CHAPTER 18

WIKISOURCE AS A TOOL FOR OCR TRANSCRIPTION CORRECTION: THE NATIONAL LIBRARY OF SCOTLAND'S RESPONSE TO COVID-19

Gavin Willshaw[1]

[1] The University of Edinburgh

Abstract

This chapter focuses on the National Library of Scotland's Wikisource transcription correction project, an organization-wide effort during lockdown that generated 1,000 fully accurate transcriptions of 3,000 Scottish chapbooks, which the Library had uploaded to Wikisource, Wikimedia's online library of digitized, out of copyright works. The project, which contributed to the Library being awarded Partnership of the Year 2020 at the Wikimedia UK AGM, is thought to be the largest ever staff engagement with Wikimedia, and has had significant benefits to the Library and staff well beyond the original aims of the project. Initially set up to improve the quality of optical character recognition (OCR) transcriptions in order to make the chapbooks more discoverable and searchable, the project gave staff a purpose and sense of belonging during lockdown, provided an opportunity to work with a varied and fascinating collection, and enabled them to develop new skills in editing Wikisource, drafting guidance documentation, and managing projects. Further to this, the initiative greatly increased library staff engagement

DOI: https://doi.org/10.3998/mpub.11778416.ch18.en

with Wikimedia, led to the formation of a Wikimedia Community of Interest, and resulted in the embedding of Wikimedia activity in staff work.

Keywords

Wikisource, Crowdsourcing, Scottish chapbooks, National Library of Scotland, Staff engagement, Digital skills.

Introduction

Like many cultural heritage organizations, the National Library of Scotland faces a significant challenge when digitizing texts: how to efficiently generate accurate transcriptions that meet users' needs, not just for search and retrieval but also for computational analysis using text and data mining (Europeana Pro, 2019). The Library runs typed and printed text through OCR software to generate transcriptions automatically and makes these available online alongside digital images on its Digital Gallery (National Library of Scotland, 2020a). Unfortunately, these often contain spelling mistakes and errors as the software struggles to deal with issues such as faint text, hyphenation, and archaic letters including the long-s (ſ) (Alex, 2012). Such issues require human intervention to correct but the Library lacks the staff resource to undertake this work. One area that the Library has been interested in exploring is whether corrections could be crowdsourced using Wikisource, Wikimedia's online library of out of copyright, digitized books. When a book is added to Wikisource, a community of thousands of editors work together to improve transcriptions using the platform's in-built error correction module and then publish the book on Wikisource ("Wikisource," 2020). Recent developments in functionality mean that the completed books can be exported not only in PDF or ePUB format but also as TXT files ("Wikisource:WSexport," 2020). The Library wanted to explore whether out of copyright, digitized books from its collections could be uploaded to Wikisource, where transcriptions would be improved in collaboration with the Wikisource community and then reimported back into the Library's image repository to improve the quality of search on the Digital Gallery.

An opportunity to explore this workflow suddenly arose in March 2020, when the Library closed its doors as the United Kingdom entered lockdown in response to the COVID-19 crisis. The Library's ten-person digitization team, whose work almost exclusively required them to be on-site using cameras and scanners to digitize books and other items from the Library's collections, needed work that would have impact and advance projects, would be large enough to keep them occupied throughout lockdown, and would require minimal access to the Library's network or physical building space. This unique situation allowed the Library to test Wikisource at scale; within twenty-hour hours of lockdown the entire team was editing transcriptions for the Library's recently digitized Scottish chapbook collection (Hagan, 2016) on the platform.

Chapbooks are small printed booklets that were sold for a penny or less on the streets, at fairs, and at markets. Typified by the use of woodblock illustrations and covering a range of subjects such as murders, disasters, love stories, and biographies of famous people, they were staple reading material for much of the population in a time before modern communication systems were invented (Hagan, 2019). This depth of content makes them a particularly useful primary source for social historians of the period, while their engaging content matter and size—3,000 separate books ranging from eight to twenty-four pages in length—meant it was an ideal collection for exploring Wikisource as a tool for OCR correction. Within a few days of lockdown, it became apparent that there were several other members of Library staff who also had time to work on the Wikisource project. Like the digitization team, staff in roles that were public facing or required access to the Library building were also limited in what they could do while working from home. By the end of March there were over fifty members of staff taking part in the work; this number increased to seventy, which was over 20 percent of the entire Library staff (National Library of Scotland, 2020b), at the project's peak.

It was exciting to have this unexpected, possibly once in a lifetime, opportunity to focus staff resources on a Wikimedia project, but rolling it out to several dozen staff in just a few weeks created many challenges. Arguably the greatest of these was that the Library, and the

country as a whole, was going through a huge change and staff had to adjust mentally and physically to their new reality. Most were using their own devices rather than Library laptops or PCs, so many were reliant on machines that were slow, out of date, or were shared with other members of their households. Furthermore, the level of Wikimedia knowledge and understanding across the Library was quite low. The layout of Wikimedia sites, the concept of Talk pages and etiquette of communication, and the use of basic HTML tags when editing source code were new to a lot of people, especially nondigital natives and those who didn't use digital technologies in their work.

In order to overcome these issues, the project team developed clear step-by-step instructions and guidance that covered all aspects of the workflow including how to set up a Wikimedia account and how to communicate with other users ("Wikisource:WikiProject NLS," 2020). A project support group was set up on Microsoft Teams where staff could discuss the issues they faced; additionally, all documentation was stored on Office 365, so staff could read and access documents without requiring VPN access to the Library's internal network. Rather than being assigned specific books to work on, staff were pointed to a Microsoft Excel spreadsheet that listed the books and the different workflow stages they were at; this allowed staff to select suitable items and to work on the project at a pace and time that suited them.

As mentioned above, the project team had very little time to plan this work in advance; one area that was initially overlooked in the rush to get the project started was codifying standards for the work being done. Wikisource has a Manual of Style ("Wikisource:Style guide," 2020), which outlines guidelines and recommendations to ensure consistency and standardization on the platform. These recommendations mainly focus on trying to achieve an accurate representation of the original item in the transcription. For example using the {{center}} tag to bring a heading into the center of the page, or the {{text-indent}} tag for indented paragraphs, while also ensuring that the spelling in the transcription matches that of the original image. While this works well for straightforward texts, for more complex items it can be extremely time-consuming to ensure that the transcription accurately matches

the original item. For example, an incorrectly spelled word in a book should use the {{SIC}} tag to show both the incorrect spelling used in the original for historical accuracy, as well as the presumed correct spelling to aid with keyword search. It became apparent early in the project that to fully meet all of Wikisource's guidelines, the proofreader would have to spend a lot of time on each book. Bearing in mind that the Library's chief motivation for engaging with Wikisource was to generate and extract high-quality transcriptions, working in complete compliance with the existing standards would slow the process so much as to make it infeasible. Instead, the project team worked with key members of the Wikisource community to develop new standards that would allow a better balance between transcription quality and throughput. The agreed approach involved a focus on correct spelling and layout, while using some of the more common tags to ensure transcriptions aligned closely to the original text.

Discussion around standards helped the team develop the project workflow, splitting the work into five discrete tasks, which are outlined below ("Wikiproject NLS Workflow," 2020).

1. Upload multipage PDFs of digitized chapbooks and their associated metadata to Wikimedia Commons using the Pattypan bulk upload tool.
2. Create Index pages on Wikisource, link these to the files on Wikimedia Commons, and add another link and information to the project Excel spreadsheet.
3. Generate initial automated transcription using Wikisource's Google OCR engine then proofread to correct errors.
4. Validate the proofread transcription, publish to Wikisource (transclude), and link to author pages and Wikidata.
5. Export transcriptions as multipage PDFs, convert to single-page TXT files, remove header and footer information and tags, and reupload into the Library's Digital Gallery.

In developing the steps outlined above, the project showed it is possible to set up an end-to-end workflow to successfully improve

transcription errors using Wikisource and reupload those transcriptions back into the Library's repository, which improves the search function of its digital collection. The project can be deemed a success: it has been estimated that at its peak there were more National Library of Scotland staff working on the platform than all other Wikisource editors combined; the project is also thought, anecdotally, to have been the largest ever staff contribution to a Wikimedia project. As of November 27, 2020, 1,064 of the 3,000 Scottish chapbooks in the collection had been fully transcluded, with an additional 535 books fully proofread ("Wikiproject NLS Progress").

However, despite all the effort in overcoming the challenges outlined above, the actual number of books fully transcribed has been quite low, and far lower than had been anticipated when the project started in March 2020. By the time the Library reopened in late July 2020, approximately 16,000 Scottish chapbook pages had been fully transcribed. Considering that approximately seventy staff had contributed to the project over a twenty-week period, the actual number of pages transcribed per person was only around ten per week. Based on this progress, it seems fair to conclude that if an organization's sole reason for engaging with Wikisource is to improve the quality of transcriptions from their digitized books, rather than using Wikisource they would probably be better building their own OCR correction module or buying one of the various commercial transcription packages or services that exist. The different stages of the Wikisource workflow take a lot of time: each Scottish chapbook was worked on by at least five different people as it progressed from upload to transcription export to the Library's gallery. Added to this, there are several manual elements to the process that are time-consuming, such as generating indexes on Wikisource by copying each individual URL from Wikimedia Commons, and frustrating, such as changing the OCR software from the default Tesseract engine to the far superior Google engine for every page. What is more, by adding books to Wikisource, there is an associated responsibility to manage the book once it is on the platform, to interact with and be guided by the existing community and to adhere as far as possible to their standards.

For the National Library of Scotland, however, the real benefit has been in areas far beyond correcting the quality of transcriptions, areas that had not even been considered at the start of the project.

The first of these was that Library staff experienced a huge amount of satisfaction and enjoyment from taking part in the project, with colleagues regularly tweeting about the interesting books they were working on and expressing their pride at being involved in the project in internal Library and Union lockdown surveys. For many, taking part in the project provided them with something tangible to work on, something to achieve and contribute to when their work environment had changed so dramatically. It gave people who didn't work with the collections on a day-to-day basis a much better sense of engagement with the Library and the materials it holds for the nation, and also helped them feel more connected with colleagues they hadn't seen since before the national lockdown.

Added to this, the Wikisource project has also given people opportunities to develop their digital literacy and skills by introducing them to Wikimedia projects and teaching them how to navigate the sites, contribute to open-knowledge initiatives, and learn how to use basic HTML and mark up for the first time. For the digitization team, who ran the project, the work had a secondary effect by giving them opportunities and responsibilities in a completely new area. Staff whose normal role was to digitize books—an important but often repetitive task—were given responsibility for different parts of the new workflow. One person, for example, became responsible for developing the internal guidance and training new staff as they joined the project, while another was responsible for liaison work with the Wikisource community on standards and workflow. A third member of the team took on the responsibility for file upload, which allowed them to learn to use the Pattypan tool to bulk upload files to Wikimedia Commons. These new responsibilities gave staff in the digitization team more confidence in their ability and raised their awareness of the wider digitization workflow and the impact digitized collections can have.

Furthermore, the project has helped to increase staff awareness of and engagement with Wikimedia projects more widely. All staff who

have taken part now have Wikimedia accounts, have been trained in the basics of editing Wikimedia sites, and have received a bespoke Wikisource overview from Sara Thomas, Wikimedia UK Scotland Programme Coordinator, and Ewan McAndrew, Wikimedian-in-Residence at the University of Edinburgh. Several Library staff also attended the Wikipedia and archives webinar by Kelly Foster in April 2020 and the Wikidata and cultural heritage collections session run by the Science Museum in June 2020. All of this means that the Library now has a better educated staff about the value of engagement with Wikimedia projects and there is now a strong base to develop future Wikimedia-related work. Members of staff who are now back to work at the Library are still working on this project during quiet periods, meaning Wikimedia work is embedded in staff roles for the first time. Following the success of this project, a Wikimedia Community of Interest was set up at the Library, which has already had a significant internal impact. There has been more staff engagement with the 1Lib1Ref campaign and Wiki Loves Monuments, for which the Library uploaded over 100 images in the 2020 campaign (Wikimedia Foundation, 2020), an intern was employed to write articles about the Library and its collections, a member of staff attended the Wikidata Summer Institute, and plans are in place to run a Bannatyne Manuscript Wikipedia edit-a-thon at the end of the year in collaboration with the University of Saskatchewan. This project has been a catalyst for far greater Wikimedia activity at the Library and collaboration with the Wikimedia community over the coming years.

And finally, adding digitized books to Wikisource appears to have improved access to and use of the Library's digitized collections. Although no large-scale analysis has been undertaken, a sample of five randomly chosen Scottish chapbooks on Wikisource was viewed an average of five times per book, which, if expanded to cover the 1,000 chapbooks already published on Wikisource, would suggest around 5,000 page views per month for the entire collection, a figure that adds to views on the Library's Digital Gallery site, and may well be higher because of Wikisource's superior search engine optimization (SEO). Indeed, another quick, internal test of five different random books

showed that the Wikisource version of a book appeared higher than the same item on the Library's Digital Gallery, with the Library's record sometimes not even appearing on the first page of Google results.

The National Library of Scotland's experiments with Wikisource have shown that the platform is not necessarily the best medium to use if an organization is solely interested in improving the quality of its transcriptions through OCR, as it is a slow process, requires significant manual input, dialog and engagement with the Wikimedia community and agreement on standards. For the National Library of Scotland, however, the benefits of running this project have far outweighed these issues, and the improved transcriptions the Library has received have been more of a side benefit rather than the main reason for engaging with Wikisource. The project has brought a lot of positives to the Library, including raising the awareness off Wikimedia platforms among staff, kick-starting an internal Community of Interest, and building relations with the wider Wikimedia community, all of which should make activity in this area a more sustainable element of the Library's work in the future. This initiative has developed staff skills and empowered them to grow in new areas during a difficult and traumatic time in their working lives, it has helped to bring an important digitized collection to a wider public audience, and it has created a workflow that will allow work to ramp up again in the event of future crises and lockdowns. Aside from anything else, the Library's Wikisource transcription project has given a glimpse of what can be achieved when considerable staff resources are committed to an open-knowledge project.

References

Alex, B., Grover, C., Klein, E., & Tobin, R. (2012). Digitised historical text: Does it have to be mediOCRe? In *Proceedings of KONVENS 2012* (pp. 401–9). www.oegai.at/konvens2012/proceedings/59_alex12w/.

Europeana Pro. (2019, July 31). Issue 13: OCR. https://pro.europeana.eu/page/issue-13-ocr.

Hagan, A. (2016, March 4). Scottish chapbooks now online! *National Library of Scotland Blog*. https://blog.nls.uk/scottish-chapbooks-now-online/.

Hagan, A. (2019, August 27). Chapbooks: The poor person's reading material. *Europeana Blog*. https://blog.europeana.eu/2019/08/chapbooks-the-poor-persons-reading-material/.

National Library of Scotland. (2020a, September 24). Digital Gallery. https://digital.nls.uk/gallery/.

National Library of Scotland. (2020b, September 24). Meet our staff. www.nls.uk/about-us/working-at-the-library/meet-our-staff.

Wikimedia Foundation. (2020, September 24). Wiki Loves Monuments 2020. https://outreachdashboard.wmflabs.org/courses/National_Library_of_Scotland/Wiki_Loves_Monuments_2020.

WikiProject NLS. (2020, September 18). In Wikisource. https://en.wikisource.org/wiki/Wikisource:WikiProject_NLS.

WikiProject NLS Progress. (2020, September 18). In Wikisource. https://en.wikisource.org/wiki/Wikisource:WikiProject_NLS#Progress.

WikiProject NLS Workflow. (2020, September 18). In Wikisource. https://en.wikisource.org/wiki/Wikisource:WikiProject_NLS#Workflow.

Wikisource. (2020, September 15). In Wikipedia. https://en.wikipedia.org/wiki/Wikisource.

Wikisource:Style guide. (2020, August 4). In Wikisource. https://en.wikisource.org/wiki/Wikisource:Style_guide.

Wikisource:WSexport. (2020, July 16). In Wikisource. https://wikisource.org/wiki/Wikisource:WSexport.