Jahrbuch für Computerphilologie 10 (2008), Cummings "The William Godwin's Diaries Project" <http://computerphilologie.de/jg08/cummings.pdf> (29. April 2009)

James Cummings

The William Godwin's Diaries Project: Customising and transforming TEI P5 XML for project work

Abstract:

This article reports on a project in progress at the *University of Oxford* that is editing the forty-eight years of detailed diaries written by William Godwin from 1788 to 1836. Support has been provided in the form of TEI training, schema customisation, website development, and general technical support. This work so far has included tutorials on TEI (and related technologies), the creation of a Godwin-specific TEI ODD in order to customise the TEI to their needs, and numerous XSLT stylesheets for producing various lists and statistics to help them proof-read and understand the markup that they've already applied and prepare for the next phases of encoding they are undertaking. This article is concerned with some of the the new features of TEI P5 they have used, how these have been transformed (with XSLT2), and the benefits this has accrued to their project methodology. The project, while it is an inter-departmental one, is not located in the History or English faculties as one might expect, but the *Department of Politics and International Relations*. Hence, the central concerns of those working on the project are the networked relationships between people, the meetings they attended, and prosopographical networks. For example, in addition to just marking up the very regular diary entries, they are also pointing from each of the over 64,000 instances of a person's name to a TEI <person> element in a separate file where metadata is stored about that person. This form of URI-based pointing combined with prosopographical data is one of the new areas in TEI P5. The article uses this as an example to look at how the particular needs of a scholarly editing project were reflected in the TEI P5 markup chosen and how that has been chosen to be processed for a variety of different purposes, before discussing the directions the project will be going in the future.

Jahrbuch für Computerphilologie 10 (2008), Cummings "The William Godwin's Diaries Project" <http://computerphilologie.de/jg08/cummings.pdf> (29. April 2009)

Introduction

This article introduces an inter-departmental project at the *University of Oxford* to digitise and create an electronic edition of the diaries of William Godwin. Instead of being a report upon a completed project demonstrating a *fait accompli* or the reverse, a desired project yet to be undertaken, this instead concerns a project that is part way through its timetable. Thus, instead of concentrating on a flashy frontend which might deliver the material, this article is examining the process by which that material is encoded and transformed for the benefit of ongoing project work. The project is being undertaken primarily by the *Department of Politics and International Relations of the University of Oxford*, but with significant contributions by the Statistics Department, Computing Services, and Bodleian Library. The full title of the project is »William Godwin's Diary: Reconstructing a Social and Political Culture 1788-1836«, and it has been funded by the Leverhulme Trust and the John Fell OUP Research Fund. The principle investigator of the project is Dr Mark Philp.[1] The *Research Technologies Service*, the part of the *Oxford University Computing Service* that I work for, is contracted to provide ongoing project technical support, training, and the eventual creation of the frontend website by autumn 2010.[2]

William Godwin, 1756-1836 was a philosopher, writer, and political activist. He is perhaps most commonly known as the husband of Mary Wollstonecraft and the father of Mary Wollstonecraft Shelley, the author of *Frankenstein*. Godwin faithfully kept a diary from 1788 until his death in 1836; the diary is now preserved in the Abinger collection in the Bodleian Library. It is an extremely detailed resource of great importance to researchers in fields such as history, politics, literature, and women studies. The concise diary entries consist of notes of who Godwin ate or met with, his own reading and writing, and major events of the day. The diary gives us a glimpse into this turbulent period of radical intellectualism and politics, and many of the most important figures of this time feature in its pages, including S.T. Coleridge, Richard Brinsley Sheridan, Mary Wollstonecraft, William Hazlitt, Charles Lamb, Mary Robinson, and Thomas Holcroft, among many others.

---

1 For more information see also, *William Godwin's Diary: Reconstructing a Social and Political Culture 1788-1836* [1].

2 See also ›*Research Technologies Service* [2].

The basic objectives of the project are:

- to undertake systematic research on the diary to identify those referred to in it and through this to construct a picture of London's literary and extra-parliamentary political life between 1788 and 1836;
- to develop a full scholarly apparatus of indexing, annotation and cross-reference to enhance the intelligibility of the material and allow its systematic searching;
- to augment the resource further by linking it directly to related electronic material;
- and to provide a reliable, searchable, online transcription of the text, alongside a scanned version of the original manuscript.

Godwin's Diaries

From a technical perspective, in many ways the project benefited from its commencement date in October 2007. The P5 version of the *Guidelines for Electronic Text Encoding and Interchange* was released as stable by the *Text Encoding Initiative Consortium* (TEI) at the beginning of November 2007.[3] This meant that the project could start using TEI P5 from the start, and reap the rewards of several years of systematic development work. Coincidentally some of the areas the project was most interested in had been added or substantially revised for the P5 release of the TEI Guidelines.

Godwin's diaries are simultaneously immensely detailed (recording the names of almost everyone he ever met with) and frustratingly concise (he only rarely gives details of what they talked about). As the project was being run by the politics department their interests were less to create a simple electronic edition of the diary, and more to create a fully cross-referenced resource which enabled exploration of the social network relations of those mentioned in the diary. Unlike those working in History or English, the project members are perhaps more interested in the statistics one can generate from the encoded texts than the textual phenomena present in the diary itself. The concentration in the encoding has been on the intellectual structure of the diary entries, meetings, and the relating of instances of named people to metadata recorded about people

[3] TEI Consortium, eds. Guidelines for Electronic Text Encoding and Interchange. (09.07.2008) [3].

as individuals. This has been evident in some of the working lists they wish me, in my role of supplying ongoing technical support for the project work, to provide. While regular support has been given to the project when they encounter difficulties (whether of an encoding nature or simply in use of software), those creating the intellectual content have progressed rapidly without significant hand-holding leaving the technical side of the project to evolve as necessary in response to their occasional desires.

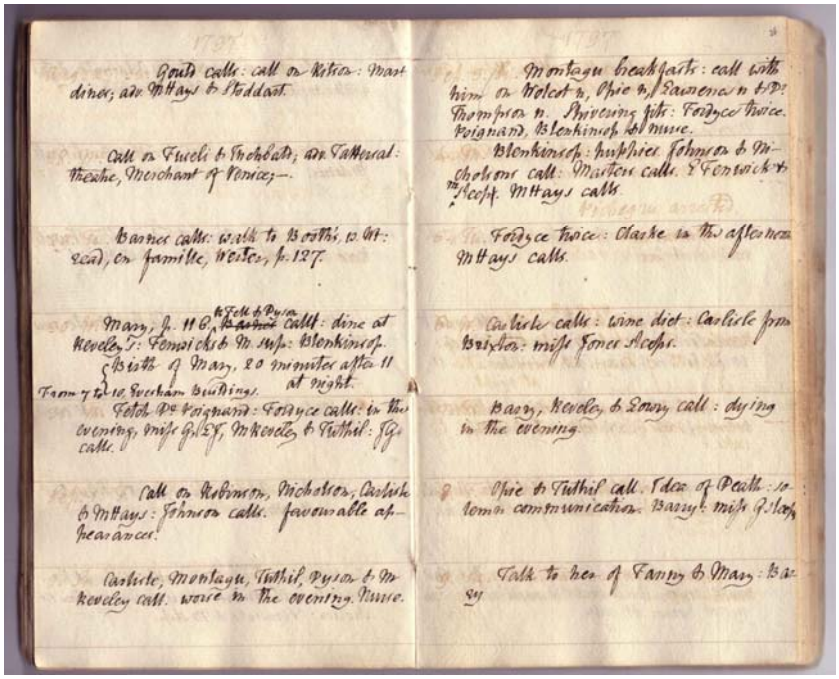

Figure 1: William Godwin's Diary

As you can see from the figure above, Godwin's diary is quite neatly written and easy to read. The dates, here in a much lighter ink, are usually given (and given correctly) and generally a day's entry forms the basic structural unit of the diary. In only a very few instances do the notes from one day stray into the page area already pre-ruled for the following day. Occasionally there are marginal notes to provide more information, but in most cases the textual phenomena are quite predictable – mostly

substitutions and interlinear additions.[4] In many ways the hierarchial nature of a calendrical diary entry makes it ideal for encoding in XML. There is some indication that Godwin may have returned to certain volumes at a later date to rewrite them. And yet, it is certainly impressive that there are entries for most days, and that whatever minimal information is given, the names of those attending the frequent meetings Godwin had with those in his circle are recorded. The majority of his diary entries were seen to be able to be broken down into several categories and subcategories. These include his meals, who he shared them with, who he met, sometimes what they talked about, and what works he was reading or writing at that time. The political historians, it is easy to understand, are eager to create a resource which allows them to explore which individuals might be meeting with what other friends of Godwin's at specific times. Meanwhile those exploring Godwin's writings might be interested to know what works he was reading when he was writing specific parts of some of his works.

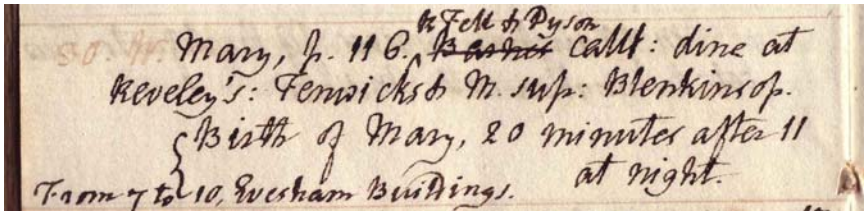

Figure 2: August 30, 1797

This entry, from 30 August 1797, records that Godwin read p. 116 of a text called ›Mary‹; that R. Fell and Dyson called; that Godwin dined at Reveley's; that Fenwicks and M. later supped with him; That he saw Blenkinsop; and that at 20 minutes after 11 at night, Mary Godwin (later Mary Wollstonecraft Shelley) was born at 7-10 Evesham Buildings. There is a substantial amount of information here in a very concise entry. In addition the textual encoding of this information forces a certain amount of interpretation on the project staff. However their familiarity with Godwin's diary and his social relationships enables them to make educated guesses at the categories of meetings and nature of some aspects of the entries. Moreover, while they have progressed through the

[4] This is not to underestimate the problems inherent in modern manuscripts, nor revision sites where interpreting the timing of substitutions can prove difficult. In Godwin's diaries, this is thankfully rare. Conf. Vanhoutte (2006).

diary working in successive phases of encoding, they have developed a set of editorial guidelines so that, as they proofread each other's work, their interpretations of entries become increasingly consistent. Here, for example, they would have to decide if R. Fell and Dyson were calling together or individually (when in doubt, they choose individual meeting entries when people are calling). It is of importance to those on the project to distinguish those who call upon Godwin (for a meal or meeting) from those that Godwin himself calls upon. Likewise, the entry ›Blenkinsop‹ is interpreted as a meeting of type ›see‹, even though Godwin just writes the name – this is where the familiarity with the diary comes into play, knowing that this is one of the common ways Godwin refers to people he's seen.

XML Markup and TEI ODD

```
<dDay>
  <date when="1797-08-30">30. W. </date>
  <dRead><title>Mary</title>, p. 116.</dRead>
  <del>Barnes</del>
  <dMeeting type="CG">
    <persName>R Fell</persName>
  </dMeeting> &amp;
  <dMeeting type="CG">
    <persName key="DYS01">Dyson</persName> call<del>s</del>
  </dMeeting>:
  <dMeal type="D">dine at <persName key="REV01">
      <placeName>Reveley's</placeName>
    </persName>
  </dMeal>:
  <dMeal type="SG">
    <persName key="FEN01 FEN03">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. <add place="subleft">From 7 to 10, Evesham Buildings.</add>
</dEventG>
</dDay>
```

Figure 3: Godwin Project Markup

Anyone familiar with TEI P5 XML will find many of these element names in the figure above confusing. The project wished to record some structured phrase-level metadata for which the TEI does not have specific elements to cope with. For example, the recording of meals and

meetings. In most cases, these elements are direct equivalents of a TEI <seg> element. However, in order to make the encoding practice easier for the project staff, these were given different names. The project faced a choice in this matter, either to use a separate namespace or to undertake a process of canonicalisation of the files. It was decided that the latter was more convenient, so instead of specifying a namespace and having a <g:dMeal> element, the element is added inside the TEI namespace. This makes all of these documents non-conformant according to the TEI Guidelines. However, the canonicalisation process reverts any Godwin elements back to their pure TEI namesakes. So, for example, <dMeal type="D"> simply becomes <seg type="dMeal" subtype="D"> and a Conformant status is returned. This use of a local encoding format is quite common and TEI customisation tools allow for this. If the documents were to be circulated in this local format then that would generally be considered inappropriate; by having these elements in the TEI namespace we are in fact falsely claiming that they are TEI elements. And yet, since the canonicalisation process returns these documents to be a pure subset of the ›tei_all‹ schema, this becomes unproblematic. The figure below demonstrates what this passage looks like in ›pure‹ TEI P5 XML.

```
<ab type="dDay" xml:id="g1797-08-30">
  <date when="1797-08-30">30. W. </date>
  <bibl type="dRead"><title>Mary</title>, p. 116.</bibl>
  <del>Barnes</del>
  <seg type="dMeeting" subtype="CG">
    <persName>R Fell</persName>
  </seg> &amp;
  <seg type="dMeeting" subtype="CG">
    <persName ref="/people/DYS01.xml">Dyson</persName> call<del>s</del>
  </seg>:
  <seg type="dMeal" subtype="D">dine at <persName ref="/people/REV01.xml">
      <placeName type="venue">Reveley's</placeName>
    </persName>
  </seg>:
  <seg type="dMeal" subtype="SG">
    <persName ref="/people/FEN01.xml /people/FEN03.xml">Fenwicks</persName>
    &amp; <persName>M.</persName> sup
  </seg>:
  <seg type="dMeeting" subtype="See">
    <persName>Blenkinsop</persName>
  </seg>.
<seg type="dEventG" subtype="PersEvent">Birth of
  <persName ref="/people/SHE02.xml">Mary</persName>, 20 minutes
    after 11 at night. <add place="subleft">From 7 to 10, Evesham Buildings.</add>
</seg>
</ab>
```

Figure 4: Canonicalised TEI Markup

This form of customisation is done with what is referred as a TEI ODD (one document does-it-all) document. This is a special form of TEI XML, including elements from the ›tagdocs‹ module, which details how a given schema will differ from the full TEI schema itself. From this document a choice of schemas (Relax NG, W3C, DTD) can be generated, in this way this document functions as a meta-schema. However, it is more than that. From a TEI ODD file not only can a schema be generated, but it can also be internationalised to provide element descriptions in the local user's language where available. Moreover, project reference documentation can be generated as well, which takes into account any of the changes the project has made to the TEI. These changes might include the removal of elements, attributes or modules not needed for this project, the addition of new elements or attributes, the renaming of existing elements or attributes, or changing of class memberships and content models. A TEI ODD can also contain other documentation (or indeed anything else from the TEI), and thus can be used for creating detailed project documentation. The TEI itself uses this format for the production of the TEI Guidelines, and using the same tools projects can generate customised documentation in HTML, XML, or PDF formats.

For the Godwin project I created a TEI ODD file which removed a significant number of elements and modules available in the TEI framework. Part of the reason to do this is to limit the choices available in a schema-aware XML editor when project staff were marking up the diary files. It is generally reasonable to assume that given less choice, the encoding will be done more consistently. Equally, in the cases of a number of elements, the acceptable value lists for individual attributes has been constricted further than in the general TEI schema. This means that the project workers can only add values from a controlled vocabulary which, even more than limiting the elements they have access to, improves their overall consistency. Another benefit of using a Relax NG schema generated from a TEI ODD file, is that in the editor chosen by the project (the ›oXygen‹ editor from SynchRO Soft) they are automatically provided with tooltip-style pop-ups displaying the TEI description for an element when they hover over that element, or similar assistance when adding elements or attributes.[5]

[5] Although the benefits of TEI customisation are highlighted here, and subsetting or conformable changes to the full TEI schema pose little problem, there are downsides to an multiplicative expansion of non-Conformant TEI-based customisations, which I have discussed elsewhere. Cummings (2007: 471ff.).

In the case of the Godwin project, this meant that they could have all their special elements, which I was sure always to make only as syntactic sugar (a simple renaming without real structural changes) for existing TEI elements, which made their editing experience more comfortable. In addition, I was able to constrain the @type attribute on many of these to a set list of values to provide the distinctions they wanted later to be able to extract.

```
<elementSpec ident="dMeal" mode="add">
 <equiv filter="godwin-acdc.xsl" mimeType="text/xsl" name="segLike"/>
 <desc>Marks a segment of diary text where a meal is described,
  including participants</desc>
 <classes>
  <memberOf key="model.blockLike"/>
  <memberOf key="model.segLike"/>
 </classes>
 <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">
    <valItem ident="B">
     <equiv/>
     <gloss>Breakfasts at P (with X, (Y, Z))</gloss>
    </valItem>
    <valItem ident="BG">
     <equiv/>
     <gloss>X, (Y, Z) Breakfasts</gloss>
    </valItem>
    <valItem ident="D">
     <equiv/>
     <gloss>Dine (at P) (with X, (Y, Z))</gloss>
    </valItem>
    <!-- more valItems removed -->
   </valList>
  </attDef>
 </attList>
</elementSpec>
```

Figure 5: The Godwin ODD

This figure is one part of a TEI ODD file, specifically an element specification for an element named ›dMeal‹. In this case we are adding this element, which has a subset of the same class membership and content model as a TEI <seg> element and therefore can always be reverted into such. A description of this element is provided so that it can be used in generating documentation and editor tooltips. A TEI <equiv> element is used here to indicate the location where a filter exists. In this case the @filter attribute indicates the name of an XSLT-file which can be used

to transform any document instance back to pure TEI, and the @name attribute refers to a named XSLT template which will handle this particular element. An attribute definition, for the attribute @type is also provided which instructs that the use of this attribute is required. A description of the attribute is given along with a closed value list, with descriptions for each of the possible values given. This means that in their chosen XML editor, not supplying a @type attribute is flagged as a validation error, and if they want to use it, they only have to select the correct value from a drop down list of the options given here while simultaneously being given reminders of the value's descriptions as tooltips.

It is, of course, easy to create a TEI ODD file one does not even need to edit the XML directly. The TEI Consortium provides a web-based frontend called ›Roma‹ to simplify the creation of a TEI ODD, the customisation of the TEI, the generation of schemas, and the production of documentation. This allows easy addition, removal, renaming, or modification of attributes, elements, and modules, the ability to save your customisation (as a TEI ODD file) and reload it at a later date to make further changes, generation of HTML and PDF documentation, and localisation of the interface, schemas, and documentation into a range of languages.

Prosopography

With their customised TEI schema the project members were set to work on marking up the diary files, which they did in successive passes each time making sure they were taking a year they had not seen before in order to help proofread each other's markup. After the basic structure of days and marking of dates, the next phase concentrated on instances of named individuals. One of the most interesting things for these researchers in Godwin's diaries is that he devotedly (not religiously) records the names of people he met with, whether this was during social meals or important meetings. Over a period of a couple months they managed not only to encode the structure, but to mark the meetings themselves and then in a separate pass through the files wrapped a TEI <persName> element around every single instance of a personal name. From that we can calculate that there are around 64,000 personal names with under 10,000 distinct individuals.

To record information concerning these individuals, the project uses separate TEI <person> elements, in separate files, which are later aggre-

gated together inside a virtual <listPerson> element. The more laborious step of identifying which individuals are being referred to in each instance of a <persName> was undertaken as a separate pass through the diary files. For each <persName> element the project attempted to provide a reference to the @xml:id of a <person> element if possible using a @key attribute that is later converted to a @ref attribute. In certain cases they came across single names which referred to a group of people. If, for example, Godwin was visited by a family called ›The Browns‹ he may only record that as an indication of who was there. However, historical knowledge and contextual information may tell the researcher that this is Mr and Mrs Brown and their daughter. If there are <person> elements for each of these, the project wanted to be able to refer to their individual @xml:id attributes (id est ›BR01‹, ›BR02‹, ›BR03‹). Unfortunately, in the initial release of TEI P5, you could only have one value for a @ref attribute. On behalf of the project, I submitted a bug report to the TEI Sourceforge site, and in due time the TEI Council agreed that this was indeed a corrigible error and this was corrected in a later release of TEI P5. Thus they could have a @ref attribute containing "#BR01 #BR02 #BR03" in a persName element around the text ›The Browns‹.

One of the decisions that will be required during processing is what form of user interaction is needed in case of a <persName> element with multiple @ref values. Does clicking on such a reference takes users to an automatically generated page where then they choose which member of the Brown family they are interested in? Instead it might take them to a page with information about all three that has been dynamically aggregated together? Might it be better for a user to be presented with three links that suddenly appear in the rendered view of the document as a user goes to click on them and the user have to decide before they click which of the links they want to follow? Whichever of these, or other, solutions is adopted, there ceases to be a unilateral relationship between a single name and a single person which may confuse some users. However, irrespective of how it is processed, this is a good example of both a project contributing back to the TEI (in filing a bug report) and the TEI responding to the needs of its community.

The project staff code up any section of the diary that contains a personal name with a <persName> element similar to this figure below.

```
<dEventG type="PersEvent">Birth of
  <persName key="SHE02">Mary</persName>, 20 minutes
  after 11 at night. <add place="subleft">From 7 to 10, Evesham Buildings.</add>
</dEventG>
```

Figure 6: Godwin <persName> Markup

In the canonicalisation of the files back to pure TEI, the @key attribute is replaced with a @ref attribute that points directly to the correct person record. The @ref attribute uses a URI to point to the @xml:id attribute as a target. This means that the values of the @ref attribute can point to @xml:id attributes not only in the same document, but anywhere on the internet through using a URL.

```
<seg type="dEventG" subtype="PersEvent">Birth of
  <persName ref="/people/SHE02.xml">Mary</persName>, 20 minutes
    after 11 at night. <add place="subleft">From 7 to 10, Evesham Buildings.</add>
</seg>
```

Figure 7: Canonicalised TEI <persName> Markup

In this case, the target of the @ref attribute will be the <person> record with the following @xml:id which is intentionally stored in a separate file. The @ref here could just as easily have pointed to "/people/people.xml#SHE02" where all the people are aggregated together. A benefit of the canonicalisation step is that any changes of processing logic, such as the location of the person records, can be simultaneously modified in this step. The individual file of a person record that this points to currently looks like the figure below.

```
<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">
    <placeName>England</placeName>
  </birth>
  <death when="1851-02-01"/>
  <occupation>writer</occupation>
  <note type="editorial">c.f. DNB for help in distinguishing</note>
</person>
```

Figure 8: Godwin <person> File

While the entry will eventually be expanded with more detail, this template entry gives an example of the lowest common denominator that the project is trying to create with the majority of identifiable people. When a user clicks on the instance of a person's name in the rendered view of the diary they will be confronted with a page that not only displays this information about them, but also statistically generated and

other information indicating when and where Godwin met these people, and providing links to those people this person was most often seen with.

It is worth commenting on how the project members go about creating these files. When they encounter a name instance in the diary text that they haven't coded, and believe that they are able to identify who it is (and since Godwin abbreviates this is a non-trivial task), and if the person record does not already exist, they create a new person record file. This follows a predictable pattern of the first three letters of the surname and two numeric digits. When they've created this file, they commit it to the project's subversion repository. Subversion is a version control system which stores all project's files, and also feeds into the front-end website, while allowing the project staff to work from any location they desire. As soon as they have committed one of these person records to the repository, an index is regenerated, sortable by name, id, birth and death dates, which links to a rendered view of this person record. Simultaneously, any instances of this person's name in the rendered diary files become links to this person record. Part of the benefit of this, as opposed to an internet-mounted database, is that the project staff does not need to be attached to the internet to complete their work, they simply update the repository the next time they are connected. This is not intended as a front-facing user interface, but merely an assistance to the project staff working on coding the names. Although usually TEI files must be complete, but in this case they are simply fragments of a single complete <person> element. This makes it easy for project staff to edit individual files for specific people, rather than worrying about conflicts on a single larger file, or needing to be connected to the internet to access some shared database. It also allows them to work on multiple computers, at home and the office, and know that they can always check out and entirely fresh copy or rollback to a previous version.

Transforming Godwin

The front-facing website is undergoing development and will be powered using a native XML database and XQuery. This will allow quick retrieval of a variety of views of the material and easy aggregation of relevant material. Underlying content will be made available for those re-

searchers who want to use it in ways not predicted at the time of the site's creation.[6] Unlike many electronic editions, the interests of the primary researchers here, are not those from the fields of literature or history – the project is being led by the politics department. Hence the primary interest is in the the network of relationships between people, the frequencies of contact, and other statistics which can be generated from the material they are encoding. In addition to provide them with some of this information as they've been going along, I've also created various lists and compiled information to assist their proofreading and consistency in application of their editorial standards. This includes lists of every new Godwin project element and attribute combination. For example, with the added element <dMeal> mentioned earlier, a combination of frequency, distinct-value lists will be produced for each and every possible @type attribute value. These lists are presented in a variety of manners such as split by year (and as CSV for the statistics department), and of course link back to the rendered diary text. This allows the project staff to make sure they are ascribing the same kind of editorial interpretation (through markup) to the same kind of textual instances in the diary.

While XQuery is intended to be used for the public site, XSLT2 has been used to create the majority of these lists. This has proven useful because the majority of difficulties in providing such lists are in fact simple grouping problems. XSLT2 grouping, with <xsl:for-each-group> in this case, allows for easy grouping and sorting of these elements and attributes in an efficient and quick manner. In the figure below, a fragment of generalised template takes an earlier defined ›element2Process‹ variable, and groups this element and attribute value XPath combination by the values lower-cased and with any white-space normalised, and sorts it by frequency, making an embedded list for each item, whose dates link back to the rendered view of the diary.

[6] Moreover, in the displayed final versions of texts unique identifiers will be given at sufficient granularity to allow texts to be referenced in a stand-off manner. This is partly intended so that they may be repurposed later when systems evolve for the interaction between editions. C.f. Robinson (2000); Robinson (2003); Robinson (2005).

```
<div>
 <head>Frequency List</head>
 <list type="unordered">
  <xsl:for-each-group select="$element2Process"
   group-by="normalize-space(lower-case(.))">
   <xsl:sort select="count(current-group())" data-type="number" order="descending"/>
   <xsl:sort select="current-grouping-key()"/>
   <xsl:variable name="count" select="count(current-group())"/>
   <xsl:variable name="date">
    <list>
     <xsl:for-each select="current-group()">
      <xsl:variable name="thisDate" select="./ancestor::tei:dDay//tei:date[1]/@when"/>
     <item> <ref target="{concat('/diary/', substring-before($thisDate, '-'),
      '.xml#g', $thisDate)}">
       <xsl:value-of select="$thisDate"/></ref>
     </item>
     </xsl:for-each>
     </list>
    </xsl:variable>
   <item><xsl:value-of select="concat($count, ' --  ', current-grouping-key())"/>
    <xsl:copy-of select="$date"/>
    </item>
  </xsl:for-each-group>
  </list>
</div>
```

Figure 9: An <xsl:for-each-group> XSLT Construct

This is a relatively straightforward grouping of elements, which produces a TEI list of one particular element and attribute value combination so that the project staff can see if they are being consistent in application of that combination. A rendered view of it is not necessarily pretty, but functional, as demonstrated by the figure below:

- 21 -- call on holcroft
  1788-10-01 1788-10-07 1788-10-20 1789-07-14 1791-09-07 1792-10-27 1792-11-27 1793-01-17 1793-01-25 1793-03-12 1793-03-28
  1793-10-29 1794-03-07 1794-03-15 1794-05-01 1794-05-02 1794-05-03 1794-05-05 1794-05-11 1794-11-05 1794-12-12
- 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
  1897-06-26 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
  1794-08-16 1795-06-13 1795-10-17 1795-10-23 1801-05-23 1802-04-21 1803-09-19 1803-10-24 1804-12-04 1805-01-24 1805-02-15
  1805-03-05 1805-03-20 1807-04-03 1807-07-15 1807-07-16 1808-04-25 1810-02-07 1810-04-04 1817-10-22 1828-05-28
- 21 -- wordsworth
  1798-04-16 1804-11-23 1804-12-07 1805-01-21 1805-04-24 1805-05-09 1805-05-23 1805-06-01 1805-06-21 1805-11-23 1805-12-13
  1805-12-20 1806-02-08 1806-05-07 1806-08-25 1807-01-31 1809-12-14 1811-02-23 1811-02-26 1811-06-13 1813-08-25
- 20 -- call 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 1805-06-20 1807-02-10 1812-06-11 1815-05-23
- 19 -- call on colburn n
  1826-10-24 1826-11-09 1828-07-22 1828-11-01 1828-12-16 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

Figure 10: A Simple Godwin Proofreading List

The number here is the frequency of this specific formulation of text inside the element and attribute value combination, and then provides all the dates where this is used. What is important to notice – and the reason this range of frequencies was chosen as an example – is that ›call on holcroft‹ and ›holcroft‹ appear separately. This is, of course, meant to highlight that absolutely no regularisation of any of this data is taking place. While this might initially appear as a flaw, it is in fact a benefit to those working on the project as it helps them track down typos and inconsistencies easily, especially when reading through the tail end of the list of phrases that only appear a single time.

Methodology and Training

The methodology the project has adopted in a number of instances has been exploratory and subject to change as they encounter new problems. Because of this the project decided that in the initial phases of the encoding of the diary files they would meet in a single room several times a week while they were doing their markup. These coding sessions meant they were able to consult each other easily, which lowered inconsistencies, and when new phenomena were discovered decide as a group on a single answer. This group-written internal editorial policy means that all project members feel a degree of ownership about the decisions of the project which further motivates their participation. Another effect seems to have been that they were able to work much more swiftly than many other projects, and the pressure of the scheduled coding sessions might be the reason for this. For each increasingly detailed layer of markup they applied to the diary files they undertook successive sweeps through the files as a whole. Moreover, they each took a different year (the diary files are organised by year) than they had previously which meant that they were able to proofread each other's work with fresh eyes. They have documented, in a ›blame sheet‹ who was responsible for each year during each phase of the process which makes each researcher more diligent with their assigned task. As the project has gone on, they have discovered aspects they have wanted to mark up and the TEI ODD file has been updated to allow for these and to provide new schemas to validate against. In many cases they have been impressed with the flexibility of the TEI when it is explained to them that this element they desire does indeed already exist, and simply had been removed from their original schema. The generated lists have enabled them to realise where they

could make finer distinctions in their editorial decisions. The successive sweeps through the diary files have allowed them all to become increasingly familiar with the diary texts. Overall this methodology is one that could be characterised as a sort of feedback loop, in which the output they create enables a greater precision of automated proofreading assistance which allows them further to be more precise.[7] This lather-rinse-repeat methodology improves the output, while simultaneously enabling greater possibilities, and thus more complicated demands for assistance and exploitation, from the technology involved.

One of the things that should be highlighted is the relative inexperience of the project with the processes and technologies they are using. When the project started they had a basic standard of computer literacy, able to use internet resources, read their email, and buy a book from amazon, but had little or no concept of markup, XML, or collaborative working systems. I provided them with two days of training. The first day concentrated on the concept of markup, XML, TEI XML, and then the Godwin-specific markup they were going to use, with practical exercises marking up a few diary entries. The second concentrated on refreshing some of that, but further introduction to and installation of the oXygen XML editor, an explanation of subversion and version control systems, the installation of a subversion client, and some practicing of adding and updating files to the subversion repository. What is impressive here is not the limited tuition that they were given, but that with such brief – but admittedly highly focused – lessons they were able to go away as a group and only need occasional support during the next few months, in which time they had powered through encoding several successive phases of markup layers.

As the project, at time of writing, is not yet complete a nice flashy front-facing website is not yet available. There is, obviously, a project website which is not currently password protected, simply intentionally unadvertised. When the project members commit something to the subversion repository, it is instantly available on this project website. This includes any errors they might make, and in some cases when saving files that are not well-formed they also have the power to break (parts of) this project website until they correct their error. While they could be protected from this, they have learnt more by having to be responsible about the validation of their files.

---

[7] This form of iterative methodology could also be seen in Eggert's idea of a ›work-site‹: Eggert (2005).

When the project nears completion (in autumn 2010) a more sophisticated website will be unveiled which will allow searching of the diary files in a variety of manners. The underlying XML will be made available for re-use by other researchers with a suitable license. Wherever possible open source software is being used and in most cases underlying scripts and queries will also be accessible for those who are interested in. Given the nature of the project, there are a lot of interesting possibilities for different forms of visualization of the data, especially the relationships between individuals mentioned in the diary. High resolution images of the diary are being taken to accompany the encoded text. The long-term hosting of the site will be done by the Bodleian Library.

Conclusion

This is an interesting project that is making available an intriguing set of historical material that will be a boon for researchers of the period. They have latched on to relatively new forms of encoding, and helped to expand the TEI itself, while making substantial use of the possibilities of transformation of document-like XML into data-like statistical information to glean the information they need for their own research. They've undertaken unfamiliar working practices and been rewarded with a resilient and flexible workflow that is customisable to their needs. A variety of transformation stylesheets have provided them with rich and varying views of the material they have been working on, which has further allowed them to improve their own methodology. While it is to soon to declare the project a success – that probably won't be known until significantly after it has finished – it looks like it is well on the way to providing a useful academic resource.

Bibliography

Cummings, James

2007 The Text Encoding Initiative and the Study of Literature. In: Ray Siemens/Susan Schreibman: A Companion to Digital Literary Studies. Oxford: Blackwells, p. 451-476.

Jahrbuch für Computerphilologie 10 (2008), Cummings "The William Godwin's Diaries Project" <http://computerphilologie.de/jg08/cummings.pdf> (29. April 2009)

Eggert, Paul
2005 Text-encoding, theories of the text and the »work-site«. In: Literary and Linguistic Computing 20/4, p. 425-435.

Robinson, P.
2000 The one and the many text'. In: Literary and Linguistic Computing 15/1, p. 5-14.

2003 Where We Are with Electronic Scholarly Editions, and Where We Want to Be. In: Jahrbuch für Computerphilologie 5, p. 123-143 and [4]

2005 Current issues in making digital editions of medieval texts – or, do electronic scholarly editions have a future? In: Digital Medievalist 1/1 [5].

Vanhoutte, E.
2006 Prose fiction and modern manuscripts: limitations and possibilities of text-encoding for electronic editions. In: J. Unsworth/K. O'Keefe/L. Burnard: Electronic Textual Editing. New York: Modern Language Association of America, p. 161-180.

Websites

[1] <http://www.politics.ox.ac.uk/research/projects/godwin_diary/> (23.02.2009).
[2] <http://www.oucs.ox.ac.uk> (23.02.2009).
[3] <http://www.tei-c.org/P5/> (23.02.2009).
[4] <http://www.digitalmedievalist.org/journal/1.1/robinson/> (23.02.2009).
[5] <http://computerphilologie.tu-darmstadt.de/jg03/robinson.html> (23.02.09).