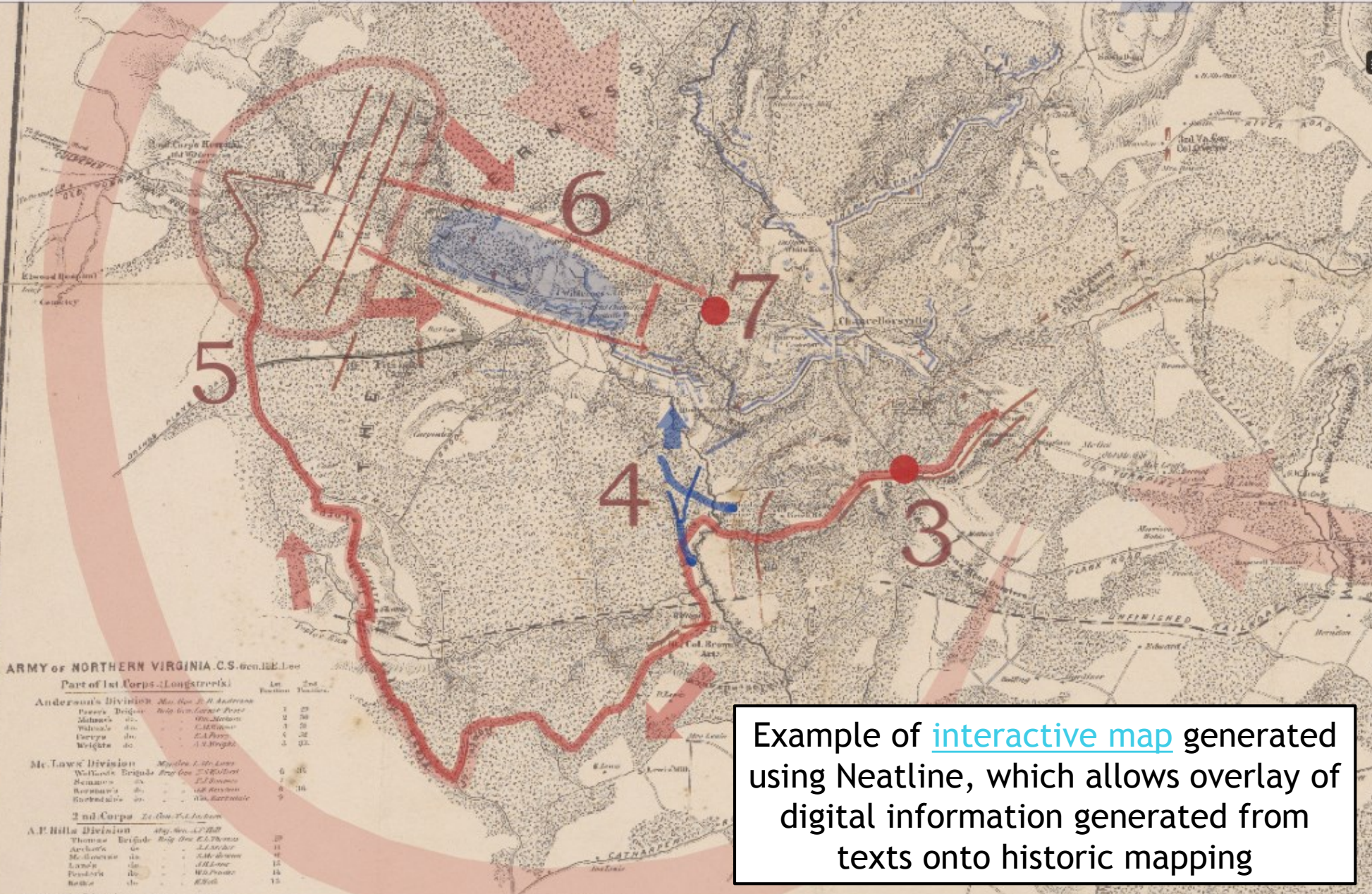
April 30

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



ARMY OF NORTHERN VIRGINIA, C.S. Gen. R.E. Lee

Part of 1st Corps - Longstreet's

Division	Commander	1st Position	2nd Position
Anderson's Division	Major Gen. D. H. Anderson	1	27
Parson's Brigade	Brig. Gen. Luther Parson	2	30
Milroy's Co.	Col. Milroy	3	30
Wilcox's Co.	Capt. Wilcox	4	32
Gregory's Co.	2d Lt. Gregory	5	32
Wright's Co.	1st Lt. Wright	6	32
Mc Law's Division	Major Gen. J. M. Lee	7	35
Wellons' Brigade	Brig. Gen. J. M. Lee	8	35
Neuman's Co.	2d Lt. Neuman	9	36
Barber's Co.	1st Lt. Barber	10	36
2nd Corps	Major Gen. F. L. Johnston		
A.P. Hill's Division	Major Gen. A.P. Hill		
Thomas' Brigade	Brig. Gen. E. S. Thomas	11	
Archibald's Co.	2d Lt. Archibald	12	
McGowan's Co.	1st Lt. McGowan	13	
Lawson's Co.	1st Lt. Lawson	14	
Proctor's Co.	1st Lt. Proctor	15	
Baker's Co.	1st Lt. Baker	16	
Trimble's Division	Major Gen. R. E. Colston		
Nichols' Brigade	Brig. Gen. P. T. Nichols	17	

Example of [interactive map](#) generated using Neatline, which allows overlay of digital information generated from texts onto historic mapping

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

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

Gary Stringer

Assistant College Manager (Infrastructure & Technical Services)

College of Humanities, University of Exeter, UK

G.B.Stringer@exeter.ac.uk

[Web profile](#)