Digitisation and Digital Library Presentation System – A Resource-Conscientious Approach

Tuula Pääkkönen[1[0000-0003-3958-9732]] and Jukka Kervinen[1] and Kimmo Kettunen[1[0000-0003-2747-1382]]

[1] National Library of Finland, Finland
tuula.paakkonen@helsinki.fi

Abstract. The National Library of Finland (NLF) has done long-term work to digitise and make available our unique collections. The digitisation policy defines what is to be digitised, and it aims not only to target both rare and unique materials but also to create a large corpus of certain material types. However, as digitisation resources are scarce, the digitisation is planned annually, where prioritisation is done. This involves the library juggling the individual researcher needs with its own legal preservation and availability goals. The digital presentation system at digi.nationallibrary.fi plays a key role, since it enables fast operation by being next to the digitisation process, and it enables a streamlined flow of material via a digital chain from production and to the end users.

In this paper, we will describe our digitisation process and its cost-effective improvements, which have been recently applied at the NLF. In addition, we evaluate how we could improve and enrich our digital presentation system and its existing material by utilising results and experience from existing research efforts. We will also briefly examine the positive examples of other national libraries and identify universal features and local differences.

Keywords: digitisation process, digital presentation system, research and development projects, research data

1 Introduction

As part of its preservation and accessibility goals, the National Library of Finland (NLF) has been digitising its collections from the year 1998, when the Centre for Digitisation and Preservation was created as a unit to Mikkeli. Since then, based on the resources available, approximately 1 million pages have been digitised annually and put into a presentation system for the public to view. At the moment, for example, the number of digitised newspapers and journals has exceeded 12 million pages [1], which can be viewed at the digi.nationallibrary.fi service. In addition to the largest material selection, the doria.fi service contains books, maps and images, and audio is also part of the National Library's digitised material.

The digitisation process, which has been formed in the National Library, has defined the key steps that are crucial for a successful digitisation workflow. The library users

usually associate digitisation with only the scanning phase; however, own expertise is also needed before and after scanning to ensure a smooth digitisation workflow. In our environment, we have defined a model called a digitisation chain, which provides an overview of the digitisation process at both a high and a detailed level. It also visualises the real-life challenges faced in ensuring that each step has enough resources so that the existing quality can be maintained or improved. This led to the planning and implementation of a new digitisation process to meet the new need to speed up the process.

2 Digitisation - What and How?

The digitisation policy of the NLF was created in 2010 [2]. As stated in the policy, "accessibility, preservation and ongoing use" are the aims of digitisation. In essence, the idea is to preserve the unique collections – this is the reason, for example, that the most-used materials are to be digitised so that they are more easily accessible to those who need them. In the long run, this also minimizes the efforts of the customer-service personnel, since the physical materials do not need to be hauled out of physical storage and brought to the researchers' desks.

2.1 The Digital Chain, Renewed

The duty of the NLF is to deposit and preserve everything published in Finland. According to the Legal Deposit Act, the NLF receives a copy of each newspaper and magazine published in Finland. Received publication materials are processed in the library according to an internal concept called the digital chain. Processing of the publication material in the digital chain consists of the following five phases: 1) material deposit and return (including cataloguing); 2) preparation, scanning and conservation (if needed); 3) post-processing, which includes structural analyses; 4) microfilming from digital version; and 5) deployment, use and preservation. The digital chain is schematically presented in Figure 1.

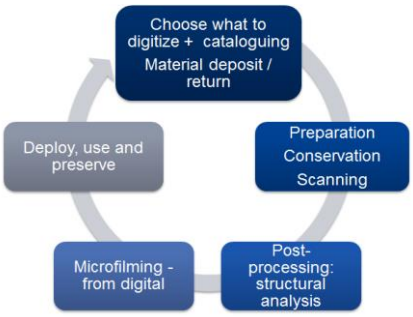

Fig. 1. The Digital Chain of National Library of Finland

Post-processing of the scanned material is the core of the digitisation chain. In this phase, the material is processed so that it can be shared with the library sector and for public use. The scanned document images are improved and run through background software and processes in the post-processing phase, which create METS (Metadata Encoding and Transmission Standard) for all the elements of the binding and ALTO (Analyzed Layout and Text Object) metadata for each page. Optical character recognition (OCR) with ABBYY FineReader is performed simultaneously to obtain text from the page images. Despite developments in OCR software and technology, there are still challenges with OCRing, since some of the materials are quite old and contain varying paper qualities, different numbers of columns, various languages and both font types both gothic and antiqua [3].

The page images and metadata are combined into a package in post-processing, where everything belonging to one issue is preserved. In the final phase, these packages are deployed so that materials are easy to browse in the online storage and retrieval system. The post-processing of digitised pages is performed in docWorks environment, which has customizable workflows for different types of digitisation materials and outputs. Digitisation operators monitor the throughput of digitisation with docWorks software, and depending on the project, they fix layouts and check and correct page numbers and issue metadata. The amount of post-processing depends on the project; sometimes, even full-text proofreading is performed, and at other times, post-processing is as light as possible, and only error cases are handled. Text-recognition software operates behind docWorks. This software produces the page text both to the ALTO files, which are used for long-term preservation, and to the database of the digital presentation system (digi.nationallibrary.fi). OCR text recognition and layout analysis are performed on several continuously running background machines, which can be monitored via separate monitoring dashboards. After these steps and the use of some additional custom-made tools, the final export package is archived, and the contents are visible on the digi.nationallibrary.fi web page.

2.2 Key Features in Digitisation

An often unknown feature in Finnish digitisation is the connection to the long-term preservation needs of the National Digital Library. The library has created a specification for mandatory (and conditional) preservation metadata[4] for each cultural history organisation that wishes to store its material using the digital preservation service.

The digitisation that the NLF conducts follows these guidelines through additional work performed over and above the digitisation processing workflow software, which, in its modular structure, enables tweaking of the post-processing flow for different material types. This is required because the NLF has many different types of materials in its collections, each of which requires varying processing steps and different scanners that all need configuration, calibration and maintenance.

The digitisation results are reproducible in the sense that the preservation image is also stored, which allows for additional enrichment of the digitised content as technology improves. Based on our research efforts, for example, with low-resource, text-recognition improvements[5] or on work done with named-entity recognition [6], it

seems that there might be a need for new process steps for research-oriented enrichments, which could be done utilizing existing tools developed by other research teams or cultural heritage institutions.

Also, with regard to long-term digital preservation, the following question arises: which version should be sent to that process – the original one from the digitisation or the more enhanced version? The digital presentation system can use either or both, depending on which features are to be developed for end users and researchers of different fields.

2.3 Digitisation Enrichment Problems and Solutions

Digitisation in the digital chain has many decisions to be made, which have impacts to the later usability of the material. As the goal is two-fold – to digitise as much as possible, with good quality – this creates constant need to find better solutions.

One example of this is article extraction, which has problems as the Finnish historical newspaper scene is quite versatile – papers have varying number of columns, different font sizes or with different paper quality, which makes detecting articles and the order of text blocks difficult [7]. One potential solution to these is seems as with case of [7] is to put extra effort on creating the article segmentation models to cover the basic "page wireframes" which a newspaper can have several over its publication period. It remains to be seen how many models would be the optimum amount to get article segmentation to production quality. In addition, the article segmentation should also take care of getting article heading and actual text in right reading order which might require redoing OCR to get the quality to better level [6].

3 Digitisation and Digital Presentation System

One key benefit, which we have realized over the years, is that keeping the digitisation and the digital presentation system close together enables faster processing of data from one step of the process to another. Regardless of need, either the digitisation or digital presentation system can be modified accordingly to meet the new needs.

3.1 Exporter and Importer

To illustrate the phases between the process steps, there are two key components in play behind the digitisation and presentation system. The digitisation post-processing produces the archival package, which contains all the results of the digitisation, as mentioned before. At that point, all the required content is in place for use with the presentation system. The export phase of the docWorks software produces the archival package based on our custom needs and places it on the storage disks.

After the export, the second part of the job is performed: importing the information about the export package to the digital presentation system. In this phase, the importer component dissects the information about the new package, and extracts needed infor-

mation to the database and finally stores the export package to correct location in storage system. The processing system of the search index then receives a new task to automatically index the new binding. After the indexing is completed the actual pages can then be found via the full-text content search.

In essence, the digital presentation system utilizes the original archive packages from the storage disks. This saves us storage space as the digitisation progresses. The quality of the scanned image also impacts the storage requirements, as depicted in Figure 2: the disk space requirements fluctuate over different weeks; however, this is directly related to the material that is digitised. With new newspapers in particular, with colour images, several supplements require storage space, in addition to the shelf space of the physical version.

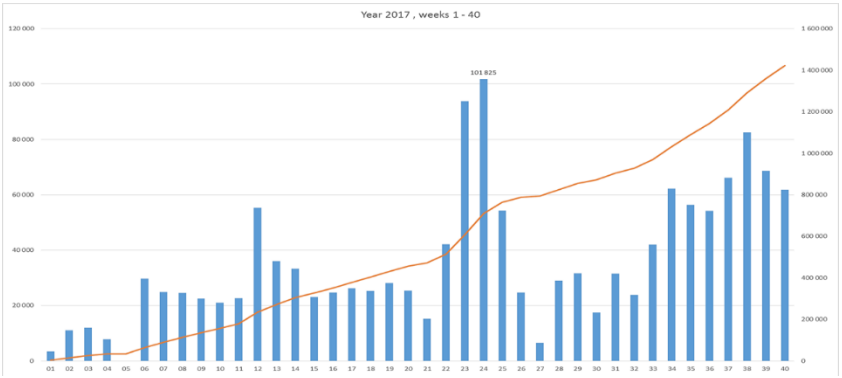

Fig. 2. Sample of the space required by importer on 2017 (weeks 1 to 40)

3.2 Digital Preservation and Content Enrichment

The changes that the digital preservation specification requires do not cause high disk space requirements as a whole. The desire is to store all of the processing steps that have been done to the digital material along with information such as who the operator was, which software and version were used, and which hardware was used. In future, this can be invaluable in opening up the digitisation process and ensuring the authenticity of the digitisation itself. The changes can be seen from the METS, which combines all of the information of an individual binding.

Adding the digitisation processing metadata to each binding is space-wise nominal; however, when implemented across all of the pages that are undergoing the digitisation process, the impact starts to increase – more so in the digitisation process, since all the steps need to be stored; however, in this case, it is possible as the information is close at hand. The space requirements can also be seen from the data export packages, which we have made for the researchers [8], where the largest data export, for one particular year, was 19 gigabytes. In future, when new enrichment methods are explored, this space requirement is one to consider, especially when defining the content that will be put into the official export packages from digitisation and determining where all the article-extraction or named-entity-recognition results should be placed.

3.3 How the Digitisation Results are Made Accessible for the End User

The digital presentation system at http://digi.nationallibrary.fi can be viewed by anyone in the world, and anyone can use the publicly available data. The ALTO XML and OCR texts are visible alongside the page image as well as in the open data export packages. NLF has created a custom XML format, which incorporates all of the key parts of digitisation, the metadata of the binding, the ALTO XML data and the recognised text by the OCR.

After the new material from 1911-1920 was opened in the beginning of year 2017, it was well received by the generic public, based both on the statistics and user feedback that we have received. The general public usage has increased based on the page view metrics, and there has been more use of the data itself.

3.4 Brief Look at the Nordic Situation in Digitisation or Presentation Systems

The situation of newspaper digitisation in the Nordic countries is described in Table 1. The progress in digitisation unfolds in each country according to priorities and with the resources available. For example, in Norway in 2006, the decision was made to digitise all of the collections of the National Library of Norway [9], where it had made an estimate in 2010 that all collections could appear in digital format by 2030, depending on funding, technological developments and collaborations [9]. The rapidly increasing data amounts via digitisation seem to be a trend in most Nordic countries recently.

In addition to Norway, Denmark has also progressed fast, especially in newspaper digitisation. There, the goal was set to digitise 32 million pages of the state newspaper collection with aim to increase the use of the collection [10].

Table 1. Digitised newspaper collection size estimates in the Nordic countries

Region	Digitised Pages (estimate, million pages)	Material Type	Reference
Denmark	33,3	Newspapers	[11]
Finland	6,0	Newspapers	[12]
Iceland	5,4	Newspapers, periodicals	[13]
Norway	22	Newspapers	[14]
Sweden	17	Newspapers	[15]

One interesting case is also the Faroe Islands (a self-governed country within the Danish realm), which has stated the goal of digitising all Faroese newspapers despite having “little to no funding” [16]. Digitisation can be resource intensive, and even if it is not, the other subsequently important process phases, such as cataloguing (the creation of initial metadata for the physical object) and creating the digital object metadata, also need to be taken into account. If some process steps should be automated, there is also additional research and development steps needed, when necessary methods are investigated, tools developed or evaluated to the particular needs. Scaling of the tools needs to be taken into consideration from the beginning, because they need to be run

both for the existing material, but also integrate to the existing workflows so that the all future materials get same benefits. Automation of various steps require then at least some initial effort.

4 In Conclusion

As noticed by [15], the digitised contents of a library create a new kind of environment in which existing data is used in different ways. This creates new challenges with regard to the copyrights, use, re-use and distribution of the materials. Despite these challenges, it is still something that national libraries will need or could choose to tackle, since they have opportunities to work as agents from the library sector towards copyright organisations and rights holders. It could be possible for the research and public libraries, with the galleries, archives and museums can join forces to impact legislation development. With new needs we can strive to find new collaborations, and to develop methods and tools for new kinds of environments and use.

Much can be done with limited resources, and with proper funding models, it is possible to do even more. The national goals have an impact towards which digitisation efforts libraries prioritize. Clearly articulated vision statements boost efforts and motivate the library personnel and make efforts visible. It would be important to hear the voice of the research community about which materials are needed, even more as resources are limited. Researchers are also focusing on different research questions, where some researchers need large corpuses and some rare and unique materials.

The additional enrichment of the materials could also be done together with the researchers. As in some of our cases with e.g. named entity recognition [6] are showing promising advancements. In the long run, in addition to the digital chain we could also create a new "enrichment chain" to include additional automatic enrichments, which are made for easing the researcher use of the materials. Also the enrichment phase can benefit from the close proximity of digitisation and the presentation system, then any improvements made can be utilized in the digital presentation system, too.

Acknowledgements

Part of this work is funded by the Academy of Finland project COMHIS – Computational History and the Transformation of Public Discourse in Finland, 1640–1910, decision number 293341.

References

[1] H. Arpiainen, 'Digi.kansalliskirjasto.fi –palvelussa jo yli 12 miljoonaa sivua aineistoja', *Kansalliskirjasto*, 02-Oct-2017. [Online]. Available: https://www.kansalliskirjasto.fi/fi/uutiset/digikansalliskirjastofi-palvelussa-jo-yli-12-miljoonaa-sivua-aineistoja. [Accessed: 15-Oct-2017].

[2] National Library of Finland, 'The digitisation policy of the National Library of Finland', 2010.

[3] M. Piotrowski, ‘Natural Language Processing for Historical Texts’, *Synth. Lect. Hum. Lang. Technol.*, vol. 5, no. 2, pp. 1–157, Sep. 2012.

[4] National Digital Library, ‘Metadata Requirements and Preparing Content for Digital Preservation’, 2017. [Online]. Available: http://www.kdk.fi/index.php/en/digital-preservation/specifications/353-metadata-requirements-and-preparing-content-for-digital-preservation. [Accessed: 04-Jan-2018].

[5] M. Koistinen, K. Kettunen, and T. Pääkkönen, ‘Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing’, in *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa*, Gothenburg, Sweden, 2017, pp. 277–283.

[6] K. Kettunen, E. Mäkelä, T. Ruokolainen, J. Kuokkala, and L. Löfberg, ‘Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910’, *Digit. Humanit. Q.*, vol. 11, no. 3, 2017.

[7] T. Palfray, D. Hébert, S. Nicolas, P. Tranouez, and T. Paquet, ‘Logical segmentation for article extraction in digitized old newspapers’, *ArXiv12100999 Cs*, Oct. 2012.

[8] T. Pääkkönen, J. Kervinen, A. Nivala, K. Kettunen, and E. Mäkelä, ‘Exporting Finnish Digitized Historical Newspaper Contents for Offline Use’, *D-Lib Mag.*, vol. 22, no. 7/8, Jul. 2016.

[9] V. M. Skarstein, ‘The Bookshelf: digitisation and access to copyright items in Norway’, *Program*, vol. 44, no. 1, pp. 48–58, Feb. 2010.

[10] Royal Danish Library, ‘Digitisation of 32 million newspaper pages — Royal Danish Library - english’, 2017. [Online]. Available: http://en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/newspaper-digitization. [Accessed: 17-Oct-2017].

[11] Royal Danish Library, ‘Mediestream - Newspaper’, 2017. [Online]. Available: http://www2.statsbiblioteket.dk/mediestream/avis. [Accessed: 17-Oct-2017].

[12] National Library of Finland, ‘Digitised pages online’, 04-Jan-2018. [Online]. Available: https://digi.kansalliskirjasto.fi/info?language=en. [Accessed: 04-Jan-2018].

[13] National and University Library of Iceland, ‘Timarit.is’, 2017. [Online]. Available: http://timarit.is/about_init.jsp?lang=en. [Accessed: 17-Oct-2017].

[14] R. Jøsevold, ‘Digitization of copyright protected newspapers in Norway’, presented at the LIBER 2015, 2015.

[15] KB.se, ‘Digitaliserade svenska dagstidningar 1758-1926 (Sida 1) — Av KB digitaliserade dagstidningar, andra tjänster — Tyck till om visning av digitala dagstidningar’, 2017. [Online]. Available: http://feedback.tidningar.kb.se/viewtopic.php?id=89. [Accessed: 17-Oct-2017].

[16] E. E. H. Kjærbo, ‘Digitization of the Faroese National Heritage’, presented at the 2016 IFLA International News Media Conference, 20-Apr-2016.

[17] B. Valtysson, ‘From policy to platform: the digitization of Danish Cultural Heritage’, *Int. J. Cult. Policy*, vol. 23, no. 5, pp. 545–561, Sep. 2017.