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

Quant en moy vint precursor
Amours, si tres doucement
me vost mon cuer enamour

4 que d’un resgart me fist present,
et tres amoureus sentement
me donna avec douiz pensar,
espoir d’avoirmerci sans refuser.

Mais ounez en tout mon vivant
hardement ne me vost donner;
et si me fait en desirant
penser si amoureusement

8 que, par force de desirer,
ma joie convient en tourment
muer, se je n’ay hardement.
Las! et je n’en puis recouvrer,

12 Amours ecores ne me vost nul prestes,

[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
encreasing,"<lb n="1769"/>Hourely ioyes, be still vpon yo
<pb n="B2"/>
<milestone unit="compo" n="C"/>
<lb n="1770"/>
<hi rend="italic">Iuno sings her blessings on you.
<lb n="1771"/>Earths increase, foyzon plentie,
<lb n="1772"/>Barnes, and Garners, neuer empty.
<lb n="1773"/>Vines, with clustring bunches growing,
<lb n="1774"/>Plants, with goodly burthen bowing:
<lb n="1775"/>Spring come to you at the farthest,
<lb n="1776"/>In the very end of Haruest.
<lb n="1777"/>Scarcity and want shall shun you,</hi>
<lb n="1778"/>Ceres <hi rend="italic">blessing so is on yo
<lb n="1779"/>
</ab>
</sp>
<sp>
<speaker rend="italic">Fer.</speaker>
<ab>This is a most maiesticke vision, and<lb n="1780"/>Harmoni
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
- 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 [Input Field] Submit

Search for [Input Field] within [Input Field] words of each other Submit

Search for [Input Field] within [Input Field] the full text
Experimental - under development
Submit

the full text
any stanza
the refrain
any single line

©2007-2009 University of Exeter - Centre for Medieval Studies
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
Automatically generated graph for 10 selected poets showing number of ‘mentions’ (citations, poem publications, reviews, etc.) per year.
Three-set venn diagram

- Cause of death: state (53)
- Censored: Yes (18)
- Exiled (99)
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](http://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.
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
Further reading...


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

- All projects cited are linked in the relevant section