Digitising Cultural Heritage Material 32LDK1 B15V3 VT2023 Project work

Introduction

This project aimed to digitise my great-grandmother's diary, which covers the years 1913-1919. Shortly after beginning her diary, the first world war broke out, and she signed up with the Scottish Women's Hospitals For Foreign Service, where she served as a nurse in Serbia, moving with the war front. Unfortunately, the original handwritten document is lost, but a typewritten transcription survives. The loss of the original makes it even more important to digitise and preserve what remains, as a historical document giving us an on-the-ground view of the first world war, the lives of one medical unit, and the personal life and thoughts of an ordinary civilian.

It's approximately 71 pages of A4, five years of single-line entries. This document is entirely text-based, the physicality of the object is much less important than the words. Accordingly, I chose to spend most of my time on text encoding and less on image capture. This was further necessitated by the family member who has possession of the diary living overseas and, while fully consenting to the project, being unwilling to send the physical item but willing to take high-quality images. This plan was discussed and cleared during the residential week. There are some spelling errors in the text, particularly in people's names, which are likely to come from the transcriber. It seems more likely that it was the transcriber, because firstly the original would know how to spell the names of people close to her, and secondly the errors are phonetic, so a team of two people with one dictating would explain this. I believe these errors make it more interesting to use, as I have to ensure not to make similar mistakes when using OCR, and because I can record the errors and possible corrections, adding context. Similar records are collected by The Royal College of Surgeons of Edinburgh Library and Archive in their Scottish Women's Hospitals collection, by the Mitchell Library Glasgow and the Glasgow City Archives, by the Imperial War Museum in their Women's Work collection, by the Women's Library Archives at London School of Economics, and by the Wellcome Collection.

Key aims of this project are preservation of the diary for the future, making it accessible and usable to more people in more ways by publishing online and using linked data, and presenting it in new ways and new contexts through choosing appropriate views and tools on the website.

The chosen strategy emphasises depth (critical digitisation rather than a broad but shallow approach, taking time to mark as many features of the text as possible), text (emphasising the content of the document rather than the document itself as a physical object), and fidelity (reflecting the document closely rather than editorialising in the process of digitisation).

The Github page, including the deployed website, is here:
https://github.com/liopleurodon90/Digitisation2024

Documentation

I chose to put the emphasis on text encoding, especially with regard to names and places and to corrections and substitutions. To this end, I selected extracts which contained all of these, as well as notable historical events in order to make clear the diary's value as a historical document. This meant reading through the whole diary to shortlist pages which included each of these features, and planning out how to allocate time and any necessary resources, like software trial accounts, for each step of the project. I also contacted family members to get their consent for this project, discuss access to the diary, and learn as much as possible about the history of the writer, transcriber, and previous personal research done, as well as looking into cultural heritage organisations which had collected similar records and how they kept and presented the material.

Image capture was done by scanning the typewritten pages, in a Canon Pixma TS6100 series. I edited the Metadata in GIMP. The IPTC Photo Metadata Standard Subject Codes have become Media Topics (IPTC Media Topics, https://www.iptc.org/standards/media-topics/, 2024), of which the topic chosen was history (NewsCode Concepts: history, https://cv.iptc.org/newscodes/mediatopic/20000747). GIMP was also used to save the image files in separate appropriate presentation and preservation copies.

For OCR, I tried both FreeOCR and Abbyy Finereader. FreeOCR separated the text from year labels, mangled some words, and didn't handle the handwritten annotations which was the reason I had selected these extracts. Abbyy preserved the structure, recognised some annotations though not all, and did still mangle some words and insert erroneous punctuation. I chose to manually correct the Abbyy version, which took less time than transcribing from scratch would but still required care to catch all the errors.

Text encoding took the majority of the time and effort. The majority of the work was done in Visual Studio Code, with the addition of extensions specific to working with the TEI, XML, XSLT, and XPath. During the final 30 days of the project work, I switched to Oxygen, for improved ease of use with the XSL transformations. I chose to repeat responsibility details in the title statement, availability statement, and source description for thoroughness. I considered using handnotes for the corrections, but decided against it because both the typewriting transcription and the handwritten corrections were done by the same people, at or close to the same time, with no potential clarification on who did what. I used when-iso for dates for maximum clarity, and to that end chose to put the full date for each day rather than just the year, as it's clear what they refer to in context and that information should be preserved and made as clear as possible, especially considering the goal to reorder the text for ease of reading in chronological order. The entries for August 12th posed what initially looked like an error, in that there are two lines for the same year, but checking the next page showed that this was in fact intentional and has therefore been preserved. The first entry looks like the record of an event like a wedding, so was likely jotted down as a record rather than being intended as a full entry for the day. Editorial corrections in the electronic transcription were handled by using a choice element, with nested sic and corrs. For example, the misspelling of politician Mr Berrill as the phonetically similar Beryl. Corrections present in the source document, such as correcting "Flowerdown" to "Fareham", were instead handled by nesting del and add in a subst element (TEI By example, 4.2 Corrections, https://teibyexample.org/exist/tutorials/TBED06v00.htm#corrections, 2020). The precise rendition of the deletions and additions was recorded in rend attributes. However, there are some non-standard spellings which were intentional on the part of the author, such

as abbreviating zeppelin to zep., and this has been treated instead as regularisation, with orig and reg nested in a choice element. I considered marking events, but judged that they would mostly be useful if the event in question was known by a specific phrase, and the only example in these extracts would be the armistice, which was not mentioned on the actual day it happened (TEI-C, eventName, https://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-eventName.html, 2023). Accordingly, I considered it a low priority for this project, though a possible option for a larger scale version. For URIs of places, I considered The Getty Thesaurus of Geographic Names (https://www.getty.edu/research/tools/vocabularies/tgn/), the GeoBlacklight At Coordinates . Info (https://atcoordinates.info/2020/07/07/geospatial-metadata-with-dublin-core-and-geoblacklight-standards/), and GeoNames . org (https://www.geonames.org). I chose GeoNames to work with because it dealt with outdated place names and names shared between several places more smoothly than the others, which was a particularly important consideration given the age of the text and the significant updates in placenames since it was written. Updating place names to the modern equivalents would not be a faithful reflection of the text and risks causing the same problem for future work, since it's entirely possible for place names to change in the future as well as the past. The planning stages of researching the Scottish Women's Hospital and the life of my author proved valuable for disambiguation at this stage. To allow for errors, I added notes to places that required a judgement call to explain my reasoning. Due to time pressure, I chose not to complete adding xml-id references to the places mentioned in the text: the same work applied to people named acts as proof of concept for this task, and I judged that getting other parts of the project to a more functional state would be a better use of time.

Publication also involved many dead ends, and ultimately the goals planned for this project proved too technically challenging for my current level of expertise within the timeframe for the work. Despite my intentions for presentation to be an important feature of the final project, my attempts to modify the XSL templates were unsuccessful. Uncompleted attempts can be found in the files, though not in the deployed website. I used the templates provided, and judge that the bare-bones website styling is adequate for this purpose. The focus is, after all, on the text. I also began making regular use of Github once the whole project had reached a point of having deployable HTML pages, despite seeing the utility of version control with relevance for all stages of the project work.

Copyright required thought: the original writer began and finished the diary in the UK, so I started by checking their laws. Copyright Notice Number 2/2014 (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/379040/c-notice-201402.pdf) stated that copyright was heritable and lasted 70 years after the author's death. All three of the author's children are dead but had children of their own, which include my mother and the cousin, Martin Fricker, who currently has the physical copy. Since I haven't yet inherited the rights from my mother as she's still alive, I sought explicit permission from available members of that generation. The agreement was that I am free to use the diary for this course and for any purpose that improves accessibility to the diary and the ability to create links with other relevant cultural heritage resources, but must not claim or transfer ownership of the diary for myself or for University of Borås. An appropriate licence therefore was Attribution-NonCommercial-NoDerivs 4.0 International (https://creativecommons.org/licenses/by-nc-nd/4.0/). Interestingly since this is an academic project, in 2014 the UK amended the Copyright, Designs and Patents Act 1988 Implementing the Information Society Directive (2001/29/EC,

https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001L0029:EN:HTML). The updated Research, Education, Libraries and Archives regulation extends the copyright exception for students and libraries from just literary and artistic works to all forms of copyright works. Fair dealing still applies. For works that need to be preserved, cultural works can be digitised by libraries, archives or museums for users to view at dedicated terminals for private study or personal research. I am a student so this is helpful. However, the work of this digitisation project is done in Sweden, so we need to confirm how that applies. The World Intellectual Property Organisation (WIPO) says that copyright valid in one member country is protected in each member country of these copyright organisations (https://www.gov.uk/government/publications/protecting-your-uk-intellectual-property-abroad/protecting-your-copyright-abroad). Sweden and the UK are on both the WIPO list. This is all somewhat moot as 70 years have actually passed since Helen's death, making the diary out of copyright, but treating the question seriously bought greater trust and good faith from the family and is therefore valuable for access. My judgement was that though I did legally have the right to use the material in any way, it's no good being technically correct if that means making people angry and less willing to work with you, and that too is an investment in long-term preservation, as it means further work can be done on this material in the future, that work can benefit from other resources like memories, photographs, and previous amateur historical research on the people, organisations, locations, and events that feature in the diary, and more people invested in the project means more people willing to store backups of the data, reducing points of failure.

On that point, maintenance is another valuable feature of Github. I made sure that the setting "preserve this repository" was turned on. I did not have particularly strong concerns about privacy as related to the necessity of making the repository public or to entrusting it to Github and thus Microsoft, especially since the diary does not contain information on any living persons, and the only living persons named in the text encoding have given explicit permission to be mentioned in this project.

Division of labour

Phase	**Time (hours)**
Selection of document	28
Document analysis and preparation	36
Text and image capture	2
OCR	3
Text encoding	84
Metadata	4
Publication	55
Testing and validating	24

Evaluation and reports	24

Relevance to cultural heritage

The materials are of interest to a broad range of cultural heritage institutions. Similar records are collected by The Royal College of Surgeons of Edinburgh Library and Archive in their Scottish Women's Hospitals collection, where the collection is relevant to both their subject areas, of the history of medicine and surgery and of Edinburgh (https://library.rcsed.ac.uk/archives): photographs, correspondence, recollections, a diary ("Little grey partridge: First World War diary of Ishobel Ross who served with the Scottish Women's Hospitals Unit in Serbia. Introduced by Jess Dixon. 1988"), and a biography, "In the service of life: The story of Elsie Inglis and the Scottish Women's Hospitals. 1994". Records relating to the Scottish Women's Hospitals are also found in the Mitchell Library Glasgow and the Glasgow City Archives, such as minutes from meetings, recruitment resources, biographies, including "Between the lines : letters and diaries from Elsie Inglis's Russian unit / arranged and edited by Audrey Fawcett Cahill, 1999" (https://libcat.csglasgow.org/web/arena/search?p_p_id=searchResult_WAR_arenaportlet&p_p_lifecycle=1&p_p_state=normal&p_r_p_arena_urn%3Aarena_facet_queries=&p_r_p_arena_urn%3Aarena_search_query=Scottish+Women%27s+Hospitals&p_r_p_arena_urn%3Aarena_search_type=solr&p_r_p_arena_urn%3Aarena_sort_advice=field%3DRelevance%26direction%3DDescending), which is mostly relevant to their resources on local and family history. The Imperial War Museum features similar materials in their Women's Work collection, on the work women have done during wartime and their perspectives on events, (https://www.iwm.org.uk/collections/search?query=scottish%20women%27s%20hospitals), including 38 objects classed as "private papers" of which include 15 diaries, and a great deal of correspondence. In addition, the Women's Library Archives at London School of Economics, devoted to the history of women's campaigning and activism, features 98 records related to the Scottish Women's Hospitals, (https://www.lse.ac.uk/library/collection-highlights/women-and-work), including personal papers and diaries such as "Papers of Elsie Edith Bowerman: The archive consists of diaries, photographs of work with Scottish Women's Hospitals, Bowerman's passport with portrait photograph, and personal correspondence (1909-1948), mainly with her mother during (1910-1911) and during her time with the Scottish Women's Hospitals unit in Romania and Russia (1916-1917) during the First World War." (https://archives.lse.ac.uk/CalmView/Record.aspx?src=CalmView.Catalog&id=7ELB&pos=52). The Wellcome Collection, focusing on medical history, has some records on the organisation, (https://wellcomecollection.org/search/works?query=scottish+womens+hospitals) including "Typed copies of Lilias Mary Grant's diary and Moir's letters, friends from Inverness, while they were nursing orderlies with Dr Elsie Inglis's Scottish Women's Hospital Serbian-Russian Unit in Romania and Russia, 1916-1917 "(https://wellcomecollection.org/works/n9fdzmac). Many of these resources are pointed to by the National Archives UK (https://discovery.nationalarchives.gov.uk/results/r?_aq=scottish%20womens%20hospitals&_dss=range&_ro=any&_st=adv_) for a quick overview.

Critical analysis/argumentation

The choices made and methods chosen in this work were guided by the strategy selected early in the process, which was itself guided by the aims identified as priorities for this project and constrained by the methods I had access to.

This did mean departing from best practices in some ways, most notably by not having control over the image capture portion of the work and not having reliable access to the physical object. It is this very same restricted access to the diary which motivated the decision to digitise it. If the document had featured more handwriting, images, or other features apart from printed text, this would have involved giving up an even more significant amount of control or the ability to make choices at an early stage that influenced the information available in the item at all later points in the digitisation chain (DGF Practical Guidelines 2013). In this matter, I was reassured by the Cornell guidelines's principle to "to match the conversion process to the informational content of the original, and to scan at that level--no more, no less", as well as Deegan and Tanner's discussion of pressures faced by scholars in digital humanities in real life projects, particularly that ideal technical resources may not be available and the need to consider what exactly a given digitisation project anticipates in terms of users and their needs (http://preservationtutorial.library.cornell.edu, 2003, Deegan and Tanner, 2001). This diary is an example of a case where capturing the content of a source is more important than capturing the form, therefore the methods chosen for the work involve capturing images only to a standard necessary to allow working with the textual content.

Although the intent for this work was to reflect the source document as closely as possible, Renear suggests that a level of editorial decision is deeply interwoven with working with TEI-XML, as it offers ways for "the encoder-scholar… to communicate a particular theory of what the text is." (Renear, 2004) The cultural heritage institutions mentioned above have privileged the selection slightly different materials over others and assigned, for example, keywords in slightly different ways depending on what they consider the relevant topics of the materials to their thematic collections: historical figures of local importance, military history, medical history, or the history of women's suffrage movements (Palmer, 2001). My work, similarly, makes judgements on what information is needed and what uses I expect to be made of it, prioritising mention of people partially because that's important to the stakeholders who hold the source document, and deprioritising exact line breaks because of this document's nature as a reproduction of another, missing original text. The choice to focus on text encoding is strongly connected to the aim that this project be used to preserve the text and increase accessibility so that being in the same location as the source document would no longer be needed in order to read the text, and allow ease of searching within what is a fairly long text and out of order by its very nature. Through the use of shared namespaces and encoding people and places with ids and URIs, this project makes use of hypertext, about which with the benefit of time I favour Palmer's stance on its value for adding context and thus information to resources, rather than Willet's scepticism of this comparatively new use of technology -to form a collection is already to allow objects to reflect on one another by putting them in a shared context, digital collections just increase the variety of options for doing so (Palmer, 2001, and Willett, 2001).

Conclusions

Scalability was of paramount importance in my decisions with this project. I am confident that these pages serve as proof of concept, and that having made careful decisions about how to represent and present the text, it would be achievable to apply those to the entire diary, and even to other diaries or writings which are similar.

I learned that text encoding took much longer than anticipated, but that most of the time was in deciding how to handle particular features: having made the decisions, applying them to the next extract went much faster. This is encouraging in light of the desire to scale this work to cover the whole diary. Getting to grips with XSLT also took much longer than anticipated, leading to less work done on that aspect than had been aimed for. I should also have allotted more time to documentation and reporting.

As far as future work goes, I would have liked to offer a view with the days sorted chronologically, to be read in order, but unfortunately I ran out of time to fully grapple with implementing the XSL options. I would also want to add a page of indexes, in order to link to each of the appearances of each person and notes on anything else we know about them. The useful nature of the tei is that the same xml-id label can be used to point to an entity even if they're referred to by several different names, and over the course of the diary this does happen (nicknames, misspellings, change of professional title, change of name in marriage). This would be especially interesting for public figures, which motivated my choice of extracts to include some. With public figures, and with place names, this presents an opportunity to incorporate linked data, increasing the information available and the uses that can be made of it. This also helps disambiguate between similar names. Where this occurs, I have included a note in the listPerson and listPlace entries explaining the decision. In a larger project, the places and people currently in a standOff would be moved to separate Personography and Geography xml files for practicality. Future work could also expand to marking events. One potential advantage of both publishing online and using ids and URIs is the ability to link this work to the materials held by cultural heritage institutions that also deal with the Scottish Womens Hospital and associated organisations, as there are individuals and events mentioned in the diary which come up in those other materials and may be of interest to both professional historians and personal descendents. Also in XSL, the corrections, deletions, and additions are currently all rendered as strikethrough and superscript, but ideally that would be improved to more accurately reflect the varying placement of these corrections.

Bibliography

Björk, L. (2015). *How reproductive is a reproduction? Digital transmission of text-based documents*. Borås: University of Borås.

Conway, P. (2015). Digital transformations and the archival nature of surrogates. Archival Science, 15, 51-69.

Cornell University Library (2000-2003). Moving theory into practice: digital imaging tutorial.

Dahlström, Mats (2011). Editing Libraries. C. Fritze, F. Fischer, P. Sahle & M. Rehbein (Hrsgg.), *Bibliothek und Wissenschaft. Vol. 44: Digitale Edition und Forschungsbibliothek*. Harrassowitz.

Dappert, A., R. S. Guenther & S. Peyrard (Eds.) (2016). Digital preservation metadata for practitioners: Implementing PREMIS. Cham: Springer.

Deutsche Forschungsgemeinschaft (2013). Practical Guidelines on Digitisation. https://www.dfg.de/foerderung/programme/infrastruktur/lis/lis_foerderangebote/digitalisierung_erschliessung/formulare_merkblaetter/index.jsp

Hirtle, Peter B., Emily Hudson, & Andrew T. Kenyon (2009). *Copyright and Cultural Institutions, Guidelines for Digitization for U.S. Libraries, Archives, and Museums* (2009).

Electronic Texts in the Humanities: Theory and Practice, by Susan Hockey (Hockey 2001)

Minerva (2008). *Intellectual Property Guidelines*. Version 1.0.

Ruthven, I. & Chowdury G. G. (eds.) (2015). *Cultural heritage information: access and management*. London: Facet.

Sahle, Patrick. 2016. “What is a Scholarly Digital Edition?” in Digital Scholarly Editing: Theories and Practices, ed. Matthew James Driscoll and Elena Pierazzo, 19-40. Cambridge: Open Book Publishers.

Schreibman, S., Siemens, R. & Unsworth, J., eds. (2004). *A Companion to digital humanities*. Oxford: Blackwell.

Sutherland, Kathryn & Deegan, Marilyn (2009). *Transferred illusions: Digital technology and the forms of print*. London: Ashgate.

Tanner, Simon (2004). *Deciding whether Optical Character Recognition is feasible*. London: King's College.

TEI P5: Guidelines for Electronic Text Encoding and Interchange (2014). Oxford: The TEI Consortium, Technical Council.

Terras, Melissa (2008). *Digital Images for the Information Professional*. London: Ashgate.

van Branden, Ron, Melissa Terras & Edward Vanhoutte. *TEI by Example* Links to an external site..

Vanhoutte, Edward (2004). An introduction to the TEI and the TEI consortium. *Literary & Linguistic Computing*, 19(1): 9-16.