TEI@Oxford Projects

TEI@Oxford

July 2009
Digital humanities at Oxford

• The Bodleian Library digitize texts, make manuscript catalogues, and host a Google team scanning large parts of the library holdings

• The Humanities Faculty has a large number of research projects which use digital texts and resources in history, archaeology, linguistics, and literature

• The Oxford eResearch Centre hosts a number of research projects in the humanities, e.g. in classical art, in machine vision, and in development of a Virtual Research Environment for the Humanities

• Oxford is active in European infrastructural initiatives such as CLARIN and DARIAH, and of course the TEI

• At the Computing Services (OUCS) we provide support and expertise in areas including text encoding and analysis, corpus linguistics, digital archives and open source licensing
OUCS Themes

I will look at a few projects which are representative of the work we do, focussing on tools we have developed for them

• A text/database hybrid: William Godwin
• Text comparison: the Holinshed project
• Cross-project searching: the Claros project
• Born digital documentation: managing translations to and from Word
• A bit of fun with mapping

Other projects, including

• Greek names: LGPN
• Manuscript description: ENRICH
• A linguistic corpus: the BNC and XAIRA search engine

we cover elsewhere in this week.
Digital Diaries

- William Godwin:
  - 1756-1836, philosopher, writer, political activist,
  - husband of Mary Wollstonecraft, father of Mary Godwin (aka Wollstonecraft Shelley).

- Inter-Departmental:
  - Politics, Statistics, Computing Services, Bodleian Library, etc.

- Objectives:
  - research to identify people mentioned in the 48 years of diary;
  - provide a searchable cross-referenced electronic edition alongside digital images of the diary

- Started in October 2007, it coincided with initial release of TEI P5 and immediately benefited from new features.
Montagu breakfast, call with
him, at Holcot, Mrs. Chirnlow & D.
Thompson, Mrs. Shovring, Mr. Fosdyke twice.
Fosdyke twice.
Fosdyke twice.
Fosdyke twice.

Montague, by chance. Johnson & Mr.
Chisholm calls. Masters calls. Mr. Fosdyke
calls. Montague calls.

Montague calls. Clarke in the afternoon.
Montague calls.

Montague, by chance. Clarke from
Bristol, Mrs. Fosdyke's.

Montague calls. Wine diet. Carlisle from
Bristol, Mrs. Fosdyke.

Montague, by chance. Carlisle from
Bristol, Mrs. Fosdyke.

Montague, by chance. Carlisle from
Bristol, Mrs. Fosdyke. Carlisle from
Bristol, Mrs. Fosdyke.

Montague, by chance. Carlisle from
Bristol, Mrs. Fosdyke.

Montague, by chance. Carlisle from
Bristol, Mrs. Fosdyke. Carlisle from
Bristol, Mrs. Fosdyke.
<dDay><date when="1797-08-30">30. W. </date>
<dRead>Mary, p. 116.</dRead>
<del>Barnes</del>
<dMeeting type="CG"><persName>R Fell</persName></dMeeting> &amp;
<dMeeting type="CG"><persName key="DYS01">Dyson</persName> call</dMeeting>:
<dMeal type="D">dine at <persName key="REV01">Reveley</persName>’s</dMeal>:
<dMeal type="SC"><persName key="FEN01">Fenwicks</persName> &amp;
<persName>M.</persName> sup</dMeal>:
<dMeeting type="See"><persName>Blenkinsop</persName></dMeeting>.
<dEventG type="PersEvent">Birth of <persName key="SHE02"
>Mary</persName>, 20 minutes after 11 at night. <note>bracket pointing
    to following:</note> From 7 to 10, <placeName>Evesham
    Buildings</placeName>.</dEventG>
</dDay>
The Godwin ODD

• Created a customised TEI ODD for the Godwin project, this included:
  • removal of many TEI modules and elements
  • addition of new syntactic sugar elements
  • providing closed attribute value lists

• ODD provides:
  • schemas (DTD, RelaxNG, W3C Schema)
  • project specific documentation
  • option for internationalisation
  • useful assistance in chosen editor
Godwin ODD

<!-- Godwin dMeal Element -->
<elementSpec ident="dMeal" mode="add">
  <equiv filter="godwin2TEI.xsl" mimeType="text/xsl" name="segLike"/>
  <desc>Marks a segment of diary text where a meal is described, including participants</desc>
</elementSpec>

<content>
  <rng:ref name="macro.paraContent"
    xmlns:rng="http://relaxng.org/ns/structure/1.0"/>
</content>

<attList>
  <attDef ident="type" usage="req">
    <equiv/>
    <desc>provides a sub-categorization of the meal marked</desc>
    <valList type="closed">
      <vallItem ident="B">
        <equiv/>
        <gloss>Breakfasts at P (with X, (Y, Z))</gloss>
      </vallItem>
      <vallItem ident="BG"/>
    </valList>
  </attDef>
</attList>
<person>, <persName>, and people

- Objective is to link c. 64000 instances of <persName> elements (with around 10000 distinct values) to the right <person> elements, stored in separate files
- several researcher collaborate in the process
- Problems of ambiguity and multiple @ref values..

<persName ref="#BR01 #BR02 #BR03">The Browns</persName>

- Person records follow a strictly limited template
- Defined person files are available from a SVN-managed project website via predefined indexes
- Similar system is planned for identifying titles cited and places
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE person[<!ENTITY HOLcroft "Holcroft">]>
<person xmlns="http://www.tei-c.org/ns/1.0" sex="1" xml:id="HOL01">
  <forename>Thomas</forename>
  <surname>Holcroft</surname>
  <addName></addName>
  <birth when="1745-12-10">England</birth>
  <death when="1809-03-23"/>
  <occupation>playwright</occupation>
  <occupation>novelist</occupation>
  <note type="editorial">DNB. Also, HOLcroft is also used for Holcrofts – as H
  marries a couple of times during the period of the diary,
  be careful to distinguish between different wives</note>
</person>
<dEventG type="PersEvent">Birth of <persName key="SHE02">Mary</persName>, 20 minutes after 11 at night. <note>bracket pointing to following:</note> From 7 to 10, <placeName>Evesham Buildings</placeName>.</dEventG>

<person sex="2" xml:id="SHE02" xmlns="http://www.tei-c.org/ns/1.0">
<persName>
<forename>Mary</forename>
<surname>Shelley</surname>
<addName>Godwin</addName>
</persName>
<birth when="1797-08-30">England</placeName></birth>
<death when="1851-02-01"/>
<occupation>writer</occupation>
<note type="editorial">DNB</note>
</person>
Transforming Godwin

- Unlike many editions, researchers interested more in social relationships, frequencies of contact, and statistics
- Lists of element/attribute combinations produced to assist proofreading and editorial standards
  - lists by frequency
  - lists by distinct-value
  - lists by year
- XQuery in eXist will be used for front-facing site
- For project site: XSLT2-based grouping with xsl:for-each-group to create statistical lists
- Transformations to CSV for people in Statistics Department interested in networked relations of contacts in meetings
Digital Diaries

Rendered Generated List

- 21 -- call on holcroft:
- 21 -- fenwick
  1794-05-06 1794-06-20 1794-12-15 1795-01-30 1795-02-11 1795-04-08 1795-06-13 1795-08-07 1797-02-20 1797-04-12 1797-06-21 1807-06-26 1797-07-06 1798-03-15 1798-04-02 1798-04-05 1799-06-28 1799-07-10 1800-10-02 1801-12-20 1804-07-09
- 21 -- taylor
- 21 -- wordsworth
- 20 -- cell on perry
  1794-02-04 1794-12-30 1795-06-02 1796-06-06 1798-11-16 1799-11-14 1801-06-16 1801-07-03 1802-03-01 1804-08-27 1809-04-06 1810-01-03 1814-01-16 1815-09-10 1815-09-13 1816-03-16 1816-08-26 1818-06-05 1818-07-16 1819-03-10
- 20 -- davis
  1792-09-07 1792-10-27 1792-10-29 1792-12-10 1793-01-08 1793-01-25 1793-01-29 1794-01-15 1794-07-23 1794-12-23 1795-01-20 1795-04-03 1795-09-12 1795-10-07 1795-12-17 1796-01-11 1796-03-05 1796-06-22 1797-02-09 1797-05-13
- 20 -- holcroft
  1789-01-30 1792-06-14 1792-10-29 1793-02-04 1793-02-07 1793-02-20 1793-04-04 1793-04-08 1793-04-12 1793-05-08 1793-12-02 1794-01-11 1794-04-17 1794-05-17 1794-05-21 1794-08-14 1794-11-08 1794-11-12 1794-12-17 1795-01-30
- 20 -- phillips
  1801-04-08 1801-04-24 1801-12-28 1802-02-05 1802-03-03 1802-04-21 1802-11-12 1803-01-13 1803-09-14 1803-10-28 1803-12-03 1804-02-08 1805-02-15 1805-03-22 1805-04-13 1805-04-24 1805-06-26 1807-02-10 1812-06-11 1815-05-23
- 19 -- call on colburn
  1826-10-24 1826-11-09 1828-07-22 1828-11-01 1828-12-15 1829-02-04 1829-02-05 1829-06-03 1829-06-04 1829-08-20 1829-08-21 1829-08-26 1829-12-23 1829-12-27 1830-02-27 1830-05-20 1831-04-24 1832-09-05 1832-12-02
Holinhshed and the TEI Comparator

• The project aims to produce a print, old-spelling, annotated critical edition of Holinhshed's *Chronicles of England, Scotland, and Ireland*

• Two original source editions exist: 1577 and 1587 (revised and significantly expanded)

• An electronic full-text copy of 1587 exists in EEBO-TCP

• EEBO-TCP was commissioned to create an electronic full-text version of 1577 edition

• We helped to to fuzzy-match paragraphs from one to another, using stand-off linking, to assist them in making their comparisons
Scoping the problem

- Getting clean texts
- Deciding what to compare
- Establishing a notation for storing results of comparison
- Working out a system to make comparisons
- Providing an interface to allow links to be made and unmade
- Showing the results nicely for scholars
These are not modern books!
Getting clean texts

The EEBO project provided the two texts in SGML markup. We converted them to be XML fully-compliant against version 5 of the *Text Encoding Initiative* Guidelines.

The result includes some notion of typeface change, the marginal notes and figures, and indications of illegibility in the source:

```xml
<p cid="sid-ce7f3bd0-f920-4476-9977-43d499fb9c6b">But to procéede, when the fayde <hi>Albion</hi> had gouerned here in this ....

to be re|membred.<note place="marg">Leftrigo.</note> It happened in tyme of
<hi>Lucus</hi> king of the Celtes, that Leftrigo and his iſſue ...
conceyued this opinion, that if they had once gotten foote into any Re|gion whatfoeuer, it woulde not be long ere they did by fome meanes or
other,<note place="marg">
<hi>Ianige</hi><gap extent="+1" unit="letters">
<desc>illegible</desc>
</gap>
</hi> the po<gap extent="+1" unit="letters">
<desc>illegible</desc>
</gap>|ty of <gap extent="1" unit="word">
<desc>illegible</desc>
</gap> lying in Italy.</note> not onelye eftablifhe their feates, but
</p>
```
Deciding what to compare

Luckily, Holinshed is relatively simple in structure. We can take paragraphs as the main unit of comparison, leaving aside things such as headings:

although this leaves the problem of one paragraph in the first edition being split into many in the second.
Establishing a notation for storing results of comparison

We had two choices

1. Maintain one master text with embedded links to the other text

2. Keep the two texts separate, and store all the linkages in a third text, referring to uniquely-identified points in the two main texts

<linkGrp type="TEI-Comparator">
  <link
    id="sid-426e7975-b861-4368-b0fe-321612c97171"
    targets="1577.xml#sid-06ed72ca-cfb3-4b8d-999c-6350c43065ae1587.xml#sid-23f5aec2-e996-4e00-ad5d-f67199c93170"
    type="match"/>
  <link
    id="sid-01cab7e2-035a-46c6-b8d5-4fe6bf89ad33"
    targets="1577.xml#sid-381bfa62-436f-4d98-95da-e0471ccc92181587.xml#sid-e52eab7e-9fd4-4672-8d02-875e292a2bf2"
    type="match"/>
  <link
    id="sid-3afffe4d-c974-49ec-b776-5e876aaf1dc5"
    targets="1577.xml#sid-9739786a-798f-453b-8b5b-fac15d2be8151587.xml#sid-9972ad3d-5b66-46d1-a3f2-afd8f1ce243b"
    type="match"/>
</linkGrp>
Working out a system to make comparisons

First, we simplify the input to remove variability between texts:

- replace long s with short s
- replace vv with w
- replace ~ with n
- replace non-letters with spaces
- remove vowels except for u
- transform u into v
- remove double t, p, s, r, n, m
- replace sc with s

Then we take a text, walk over each identified unit, find all the words in it (using whitespace separator) and try to find ones somewhere in the other text which use a reasonable number of the same words.

Remembering that we have to search the whole text each time, so the processing is quite expensive!
Providing an interface to allow links to be made and unmade

The application lets you

- navigate the structure of either edition
- for a unit, ask to see matches from the other edition
- confirm and annotate one of those matches
- join two units by manual search
- see page image

One of the more problematic issues is the navigation by structure rather than the more familiar (but not formally recorded) reference by year.
Showing the results nicely for scholars

The Comparator shows one fully formatted text beside overlapping passages in another text.
The web static display shows one fully formatted text, with the ability to toggle between versions.

their whole race must needs have sustained the uttermost confusion, and thereby the memory of the Britons utterly has perished.

1.5. Whether it be likely that there were ever any Gyaunts inhabiting in this Ile or not. Cap. 4.

BEfides theses aforefayde nations, which have crept as you have hearde into our Ilande, we reade of sundry Gyaunts that should inhabit here, which report as it is not altogither incredible, fith the poorteries of divers princes were calle by the name: so vnto some mens eares it feemeth so strange a rehearulfall, that for the same onely they suppose the credite whole hystorie and reiect it as a fable, vnworthy to be read. For this caufe therefore I have nowe taken upon me to mak breife discourse infuing, thereby to prooue, that the opinio of Gyaunts is not altogither grounded vpon vayne & fabulous narrations, inuented only to deile the eates of the hearer [...] with the report of marvellous things. But that there have such men in deede, as for their hyseneffe of person have resembed rather highe towers then [...]etall men, although the poorteries are now confumed, and their monftruous races vnterly worsed out of knowledge.

A doe not meane herin to difpute, whether [Page 12] this name was giuen vnto them, rather for their tyrannie and oppriffion of the people, then for their greatnesse of bodie, or whe[ther the worde Gygas doeth onely signifie Indigenas, or homelinges, borne in the lande or not, neyther whether all men were of like quanti.flature and farre more greater in olde tyme then at this present they be, and yet absolutely I denie neyther of these, fith probable reasons may be brought for eche of the~, but especially the laft rehearfed, whose confirmation dependeth vpon au[thorities of sundrie auncient writers, who make diuers of Noble race, equall to the Gyaunts in strengthe, and manhood, and yet doe not gyue the same name vnto them, by caufe their quarles were iuft, and commonly taken in hande, for defence of the oppressed. Example hereof, alfo we may take of Hercu|les and Antheus, whole wreftling declareth that they were equall in flature & stomacke, such alfo was the courage of Antheus, that being often overcome, and as it were vttely vanquished by the fayde Hercules, yet if he did eftfoones returne agayne into his kingdome, he furthw^t recovered his force, returned & helde Hercules tace, till he gate at the laft betweene him & home, fo cutting of the farder hope of the reftoring of his army, and killing finally his aduerarie in the field. The like doe our his report of Corineus and Gomagot, who fought a combate hande to hande, til one of them was flayne, & yet for all this man reputeth Corineus for a Gy|aunt. But fith I faye it is not my purpose to fande vpon thefe pointes, I passe ouer to speake any more of them, and where as alfo I might haue proceeded in such order, that I should finde firt let downe by many circumfiances, whether any Gyaunts were, then whether they were of [such huge &
Questions

Is this on the desktop or the web? It is a web application, to avoid any platform or installation issues.

Does it work with other texts? Yes, it should, though we have not done much testing. Each new set of texts would need a new configuration explaining how the text is organized.

Is the software available? We have not yet considered how to generalize, document, or release this work.
The problem is to combine information about the classical world in Europe from:

- Beazley Archive, Oxford
- German Archaeological Institute, Berlin
- Lexicon Iconographicum Mythologiae Classicae - LIMC Paris
- Research Archive for Ancient Sculpture, Cologne
- Lexicon of Greek Personal Names
Methodology

- Agree on common data categories
  - place names
  - dates
  - object classifications
  - personal names
  - (possibly) data sources

- Each project exposes data subset in common format (RDF against CIDOC CRM format)
- RDF is aggregated and searched
- Results contain pointers back to original providers
Some of the problems

- CIDOC CRM is a large and confusing ontology susceptible to many different interpretations
- The CRM does not define the details (e.g., geographical coordinates, how to record dates)
- Leaving aside encoding problems, projects use different anglicisations of Greek names
- Fundamental problems over place names and/or locations
- Even if all the data maps to the CRM, there is not necessarily any overlap
A major objective in making large collections of TEI texts is to *analyse* them:

- searching for words and phrases as objects of interest in their own right (cf Google)
- identifying patterns of linguistic usage
- ‘intelligent search’ paying attention to associated markup and metadata
- but not *requiring* complex markup
People need the ability to display TEI-encoded texts as well as search or analyse them

• Hence we are interested in converting between TEI XML and other formats, including
  • HTML (relatively easy, using XSLT)
  • PDF (converting to XSL FO or LaTeX and then running a formatter)
  • OpenOffice (plug in filter using XSL)
  • Word 2007 (external XSL scripts and Java library)
TEI ISO

- Project with ISO, OUCS, TEI, Brigham Young University, and the Max Planck Digital Library
- International Standards Organization: production of standards documents
- TEI ODD for documenting a schema to store these standards as TEI XML
- A suite of XSLT to allow lossless conversion to and from various word-processing systems (e.g. MS Word)
- This ability to round-trip from presentational markup may also be available to other users
Vesta, a desktop processor

A Java application which can

- process TEI ODD schema definitions (like the Roma web application)
- convert TEI XML files to — and from — supported formats
- support multiple profiles