DCHM Project Report

Introduction

The purpose of this assignment was to deliver a digitisation project – choose material to digitise and make it available through a website to users. The target users were anyone interested in Victorian children's books. This project is a requirement for the completion of the course entitled Digitising Cultural Heritage Material (DCHM), which is part of the Master's programme: Library and Information Science, Digital Library and Information Services at the University of Borås.

To carry out this task several steps needed to be completed, including: scanning the document, saving a master copy, OCRing it, transcribing the text, creating a TEI document, and modifying the HTML, CSS and XSL files provided by Wout Dillen on GitHub, to suit this particular project. The final step would be publishing this work on the internet through GitHub Pages.

This task was meant to help me learn several skills linked to digitisation and gain experience in the processes involved. As aforementioned, it was also supposed to lead to the delivery of a digitisation project, completed and documented. This report constitutes the documentation part of this task. It describes the steps followed, and challenges and issues addressed and overcome during the project work.

The goal of digitisation is to recreate the original material as accurately as possible in accordance with scientific guidelines (Altenhoner et al., 2023, p. 7). In this project, this has been attempted, yet respecting the limited resources available and the lack of experience of its creator regarding digitisation projects.

Choosing the material to be digitised

While the technical components of digitisation may be planned rather well, the mental work necessary to choose the appropriate objects is difficult to estimate (DFG, as cited in Björk, 2015, p. 146). This is true for this project in that the exact effort and time taken to make this decision is impossible to be established.

However, it is possible to say that the choice of the document to be used has taken weeks to be made. It has involved careful thinking and lots of criteria had to be considered. The first criterion was looking for a book that was out of copyright, so I did not need to ask for permission to use it. The second criterion was looking for something with historical and social relevance, interesting and attractive. The third criterion was originality, that is, choosing a book that had not been transcribed before. The fourth criterion was affordability, that is, as the material would have to be purchased, it needed to be affordable.

During initial searches for subjects, the above criteria led to Victorian children's books appearing prominently in search results. The shortlist consisted of about six books – some of them required MEI encoding – skills beyond the ones required for this assignment. Some were in very poor condition and, therefore, were not suitable candidates for scanning. Finally, the book chosen was *A House to Let* by Mrs Molesworth.

The chosen book's year of publication was only estimated as being around 1889, but due to the date of death of its author, research showed that it was out of copyright. In addition, reading the book in question after purchasing it, I confirmed that it was a good candidate for this project.

The book in context and its historical and social relevance

Mrs Molesworth was born Mary Louisa Stewart, in 1839 and spent her childhood in Manchester, United Kingdom (Green, R., 1964, p. 12-13). She is a well-known author, who wrote many children's books, including the one digitized for this project, entitled *A House to Let*, published around 1889. This book is particularly well written, and it really captures Victorian values in an engaging story. W. J. Morgan's beautiful illustrations are reproduced in colour, and considerably enhance the book.

Regarding the item's historical and social relevance, its narrative is a very clear portrayal of contemporary Victorian society – it clearly shows Victorian values, such as the importance of hard work and the idea that the poor should help themselves (Walvin, J., 1988, pp. 99-103), which is reflected in the beliefs and behaviour of the protagonist's mother – she works very hard throughout the story and takes pride in a job well done. Similarly, her mother does everything to the highest standard and imbues her daughter with the idea of hard work, to achieve what she wants to achieve.

Along this narrative, there is also the story and illustration of the beliefs and behaviour of the people they work for, and the reader also sees the importance of charity in Victorian times (Walvin, J., 1988, pp. 96-103), which is highlighted when the better off characters provide the protagonist and her mother with a place to live when they are about to be made homeless. However, it is emphasised that the protagonist's mother is deserving of help because she is honest, humble, and hardworking, thus referring back to the Victorian belief that the poor should help themselves, as aforementioned.

According to Walvin, J. (1988), British civility was measured in the compassionate traits that gradually came to define British social life (p. 98). In Mrs Molesworth's book, civility is shown in the courteous and polite behaviour of the protagonist and her mother and in the sympathetic actions of the family who prevent them from becoming homeless.

There are other children's books that portray Victorian life and values. However, in the research for this project, no other book has been so eloquent in doing so through both text and illustrations. In addition, Mrs Molesworth was not only an author whose texts portrayed Victorian values. She was born and lived during the Victorian era. Therefore, her knowledge of that time was more considerable than that of a writer from another era.

Equally important, many of the values held during the Victorian era are still upheld in the present day, which reveals how historically and socially important that era was and, therefore, makes this book worth preserving.

Scanning

First, I tried using a flatbed scanner. However, I encountered two issues. Firstly, although the image was uniform in terms of lighting, I would have to press the book down to scan it, which would have caused damage to the book. However, I did not want to cause any further damage to the book, as it was already in bad condition. Another issue was that there was no option to save the scanned files in the TIFF format. I wanted to save my images as

TIFF, not convert them from jpg or another extension to TIFF, as this could cause loss of picture quality. So, I decided not to use the flatbed scanner.

I do not have access to a V-shaped overhead scanner, used in professional digitisation projects, which would have kept the book open, but not straight, and which would have distributed light evenly through the whole book. Therefore, the book was scanned using an overhead scanner (but not V-shaped), a commercially available scanner, called CZUR Scanner (ET18 Pro). Please, see appendixes 1a and 1b for photos of the scanner used.

The next step was to decide whether to scan one page at a time, or two pages at a time – that is, facing pages. I chose to scan two pages at a time, since the page flattening feature in my overhead scanner does not work with single pages. This feature is very useful for giving an even and clear image, especially near the stitches, where the pages join, which can be a little blurry in single page scans. In other words, by using this function, the software flattens the pages during the scan, which avoids distortion of the words in the text (See appendixes 3a and 3b).

In the overhead scanner, the light source is quite central, that is, it does not cover the whole document, illuminating some parts of the document more than others. As this was not made better by scanning one page at a time, I chose to scan two pages at a time. The light spots were reduced by adjusting the level of light directed at the pages. In sum, I obtained the best quality images possible with the equipment I had available.

According to DFG the resolution of the scanned images should be at least 300dpi (DFG, 2013, p. 8). The DFG guidelines also suggest that the colour depth for the master copy be 8 bits per channel – that is, 24 bits for RGB colour (DFG, 2013, p. 10). My master copies have followed this guidance when it came to the bit depth. Regarding the resolution, I chose to use 600dpi because the images were crispier with this resolution (See appendixes 2a and 2b).

My experience was that the scanner took about 2 seconds to scan two pages. What took longer was to centre the book, to make sure that the dotted lines were exactly where the stitching was and that the light level was the most appropriate. If the light level was too high, the letters started fading; if the light level was too low, the whole page looked too dark. The decision was to choose the light level that best represented the colours in the original book and that showed the text clearly. Afterwards, the image masters were saved in a folder with the extension TIFF, uncompressed.

This scanner also had the option to save the scanned image file in other formats, including searchable PDF, “Word (OCR)” and txt. So, I have also saved my scanned images as “Word (OCR)”, PDF, searchable PDF, and txt. This gave me more options of files to work on later in the project.

Subsequently, the pages were cropped when necessary. However, care was taken, so the pages were not over cropped. The intention was to have neat images, but that still showed what the original images looked like. Therefore, the digitized images show the age of the pages, with some damage to the stitch area, grime, foxing, etc.

Both JPEG and PNG are suggested for web publication (DFG, 2013, p. 15). However, due to a lossy compression method, which is used by JPEGs, some image data is permanently lost when the image is reduced in size. Consequently, long-term quality issues could result from this since each time a file is updated and saved, more data is lost. As I needed to reduce my images to thumbnail size, for my reading and top layer views, to be put in the HTML file and I did not want the image quality to be degraded, my TIFF file was subsequently converted into PNG format, creating a derivative file.

OCR Process

Since the book had nearly 100 pages, this amounted to a lot of text to transcribe. Hence OCR was applied to the scanned images, so the text could be copied to the TEI editor for preservation and used in the HTML files. I attempted to use ABBYY FineReader because I had used it in a previous assignment, and it was quite accurate for English text recognition. In addition, this software has a function that makes it easier for its user to correct errors in text recognition, post OCR. However, when I finished the correction process, the software would not allow me to save the corrected text. Therefore, I had to find an alternative solution.

I had already saved the book's scanned images in the "Word (OCR)" format, as that was a feature of the scanner I used in this project. So, I copied the text from the "Word (OCR)" file into my TEI editor and corrected any errors manually.

TEI XML Encoding

As Lou Burnard (2014) explains, there are three main reasons to use TEI XML encoding. The first is that TEI XML prioritises the meaning of text over its look. To use Burnard's example, if an individual uses a word processor, for instance, Word, to search the many instances of the term "London", this processor will not distinguish between "London" the place in Canada and the place in the UK. Neither will it distinguish between "London", the place name and an author's surname. (pp. 7-8). This is relevant to this project because the preservation of its content is more important than preserving its appearance. The second is that no specific software environment is required for TEI XML to function. To clarify, every piece of software that uses a TEI XML document sees the identical information captured in it. This is very useful if a person would like to share the written materials they produce with others or access them again in a few decades (Burnard, L., 2014, pp. 8-9)., and, therefore, an essential feature in this enterprise. Thirdly, TEI XML was created by and for the academic research community, who are also in charge of its continuous improvement. As its user community makes adjustments that they deem useful, TEI adapts by disregarding those that do not (Burnard, L., 2014, pp. 9-11).

The digital representation of the source's pages might suffice for the knowledgeable human reader to get any value from it. An encoded transcription, however, will be a necessary finishing touch for the image for a larger audience and most definitely for any type of automated analysis or search. (Burnard, L., 2014, p. 9). The TEI offers several tools to make it easier to produce digital editions of all kinds.

In TEI files, tagging is added to a document in order to categorise and arrange it for automated processing. For hundreds of tags, the TEI offers definitions, names, and guidelines for possible combinations. In addition to the actual text, the encoding describes the headings, text emphasis, etc. in the document. In other words, you have the text and encoding to describe not only what is in the text, but also its meaning. This is very important for preservation, as it will enable the reader to visualise the text even if its physical copy is no longer available.

In this project, I used TEI to encode the text. Then I used the tool Oxygen as the editor, as it suggests certain elements and attributes while the document is being typed. In addition, it shows when the encoding is invalid – indicated by the red square on the top right-hand corner of the screen, which turns green when the code is valid. This is very helpful. I watched the

lectures provided in the course to learn how to use TEI. Moreover, I studied the guidelines provided by the TEI Encoding Initiative, to learn how to use this encoding technique.

The TEI document for *A House to Let* comprises the following parts:

1-TEI header: this contains document-specific metadata, and it is signified by a <teiHeader> element.

2-<text> element. The text element is then divided into front and body, represented by its corresponding elements <front> and <body>. I have not used the <back> tag, as there are just blank pages after the body of the text.

The <front> element includes a dedication, the title page, the printing company, and the list of contents.

The <body> element includes the whole body of the text.

HTML Coding, CSS, and XSL files

I have used the template provided by Wout Dillen, to create my HTML, CSS, and XSL files. I have modified the template according to the needs of my project. I have worked with a printed book, and not a manuscript, so I did not need to use elements such as del and add. However, my book had illustrations and they had to be added to the diplomatic copy, which was quite challenging because the text would wrap around regular shapes, but not irregular shapes. I also wanted to indent my paragraphs, to make this view as close in look as possible to the physical object. This was straightforward. However, the paragraphs were still separated. I found a way of placing them together and indenting the text at the same time. However, that way would have to treat all the paragraphs on the same page as one and I would have to add page breaks and an indentation element to make the look of the page similar to the one in my book. I did not choose this method because I wanted to preserve the paragraphs in my HTML file.

Some CSS file settings were modified so that the style of the text would be closer to that in the physical book. Regarding the XSL files, I made just small changes to them, like replacing what was inside the tag <title> with the author and title of the book I have transcribed.

Kinds of transcription (Views to display the transcribed book)

For this project, three transcription kinds were performed, which can be accessed from the project's main page, namely diplomatic, reading, and top layer. The images used for my diplomatic view were in the following extension: PNG. The images used for my reading and top layer views were also in the following extension: PNG. These had to be turned into thumbnails. To do that, I used Gimp, to batch reduce their size.

In strictly diplomatic transcriptions, every detail that might be accurately printed is preserved. These elements include capitalization, word division, alternative letter forms, spelling, and punctuation. The page's layout—which includes big initials, line breaks, and other elements—is likewise preserved. The text will not contain any expansions for acronyms, and

even in the most formal transcriptions, typographical errors will not be fixed (Driscoll, M. J., 2007).

In this project, the diplomatic transcription has followed these rules as close as possible. However, in cases in which extensive research has been made, to find suitable HTML and CSS elements to create the diplomatic transcription, but with no success, I have not achieved a strictly diplomatic transcription. To illustrate, weeks have been spent in the search for a way to wrap the text around irregular shapes, that is, around irregularly shaped illustrations. However, every option encountered has not returned the result expected. Therefore, due to time constraints, I have had to abandon the idea of wrapping text around irregular shapes.

Regarding the reading view for my project, there was an increased emphasis on the text. For example, line breaks were eliminated to improve readability. As for the top layer view, I attempted to improve the text's readability further by removing the tags from the transcribed text.

Publishing-About GitHub, GitHub desktop and GitHub pages

In this project, I have used GitHub.com and GitHub desktop, to link the work I had done on my laptop to the GitHub desktop and, then to GitHub (online). I did most of the work before I was able (learnt how) to use GitHub.com and GitHub desktop to synchronise my work, so I did not benefit much from Git's version control capability. However, GitHub and GitHub desktop were useful in other instances, including to use Wout Dillen's template while creating my project's repository; to clone the same template so I could work on it on my laptop, to suit my project; and to commit changes through GitHub desktop and push them to GitHub.com. In addition, I was also able to make changes on GitHub.com first, for instance, when I updated the README file, and then commit changes, fetch the origin" and, pull the origin on the GitHub desktop, to update the files on my laptop as well.

Learning how to use GitHub.com and GitHub desktop for my project has been challenging, that is, I have watched both videos created by Wout Dillen more than three times (each) and watched other videos available online. Yet each time I tried using the GitHub service I encountered challenges, including issues with their system, which crashed a few times. For instance, initially, the GitHub desktop was not able to find my repository. So, after doing some research, I learnt that I had followed the procedure correctly and that other people had encountered the same problem. In my case, I solved the problem by deleting all repositories from my GitHub desktop, logging out of GitHub and then logging in again. After doing this, the GitHub desktop was able to find my project's repository. Another issue encountered was that, at some point, GitHub stopped updating the report changes I made in my local file. After doing some research online, I decided to remove the report file from GitHub.com, finish my report locally, save it as PDF instead of as Word, and try committing and pushing to the origin again with this new file extension.

In order to transform my project into an actual website, I used GitHub Pages.

Challenges

In addition to some issues already mentioned above, I have encountered other challenges during the development of my project, including regarding the following:

Choice of material – It was difficult to find a book that was out of copyright, had historical and social relevance, and had not been digitised before. Meeting all three criteria was challenging, therefore, taking me a few weeks to find the right item.

Choice of scanner – I had to do some test scans first, to discover what limitations I would encounter in practice. I did this for both my flat and overhead scanners. One issue was that there was not much information in the overhead scanner user guide about how to scan items. What is more, the overhead scanner was not very intuitive. For instance, it was not clear how one could choose the appropriate resolution and colour depth. Moreover, how to save the scan in different formats was not evident. Similarly, the terms used to refer to the different saving modes were not self-explanatory.

TEI encoding- Because of a previous assignment I had a basic idea about how to encode a text with a common layout, but encoding a whole book was much more difficult. Therefore, I needed to do a lot of research, and study the TEI guidelines in depth. I even attended a workshop run by the TEI Encoding Initiative at the University of Cork. Moreover, I joined the TEI encoding Facebook group and learnt from specialists in the field, that, even for them, some elements' purposes are still open to discussion. I also learnt that I needed to think about what I wanted to represent with the TEI encoding and what was relevant to encode for this specific project.

The Scope of the project and time constraints- The digitisation work has taken longer than anticipated and significantly longer than the 260 hours suggested in the project instructions. The size of my project is most likely one of the causes. Because of its page count and intricacy, the book that was selected took longer than anticipated to digitise. Despite not being handwritten and being in legible condition, there were various features to encode, such as chapter numbers, names and subheadings, recurrent page headings, page numbers and folio letters, and illustrations.

Estimated Record of days and hours worked

14th June 2023-studying the DFG digitisation guidelines.
15th June 2023-studying for the digitisation project.
From 15th June to 17th June 2023-scanning
21st June 2023-ABBYY FineReader-OCR 1
22nd June 2023-Trying to convert PDF into text with ABBYY FineReader
22nd June 2023-Oxygen TEI document started.
From 22nd June to 24th August-Working on the TEI document
30th June 2023-OCRd PDF saved as an editable copy in both rich text format and Word.
1st to 3rd July-Watching Wout Dillen's videos on using GitHub.com, GitHub Desktop and publishing a website on GitHub; doing the pizzaParty exercise as a way of practising.
31st July 2023-studying HTML again to remind myself of the document structure and rules.
1st August 2023-DCHM template-main-downloaded as HTPPS from Wout Dillen's page
1st August 2023-DIY-frankensTEIn-main downloaded

1st Aug 2023-Book illustrations saved in Word document to be placed within the transcribed text.
2nd Aug 2023-Book illustrations saved as JPEG.
3rd August 2023-Working on the HTML, CSS and TEI files.
31st August 2023-I started the digitisation project report.
From 31st August 2023 to 30th November 2023-Working on the digitisation project report and the digitisation files, especially on making improvements and solving issues encountered.
From 1st Dec to 14th Dec 2023-double checking if there were any mistakes and fixing them if possible; finishing the report; and attempting to add the work to GitHub.com.
From 14th Dec 2023 to 12th Jan 2024-trying to add my project folders and files to GitHub.com; watching Wout Dillen's videos with instructions on the topic again; solving issues encountered while working with GitHub; and updating my project report.

Estimated number of hours worked: 380 hours.

Conclusion

This digitisation project has taken more than the expected and required number of hours to complete. One of the reasons is probably the scope of my project. The book chosen required more hours than expected to be digitised due to its page number and its complexity. Although the book was easy to read, as it was not handwritten and it was in readable condition, there were a lot of different details to encode, including chapter numbers, names and subheadings, recurrent page headings, page numbers and folio letters, illustrations, just to name a few.

In addition, I worked on this project on my own, which meant I did all the work by myself. On one hand, it would have been helpful to have a partner if they had access to resources I did not have and if they shared the workload with me. On the other hand, doing the work by myself meant that I learnt about all aspects of a digitisation project, thus better preparing me if I ever work on such a project in a professional setting.

If this project had been created for a professional setting, I would have never started it without having available the required resources, such as a more appropriate scanner. However, as this project was a requirement for the Digitising Cultural Heritage Material (DCHM) course, I had to deliver the best results possible with the resources I had available. Looking ahead, if I decide to continue working on this project after I receive my grade, I will try finding a more appropriate scanner and scan the book again, subsequently, replacing the current scanned images with clearer ones.

References

Altenhoner, R. et al. (2023). DFG Practical Guidelines on Digitisation. Updated version 2022. *Zenodo*. https://zenodo.org/records/7561148

Björk, L. (2015). How reproductive is a reproduction? Digital transmission of text-based documents. *DiVA*. https://www.diva-portal.org/smash/record.jsf?dswid=-7983&pid=diva2%3A860844&c=1&searchType=SIMPLE&language=en&query=how+reproductive+is+a+reproduction&af=%5B%22publicationTypeCode%3AmonographDoctoralThes

is%22%5D&aq=%5B%5B%5D%5D&aq2=%5B%5B%5D%5D&aqe=%5B%5D&noOfRows=50&sortOrder=author_sort_asc&sortOrder2=title_sort_asc&onlyFullText=false&sf=all

Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*. Open Edition Press.

DFG. (2013). DGF Practical Guidelines on Digitisation. *Studylib*. https://studylib.net/doc/18624948/dfg-practical-guidelines-on-digitisation

Driscoll, M. J. (2007). Electronic Textual Editing: Levels of transcription. *Tei-c.org*. https://tei-c.org/Vault/ETE/Preview/driscoll.html

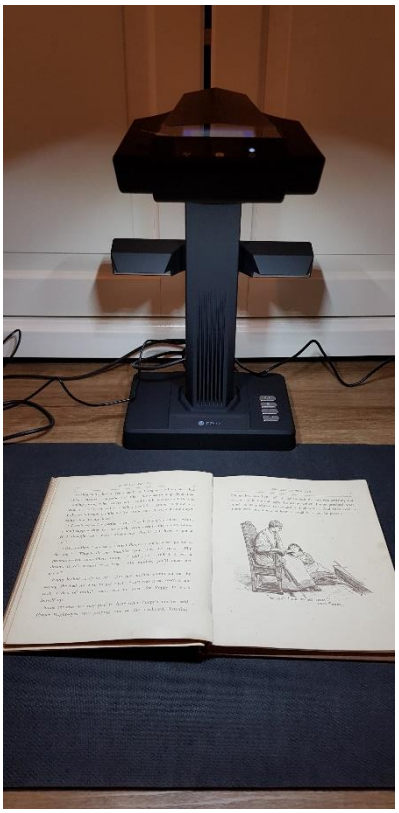
Molesworth, Mrs. (c. 1889). *A House to Let*. Richard Clay and Sons Ltd.

Green, R. (1964). *Mrs Molesworth: A Walck Monograph*. Henry Z. Walck, Incorporated.

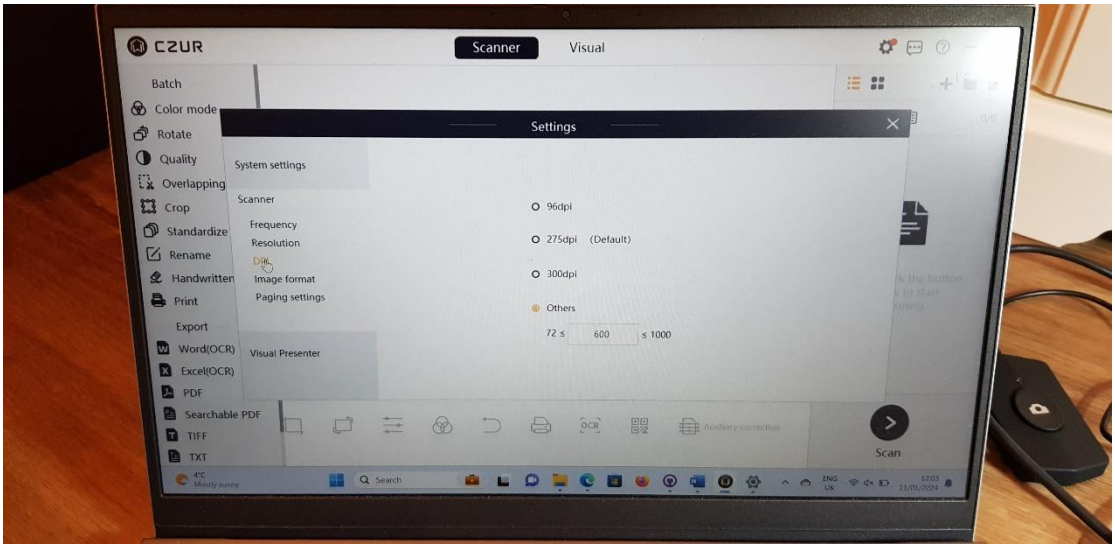
Walvin, J. (1988). *Victorian Values*. Cardinal.

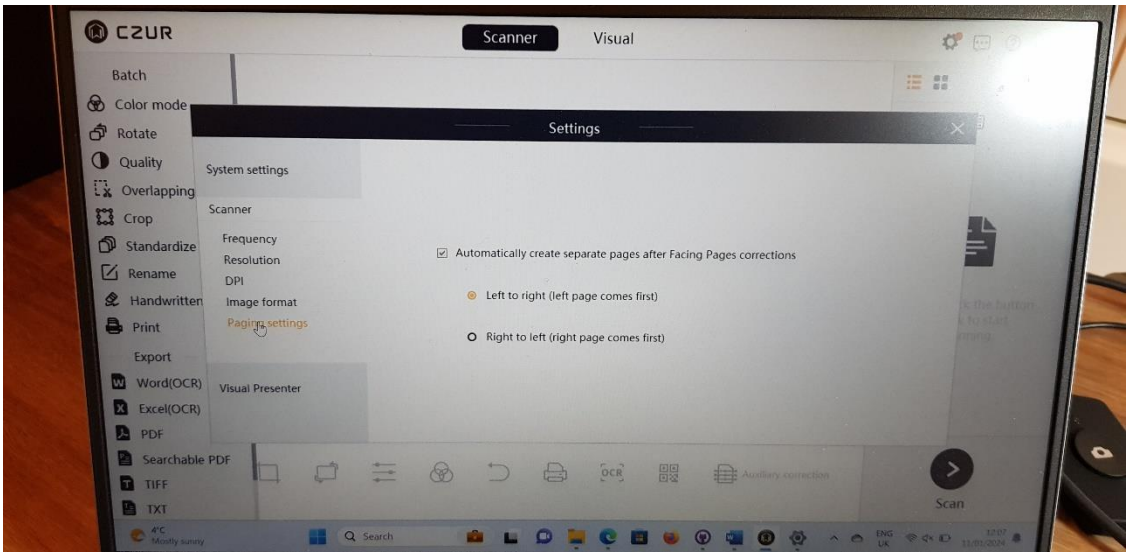
Appendix

Appendixes 1a and 1b (Scanner used)



Appendixes 2a and 2b (Scanner Settings used)





Appendixes 3a and 3b (Positioning of the book and flattening feature respectively)