Preview
Volume Number

Academics Retire and Servers Die: Adventures in the Hosting and Storage of Digital Humanities Projects

James Cummings <James_dot_Cummings_at_newcastle_dot_ac_dot_uk>, Newcastle University

Abstract

This article examines the technical development and afterlives of two projects, the *CURSUS* project (2000-2003) and the *William Godwin's Diary* project (2007-2010) to undertake case studies in problems relating to hosting and storage of digital humanities projects. In both cases a combination of outside events or project decisions negatively impacted the project. This was discussed as part of a symposium for the Endings Principles for Digital Longevity and reflects on whether following these principles would have benefited these projects. Overall, the case is made that we should always be planning for events that could affect the sustainability of digital research projects.

Introduction

This article looks back at the technical development of two projects, The *CURSUS* project (2000-2003) and the *William Godwin's Diary* project (2007-2010), as case studies in which outside events or internal decisions negatively impacted the project.[1] The *CURSUS* project is a collection of XML editions of medieval liturgical texts, and The William Godwin's Diary project makes available 32 volumes of Godwin's diaries covering 1788 to 1836. Although both websites are still available, they have had difficult histories and both should be considered at risk of disappearing from the web at any moment. The *CURSUS* project faced challenges partly because the principal investigator passed away. The *William Godwin's Diary* project suffered because the people maintaining the project were not given the resources to do so, and because all stakeholders left the institution. As the main technical developer for both of these projects, I have direct insight into the technical background and general lessons learned, which are useful to examine as case studies for each project before contrasting them with the *Endings Principles for Digital Longevity ([Endings 2021]; hereafter "Endings Principles")*. Although the projects pre-date these *Principles*, they are worth comparing to each other given that both of these projects are TEI-based editions and relate directly to the issues the *Principles* tackle. It is easy, of course, to criticize the technical decisions of a project decades afterwards; seeing in hindsight how the digital landscape has evolved in the meantime gives us an immediate sense that we would certainly have done things differently and definitely have made different choices. In these cases, since I was the developer and technical consultant on these projects, I do not need to guess. While I would like to think I would make better decisions now, I know that I would not unless I were time-traveling back with my current knowledge and experience. These two projects are useful proxies for examining many issues of the longevity of our research outputs. 1

Part of the overall argument I am making through these case studies is that we should always plan for events that affect the sustainability of digital research projects. Some of these events are obviously unpredictable, while others are of our own making. While we pay lip service to doing so in funding applications, once a project is under way, unexpected challenges present themselves and our solutions to them should adhere to an underlying set of principles. We need to plan for failure, plan for the project being cut short, plan for staff disappearing, and plan for all kinds of threats to the project by adopting and following principles such as those recommended by the Endings Project in the *Endings Principles*. 2

CURSUS: An Online Resource of Medieval Liturgical Texts

From 2000-2003 I was fortunate enough to be employed by the AHRB-funded *CURSUS* project as a postdoctoral research associate. This project was the brainchild of Professor David Chadd of the School of Music at University of East Anglia (UEA) and sought to produce editions of medieval Benedictine Latin liturgical texts online, and experimentally to explore the use of XML publication for such materials. The main purpose was to enable investigation of the order of antiphons, responds, and prayers in these liturgical manuscripts. The project proposed that this order of service in many ways gives a fingerprint of the liturgy in different places in England. We edited these liturgical items with the utmost care, but for other aspects such as the biblical readings transcribed only the first few and last few words, since the point of the project was not to spend time making a critical edition of the Latin Vulgate Bible. 3

In the end, the project produced editions of twelve main texts and a number of ancillary works. These included an XML conversion of the Latin Vulgate Bible, derived lists of individual incipits, and a repository of liturgical items (antiphons, responds, and prayers) based on the *Corpus Aniphonalium Officii* (CAO). The approach the project used was not merely to edit a liturgical item in its context in the edition of a particular manuscript, but also to create a textual-critical apparatus in a repository with all the other textual variants from CAO and any other manuscripts edited by the project. In this way each liturgical item in the edition of a manuscript, instead of existing in the source file for that manuscript edition, is in reality a pointer to a specific antiphon, respond, or prayer in the project's CAO repository file. Pointers are not a URI-based system in TEI P4, so extraction involves looking up the correct reading for that manuscript to display and importing it to this point through an XSLT stylesheet. This makes for a dense CAO repository file but for very light editions where most of the file is filled with pointers. 4

CURSUS Technical Background

A worked example, looking at one antiphon to which the *CURSUS* project gave the ID 'c5111' and how this and references to it are encoded, will help to explain some of the technical details of the project.[2] The c5111 antiphon appears, among other places, in the Peterborough Antiphoner[3] during the Vespers service for Maunday Thursday. Most manuscripts have a different date for this antiphon; all the other manuscripts edited on the project, for whatever reason, use it on the Tuesday or Wednesday of the medieval Christian Holy Week. In the underlying XML of the manuscript the pointer to this specific antiphon is encoded as in Figure 1 below. 5

```
<xptr href="Antiphons" type="aBody" from="ID(c5111)"/>
```

Figure 1. An xpointer to the *CURSUS* project antiphon c5111.

What becomes instantly obvious (to those familiar with XML markup standards at least) is that this does not follow the current TEI P5 Guidelines. Indeed, the *CURSUS* project predates TEI P5, which was not released until 2007, and uses a project-specific extension of TEI P4 XML for its markup.[4] In Figure 1 a TEI P4 `<xptr/>` element points to the ID 'c5111' in the CAO repository of antiphons, responds, and prayers, and the markup instructs the processing to retrieve the content of the `<aBody>` element there (denoted by the value of the type attribute). A mere examination of the form of this `<xptr/>` in isolation does not explain much, but looking at it in progressively greater context shows more of the liturgy-specific elements that the project had added to its use of the TEI. The (custom) `<antiphon>` element that surrounds it is seen in Figure 2. 6

```
<antiphon> &AE;<add place="at top of music stave" resp="DC" hand="later">
  <rubric type="general">Cantabitur cum neuma tribus vero diebus sequentibus versiculi debent
      pronuntiari sine n<add place="interlined">e</add>upma nisi solummodo in vigilia
      paschae ad completorium</rubric></add>
  <hi rend="music">
    <xptr href="Antiphons" type="aBody" from="ID(c5111)"/>
    &Ev; <add place="on music stave below notation of differentia">Magnific<supplied>at</supplied></add>
  </hi>
</antiphon>
```

Figure 2. An `<antiphon>` element from the *CURSUS* project.

In Figure 2 the mixture of TEI P4 elements (such as “add,” “hi,” “supplied,” and “xptr”) and the project’s own additions (such as “antiphon” and “rubric”) are evident. Once the project principal investigator learned of the extensibility of the TEI as a system, he preferred discipline-specific terminology for important aspects of the text. Although there are differences between this markup and what one might see today in TEI P5, the general intentions and interpretations of the markup are still relatively clear. 7

```
<Day num="5" code="07065000">
  <rubric type="dayTitle">In cena domini</rubric>
  <!--***   First Vespers -->
  <service name="vespl"><rubric type="serviceTitle"> Ad vesperas&punctus;</rubric>
    <antiphon> &A; <add place="on music stave, above initial notation">
      <rubric type="general">Ebdomadarius cantor</rubric></add>
      <hi rend="music"><xptr href="Antiphons" type="aBody" from="ID(c4570)"/></hi>
    </antiphon>
    <incipit type="psalm">&P;
      <hi rend="music">Laud<expan abbr="&hook;" resp="DC">ate</expan>&punctus;</hi>
    </incipit>
    <incipit type="respond" init="C"><rubric type="general">duo ad gradum&punctus;</rubric>&R;
      <add place="on music stave, below notation" resp="DC">duo ebdomadarii&punctus;</add>
      <hi rend="music"> Circumdederunt me&punctus;</hi></incipit>
    <incipit type="versicle">&V; Homo pacis meae&punctus;</incipit>
    <antiphon> &AE;<add place="at top of music stave" resp="DC" hand="later">
      <rubric type="general">Cantabitur cum neuma tribus vero diebus sequentibus versiculi debent
          pronuntiari sine n<add place="interlined">e</add>upma nisi solummodo in vigilia
          paschae ad completorium</rubric></add>
      <hi rend="music">
        <xptr href="Antiphons" type="aBody" from="ID(c5111)"/>
        &Ev; <add place="on music stave below notation of differentia">Magnific<supplied>at</supplied></add>
      </hi>
    </antiphon>
    <add place="bottom margin" resp="DC" hand="later">Notandum quod si aliquid festivitas
      albarum vel duodecim lectionum in crastino contigerit fiet de ea commemoratio ad istas
      vesperas&punctus; si ferialis nulla processio fiet propter passionem quod
        <del>legend</del> lecturus est ad magnam missam&punctus; </add>
  </service>
  <!-- More services  -->
</Day>
```

Figure 3. The entire *CURSUS* project encoding of the Vespers service that contains antiphon c5111.

Even in the full context (Figure 3) of the entire Vespers service that contains antiphon c5111, there are only a few additional non-TEI P4 elements (such as “Day,” “service,” and “incipit”). What starts to become evident through these examples, at least to those familiar with TEI P4 markup, is that this project took full advantage of a TEI DTD-based feature allowing the creation of custom DTD entities for repeated formulaic text, punctuation, and markup that are very common in the highly repetitive liturgical documents. This means that the encoders did not need to include repetitive portions of markup, and merely used a smaller “entity” to stand in for that markup. Even in the context shown through the markup in Figure 2, the use of the entity “`&AE;`” at the beginning of the “`<antiphon>`” element demonstrates this exploitation of a technical method of providing formulaic text and markup. The cursus.ent file listing the project’s DTD entities expands this as seen in Figure 4. 8

```
<!-- Symbol or Characters for Antiphon, use as &A; -->
<!ENTITY A '<rubric type="antiphon">Ant.</rubric>'>
<!-- Symbol or Characters for Antiphon with In Evangelio, use as &AE; -->
<!ENTITY AE '<rubric type="evangelio">In Evangelio</rubric> &A; '>
```

Figure 4. The *CURSUS* project DTD entities file showing the “A” and “AE” entities.

In this file the “AE” entity is shown to be expanded to a string of textual markup showing the rubricated text “In Evangelio.” Moreover, this formulaic text itself recursively includes an additional custom entity “A” which is replaced with the rubricated label “Ant.” While there are a number of drawbacks to such a system that mean it would not be recommended today, it was an imaginative exploitation of the ability of DTD-linked documents to provide re-used bits of text and markup.[5] However, this approach does introduce a significant fragility based on all files’ dependence on the DTD and entities file being present and accessible at the time of processing. If the files were to be normalized to TEI P5 XML, then a conversion process would have to begin with a basic identity transform which would expand all of these entities into their non-entity form. 9

This discussion explains what the encoding of the xpointer only for antiphon c5111 looks like solely in the context of one of the manuscript edition files, but the content of the antiphon is stored with all its textual 10

variants in the CAO repository file. Figure 5 gives an example of this file.

```
<ant previd="c5110" nextid="c5112" id="c5111">
    <header>
        <usage ms="Cdm" code="07064000">[77r] prime</usage>
        <usage ms="Cht" code="07063000">[58v] vesp2</usage>
        <usage ms="Ely" code="07064000">[88r] sext-a1</usage>
        <usage ms="Evm" code="07063000">[146r] vesp2</usage>
        <usage ms="Glo" code="07063000">[78r] vesp2</usage>
        <usage ms="Hyd" code="07064000">[93v] nones</usage>
        <usage ms="Pet" code="07065000">[88r] vesp1</usage>
        <usage ms="Wcb" code="07063000">[114r] vesp2</usage>
    </header>
    <aBody wit="CAO-C CAO-G CAO-B CAO-E CAO-M CAO-V CAO-H CAO-R CAO-D CAO-F CAO-L Ely Cht Hyd Pet Evm Wcb Glo Cdm">
        Tanto tempore vobiscum eram
        <app>
            <rdg wit="CAO-G CAO-B CAO-M CAO-V CAO-H CAO-R CAO-D CAO-F CAO-L Ely Cht Hyd Pet Evm Wcb Glo Cdm">
                docens vos in templo</rdg>
            <rdg wit="CAO-C CAO-E">docens in templo</rdg>
        </app>
        <app>
            <rdg wit="CAO-C CAO-G CAO-B CAO-E CAO-V CAO-H CAO-R CAO-D CAO-F CAO-L Ely Cht Hyd Pet Evm Wcb Glo Cdm">
                et non me tenuistis</rdg>
            <rdg wit="CAO-M">non me tenuistis</rdg>
        </app>
        <app>
            <rdg wit="CAO-C CAO-G CAO-B CAO-M CAO-V CAO-H CAO-R CAO-D CAO-F CAO-L Cht Hyd Pet Evm Wcb Glo">
                modo flagellatum</rdg>
            <rdg wit="Cdm">
                <add place="interlined">et</add>
                modo flagellatum
            </rdg>
            <rdg wit="CAO-E Ely">et ecce flagellatum</rdg>
        </app>
        ducitis ad crucifigendum
    </aBody>
</ant>
```

Figure 5. The *CURSUS* CAO repository file antiphon entry for c5111

New elements that the project introduced here (including ant, header, usage, and aBody), like most of the additions made by the project, could have been modeled in other ways with TEI elements. For example, the <ant> element could have used a TEI <ab> (anonymous block) element. However, in customizing the TEI, the project used element names that made sense for its encoding needs. The bespoke markup and TEI P4 XML of this project could, should there ever be a need, be converted to TEI P5 XML to be used in a modern production environment. Indeed, it is because of the adoption of the open international standard of the TEI that the markup would be straightforward to convert, even in the case of this extensive customization. The entries in the header are given as standardized numerical codes for all days of the medieval liturgical calendar created by the *CANTUS* project and are used to generate the links into the manuscript editions.[6] Similarly, the website generated version of the files automatically adds in the IDs of the previous and next liturgical items to the enriched form of the antiphon during the process of extraction by XSLT stylesheets into individual files. These are pieces of convenience data to enable easier processing of the underlying XML for online browsing. 11

Figure 6 shows the HTML view of this CAO repository entry – the entries are all browseable online and are dynamically transformed to HTML pages on the fly. The textual variants are laid out in parallel boxes showing their mutual differences with the witness sigils discreetly present, mimicking the underlying TEI parallel segmentation markup. In modern websites, these variants would likely be presented in a very different manner – the interface certainly shows its age, having been created in 2002. 12

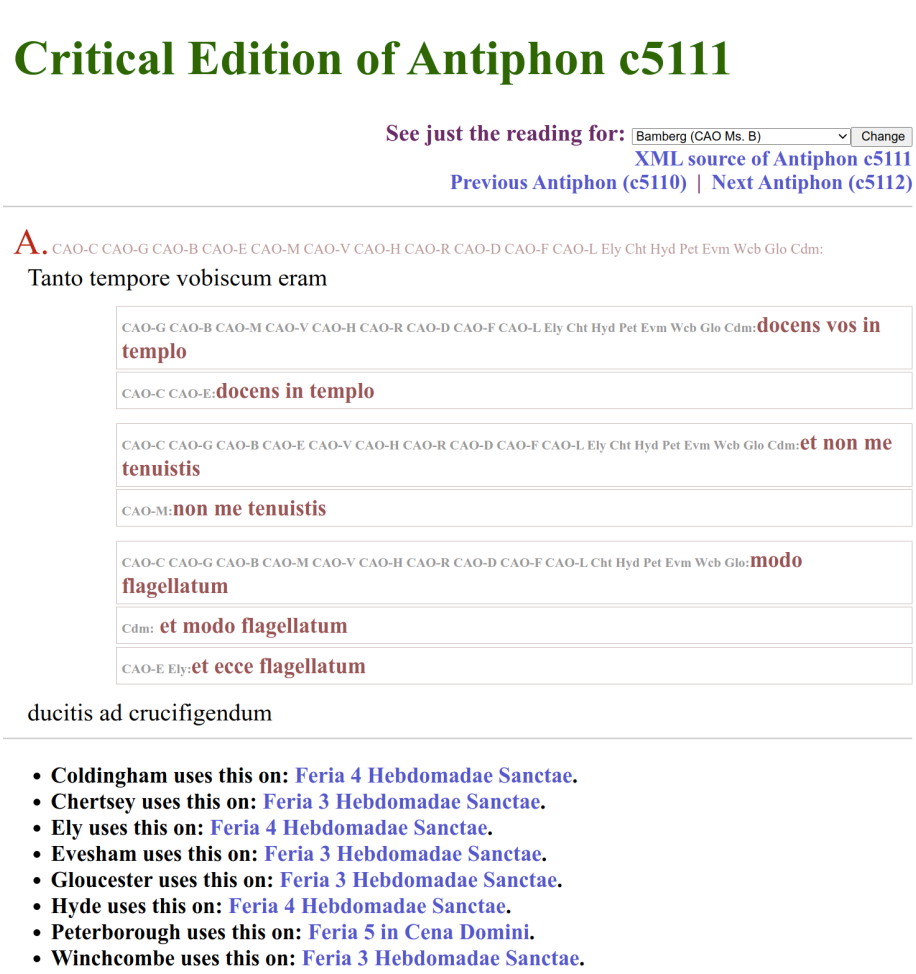

Figure 6. A *CURSUS* antiphon web page for antiphon c5111.

This form of interlinking between manuscript edition and repository entries, which then link through to all the other places this antiphon is used, promotes a circular and generally beneficial form of user experience in the navigation. While the XML files store most of the intellectual output of the project, preserving only the XML files would mean that we would lose the argument presented by the interface for how we should interact with such editions. The aspects of interface as edition might be lost if we look only at the underlying files; the edition itself is a publication that includes not only the underlying data but the manner in which it was presented. The output of projects is not only the data they produce but the way in which those products are presented to the end user — when research is published in print, the book is the product, not the research notes or data that underlie it; in other words, the interface is the means of preservation. However, as with many digital resources, perhaps in a demonstration of an appreciation of the much vaunted but seldom realized re-use of materials, we concentrate on the underlying data as the primary output to the detriment of other aspects of the outputs. As a historical artifact representing digital editions from this period of our development in thinking about how to present such editions online, the CURSUS interface deserves preservation or conservation, in the same way one might argue the primary source documents do. 13

CURSUS Project Afterlife

By 2003 the main *CURSUS* project was completed. Professor David Chadd and I had edited more manuscripts and produced more additional outputs than promised in the AHRB funding bid — but my contract came to an end, so I left UEA for a post at the Oxford Text Archive, University of Oxford. However, Professor Chadd continued work and did not really consider the project concluded, only its funded portion. He occasionally continued to update the website, and, in order to simplify some of the underlying publication technology and assist with technical aspects generally, he later employed Dr Richard Lewis (then a departmental postgrad).[7] This ongoing work meant that the project never reached a point at which Professor Chadd felt that working files could be canonicalized in their final form. 14

Sadly, in late 2006 Professor Chadd died after a short illness. Although his passing did not immediately affect the website, it did truly bring the project to a close. Professor Chadd continued to edit manuscripts as and when he could up until his death, but, because he was not using a version control system or frequently uploading these to the website, his final work has been lost.[8] But even death was not the most impactful event in the *CURSUS* project afterlife. 15

More dramatically, an event totally unrelated to the *CURSUS* project in November 2009 had a severe detrimental impact upon it: a hacker illegally obtained over 1,000 emails spanning 13 years, along with 3,000 other documents, from the Climate Research Unit at UEA. A combination of ignorance and willful misunderstanding meant that even any unprofessional or confusing comments in them were used by climate change deniers to spread misinformation. This event became known as “ClimateGate” and as a result UEA temporarily closed all off-campus access to its servers.[9] 16

Originally, we had “set up one of the project’s desktop machines as a Debian Linux server” [Cummings 2006] but shortly before leaving I suggested it might be best to make it more official, and in my mind more stable, on a departmental server. Throughout all the tumultuous events above, even though the PI of the project had passed away, the *CURSUS* website continued to run unproblematically on the School of Music departmental server. The site had begun to be cited in journal articles, and not just by those directly concerned with musicology or digital publishing [Licence 2006]. 17

However, sometime in 2010 a software upgrade on that server caused the website to go down, and it needed a configuration change and a restart. With the continuing ban on off-campus access and the fact that neither Dr Lewis nor I were in Norwich or had worked for the university for a few years, we had no easy way to restart the server. By this point the School of Music had little IT support, and UEA’s central IT had no ability to take on these extra duties. Indeed, later in 2010 the School also replaced their departmental server with a new one – having no local champions or even people who knew much about this legacy site, the *CURSUS* website at the University of East Anglia finally disappeared.[10] 18

We had backed up the latest version of the data on the server shortly after Professor Chadd’s death, of course, and it was around this time that I first contacted UEA to discuss getting the rights to put the website up again or hosting it elsewhere. Starting with the School of Music, who disavowed any responsibility for the site, I was eventually put in touch with UEA’s Commercialisation Manager from Research and Enterprise Services. A slow process of back-and-forth explanations resulted in 2016 in a Creative Commons Attribution Non-Commercial license for the intellectual property – after six years of negotiation – for a project that always had intended its data to be open but hadn’t explicitly licensed it as such.[11] Indeed, UEA had eventually closed its School of Music in 2014, after hard-fought campaigns to keep it, so even the academic department that had created and hosted this project was now gone [Cunnane 2011] [BBC 2011]. The closure of the department caused additional confusion as it was unclear to the university who in authority might give permission to license this data. But perhaps we did not even need permission: UEA as an institution did not even know it owned this data and the project’s intention was always to release it openly. An alternative, more assertive approach, would have been to put the site up while simultaneously pursuing the permission to do so.[12] While the website is still up, the problems noted above concerning sustainability of such resources, indicate that even to preserve the project as a working resource has an inherent fragility. If it is not practical or feasible to preserve such resources, recording a screencast video using the website might be one solution at the very least to document the ephemerality of user experience and interaction for the future.[13] 19

Since 2016, the site has been hosted on a small personal VPS run by Dr Lewis and the cursus.org.uk domain name (paid for by me), and the underlying data and code stored in GitHub.[14] This arrangement is contingent and precarious; Dr Lewis might decide to stop hosting the site or I might not continue to pay the domain name registration fee. Although people still do use the site and the data, it raises the question of when and how to retire websites and merely preserve the data in case someone wishes to re-use it at a later date. Part of our reason for hosting it is a feeling that the main constituency of users would find the data hard to use in its underlying format. A compromise would be to undertake a project to flatten the website, removing any need for server-side processing and make this available online in a variety of forms. As the website was mostly static (only the XML to HTML transformations being dynamic), it would be easy to flatten it. The XSLT 20

stylesheets and an HTML copy of the website as currently served (with relative links) was added to maintain a coherent version in the zip archive uploaded to Zenodo.[15] Updating the data to be more usable, for example converting to TEI P5 XML, is not a large project but would take a bit more work to rationalize the bespoke elements (only lightly documented in the DTD — these days we would document these extensions fully with a TEI ODD). [16]

Lessons from the *CURSUS* Project

21 Numerous red flags and warning signs indicated that the *CURSUS* project was at-risk and not properly preserved. Two decades on, it would be gratifying to be able to say that digital humanities projects no longer do such things. Alas, we cannot. However, there have been significant changes that make the preservation of project outputs more likely. For example, instead of keeping the master copy of data on the PI's laptop, digital humanities projects now use GitHub routinely to store both versioned code and data while enabling collaboration.[17] Instead of using departmental servers stored under someone's desk to host funded research projects, we now use institutional or virtual machines hosted in the cloud. Major institutional and international data repositories now exist for depositing copies of research outputs but did not at the time.

22 The majority of problems in the project's afterlife are those that we did not reflect upon at all while the project was running. For example, we never considered the so-called "Bus Factor" – the minimum number of people who have to vanish from a project before it completely stalls due to the lack of necessary knowledge to sustain it. In this case it was Professor Chadd's sudden death and the fact that legacy planning for the project site had not been undertaken. Since no other project staff were still employed by UEA, the intellectual content of the project could easily have slipped away. It is only because Dr Lewis and I wanted to preserve it, partly as a tribute to Professor Chadd's legacy, that it survives at all. We should also recognize the *CURSUS* project's strengths – in being fairly simple TEI-based XML with XSLT 1.0 conversion to HTML (albeit dynamic) without any JavaScript or additional libraries, it could easily be migrated and preserved. It is, I would argue, the relative simplicity of the site that has enabled its preservation. Just as a physical object in good condition might be preserved by benign neglect, this website's minimal footprint, structured text-only content, and generally uncomplicated needs, enabled it to survive relatively well despite being subject to benign digital neglect. More active interventions and conservation might have resulted in the use of more advanced frameworks, features, or libraries, as these became popular, causing more long-term harm for its sustainability than otherwise (see Holmes and Takeda in this issue).

23 One of the reasons for describing the technical background of this project, and giving the detailed explanation of their use of the TEI Guidelines, is that it demonstrates the kind of complexity that could be achieved within the *Endings Principles*. To be clear, the *CURSUS* project does not meet those requirements since, in its current form, it still requires bespoke dynamic pipelined conversion of the underlying XML to HTML. While the experimental DTD-based markup entities were interesting, the fragility they introduce means such an approach should have been used only for the development of materials, not the publication copies. It would have been better to expand all the custom entities in all the files (as well as normalize the markup vocabulary) for a publication version, but with the ongoing nature of the project as Professor Chadd worked towards his (sadly unachieved) retirement, the curse of the perpetual beta meant that it never quite seemed "finished." It would have been better if regular fixed releases were made, where the production system files (as opposed to the development version) were always in their final canonicalised forms. The server-side dynamic transformation does not add much that, with hindsight, could not have been accomplished with pre-generated flat files. The project was experimenting with various methods of digital publication of TEI-based XML and relished the idea of being "able to have virtual URLs, allowing our users to create dynamically-assembled pages from a variety of XML source files" [Cummings 2006]. It would have been better to pre-generate output files with massive redundancy of every possible view as the *Endings Principles* suggest.

24 A lesson for encoding and technical expectation management might be drawn from the number of new custom entities and bespoke elements the project created. Once Professor Chadd discovered the ease with which we could create new bespoke elements, his default approach when a new textual phenomenon or encoding problem was encountered was to create a bespoke element. Taking the easiest approach in any

digital project always results in trade-offs and compromises for longevity. This approach sometimes meant that we were encoding things with new project-specific elements rather than looking slightly harder in the community for standardized approaches and solutions. While the *CURSUS* project was groundbreaking in some of its experiments in the publication of digital medieval resources, hindsight reveals some problems which should have been foreseen at the time.

William Godwin's Diary

A similar story is that of the *William Godwin's Diary* project at the University of Oxford, which ran several years after the *CURSUS* project. Funded by the Leverhulme Trust from 2007-2010, this project coincided with the Bodleian Libraries' receipt of funding from the National Heritage Memorial Fund and various donations to buy the Abinger Collection which, among many other things, includes William Godwin's diary. Godwin "was the founding father of philosophical anarchism and was also a major novelist, although he is perhaps better known today as the husband of Mary Wollstonecraft and the father of Mary Shelley" [Philp 2021]. His diary contains 48 years of records in 32 octavo notebooks, written in often highly abbreviated entries that the project had to decipher and disambiguate. People's names are often given only as initials and there is little detail concerning the substance of any meetings and the meaning thereof can be hard to decipher.[18] Initial transcriptions provided as MS Word documents were converted to TEI P5 XML. 25

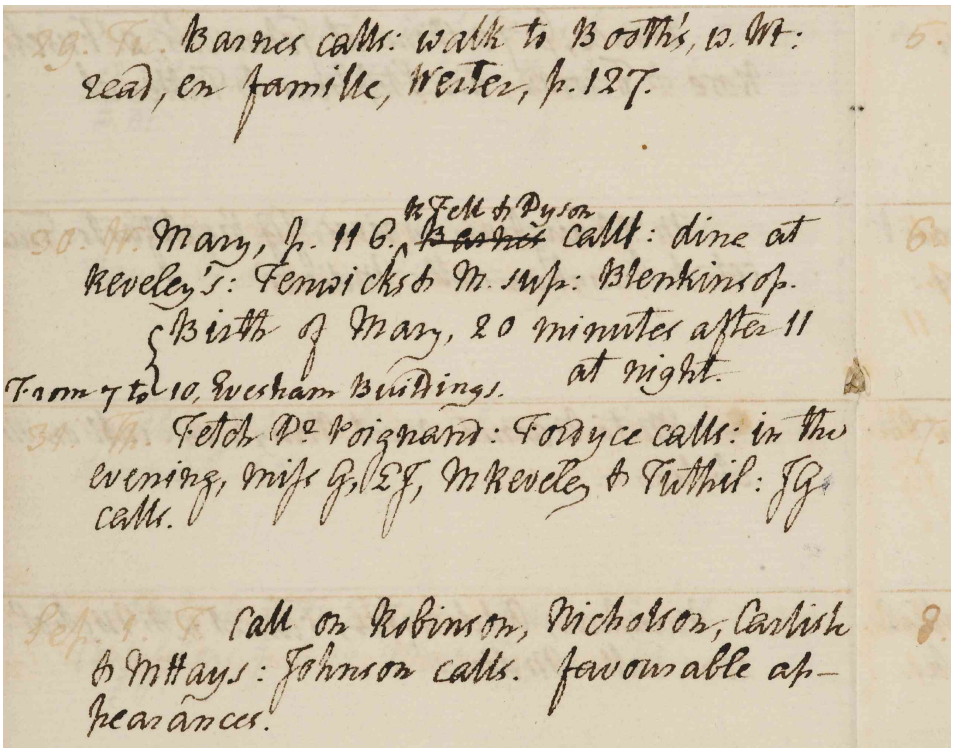

Figure 7. William Godwin's diary entries including that of 30 August 1797 where he, concisely, notes the birth of his daughter Mary (later Mary Shelley) at 11:20pm.

The project was interested in highlighting relationships between Godwin and other people and extracting datasets of information from the diaries. To do this, I trained a team consisting of the PI (Professor Mark Philp), a postdoctoral research associate (Dr David O'Shaughnessy), and two DPhil students (Kathryn Barush and James Grande) in a project-specific TEI P5 customisation that was very reduced and used renamed elements to make encoding these entries easier.[19] The XML they were hand-encoding in successive passes consisted of fewer than 20 separate elements in total but was automatically expanded to full TEI on the website each evening. So while the project might encode using a non-TEI `<dMeal>` (diary meal) element, the renamed element was converted to a pure TEI `<seg type="dMeal">` element in the production XML [Cumming 2008]. Similarly, the project workflow meant checking work into an institutional Subversion version control system; unlike the *CURSUS* project, the real risk of data loss, while never zero, 26

was minimal. The overall training took just over a day, but the project benefited by having me as their on-call technical support when they needed it.

William Godwin's Diary Technical Background

27 The website was built on top of an early version of eXist-db (a native XML database) that at the time used Apache Cocoon for URL-based pipelined conversions, with which I was familiar from the *CURSUS* project. I had experimented with eXist-db during the *CURSUS* project; an XML database was desired because of the nature of the queries to be done — again on the fly — against a fairly complicated XML dataset. At the time, the popular idea that it "may not be possible to achieve one input — *all* outputs, but surely one input — many outputs is an entirely practical goal" [Walsh 2002] was brought from the *CURSUS* project and taken to heart. The project leads viewed it as not only feasible but also desirable to generate many of the project outputs from a single or at least small number of input datasets. Indeed, the project tried to create a network of interlinked inputs that generated a network of intertwined outputs that would decentralize any starting point and thus be open to exploration. The *William Godwin's Diary* site, a "finely engineered architecture of XML tagging[,] allows the user to trace acquaintances across the decades, books through his library catalogues, and the author himself through the streets of the metropolis" [Bullard 2013, 752].

28 The project methodology of working in phases (first adding structural markup, then adding markup recording meetings, then recording and eventually identifying people, places, and events) meant the encoders always moved on to a new year of the diary they had not seen before. This iterative but distributed process meant that fresh eyes saw each entry several times, thus acting as additional proofreaders to reduce human error. From the point of view of the *Endings Principles*, this methodology ensured that after each phase the content was coherent and complete (as far as it went).[20] This methodological approach is something that many DH projects use to ensure a phased production of output work. The encoders also had diagnostics using a local XSLT stylesheet to transform their work into a debugging "proofreader's" view that allowed them to spot mistakes through formatting realized in a manner that could never be acceptable in a user-oriented front end. This view highlighted any elements that should have had content but didn't, and coloured things in a way to emphasize aspects that had not yet been completed.

29 The website infrastructure was fed directly from the Subversion repository and automatically produced a number of views on the website when updated each evening. The nightly jobs would transform any updated (but well-formed and valid) files into pure TEI P5 XML, load these into the eXist-db database and regenerate a wide variety of tables of information. These included not only lists of plays Godwin went to or books he was reading/writing, but also detailed tables of meals he had (and whether he was dining with the person or the person was dining with him), and other forms of meetings. These were displayed using the external jQuery library DataTables plug-in to provide a filterable, pageable, sortable view of the data that would certainly be against the *Endings Principles* if they had existed at the time.[21] The level of detail of the encoding and cross-references throughout are what enable the site to provide detail to researchers using it. As one reviewer wrote,

> What makes the site an amazing research tool is the level to which Godwin's meticulous (but brief) notes are cross-referenced against one another, creating a vast web of information that not only fleshes out the skeleton of the author's life, but also provides a wealth of information about Romantic-era social networks and day-to-day life in the London of the period. [Thomas 2018, 603]

Figure 8. DataTable of meetings where Godwin was calling on a person

As one might expect, the diary data itself is organized chronologically. Each diary day is represented by a TEI `<ab>` element with a @type attribute of “dDay” and is required to have an @xml:id attribute based on the date. Each diary entry is required to have a `<date>` element with a @when attribute provided in W3C format, but the element will have transcribed text content here only if it existed in the diary. The production server markup uses only standard TEI elements, with arbitrary segments denoting meals, meetings, or similar concerns of the project using the `<seg>` element. 30

```
<ab type="dDay" xml:id="g1797-08-01">
   <date when="1797-08-01">Aug. 1. Tu.</date>
   <ref type="dWrote" subtype="write" target="/works/life01.html">Life, 4
   pages</ref>.
   <ref type="dText" subtype="read" target="/bibl/te0170.html">Nouvelle
      Eloise, p. 16</ref>.
   <seg type="dMeal" subtype="D">Dine at
      <persName ref="/people/JOH01.html">
         <placeName type="venue">Johnson's</placeName>
      </persName>, w.
      <persName ref="/people/FUS01.html">Fuseli </persName>,
      <persName ref="/people/BON01.html">Bonnycastle</persName> &amp;amp;
      <persName ref="/people/GRE04.html">Gregory</persName>
   </seg>:
   <seg type="dMeeting" subtype="C">call on
      <persName ref="/people/DAV02.html">
         <placeName type="venue">Davis</placeName>
      </persName>
   </seg>:
   <seg type="dMeal" subtype="SG">
      <persName ref="/people/OPI01.html">Opie</persName> &amp;amp;
      <persName ref="/people/ALD02.html">A A </persName>sup
   </seg>.
</ab>
```

Figure 9. The *William Godwin’s Diary* XML entry for 1 August 1797.

One of the major intellectual contributions of this encoding work — and a significant demonstration that good text encoding really is research in itself — is the identification and deduplication of 50,000 of the 64,413 instances of names in the diary.[22] In Figure 9, they are mostly given as their surnames, but even “A A” is identified and points directly to the website’s ALD02.html file where information about Amelia (Opie) Alderson and her 119 appearances in the diary is provided. Arguably, the data should have pointed to the underlying XML file (ALD02.xml) but this file URI was provided as part of the transformation to pure TEI for the website as a processing convenience. Similarly, the decision was taken early on that when a person’s name was 31

mentioned as a venue where a meeting took place, the encoders would denote this distinction for various project needs, by embedding a `<placeName>` of `@type` “venue” inside the `<persName>` content. While this might have made more sense the other way around (a place name which had a personal name as its content rather than a personal name that had a place name as its content), this method was chosen for consistency with other decisions on the project.

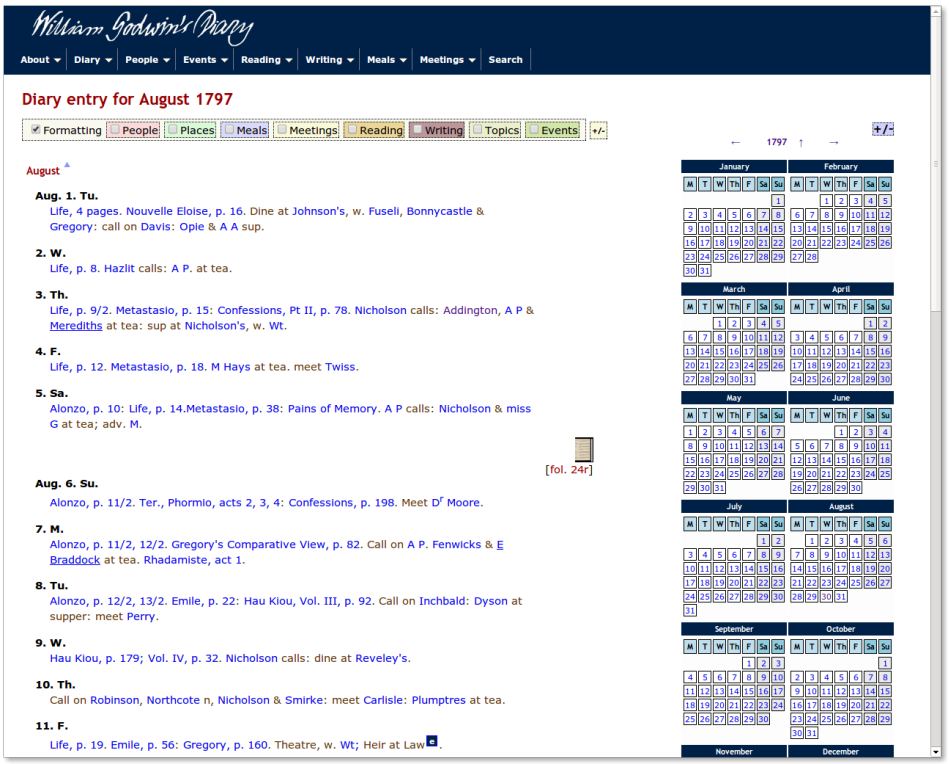

Figure 10. The *William Godwin's Diary* website showing the beginning of August 1797.

The display of the diary entries themselves is as running text based on calendar entries. One can view a single day, a particular month (as in Figure 10), or a whole year at a time, all extracted dynamically from the underlying XML year file. Clicking on the year, month, or day in the accompanying calendar provides an easy form of date-based navigation for users. All of the transformations are done dynamically, converting underlying XML with XSLT to HTML for display. However, in some cases the query to discover, extract, and then convert the XML took so long that pre-cached static copies of the XML were created.[23] As one can tell from Figure 10, almost all the content of the diary entries links to more information. Any mention of people, places, meals, meetings, texts Godwin was reading or writing, topics discussed, or events links to more information about that named entity. In the display of the website, jQuery is used to toggle on/off highlighting of these to enable people to discover them more easily. If one toggles on highlighting of people, for example, then all the names of people are highlighted with a light pink box.[24] If one clicks on a name — for example, in Figure 10, the name “Johnson’s” in the “Dine at Johnson’s” in the entry for the 1 August 1797 — the name links to a page assembling not only the project-provided biographical details, editorial notes, and bibliography concerning the identified Joseph Johnson, but also an automatically-generated chart and details of all of his appearances in the diary that can be used to navigate back through the diary (Figure 11). This page includes some basic statistics, here that he was mentioned 413 times, not at home when Godwin called in 23 times, and listed as a venue 324 times. The statistics clearly indicate that Godwin called on Johnson much more than the reverse and that he also liked to host gatherings of all sorts that many attended – the data providing rich views of their interactions over the course of Godwin‘s life. Through such techniques, the site navigation still creates a richly encoded endless loop of exploration for researchers over a decade later, as noted by Thomas:

> Part of the intent behind the site’s extensive interlinking is to allow readers to navigate the diary in any order that they choose, permitting a type of exploration that facilitates

serendipitous insights (and, for enthusiasts, affords the simple amusement of searching for whatever the author was doing on any given day). [Thomas 2018, 603]

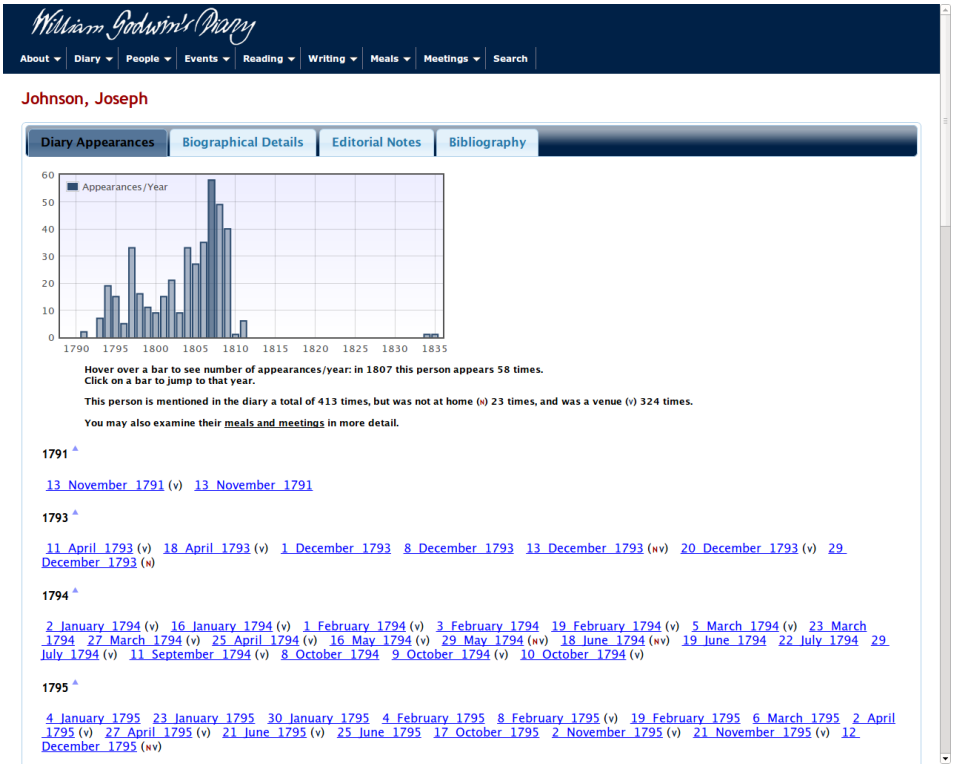

Figure 11. The *William Godwin's Diary* person page for Joseph Johnson.

Another direction of travel from the diary pages is to click on the thumbnail for that week's page image. As part of the agreement in receiving funding to purchase the Abinger Collection, the Bodleian Library imaged all of Godwin's diary. The project PI decided strongly against the facing-page Text/Image that is common in many digital scholarly editions. Instead, he wished to privilege the edited text, only providing links to the images as thumbnail images that are placed alongside each page of text as a way of making these linked resources available. However, linking through to these images has proved very useful to those undertaking Godwin studies who now routinely include snippets of them in their slides and articles.

Figure 12. The *William Godwin's Diary* pan/zoom Google Maps-driven image viewer.

The images were taken and tiled, and a pan/zoom image viewer was built using the Google Maps API, long before collaborations such as IIIF were available (Figure 12). Today, a project would simply embed a IIIF viewer for the digitized images that the Bodleian makes available on its Digital Bodleian website, even though doing so introduces a dependence on an external service. Indeed, one of the reasons the images proved so popular with researchers was not the easy interface, but the fact that we provided a link to the full high-resolution image and had (after much effort) convinced the Bodleian to license the images with a Creative Commons Attribution license. With this full image, researchers could crop the portions they were discussing and use them in their conference presentation slides. However, the necessity of having both the full high-resolution image and pre-generated tiles for each of the page images at each level of magnification added significantly to the size of the virtual machine to be requested, which in turn may be partly why it was only reluctantly hosted by the Bodleian Library. For a project following the *Endings Principles,* there is a calculation to be done concerning any additional storage costs for hosting its images locally versus the fragility of pointing to an external institutional image server. 34

Project Afterlife

The project ran successfully and completed on time and budget in 2010, and in 2012 won an award for Digital Resources from the British Society for Eighteenth-Century Studies.[25] Shortly afterwards, the DPhil student assistants completed their theses, and the postdoctoral research associate moved on to another university, and within a few years the PI of the project moved on to the University of Warwick. The project Twitter account continued to tweet out diary entries every few hours (until a Twitter authentication change meant the script started to fail). Nevertheless, although the project was over, the website continued to function, relatively unproblematically. And yet, one of the aspects of hosting that was problematic with the *William Godwin's Diary* project was the question of where institutionally the website was to be hosted. As the Bodleian Libraries did not yet have a centralized media server (now https://digital.bodleian.ox.ac.uk/), and the images were hosted locally on the virtual machine (VM) that housed the website, the VM would need to survive with all it needed, but this need meant a fairly large (for the time) VM. The availability of resources always affects projects' afterlives. 35

36

One of the reasons for mentioning (above) the sources of funding for the Bodleian's purchase of the rest of these (and other) materials of the Abinger Collection was that as part of the purchase agreement the library had promised that web hosting for this project, like the imaging, would be provided pro bono. At the beginning I brought together representatives of the project with those from Bodleian Digital Library Systems and Services, who agreed to provide a suitable VM inside the Bodleian Libraries' infrastructure. When it came to it, however, since I was the sole technical developer on the website and happened to work for the Oxford University Computing Services (OUCS), the Bodleian decided that I should just set up a VM in Computing Services, and that they would transfer it over to their infrastructure at the end of the project. At the conclusion of the project, the Bodleian then made the argument that the OUCS VM was working well, and so even though it had a Bodleian URL there was no pressing reason to move the underlying VM to their infrastructure. It was only several years later, when the Computing Services (by then IT Services) and Bodleian Library were both having major upgrades to their VM infrastructure, that I finally convinced them that the VM should be moved under the Bodleian Library's care. The operating system itself was updated to the latest long-term support version, but as I had only limited time to donate to the project, the version of eXist-db (now quite ancient) was not upgraded and just copied across. Even so, the VM remains one of those VMs on an older infrastructure on the outskirts of the purview of Bodleian's overworked infrastructure team.[26] As it is, the website occasionally needs a restart; when the VM is rebooted, the website does not automatically start, and occasionally a small partition fills up with log files.[27] Both of these are problems that would have been addressable while I had access to the server, but in an immediately post-project setting, it never seemed important enough to ensure such access. While I would restart the website maybe once or twice a year while working there, I have had no local access to the server since I left the University of Oxford in 2017; as there is no one connected to the original project left at the institution, restarting falls to the busy infrastructure team of the Bodleian Library. This need to care for legacy websites and their compounding maintenance burden will become only more pressing as we continue to produce digital projects that are deemed too important to just "turn off" after a discrete period.

Lessons Learned from *William Godwin's Diary*

37 It may sound as though I am faulting the Bodleian Libraries, and especially the Bodleian Digital Library Systems and Services section, for not providing a VM at the beginning of the project as they had promised to do. And while they should have done so and enabled development or at the very least a production server to be hosted on their infrastructure, I certainly have sympathy for their point of view. Before the start of the project, the upper levels of the library administration committed them to providing support from system administrators but without providing any extra resources to the staff to enable this additional work. The necessary imaging of materials was provided, pro bono, and they promised to host the website but, when the development of the website had finished, hosting was understandably not a priority task for them. If I had been working directly for the Department of Politics and International Relations (where the project originated) and had not had sophisticated local IT resources at the time, then there would have been no question that the Bodleian would have provided the necessary infrastructure — the project would have been impossible without it. However, I happened to work for the Oxford University Computing Services, and so I had easy access to VMs and it seemed natural to them that I should provide these at no extra cost or immediate effort from them. This kind of approach is dangerous, and projects should be sure to get formal agreements of in-kind support where feasible.

38 This project relied on a single developer (me) to provide some degree of unpaid support long after the project had finished. This is why the *Endings Principles* are so important – they encourage discussion of the project afterlife. In the case of the *William Godwin's Diary* project, the project team gave little consideration to what would happen to the site after it was handed over to the Bodleian. Much like a physical output such as a book, everyone seemed to think it would be handed over and that would be that. But digital research outputs are not like books; they demand care and attention, however fleeting, irregular, and inconsistent. This remains true, though to a much lesser degree, of static websites as well which at very least need servers to run on. While of course I had backups of the *William Godwin's Diary* data, and it was deposited in the institutional repository, after my departure in 2017 I no longer had access to the server or subversion repository. I only got around to uploading the underlying data and code to an open GitHub repository in

November 2019.[28] I had also previously provided a full copy of the data to the *Shelley-Godwin Archive* at the Maryland Institute for Technology in the Humanities, who have had their own ongoing maintenance burden for the sites they create [Muñoz & Viglianti 2015].

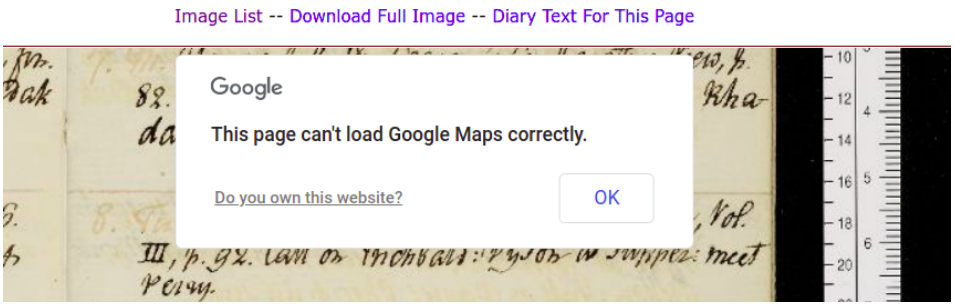

Figure 13. Google Maps API error on *William Godwin's Diary* website.

Using freely-available but proprietary systems like the Google Maps API to build a pan/zoom image browser was always going to be a compromise.[29] The API has had major backwards-incompatible updates at least twice since the project adopted it. Indeed, I am amazed that it still works at all. There is an error message notifying the user that it cannot load Google Maps, but the message can be easily dismissed and the viewer functions as before, since all of the image tiles needed are on the local server. As I mentioned earlier, if built today the project would certainly have used the IIIF media server that the Bodleian has set up in the meantime. With archival digital image resources there is an open question on how best to interact with them and still adhere to the *Endings Principles*. In discussing these issues as part of the *Endings Project* Symposium, Martin Holmes pointed out that the use of IIIF might imply a viewer (such as Mirador) with server-side requirements and the consequent longevity problems inherent in any such software. This is true if the IIIF is hosted as part of the project website, but one could design the site so that only the remote media server's viewer were used (embedding or merely linking to it) and if one day the images disappeared, the site could degrade gracefully, presenting only the text view. For some projects, a fine balance must be struck between treating whole remote collections of images as research objects and hosting these resources locally [Fenlon 2019]. A centralized institutional image store may make sense, especially where the same set of images may be used for multiple projects, but it needs to be carefully incorporated into the site in a manner that prepares for graceful degradation of any linked resources.[30] 39

Comparison with *The Endings Principles*

There are certainly a number of commonalities between the *CURSUS* and *William Godwin's Diary* projects (other than my involvement) from which we might learn. The latter project, having been developed some years later, benefited from some of the lessons learned from CURSUS. 40

Technology is bound to change — some of what now seems impossible will become plausible and eventually a reality and developing the resources of the future will inevitably be different. Indeed, in archive-based research projects digitizing and exploiting textual sources the development and significant improvements made to handwritten text recognition (a machine-learning-based technique for transcribing handwritten documents, though also applicable to print) promises the glimmering hope of an age where whole archives (or those materials in similar scripts at least) will be able to be transcribed (c.f. [Terras 2021] [Muehlberger 2019]. With a wealth of text available, some will exploit these corpora through programmatic analysis, but others will want to edit and up-convert the material to provide interpretation and make it more accessible to readers. As expected, the *Endings Principles* for most projects do a good job in suggesting some ground rules that will always result in easily archivable outputs.[31] Looking at these Principles in comparison to the projects above will highlight some of their many flaws but looking at these points of failure gives us a method by which we can improve. 41

In the creation of the data, the *CURSUS* project did not conform to open standards. While based on TEI P4, the encoding diverged from P4 significantly and did not put the effort into canonicalizing the data before 42

release.[32] Similarly, the files were not really subject to version control.[33] The *William Godwin's Diary*project was better in sustaining itself these areas, by using standard TEI P5 and a subversion repository. Both did employ validation and diagnostic analysis of a sort.[34]

Neither site truly meets the principles for documentation, in that while they provided some high-level project documentation about the content, neither documented the technical infrastructure very well.[35] Although the *CURSUS* project had a commented DTD, it merely noted the element that was being added; it didn't explain any rationale or give a description of it. With full hindsight, CURSUS should have followed the *Endings Principles* on licensing; although the original web page stated that the outputs were released "openly", establishing the rights after the fact was a major hurdle.[36] This lesson had truly been learnt by the time of the *William Godwin's Diary* project, which licensed all of its materials with a Creative Commons Attribution Non-Commercial 3.0 license. 43

In the websites' processing, there were also good and bad aspects. As they were XML-based systems, the source files were always valid against a defined schema, though the "relentless validation" did not always extend to the HTML and CSS.[37] While these were both validated at one point, the validation requirements have also changed in the intervening decades, as guidelines for web accessibility have improved. Neither site had true continuous integration, but both had some degree of automation for the generation and testing of the sites.[38] The Godwin project had a variety of behind-the-scenes automated testing, proofreading, and checking before any file was copied into the database. The processing code was treated similarly, though should have been recognized as being more contingent.[39] The benefit of a static site for some projects in conforming to the *Endings Principles* (long after these two projects) is the reduction in the burden of server-side processing, that is, making sites "untethered from the processing that created them" (Holmes and Takeda in this issue). Both of these projects have benefits resulting from those aspects that are static, but fragility still exists. 44

The main fragilities of the two projects lie in their production of outputs/products. Both were dependent on server-side software, *CURSUS* merely for searching and pipelining dynamic transformations to HTML, but *Godwin* for its entire XML database-backed infrastructure.[40] The *Godwin* project, in using jQuery for a number of aspects, fails at the *Endings Principles'* caution against fashionable technologies and external libraries or services.[41] CURSUS is better here, as many of these technologies, which either did not yet exist or were too complex to use, but suffers the opposite problem in having later created bespoke server-side software (e.g. PyCoon). I suspect that these will be some of the hardest Principles to convince developers to follow since they often love fashionable external libraries and services that appear to make their urgent tasks easier in the short-term.[42] 45

Neither site has URLs that contain query strings and both have straightforward URLs for all individual entities on the site.[43] The two sites also provide all the underlying data for download, though both could do better at documenting it.[44] While both sites do provide all the necessary data in order to function in the page, they do not truly meet the 'massive redundancy' that the *Endings Principles* suggest.[45] In both projects, while an individual page might have all the text necessary for its own functions, it might link to a shared liturgical item as a source that can take users to other pages, or provide additional prosopographical information never intended to be embedded in that individual page. 46

Both sites meet the principles of graceful degradation reasonably well. As primarily text-based sites, they function fairly well, though in an even uglier way, with JavaScript and CSS turned off.[46] With *CURSUS* this functionality is partly because it predates the mass adoption of JavaScript and uses only fairly simple CSS. With *Godwin,* conscious attempts were made to enable graceful degradation; in reality, this approach only means mitigations such as a system whereby clicking on the main menu item with JavaScript disabled leads to a page containing the menu sub-items as a nested list. Happily, this choice is also beneficial for user navigation with touch interfaces, whose popularity was only beginning at that point, and is something that is still sometimes neglected even in responsive mobile sites. 47

Two concessions in the *Endings Principles* [47] (that good static websites generated from data may be enhanced with server-side tools like eXist-db or external libraries) might justify some aspects of the *Godwin* website, but not convincingly, since these tools were the foundation of the site rather than an added extra. This concession might have justified adding a Swish-E index to the *CURSUS* site, although Swish-E has since been removed because it failed when the site was rehosted. 48

Neither site fully met the principles for release management. In both cases the website was updated as and when particular data was available or software improvements were complete, with no news items, warnings, or even public version numbering.[48] While all files are valid, the project release could not be said to be both coherent and complete, or have an edition/version number.[49] Neither site gives unambiguous information on how to cite any specific page of the resource, although providing this information would have been easy. [50] Similarly, changes made to the site during the lifetime of the project meant, to our shame, that resources at particular URLs were not persistent.[51] Although neither site fully meets the ideals of the *Endings Principles*, it is surprising how little might need to be done to upgrade, flatten, and prepare these projects for a more archival afterlife. 49

Conclusion

In the funding bids for academic projects, Data Management Plans often extol the standards used and the long-term preservation benefits resulting from the way their data and websites will be constructed. But, as with the agreements for in-kind support or partnerships with external partners, projects should also plan their afterlife with more than just funding bid fictions. However good a digital research project's sustainability plan is, it is still very rare for the lifespan of most digital projects to outlast their creators for very long. Just as humans should not leave it up to the grief-ridden survivors to guess at what should be done with their effects after they pass away, neither should digital humanities research projects leave it to librarians or technical teams to decide what should be done with their outputs. Instead, we should all be clear in advance what the plan is for the eventual sunsetting of projects, having already archived our well-documented data long in advance, and not rely on the best efforts of those left to interpret what should be done with them. Some websites will be rejuvenated and preserved as the front-facing access to their data is seen as too important (or costly) to merely archive, while others will become nothing more than a ZIP archive downloaded by those who really want to explore the data. But we should plan for failure, and we should plan for the project being cut short for any reason with no notice. To realize these plans, we should make the data we produce as transportable and transparent as possible. We should simultaneously recognise that all aspects of an archival research project's website may be important, not only the underlying data but the choices made in presenting it. Some may be inconsequential, as with both of these two websites, but we should be mindful that the interface through which the data is presented also forms a part of the editorial argument and needs to be preserved if possible. Limiting oneself to minimal technologies may facilitate this.[52] 50

Many of the project flaws identified in this article are partly down to inexperience (as many of the technologies were only just emerging), and the eternal problem of busy people having too much to do and relying on shortcuts. Indeed, there were so many moments along the way when these websites could have been destroyed through policy change or simply vanished through neglect. It is easy to see now in hindsight what should have been done — and by and large what should have been done was to follow the then not-yet-created *Endings Principles*. 51

Notes

[1] The article started as a brief video and then symposium talk for The Endings Project Symposium and the slides initially used are available at https://slides.com/jamescummings/endingsproject2021/. Although there is substantial discussion of TEI Markup that is useful in recording for posterity the nature of the projects, in the end it is this portable markup that helps in their preservation. I hope those unfamiliar with TEI markup will bear with those parts in reading this article. While it is necessary to give the detailed background for those readers who are interested in these TEI aspects, the lessons learnt are also equally applicable to non-TEI projects.

[2] For additional material on the *CURSUS* project technical decisions see Cummings, 2006.

[3] *The Peterborough Antiphoner*. Cambridge, Magdalene College Ms F.4.10. An Antiphoner of the fourteenth century from the Benedictine Abbey of St Peter, St Paul and St Andrew, Peterborough, Northamptonshire.

[4] For the now deprecated TEI P4 Guidelines see https://tei-c.org/Vault/P4/doc/html/index.html [Sperberg-McQueen & Burnard 2002].

[5] The amount the *CURSUS* project used this feature of DTD-based document processing to include any repetitive portions of text should not be underestimated. Not only rubricated labels such as 'Ant.' but accents, unusual punctuation, character abbreviation markers, books of the Bible, portions of the `<teiHeader>`, and manuscript witness information were also included. The use is more akin to how one might use XInclude or similar these days. For the full horror see https://github.com/jamescummings/cursus/blob/master/dtd/cursus.ent.

[6] See http://cantusindex.org/ for more information about *CANTUS* and related projects.

[7] Dr Lewis replaced very early experiments using eXist-db 1.0b1 to index and search with Swish-E, and created a replacement for the Apache Cocoon system that we used for pipelining the dynamic conversions with a tool called "Pycoon" (because it was written in python). See https://github.com/ironchicken/pycoon for more information about Pycoon.

[8] c.f. Endings Principles, 1.2: "Data is subject to version control (Subversion, Git)."

[9] This hack was partly responsible for catapulting climate change denialism into the public consciousness [Raman & Pearce 2020]. For more information on ClimateGate, the Wikipedia article contains a fairly in-depth explanation https://en.wikipedia.org/wiki/Climatic_Research_Unit_email_controversy. Fact Check (a project of the Annenberg Public Policy Center that focuses on debunking political misinformation) also has a good summary https://www.factcheck.org/2009/12/climategate/.

[10] Fortunately, it did have several snapshots taken by the Wayback Machine of the Internet Archive. See for example https://web.archive.org/web/20061012165859/http://www.cursus.uea.ac.uk/ed/c5111 for antiphon c5111 from October 2006 shortly before Professor Chadd's death. Until we reinstated the *CURSUS* website at http://cursus.org.uk/, we directed people to the Wayback Machine instead.

[11] c.f. Endings Principles, 2.2: "All rights and intellectual property issues should be clearly documented. Where possible the Data and Products should be released under open licenses (Creative Commons, GNU, BSD, MPL)".

[12] As the pioneer of early computer science Rear Admiral Grace Hopper has been quoted as saying: "It is easier to ask forgiveness than permission." The *CURSUS* Creative Commons licensing discussions certainly reinforce that lesson. Under the intellectual property policy of UEA at this date, the institution and not the individual academic owned the IPR. The AHRB encouraged the production of open access outputs but did not mandate it at this point. In preparation for the Endings Project Symposium Martin Holmes queried the small amount of risk that putting the site up would have entailed, and while it crossed our minds, we were also busy with other projects.

[13] The idea of recording interactive sessions with digital editions to preserve a sense of their functionality came up in discussion during the Endings Project Symposium. Of course, having videos helps but the full interactive user experience is lost,, and any video will need to be preserved with the resource in a standard accessible format.

[14] The original URL, http://cursus.uea.ac.uk/, ceased to function in 2010. The http://cursus.org.uk site holds a legacy copy of what was on that site at the time.

[15] As part of writing this paper, a copy of the data and code, stored in GitHub at https://github.com/jamescummings/cursus, was archived and put into the international open repository Zenodo https://doi.org/10.5281/zenodo.5090613 ensure its survival.

[16] TEI ODD is the machine-readable meta-schema documentation and customisation format for the TEI Framework that most projects employing TEI should use to record their project's schema and encoding guidelines.

[17] c.f. Endings Principles, 1.2: "Data is subject to version control (Subversion, Git)."

[18] For more information about the diary of William Godwin see Mark Philp's "William Godwin (and his diary)" https://www.digitens.org/en/notices/william-godwin-and-his-diary.html.

[19] These were the project members with whom I interacted, but I should also note the involvement of the co-editor Dr Victoria Myers from Pepperdine University who provided many of the initial draft transcriptions as Word documents before their conversion to TEI P5 XML. For a full acknowledgements list see http://godwindiary.bodleian.ox.ac.uk/team.html.

[20] c.f. Endings Principles, 5.2: "A release should only be made when the entire product set is coherent, consistent, and complete (passing all validation and diagnostic tests)."

[21] For more information about DataTables see https://datatables.net/. Using DataTables would contravene the *Endings Principles* 4.3: "No dependence on external libraries or services: no JQuery, no AngularJS, no Bootstrap, no Google Search."

[22] For more general numerical statistics of the diary content see http://godwindiary.bodleian.ox.ac.uk/stats.html.

[23] An example of this is in the source XML for the DataTables of information for both identified and unidentified people. Either the query was composed in an inefficient manner, or the XML poorly indexed, but the result took several seconds to retrieve, even locally. So the decision was made to pre-generate the results of the query and merely transform this to HTML when required. This approach is halfway along the route to the creation of a fully static website that would be more compatible with the *Endings Principles*.

[24] If one toggles on all of this named-entity formatting, the site becomes an unreadable fruit salad of inaccessibility. The PI was challenged several times on his insistence for this feature. Using colours to denote semantics is usually a poor choice for reasons of accessibility.

[25] See https://www.politics.ox.ac.uk/news/william-godwins-diary-wins-award.html for information about this award.

[26] Those whose institutions have entirely centralized IT provision may find such parochial outlooks and redundant duplication inside a highly collegiate university confusing; so do many who have worked there.

[27] In a moment of deep irony, on the weekend during which I initially wrote these precise paragraphs the website was down. Given its legacy position, out-of-date software, and operating system, it should be considered at-risk, and yet is still frequently consulted and cited by those studying Godwin who likely have little sense of its precarity. Scholars may reasonably expect that a

site hosted by the Bodleian Libraries is stable. But this expectation places additional maintenance burdens on their technical support teams.

[28] Although this GitHub repository https://github.com/jamescummings/godwindiary acts as another copy for preservation means, it would make sense to deposit a copy of this with the images into an international repository like Zenodo. I have not yet done this, mea culpa.

[29] Other pan/zoom browsers of the time (such as OpenLayers which was quite popular) were tested and the user experience of the Google Maps version was preferred by the PI.

[30] c.f. Endings Principles, 4.7: “Graceful failure: every page should still function effectively even in the absence of JavaScript or CSS support” and 4.9 “The use of an external library may be necessary to support a specific function which is too complex to be coded locally (such as mapping or cryptography). Any such libraries must be open-source and widely-used, and must not themselves have dependencies.”

[31] The *Endings Principles* are available at https://endings.uvic.ca/principles.html.

[32] c.f. Endings Principles, 1.1: “Data is stored only in formats that conform to open standards and that are amenable to processing (TEI XML, GML, ODF, TXT).”

[33] c.f. Endings Principles, 1.2: “Data is subject to version control (Subversion, Git).”

[34] c.f. Endings Principles, 1.3: “Data is continually subject to validation and diagnostic analysis.”

[35] c.f. Endings Principles, 2.1: “Data models, including field names, descriptions, and controlled values, should be clearly documented in a static document that is maintained with the data and forms part of the products.”

[36] To be fair, Creative Commons had just recently launched during the years of the *CURSUS* project and the number of academic research projects in the humanities using CC Licences was still tiny. c.f. *Endings Principles*, 2.2: “All rights and intellectual property issues should be clearly documented. Where possible the Data and Products should be released under open licenses (Creative Commons, GNU, BSD, MPL).”

[37] c.f Endings Principles, 3.1: “Relentless validation: all processing includes validation/linting of all inputs and outputs and all validation errors should exit the process and prevent further execution until the errors are resolved.”

[38] c.f Endings Principles, 3.2: “Continuous integration: Any change to the source data requires an entire rebuild of the site (triggered automatically where possible).”

[39] c.f Endings Principles, 3.3: “Code is contingent: while code is not expected to have significant longevity, wherever possible, all code should follow Endings principles for data and products.”

[40] c.f. Endings Principles, 4.1: *No dependence on server-side software: build a static website with no databases, no PHP, no Python.*

[41] c.f. Endings Principles, 4.2: “No boutique or fashionable technologies: use only standards with support across all platforms, whose long-term viability is assured. Our choices are HTML5, JavaScript, and CSS”, and 4.3: “No dependence on external libraries or services: no JQuery, no AngularJS, no Bootstrap, no Google Search.”

[42] What is really needed, and StaticSearch is a good start, is end-to-end solutions that make it easier for projects to follow the Ending Principles. They should be using software because it is easy and as a benefit also get *Endings Principles* compliance [Cummings 2019].

[43] c.f. Endings Principles, 4.4: “No query strings: every entity in the site has a unique page with a simple URL that will function on any domain or IP address.”

[44] c.f. Endings Principles, 4.5: “Inclusion of data: every site should include a documented copy of the source data, so that users of the site can repurpose the work easily.”

[45] c.f. Endings Principles, 4.6: “Massive redundancy: every page contains all the components it needs, so that it will function without the rest of the site if necessary, even though doing so means duplicating information across the site.”

[46] c.f. Endings Principles, 4.7: “Graceful failure: every page should still function effectively even in the absence of JavaScript or CSS support.”

[47] c.f. Endings Principles, 4.8: “Once a fully-working static site is achieved, it may be enhanced by the use of other services such as a server-side indexing tool (Solr, eXist) to support searching and similar functionality” and 4.9: “The use of an external library may be necessary to support a specific function that is too complex to be coded locally (such as mapping or cryptography). Any such libraries must be open-source and widely-used, and must not themselves have dependencies.”

[48] c.f. Endings Principles, 5.1: “Releases should be periodical and carefully planned. The ‘rolling release’ model should be avoided.”

[49] c.f. Endings Principles, 5.2: “A release should only be made when the entire product set is coherent, consistent and complete (passing all validation and diagnostic tests)” and 5.3: “Like editions of print works, each release of a web resource should be clearly identified on every page by its build date and some kind of version number.”

[50] c.f. Endings Principles, 5.4: “Web resources should include detailed instructions for citation, so that end-users can unambiguously cite a specific page from a specific edition.”

[51] c.f. Endings Principles, 5.5: “URLs for individual resources within a digital publication should persist across editions. Any moved, retired, or deleted resources no longer available at a previously accessible URL should be redirected appropriately.”

[52] As part of preserving the interface, as mentioned earlier, screen capture videos should be recorded in standard formats documenting the interactivity and user experience of the website for archiving alongside the data because the videos may be useful to information historians at a later date.

Works Cited

BBC 2011 BBC. (2011) “University of East Anglia closes school of music”, *BBC Website*, 28 November 2011. Available at: https://www.bbc.co.uk/news/uk-england-norfolk-15919759.

Bullard 2013 Bullard, P. (2013) “Digital Humanities and Electronic Resources in the Long Eighteenth Century”, *Literature Compass*, 10, pp. 748–760. Available at: https://doi.org/10.1111/lic3.12085.

Cumming 2008 Cummings, J. (2008) “The William Godwin’s Diaries Project”, *Jahrbuch für Computerphilologie* 10, pp. 1–19. Available at: http://computerphilologie.de/jg08/cummings.pdf.

Cummings 2006 Cummings, J. (2006) “Liturgy, Drama and the Archive: Three Conversions from Legacy Formats to TEI XML”, *Digital Medievalist*, 2(1). Available at: http://doi.org/10.16995/dm.11.

Cummings 2019 Cummings, J. (2019) “Opening the Book: Data Models and Distractions in Digital Scholarly Editing”. *International Journal of Digital Humanities*, 1(2), pp. 179–193. Available at: https://doi.org/10.1007/s42803-019-00016-6.

Cunnane 2011 Cunnane, S. (2011) “Campaigners Battle Plans to Close UEA’s School of Music”, *Times Higher Education*. Available at: https://www.timeshighereducation.com/news/campaigners-battle-plans-to-close-ueas-school-of-music/418013.article.

Endings 2021 Endings Project. (2021) *Endings Principles for Digital Longevity*, Version 2.1. Available at: https://endings.uvic.ca/principles.html.

Fenlon 2019 Fenlon, K. (2019) “Modeling Digital Humanities Collections as Research Objects”, *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pp. 138–147. Available at: https://doi.org/10.1109/JCDL.2019.00029.

Holmes & Takeda 2022 Holmes, M., and Takeda, J. (2022) “From Tamagotchis to Pet Rocks: On Learning to Love Simplicity through the Endings Principles”, *Digital Humanities Quarterly*.

Licence 2006 Licence, T. (2006) “Goscelin of St Bertin and the Life of St. Eadwold of Cerne”, *The Journal of Medieval Latin*, 16, pp. 182–207. Available at: https://doi.org/10.1484/J.JML.2.303234.

Muehlberger 2019 Muehlberger, G. (2019) et al. “Transforming Scholarship in the Archives through Handwritten Text Recognition: Transkribus as a Case Study”, *Journal of Documentation*, 75(5), pp. 954–976. Available at: https://doi.org/10.1108/JD-07-2018-0114.

Muñoz & Viglianti 2015 Muñoz, T., and Viglianti, R. (2015) “Texts and Documents: New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive”, *Journal of the Text Encoding Initiative*, 8. https://doi.org/10.4000/jtei.1270.

Philp 2021 Philp, M. (2021) “William Godwin (and his diary)” *The Digital Encyclopedia of British Sociability in the Long Eighteenth Century*. Available at: https://www.digitens.org/en/notices/william-godwin-and-his-diary.html.

Raman & Pearce 2020 Raman, S., and Pearce, W. (2020) “Learning the lessons of Climategate: A Cosmopolitan Moment in the Public Life of Climate Science”. *WIREs Clim Change*. 11:e672. Available at: https://doi.org/10.1002/wcc.672.

Sperberg-McQueen & Burnard 2002 Sperberg-McQueen, C. M., and Burnard, L. (eds.) (2002) *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Oxford, Providence, Charlottesville, Bergen: Text Encoding Initiative Consortium. Available at: https://tei-c.org/Vault/P4/doc/html/index.html.

Terras 2021 Terras, M. (2021) “The Role of the Library when Computers can Read: Critically Adopting Handwritten Text Recognition (HTR) Technologies to Support Research” in Wheatley, A., and Hervieux, S. (eds.), *The Rise of AI: Implications and Applications of Artificial Intelligence in Academic Libraries* ACRL - Association of College & Research Libraries, pp. 137–148.

Thomas 2018 Thomas, R. G. (2018) “Review of The Shelley-Godwin Archive (S-GA), New York Public Library and the Maryland Institute for Technology in the Humanities, gen. ed. Neil Fraistat, Elizabeth Denlinger, and Raffaele Viglianti; Willi Godwin’s Diary, dir. Victoria Myers, David O’Shaughnessy, and Mark Philp”, *Eighteenth-Century Fiction* 30(4), pp. 601–603. Available at: https://doi.org/10.3138/ecf.30.4.601.

Walsh 2002 Walsh, N. (2002) “XML: One Input — Many Outputs: A Response to Hillesund”, *Journal of Digital Information*, 3(1). Available at: https://journals.tdl.org/jodi/index.php/jodi/article/view/jodi-64.