Bert Kunzmann
Mohammad Habibur Rahman

Project work report on the digitization of cultural heritage: Kristdala Socken

by Bert Kunzmann and Mohammad Habibur Rahman

1. Introduction

Digitization is the procedure to transmute the primary physical material of analogue manner to digital portrayal with the help of computer facilities like digital camera or scanner (Zhang and Gourley, 2008). The purpose of digitization of cultural heritage can be seen in the light of recent technological developments. As technological trends tend to move from the analogue to the digital realm, it has become more reasonable in recent years to digitize already available material.

Motives of digitization cultural heritage materials:

- Preservation - to reduce wear and tear on the original materials
- Making the materials more accessible, usable, and re-usable
- Developing and sharing technologies and competences
- Facilitating new research
- Presenting the materials in new ways and contexts
- Administrative, political, and economical purposes

The material mentioned in this report is referred to as foremost printed media and should have some cultural value. Printed material in this sense is either plain text (such as a book), or other texts (such as magazines or even postcards) with both text and illustrations. Digitization in the sense of this project is being referred to as “the process of creating digital representations of information resources recorded on analog carriers”, or, to put it more simply, the conversion process from an analog source to a digital medium (Xie & Matusiak, 2016, 59ff.). The overall goal of digitization is access and preservation of the given material. These two complementary aspects share the same motivation – to give users the opportunity to have access to a wider range of material, both in the present (hence accessibility) as well in the future (as preservation is set out for future use of the digitized material).

Bert Kunzmann
Mohammad Habibur Rahman

Purpose, goals, and objectives of the project work:

To digitize the selected pages of physical textbooks and images regarding cultural heritage named “Anteckningar om Kristdala socken i Tunaläns härad och Kalmar län” for long-term preservation and gaining increased introduction of and access for the book.

2. Problems and challenges

When taking on a digitization project, several problems should be addressed before any work can be done. What is the actual value of digitizing material that is written by an (perhaps unknown) author, about a place and a time that does not seem to be of interest for the common reader? Then, copyright issues should be addressed as well. Is the material still in copyright or already in the public domain? For the former, are there any persons in charge of the distribution of copyrighted material or is it impossible or improbable to get in contact with the copyright holders? For the latter, what would be the purpose of digitizing material that is in the public domain? If copyright does not exist anymore (as more than 70 years should have been passed since the author’s death), who should then be responsible for the preservation and maintenance of the material? Public institutions such as libraries and archives are the holders of the material, but there are remaining questions about the value as well as maintenance issues of said material.

Another issue is the proportionality of a digitization project. Is it fruitful or useful to digitize material that is out of copyright, with regards to readers, and the holding institution, but also concerning the amount of work that such a process entails. Within a digitization project, several steps have to be taken into consideration. Before the actual digitization process can be initiated, planning, selection, and preparation of the materials provided should be taken care of (see Xie & Matusiak, 2016, 65ff.). First, analog material (books, magazines or other media such as film or 3D sculptures) should be selected and assessed and prepared. The selection process is crucial for the whole project as its realization and outcome are dependent on the material that was first selected. In this regard, one has to take several things into consideration, such as the quality of the material (is it in good condition or damaged, or nearly deteriorated), the format and size (a book too large or heavy may not be properly scanned), the copyright status (copyright holders may block an ongoing project if they deem the material to be too sensitive to be digitized). Also, if an object is too fragile (or nearly deteriorated) to be digitized, then it either has to be conservated beforehand, or the project may not be initiated at all as its outcome would be devastating in a worst case. Digital image capture is the next step within the digitization process and can be done using scanners, digital cameras or analog-to-digital converters.

The files produced during this section are master files of high, lossless quality and can be used by future researchers for further examination and preservation, or to create derivates that can be used for digital collections or project webpages. During digital processing, the acquired master files can be edited in order to improve their quality, or to enhance their functionality. Also, with the proper software, Optical Character Recognition (OCR) can be processed on printed text to create searchable documents. After metadata has been created, the digitized material can be uploaded to digital management systems such as databases or collections, and are now ready to be used (given that the proper digital infrastructure was implemented beforehand). Maintenance in this context is important as the purpose of digitized material should not only be aimed at accessibility but also preservation for future use. As digital files do not deteriorate as analogue objects, storage space should be managed and held under maintenance so that future researchers or staff can access the material for projects, exhibitions or any purpose alike.

3. Material for digitization

The material provided for this small-scale digitization project is an excerpt (62 pages) of a book of 184 (172 + 12 introductory) pages by Swedish author Carl Alexander Carlsson (1859-1921). The contents of the book "Anteckningar om Kristdala socken I Tunaläns härad och Kalmar län" that was printed in 1900 deal with the history, topography and people and their daily work and life in the area of Kristdala socken, a small district in Northern Småland (Sweden), set at the end of the 19th century. The text is written entirely in Swedish, with some annotations, words or phrases in Latin. All material mentioned in this report is stored at the municipal library in Vimmerby, under supervision of both the head of the library, Thomas C. Ericson, and the director of the local museum *Näktergalen*, Gunilla Gustafsson, who both agreed on the digitization and the subsequent publication on the Internet via GitHub.

3.1 Selection process

The choice of material for this small-scale digitization project was motivated through a variety of factors, which shall be further explained in the following section. As described above, the selection process is crucial for the whole cultural digitization project. Which kind of material should be chosen, how much, and does the time frame for a small-scale project even allow to give alternatives in case the material that was first chosen reveals to be of less quality than expected. Then, not only the quality of the content is important but also in which condition the material is presented. The book seemed suitable for the project as the pages were readable, not blurred or deteriorated as with other printed material of that time era, and the overall quality was deemed satisfactory. The size of the book is also suitable for the project as two pages in the book are nearly one A4 page, which makes it perfect for scanning

purposes (the scanning process shall be described in detail later). Also, the fact that the material is out of copyright proves to be useful as the problem of copyright infringement will not be an issue when publishing the project on the Internet. As the author passed away in 1921 and the book was printed in 1900, the work is now in the public domain and therefore acceptable and of use for the project. Then, perhaps the most important issue has to be addressed, that being the value from the viewpoint of cultural heritage. Is the material important from a cultural heritage perspective or is the effort not worth the result? In this case, when it comes to the issue of giving value to a project, it shall be mentioned that the book about Kristdala county has not been digitized yet, neither by the National Library of Sweden (Kungliga biblioteket) or any other institution. This should not be seen as a sign of lack of interest but as an opportunity to initiate a project that first takes on to digitize a part of Southern Swedish history. The county of Småland is situated in the south, covered by woods and lakes and is known to be inhabited by people who were able to adjust to the rough environment. Looking at the material, historical as well as geographical aspects of this very region represent a certain value from a cultural heritage point of view. The material provides many insights into the history, way of living and habits of the people living in the area at that specific time, around the millennium (‘sekelskiftet’) of the 1900s. The aspects mentioned above were the main reasons why the book about Kristdala county was taken as material for the digitization project.

3.2 Unused material

Other material that was not deemed appropriate for further use shall be briefly mentioned here. Out of the unused materials is a small collection of about five booklets, 50 pages each, printed in the 1960s in the former German Democratic Republic (GDR). These are reprints of German, Russian or Slavic authors, translated into German. This is a private collection, in possession of one of this project’s authors. All of these booklets are part of a collection called *Das Neue Abenteuer*, which was primarily aimed at adolescents. One issue of this collection is no. 219, *Der Kommandant der Vogelinsel* by Sergej Wladimirowitsch Dikowski, first printed in 1952, second edition in 1962 (the one in possession). As the German translation has its own copyright and one of the illustrators who provided sketches for the booklet passed away in 1994, the material was considered not suitable for digitization due to copyright issues. Another example of unused material was one of the first editions of a local newspaper, “Wimmerby Weckotidning” from the town of Vimmerby in Småland, dated January 1865. Holding institution is the local library in Vimmerby. The material, although out of copyright and in an acceptable condition, could not be taken as it had already been digitized by the National Library of Sweden. Also, the text of the newspaper was written in fraktur (i.e. Blackletter) which can cause problems with the OCR process. Even sophisticated OCR software like Abbyy Finereader would not be able to transcribe the text in a satisfying way.

4. Digitization process

4.1 Image capture

After the proper material was carefully selected, that being "Anteckningar om Kristdala socken I Tunaläns härad och Kalmar län" (1900) by Carl Alexander Carlsson, the next step was image capture and image processing. As image capture is a crucial part of each digitization project, both the capture method and the technical equipment had to be chosen with care. The options were either photography with a digital camera or scanning via an A4 flatbed scanner. As the material was in good condition, not faded, of light weight (ca 500g), and had a practical size (one page equals an A5 size), the latter option was deemed appropriate. It has to be mentioned that this was a sound choice as a digital camera that could shoot high quality (i.e. lossless images) could not be acquired by the project members, nor could a photo studio be contacted. Also, scanning the material was deemed a good choice as it was fairly cost and time efficient, and of good quality. It was mentioned that at a later stage in the project, derivatives (PNG or JPEG) could be made from the master files (TIFF) by using image software such as GIMP. Thus, scanning the material directly to PDF would be unnecessary.

In the following section, the image capture process shall be explained in more detail, with regard to the guideline set by the Deutsche Forschungsgemeinschaft, and the publication *Practical Guidelines on Digitisation* (2013). The scanning equipment used for the project was provided by the library in Vimmerby. Initially, there were three different scanners provided, only one of them that was working properly, that being a SHARP office scanner used primarily for copying and printing issues. The scanning software was built in and could manage TIFF, JPG and PDF for up to 600dpi. As mentioned in the *Practical Guidelines on Digitisation* (2013, 13f.), the scanning resolution of plain text should be 300dpi, and be scanned into TIFF which is a lossless image standard used for image capture. The material was first scanned as TIFF (color) at a resolution of 600dpi. The scanning process proved to be difficult as the material could not be placed at a right angle due to the size of the scanning bed. This means that each book page was skewed, which can be problematic with the OCR text later in the project. Also, the quality of the scanned images was not satisfying enough to convince the project members. The result was that the images were overexposed and contrast too low, i.e. it was difficult reading the scanned pages. Therefore, image capture had to be done again, at the municipal library in Linköping, under supervision of Mia Moberg, librarian at the library and with collaboration of the DKV (*Digital kreativ verkstad*), housed in the same building. The equipment used was an A4 flatbed scanner, Epson Perfection V600 Photo, that could scan the images at a much better quality. The images were then captured at a resolution of 400dpi (color, 8 bit per channel equals 24 bit) with LZW compression. The resulting images were of far better quality, and not skewed because the book fit exactly the A4 format of the flatbed scanner. The metadata (eg. resolution, colorSpace-RGB etc.)

acquired through the scanning process was stored in the images, without any further changes. The image capturing process took about three hours (excluding one hour scanner calibration) and resulted in digitizing about 62 book pages into 31 TIFF color image files.

4.2 OCR transcription

As the image capture was completed, the OCR process had to be initiated. The process of text recognition can be described as an image analysis that, when put in an OCR program, is converted from analogue script to machine readable digital text (see Tanner, 2004, 3). Abbyy Finereader was chosen as the default software for Optical Character Recognition, as it is superior to other open software applications such as Tesseract. Furthermore, the Finereader has more options regarding file output (in .txt, .docx, .pdf and other formats) and a more sophisticated language setting. Several languages can be processed at once with Finereader, and the overall quality of the transcription is far more convincing than that of other OCR software, with a much lower error rate and a much higher accuracy performance. One has also to consider that, although the material is easily readable and the font set in Antiqua (not Blackletter which is not suited for OCR processing), text recognition is feasible and meaningful as the sheer volume of the text would be too large to write manually. With Finereader, most of the text was being transcribed accurately, with approximately 90% word accuracy, transcribing errors (such as symbols or historiated initials) not being counted. All TIFF images were transcribed into text (.txt) files, as it was considered best for the ongoing digitization process. Other formats such as .pdf or .docx would not have formatted text and thus be useless in the later project. Text transcription using sophisticated OCR software generally is time and resource efficient, as the time frame set within the digitization project was limited.

4.3 TEI encoding

After the OCR processing was finished, cultural text encoding was initiated according to the TEI (Text Encoding Initiative) guidelines. The software used for this approach in cultural text encoding was oXygen, a text editor that is specifically designed for working with TEI. The motivation behind the decision of text encoding was that the OCR transcription would not be enough for future researchers who want to work with the project and its contents. OCR transcription would be merely of interest for internet users whose sole interest lies in reading the book pages, but who do not share any interest in the implementation of XML files into databases or XSL transformations. As Renear (2004, 234) points out, the motivation behind the TEI consortium was “to develop interchangeable guidelines that would allow projects to share textual data (and theories about that data) and promote the development of common tools.” Using the TEI guidelines for text encoding, the crucial part of the work was what had to be encoded and why. From the beginning it was clear that, as the material entails historical

and geographical descriptions, as well as place names, dates and names of people, that most of these had to be encoded, and this had to be done thorough with regard to the guidelines developed by the TEI consortium. In the end, what items are to be encoded is decided by the encoders themselves, yet a more thorough encoding proves to be more convincing and useful for future researchers who want to enhance the digitization project about Kristdala socken with further material. Given that the size of the digitization was quite excessive (60 pages) for a small-scale project, the text encoding part was split between us (the project members). Further discussion about which parts were to be encoded proved to be quite challenging, but this process led to a coherent (and valid) XML file that also includes metadata about the book, its holding institution, encoders and in which context the project was made (i.e. digitization project at the University of Borås). From the resulting XML file, an XSL file was then added to apply XSL transformations. The approach with XSLT only applies to the diplomatic transcription; other web pages within the project such as the index page were not included in this process. The work with XSL proved to be difficult due to a lack of both knowledge and time. An example XSL file and a description about the transformation process was provided beforehand by the course lecturer, yet a more detailed and thorough explanation as to what template should be included in the XSL file to what purpose would have been more helpful. Thus, a more or less satisfying XSLT was the result of this approach. XSLT was done with an HTML as output. While the basic layout of the web page with two main columns was successfully transformed, the result was not sophisticated enough to be satisfying for users, as some of the text was not transcribed properly. Further improvements would have to be done in order to reach a more satisfying result. Yet, as the time frame was quite limited, no further improvements could be done with XSLT.

5. Publishing the project via Github

As the project was about to be published on the Internet using Github Pages, several aspects had to be taken care before even building a functioning website presence. First, copyright issues had to be solved. This aspect was, at least for this project, out of the question as the material provided for the project was exempt from copyright (as mentioned before). Then, derivatives of the scanned TIFF images had to be created, as master files are not meant to be published on the Internet publicly but are more a resource for future researchers (see Xie & Matusiak, 2016, 60f.). Additionally, TIFF files are rather large (ca 20MB each), so publishing them on the Internet would cause the webpage to load very slowly, plus downloading these files would be counterproductive for normal users. Instead, PNG derivatives of the image files were first created. Yet, although this file format is more suitable for digital publishing, the files were still very large (roughly the same size as the master files), so a different approach had to be taken in order to efficiently use the digital resources. The file format that was agreed on was JPEG, and was created from the master TIFF files using the Open software GIMP, a common image manipulation and editing software. The conversion of the images proved to be better, as less storage space was used (ca 2MB per file) and image

quality remained the same. An additional step had to be taken before the material could be uploaded to the web repository on Github. As the scans showed two book pages at once, it was hard reading the script from the image files as the font size was too small. Thus, the images had to be cut in two and then be uploaded separately. When accessing the website, users could then read each separate page alongside the OCR transcription, instead of two pages at once. This proved to be a more sophisticated and useful approach to publishing the content on the Internet. It was also made possible to download each image file (JPEG) by simply clicking on them. The website itself then consisted of four different pages with relative. First, an index page with information about the project, its contributors, contact data, copyright and the holding library. Second, the diplomatic transcription with the image files (JPEG) on the left and its OCR transcription on the right side. Third, a slideshow of the pages used during the project is presented, and on the fourth page users can download several files relevant for the project. When accessing the latter, users can download the project report, the OCR transcription of the book as plain text, and the TEI as well as XSL file used for cultural text encoding and XSL to HTML or PDF transformation. Thus, users and researchers will have a basis either for the enhancement of the current webpage – which is, in fact, currently a work in progress – or a useful asset for their own digitization project.

Division of labour

Activities in various phases of the project	Time used
Planning: reading literature, analysis and preparation	10 hours
Selecting, assessing material	3 days, approx. 24 hours
Image capturing (TIFF master files)	5 hours
Derivatives (PNG, JPEG)	1 hour
OCR processing (TIFF to .txt)	2 hours
TEI (Text encoding)	2 weeks, approx. 90 hours
Proofreading (OCR, TEI)	5 hours
GitHub page creation and uploading all doc (OCR text, image files, .tiff file tei xml xsl etc.)	3 weeks, approx. 120 hours
Project report (writing, proofreading)	10 hours
Total time	**267 hours**

6. Outlook and future of the digitization project

As the website is created by using a Github repository that is publicly available, the content and presentation of the digitization project can be altered afterwards. New files and additional content can be created, deleted, and adjusted according to the intentions of present or future users and/or researchers. Book pages can be added as well as TEI and XSL files updated; more master files can be added and derivatives be created. The repository should serve the purpose not only of increased accessibility but also preservation, as mentioned in Xie & Matusiak (2016, 60f.). Sustainability in a sense of making content available and then preserving for future use it is also crucial and beneficial for digital projects.

7. Concluding remarks

In order to be a fruitful project, the products and deliverables need to be fastened into the goals and objectives. Quality management is very important in a project, and it is the key element in confirming a successful project. The circumstance of building an environment for collection storage, preservation selection and physical handling of collections (books and images) should be evaluated as an initial assessment to make a report on the general preservation environment to improve the storage and environments with proper suggestions. Other problems when initiating a digitization project, either on a small or large scale, are the use of proper technical equipment and the time frame. The material used in this small-scale project was considered of good quality and out of copyright, thus making it suitable for publication on the Internet. Yet, to digitize roughly 60 pages was, in the end, too ambitious as the narrow time frame would set difficulties when working with OCR transcription and TEI encoding. As the project had two members, the work was divided into different parts so that each step in the process could be done individually, as long as communication was provided and the outcome was checked by each member. The OCR part was necessary as writing the text manually would have exceeded the time limit. Another problem included the cultural text encoding as well as XSL transformation. As TEI encoding was a new approach to both project members who had no previous experience with this kind of work (except for a short XML introduction in another course), several issues were addressed during the project. For instance, how will metadata be edited in the TEI header, and how much encoding is necessary for material that is written in prose. XSL transformation was another problem, due to two reasons. First, inexperience and lack of knowledge about the functionality of XSLT was an aggravating factor during the latter part of the project. Second, due to scheduling issues, there were not many opportunities left for a closer look at XSLT. The resulting XSL file and the transformation to HTML might be satisfying to the project's members, yet more experienced researchers may consider it imperfect. Due to the project being very ambitious because of its amount of material being digitized, the time factor was of concern.

8. Bibliography

Blanke, T., Bryant, M., &, Hedges, M. (2012). Open source optical character recognition for historical research. *Journal of Documentation, 68(5)*, 659-683.

Calhoun, K. (2014). *Exploring digital libraries: foundations, practice, prospects*. Facet Publishing. DOI:10.29085/9781783300297

Cameron, F. & Kenderine, S., (Eds.). (2007). *Theorizing digital cultural heritage: a critical discourse.* Cambridge, Mass.: MIT Press.

Chowdhury, G. (2013). Sustainability of digital information services. *Journal of Documentation*, 69(5), 602-622. DOI: 10.1108/JD-08-2012-0104

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2020). Flexible character accuracy measure for reading-order-independent evaluation. *Pattern Recognition Letters, 131*, 390-397.

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

Dempsey, L. (2020). Foreword: library discovery directions. In McLeish, S. (2020), (Ed.), *Resource Discovery for the Twenty-First Century Library: Case Studies and Perspectives on the Role of IT in User Engagement and Empowerment* (pp. xxi-xxxii). Facet Publishing.

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

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

Islam, N., Islam, Z., &, Nazia, N. (2016). A Survey on Optical Character Recognition System. *Journal of Information and Communication Technology, 10(2).*

Renear, Allen (2004). Text Encoding. In: Schreibman, S. Siemens, R. & Unsworth, J., (Eds), *A Companion to Digital Humanities*, p.218-239, Oxford: Blackwell.

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

Bert Kunzmann
Mohammad Habibur Rahman

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

Terras, M. (2015). Opening Access to collections: the making and using of open digitised cultural content. *Online Information Review*, 39(5), 733-752.

Xie, I. & Matusiak, K. (2016). *Discover digital libraries: theories and practice*. Amsterdam: Elsevier Science Ltd.

Zhang, A. B., & Gourley, D. (2008). *Creating digital collections: A practical guide*. Chandos Publishing.

Bert Kunzmann
Mohammad Habibur Rahman

Appendix 1: Technical documentation

Table 1: Metadata and material analysis

Type of object	Book
Title	Anteckningar om Kristdala socken i Tunaläns härad och Kalmar län
Author	Carlsson, Carl Alexander (1859-1921)
Format	Printed text
Language	Swedish
Publisher	Smedbergs boktryckeri
Publishing place	Stockholm
Publishing date	1900
Binding	Hardcover
Total number of pages	184 pages (xii, 172)
Pages digitized within the project work	62 pages in 31 image files (TIFF master files, see below)
Typeface	Antiqua
Repository	Stadsbiblioteket Vimmerby, Kalmar district, Sweden
Placing	Archive Smålandssamlingen (local heritage section)
Physical condition of object	- Pages: good condition without any markings or stains - hardcover, no bruises - binding intact
Size of object: l x w x d (cm)	23 x 16 x 1.4
Size of pages: l x w (cm)	22.3 x 14.5
Total number of pages	xii, 172 (total 184) pages
Languages	Swedish, some annotations in Latin

Bert Kunzmann
Mohammad Habibur Rahman

Content of the material	- plain text, no photographies - one small illustration on page 1 (top, size 9.5 x 2.5cm) - bibliography in the first 12 pages - annotations at the end of the book - library stamps at the beginning (frontpage, page ii)
Condition of the text	- text clearly visible, not faded, not stained or damaged - smaller font size in annotations and footnotes - text showing through backside of pages (whole book)

Table 2: Technical equipment (scanner, software)

Scanning equipment	Epson Perfection V600 Photo
Image capture master files	Master files scanned at 400 x 400 dpi TIFF lossless Compression : LZW RGB color: EPSON sRGB 8 bit color depth per channel = 24 bit Image enhancement: dust removal, backlight correction
Scanning Software	Epson Scan
OCR Software	ABBYY FineReader (TIFF conversion to .txt)
Graphic Editing Software	GIMP 2.1 (TIFF to JPEG at 8 bit, lossy, GIMP built-in sRGB)
	Convertio online converter (TIFF to PNG, PNG not used)
XML Editor	oXygen XML Editor 24.1
Content Management Software	GitHub
URL Project work	https://github.com/bekun0700/Digitisation_Kristdala

Table 3: Storage space requirements for further digitization

	Average file size per image	File size total	Total sum
Digitization project: book excerpt			
31 images (equals 62 pages)			
JPEG	3 MB	93 MB	
TIFF	20 MB	620 MB	
			713 MB
Digitization project: whole book			
92 images (equals 184 pages)			
JPEG	3 MB	276 MB	
TIFF	20 MB	1,840 MB	
			1,916 MB

Note: In order to successfully digitize the whole material, a storage space of minimum 2 GB is needed (excluding OCR text transcription, XML/XSD TEI encoding, HTML/CSS files).