DCHM Project Report

HÖGSKOLAN I BORÅS

Name: Emilie Lindström & Alexandra Sasha Spirkina

Course: Digitising Cultural Heritage Material 32LDK1 B15V3 VT2023

Date of Submission: 05/06/2023

Table of Contents

DCHM Project Report..1

Introduction..1

Relevance of the project to the area of cultural heritage digitisation...........................1

Digitisation strategy and aim...2

Documentation of the methodological work...3

Planning and division of labour..3

Selection and preparation of material..3

Scanning...4

Image Format and Storage...4

Optical Character Recognition...5

Text encoding..5

TEI..5

HTML..6

Publication and distribution...7

Evaluation..7

Conclusion...10

References...11

Appendix...12

Copyright...12

Scanner Information..13

Division of Labour...14

Timeline of the Project...14

Project Time Estimates...15

Introduction

This report describes the methodological work and conclusions drawn from a digitisation project performed by Sasha Spirkina and Emilie Lindström for a course in Digitising Cultural Heritage Material at Borås University in the spring of 2023. The project instructions stated the importance of the project to be relevant for the general public and not just to the student or his/hers family. For transparency, the material for this project has been chosen because of its connection to the relatives of Emilie. The book "Ammarnäsbygdens historia 1803-2008 : Nedtecknad av nybyggarnas ättlingar" was bought by her father in the local grocery store on his soul searching road trip to Ammarnäs, where his mother grew up. The copyright owner of the book happens to be his mother's sister. So you could say that the administration of this particular book runs in the family. However, there is a natural distance to the material since neither Emilie nor Sasha have been to Ammarnäs or met the copyright holder in person. With this background, we hope that we have managed to produce a project that combines both the motivated heart and the professional eye.

This report will document the work carried out during the project, including the planning and division of labor, the selection and preparation of the material, the implementation of the digitisation process, as well as an evaluation of the project's outcomes and recommendations for future work.

Relevance of the project to the area of cultural heritage digitisation

The selected book "Ammarnäsbygdens historia 1803-2008 : Nedtecknad av nybyggarnas ättlingar" is an outline of local history and accounts from a small village in the county of Västerbotten in the north of Sweden. The content was assembled by a local study group initiated by the local history society of Ammarnäs in 2005. Their work is based on previous material collected and documented by a similar study group in the 1970´s.

The content of the book itself is already an attempt to gather and preserve cultural resources for the future. To digitise it would be a continuation of this ambition.

The book itself is published by Ammarnäs local history society and Vuxenskolan (the study circle association). Considering the involvement of the locals in the making of the book, one can assume that there is a local or genealogical interest in the digitisation of the information that the book contains. It might be especially interesting for distant relatives who are looking for information about Ammarnäs and its inhabitants online.

Additionally, Ammarnäs has grown to be a popular tourist destination and could spark some public interest among visitors who want to know more about the area and its history.

By digitising the book, we can make the content accessible to future generations, tourists, and researchers interested in topics such as genealogy or local history.

Digitisation strategy and aim

Dahlström (2011) presents the concepts of critical digitisation and mass digitisation when it comes to digitisation modes and strategies. Critical digitisation features a manual process with critical selection, maximized editing and metadata, and a strong focus on the preservation of the document and its artifactual qualities. Mass digitisation is a quantitative approach defined by automatization, focusing on text rather than the source document, minimized selection, editing and metadata. Dahlström suggests that the two modes are on opposite ends of the scale along which any digitisation project places itself.

The digitisation strategy of this small scale project leans mostly towards the critical digitisation approach. We will have a selected part of the book carefully digitised, edited, and marked up for use and accessibility.

There are not many copies of "Ammarnäsbygdens historia 1803-2008" available for purchase or in public libraries, and it has not been digitised in any form before. This makes it a suitable target for digitisation, both in order to make the content more widely available and, in some sense, to preserve the book as proof of the efforts made to produce it.

However, because the physical copy of the book used in this project is not unique, the project does not aim to preserve or reproduce the copy itself. The priority is the textual content of the digitised version and not the preservation of the source media itself, i.e., the copy of the book and its physical qualities.

The aim of the project is essentially to make an accessible and user-friendly digital interface for a selected part of a book, and lay the foundation for it to potentially be digitised in its entirety in the future.

Documentation of the methodological work

Planning and division of labour

The project team consists of Emilie Lindström and Sasha Spirkina. There are some specific circumstances that have affected the planning and work processes.

One is that Emilie is the owner of the copy being digitised, so she has had the main responsibility for the scanning while Sasha took over the scanned material to perform the OCR. Being Swedish, Emilie has also been the one communicating with the copyright owner, Sylvia Berg.

Sasha attended the physical meetings in Borås and received initial training in TEI and GitHub, unlike Emilie, who could not attend. Sasha set up the GitHub page for the project and had been more inclined to take on any problems that arose with the platforms and programs used.

One important factor influencing the work process has been the distance between us, not just geographically but also in terms of time differences. Emilie and Sasha are living in Sweden and Ireland, respectively. At the end of the project timeline, Emilie was working from Mexico.

Communication has mostly been done through WhatsApp for smaller updates, pull requests in GitHub for changes in the TEI encoding, and Zoom calls for strategic planning. During one week in April, we managed to meet physically in Stockholm. During this time, we talked through the project in more detail, worked out the basics of the TEI encoding and how to divide the parts of the encoding, publishing, and writing of the report between us.

See Appendix for a detailed description of the division of labour. Note that we also made significant contributions to each other's assigned tasks, forming a very collaborative approach to our division of responsibilities. We were able to edit and incorporate changes into one another's written material using Google Docs and the encoding material using Github.

Selection and preparation of material

The book was published in 2008 and is protected by copyright so before embarking on the project, Emilie contacted the copyright owner, Sylvia Berg, who gave permission to perform the digitisation and digital publication of the material (see Appendix Copyright). The photographs used in the first chapter are not protected by copyright, and the front cover photograph has been approved by the photographer, who is the same Sylvia Berg as previously mentioned.

The material lends itself well to digitisation due to its qualitative condition and clarity of print. It also contains some black and white images. Its content can be considered to be adequately captured in digital form. In this case, the physical copy was not subject to preservation so we were able to cut out pages in order to enhance the quality of the digital version.

The book consists of about two hundred pages. For this project we have selected the front cover, title page, table of contents, preface, introduction and first chapter for digitisation which was a total of 14 pages. The first chapter has been TEI encoded according to best practices, while the front cover, title page, table of contents, preface and introduction are published for visual display only.

Scanning

Initially, Emilie made a rough scan to present the material to Sasha. From that sample, decisions could be made concerning how to enhance the quality for the final scan. By cutting out the pages from the book, the scan could be performed without shading or distortion of the text. Following the Practical Guidelines of Digitisation by Deutsche Forschungsgemeinschaft (2013), the scanned images of the content pages were captured in TIFF-format, with a resolution of 400 dpi and in grayscale to make sure the black and white photographs would come out as detailed as possible. The cover page was captured in color. The scanner used in this project is a floorstanding, flatbed Kyocera TASKalfa 5052ci, at a public library (Hornstull) in Stockholm.

Image Format and Storage

In line with the recommended practices, we made the decision to store the images on Github using both TIFF and JPEG formats. This approach aligns with the initial default choice of archiving in uncompressed TIFF, which ensures the preservation of image quality. Additionally, we acknowledged the potential benefits of JPEG as an alternative format for reducing storage space. However, considering the mentioned compatibility concerns with open source platforms, we opted for a combination of TIFF and JPEG to strike a balance between image fidelity and accessibility (Björk, 2015, p.154). The TIFF files were converted to JPEG using a free online TIFF to JPG converter (TIFF to JPG). The JPEG images were cropped using Adobe Photoshop Express to remove excess background.

Optical Character Recognition

For this project, OCR has been utilised to improve indexing, enable effective full text searching and retrieval, as well as facilitate potential text editing. The implementation of OCR technology has led to a significant reduction in the amount of time required for manual transcription, which is particularly advantageous when dealing with substantial amounts of text such as this one chosen for the project.

According to Tanner's article (2004), our chosen book was perfect for OCR. It consists mainly of clear and legible text with consistent alignment and formatting. To execute the OCR process, Adobe Acrobat Pro software was used, which achieved a high level of accuracy. Other programs, such as the OnlineOCR website, were tested; however, Adobe Acrobat Pro proved to be the most user-friendly and accurate. Only minor adjustments were needed to enhance the document's layout, such as improving paragraph spacing. Emilie took on the task of proofreading and text alignment to ensure complete accuracy.

Text encoding

TEI

The transcription of the material was encoded using Extensible Markup Language (XML) according to the Text Encoding Initiative (TEI) P5 guidelines, using the program oXygen as recommended by our teachers. The current encoded transcription is meant to provide a foundation for a more extensively encoded dataset for greater functionality and long term preservation (Björk, 2015).

Metadata plays a crucial role in the digitisation of cultural heritage material. It serves many purposes in long-term preservation, ensuring accessibility, discoverability, interoperability, and contextual understanding of the digitized resources (Dappert et al., 2016). In recognition of the role that metadata plays in the digitisation process, we dedicated a significant amount of time and effort to ensuring comprehensive and accurate metadata integration within our project.

The TEI code was tested using the Oxygen XML editor. The editor assists in verifying that the code adheres to the recommended structure, elements, and attributes specified by the TEI standards. By ensuring the code is "all green," meaning there are no errors or warnings highlighted by the editor, it guarantees adherence to TEI best practices.

HTML

The website's components, such as containers and image carousels, were primarily developed using the Bootstrap framework. W3Schools served as a valuable resource to establish the basic structure of each HTML page.

The Index page showcases an image carousel, providing a general overview of the book's appearance rather than displaying its contents in detail due to the small size of the carousel. A bigger carousel was trialed; however, due to the size of the images in the bigger container, it did not look visually appealing. The page also features a brief introduction to the book and the project.

The Transcription page presents images and transcriptions side-by-side for easy reading and efficient information retrieval.

The Gallery page provides an overview of the scanned pages of the book, with the opportunity to zoom in on each image for a closer look. On most current websites with solid UX design, you may click into the picture and exit the full view in a variety of ways, such as by using the browser's back button or by touching on any blank area surrounding the image. However, there was only one option to leave the entire view: push the back button. This may be a potential feature that could be improved in the future to increase usability.

The Resource page includes links to different project components, such as the TEI file, GitHub repository link, and PDF download. Additionally, we incorporated a contact form using Google Forms. Initially, we attempted to create a simpler form without relying on Google Forms, but encountered compatibility issues with GitHub's lack of support for PHP files. This led to a significant amount of time wasted on an unsuccessful endeavor.

Testing of the HTML was done through the W3C validator (W3C, n.d.). By doing thorough testing, we detected and resolved a number of errors and inconsistencies, leading to significant improvements in the overall project quality.

Publication and distribution

The digitisation project was successfully published and distributed using GitHub as the platform. GitHub provided an efficient and user-friendly environment for sharing the project. Using Wout Dillen's template made it straightforward to set up the project on this platform.

The presentation consists of a short introduction of the project, scanned pages in a carousel format for reading, scanned pages with a parallel transcription, the TEI and the HTML. The website has been user tested the old fashioned way - by asking our family members for valuable feedback.

Evaluation

This project has been a fruitful learning experience for us in almost every stage and aspect of it.

A short reflection on the collaboration itself seems relevant considering that we are studying a distance Master's programme in Digital Libraries and Information Services, and are most likely to end up in a similar position when working on the thesis or, even more relevant, in our future working life. To collaborate with somebody you have never met before, working from different countries and in different time zones was challenging in itself. We managed well by being so lucky that we had the same level of ambition, the same way of prioritizing issues and managed to come to consensus quickly when problems arose. We basically had no problems with trust, and all strategic decisions have been discussed and taken together. But we also managed to organize our communication in a way that moved the project forward. We have continuously presented and commented on any steps and changes made by the other team member. By doing this, we have been able to discuss and solve problems at an early stage, make sure we are on the same page, encourage one another while working separately, and ensure that we both approve the final product. So according to project management theory (as explained by, for example, Krystal Thomas, 2018) our project could have done better with better documentation of project briefings and also we overlooked the potential of using spreadsheets to keep track of the progress. But overall our collaboration was successful when it came to scoping, planning, communication and decision making. To have a good base was important when it came to troubleshooting, because of course things went wrong.

It is recommended (Cornell, 2003; Deutsche Forschungsgemeinschaft, 2013) to go through the legal aspects of a potential project before embarking on it. In this project, we have learned a bit about the challenges of the art of obtaining copyright approval. We obtained approval by the copyright owner before starting the project, but halfway through we realized that the photographs in the book might need a separate process. Sylvia Berg was the photographer of the front cover photo and could give her permission immediately. After some research on copyright of older photographs we could draw the conclusion that the photographs in the first chapter can be considered photographic images (as opposed to photographic works of art), and the people depicted are not alive anymore, so no additional approval was needed from the people depicted nor the photographers. But the timeline did not last for the people depicted on the back page image to give their approval, which meant that we had to remove it from the project temporarily. In our case, this is a school project and a learning experience, but the lesson learned is that copyright must be assigned time and competence in any digitisation project to avoid jeopardizing the accessibility of the work.

The result of the scanning was successful in terms of quality and usability in the project. But one mistake that was made in the scanning process was to not scan each page separately, converting each page to a separate TIFF-file. It turned out that it took unnecessary time and effort to split the one document into separate files, that probably would have been saved by doing the scan in a different way to begin with. However, since the scanner was not available to us at all times it was more efficient to go through the process of splitting the available pages digitally.

Dahlström (2011) expresses critique against digital libraries not taking their publishing role more seriously, by not utilizing the potential of the digital interface when it comes to exhibiting analogue media in digital form. In this project, neither of us have much experience in building websites so we used the template provided by our teacher Wout Dillen to make sure that we could present a basic and user-friendly website. However, we wanted to make some adaptations in order to do the material justice, to the best of our abilities of course. One of the features that we thought would be appropriate for the presentation of our material is the so-called "carousel" of pages that is displayed on the main page of the website. It was challenging to get it right. The conclusion is that it is worth it to create an engaging interface, but that it takes time and training to be able to execute any such type of vision.

We had not planned to user test the website, but got some very valuable feedback from family members that were asked spontaneously what they thought. Questions like: "What am I supposed to do with this?" and "Why is there a transcription if I can already look at the scanned images?" served as guidance on how to formulate the texts and tweak the interface of the website. The lesson learned is to never forget to test any service that we develop, on users who are not involved in the project, and preferably in a slightly more formal manner.

Recommendations for future work

Due to the time constraints of this project we have digitised only the introducing texts and first chapter. The reason for selecting the introductory part of the book is to make a natural starting point for any further work on digitising the rest of the book. The primary focus for future work would involve the continuation of digitising the remaining chapters of the book using TEI. It is crucial to maintain consistency in encoding practices, formatting, and overall structure across the entirety of the book.

A natural way forward when it comes to the TEI-encoding would be to develop a more detailed description of names and places throughout the book. This would make the material even more useful when it comes to information retrieval and analysis, especially considering users interested in genealogy. Specifically, we noticed that the majority of the female names were mentioned without surnames as they usually appeared in connection to their husbands who always were presented with full names. With some further research on what the full names of the women mentioned were, it could be included in the TEI.

Referring to another point made by Dahlström (2011) which is digital libraries not taking more responsibility for whether the digital products are actually accessed and used or not, one question for future work would be to facilitate that the digitized version finds its readers and users, even opening it up for further interaction with users and expansion with more documentation and accounts on the history of Ammarnäs.

Currently, the website's layout is unsuitable for smaller mobile screens and is difficult to navigate. Mobile users are the fastest-growing user group and access to information is increasingly shifting away from the desktop and into mobile environments (Xie & Matusiak, 2016; Broussard et al., 2010). Future development may involve designing a more user-friendly and intuitive interface for smaller screens that facilitates easy access and navigation of the digitised book.

Furthermore, we could explore the possibility of generating multiple output formats from the digitized content. One possibility is to create e-book formats such as EPUB and MOBI. These formats are specifically designed for electronic reading devices and offer enhanced readability and flexibility. By converting the content into e-book formats, we can reach a broader audience who prefer reading on devices like e-readers, tablets, or smartphones.

Additionally, exploring formats suitable for text-to-speech applications would enhance accessibility for individuals with visual impairments or those who prefer audio content. By offering a range of output formats, we can ensure a broader reach and better cater to the diverse needs of users.

The completion of the full book transcription is estimated to take approximately 200 hours as a starting point, encompassing the scanning and digitising process for the remaining pages. However, it is important to account for potential time extensions due to copyright considerations, especially for pages containing photographs that require additional permissions. Furthermore, the integration of text-to-speech capabilities and the creation of a more user-friendly interface suitable for smaller screens can significantly increase the overall time needed for the project. Given the complexities involved and the future goals of the project, accurately predicting the precise time required for the continuation of this project proves to be a challenging task.

Conclusion

In conclusion, the digitisation project of "Ammarnäsbygdens historia 1803-2008" has been a valuable learning experience, showcasing the successful collaboration between Sasha Spirkina and Emilie Lindström. The selected book, with its local historical significance and limited availability, proved to be a suitable candidate for digitisation. Overall, the project not only achieved its objectives but also laid the foundation for potential future digitisation and preservation of this book.

References

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

Broussard, R., Zhou, Y., & Lease, M. (2010). Mobile Phone Search for library catalogs. Proceedings of the American Society for Information Science and Technology, 47(1), 1–4. https://doi.org/10.1002/meet.14504701128

Cornell University Library (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., Guenther, R. S., & Peyrard, S. (2016). *Digital Preservation Metadata for Practitioners: Implementing PREMIS*. Springer International Publishing AG.

Deutsche Forschungsgemeinschaft (2013). *Practical Guidelines on Digitisation*.

P5 guidelines. TEI Text Encoding Initiative. (n.d.). Retrieved February 21, 2023, from https://tei-c.org/guidelines/p5/

Tanner, Simon (2004). Deciding whether Optical Character Recognition is feasible. London: King's College. 11 pp.

Thomas, K. & Southwest Florida Library Network (SWFLN). (2018, August 14). *August 14, 2018: Project Management for Digital Libraries* [Video]. YouTube. https://www.youtube.com/watch?v=3KrHfDRrNH8

W3C. (n.d.). The W3C Markup Validation Service. W3.org. https://validator.w3.org/

Xie, I., & Matusiak, K. (2016). Discover digital libraries: Theory and practice. Elsevier.

Appendix

COPYRIGHT

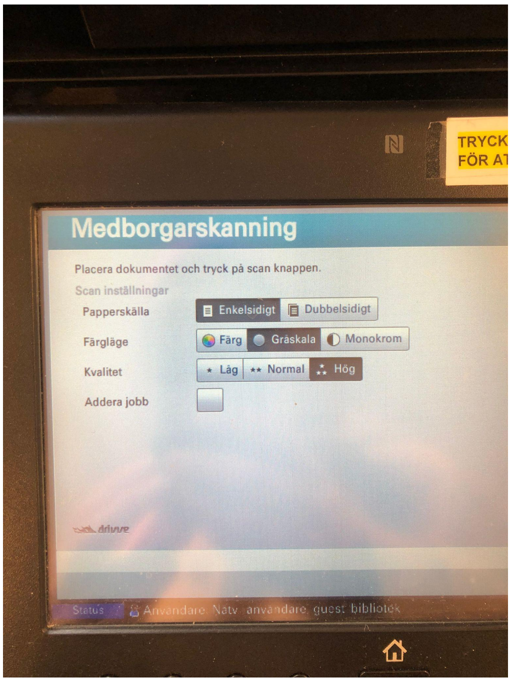
March 12 - Copyright admitted by Sylvia Berg:

”Jag godkänner att uppgifterna som finns i boken Ammarnäsbygdens historia 1803-2008 används i projektet om Digitalisering av kulturarvsmaterial. Frösön den 12 mars 2023. Sylvia Berg tel 070-3250979 mail: sberg8945@gmail.com

Sylvia Berg”

SCANNER INFORMATION

Figure 1. Scanner type: Kyocera TASKalfa 5052ci



Medborgarskanning

Placera dokumentet och tryck på scan knappen.

Scan inställningar

Papperskälla [✓ Enkelsidigt] Dubbelsidigt

Färgläge Färg [✓ Gråskala] Monokrom

Kvalitet * Låg ** Normal [✓ Hög]

Addera jobb □

Status Användare Nätv. användare guest bibliotek

Figure 2. The picture shows the scanner setting options:

Paper source: one sided or double sided

Color: Color or Gray scale or Monochrome

Quality: Low* or Normal** or High***

DIVISION OF LABOUR

Emilie Lindström	Sasha Spirkina
Obtained the book and copyright	Created the Github repository and google doc for the report
Scanned Images in TIFF	Completed the OCR
Corrected the OCR transcription	TEI: facsimile and body
TEI: Header	TEI: dates
TEI: place names and people names	HTML: Index, Gallery, Resources
HTML: Index and Transcription	Converted Images to JPG
Report: Introduction	Github: Read Me
Report: Digital strategy and aim	Report: Image Format and Storage
Report: Planning and division of labour	Report: OCR
Report: Scanning	Report: TEI and HTML
Report: Evaluation and Conclusion	Report: Evaluation and Conclusion

TIMELINE OF THE PROJECT

12/03/2023 - Approval received from the copyright owner for material usage.

13/03/2023 - Initial Zoom meeting: Emilie presents the book as the project's foundation.

28/03/2023 - 2nd Zoom meeting: Discussion on Image Capture and OCR.

18/04/2023 - 3rd Zoom meeting: Discussion on TEI.

23/04/2023 to 25/04/2023 - In-person meetings, dedicated 5 hours each day to work on TEI.

28/04/2023 - Meeting with Mikael to discuss the project.

26/05/2023 - Project work seminars.

13/05/2023 - 4th Zoom meeting: Discussion on HTML issues.

20/05/2023 - 5th Zoom meeting: Discussion on copyright matters.

26/05/2023 - 6th Zoom meeting: Finalization of the project website.

03/06/2023 - 7th Zoom meeting: Finalization of the project report.

05/06/2023 - Project submission.

PROJECT TIME ESTIMATES

Project phase	**Time estimate**
Preparation of material	5 h
Scanning	5 h
OCR	10 h
TEI	60 h
HTML	60 h
Publishing	10 h
Literature	40 h
Report	50 h
Strategic planning/meetings	20 h