Digitisation and Digital Library Presentation System – Sheet Music to the Mix

Tuula Pääkkönen
National Library of Finland
University of Helsinki
Mikkeli, Finland
tuula.paakkonen@helsinki.fi

Jukka Kervinen
National Library of Finland
University of Helsinki
Mikkeli, Finland
jukka.kervinen@helsinki.fi

Kimmo Kettunen
National Library of Finland
University of Helsinki
Mikkeli, Finland
kimmo.kettunen@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 priorisation is done annually. 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 enables fast operation by being close by to the digitisation process to enable a streamlined flow of material from production 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 present the enrichment phase of digitisation, which could be applicable with processing of musical sheets.**

***Index Terms*—digitisation process, digital presentation system, research and development projects, digital chain**

I. 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 recently exceeded 13 million pages [1], which can be viewed at the digi.nationallibrary.fi In addition, the doria.fi service contains books, maps and images and audio.

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 describes the tasks of digitisation process.

Part of this work is funded by the Academy of Finland project COMHIS Computational History and the Transformation of Public Discourse in Finland, 16401910, decision number 293341.

II. 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 brought to the researchers desks.

A. The Renewed Digital Chain

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 Fig. 1.

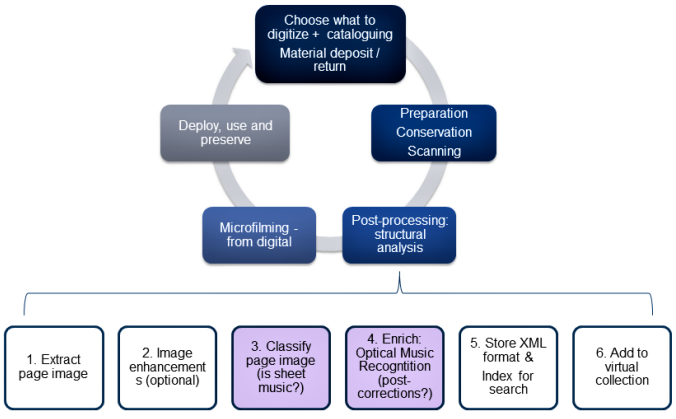

Fig. 1. The digital chain with enrichment tasks of NLF.

Processing of sheet music is a case example of the enrichment phase, which is starting to form within the post-processing phase. After we get the page image, we can classify it and based on that result take specific enrichment path. In case of sheet music, doing optical music recognition (OMR) with external tools and storing its results in standardized format. Finally indexing and automatically adding material to correct virtual collection enables end-users to find material.

B. Increasing Amount of Material Types

The realm of publication is constantly changing, which requires changes also to the digital library presentation system, by providing support to new material types. Currently we have a project for adding support to monographs (books, catalogues, and possibly maps) to the digital presentation system and music sheets could be possible next step.

There is increasing interest towards various music materials which are digitized within the NLF, so highlighting them at the same presentation system would ease processing requirements. This could include processing of the notes and scores, utilizing their metadata and processing them in a user-friendly way, i.e. by providing a separate search for musical notes. At the moment, the musical notes are recognized in the post-processing of the digitisation as illustrations. There has been initial experiments to classify illustrations via TensorFlow library [3], which in first step requires creating a training set of sheet music illustrations in order to find more of them. Now the search of the illustrations happens via the recognized text on the page. In future it would be useful to add tags to the illustrations and to get to more granular level of utilizing illustrations, and music sheets. Fig. 2 shows how the presentation system shows a page image with sheet of music on the right functions bar, then page image and on the left the recognized text of the page.

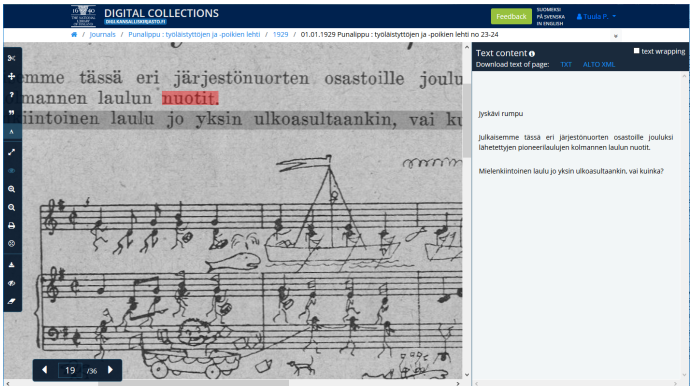

Fig. 2. Sheet of music in the presentation system. [4]

C. Planning with Resource Constraints in Mind

Adding musical sheets and notes provide intriguing challenges for the digitisation process and to the requirements of the presentation system. As with any material creating metadata of different music document types requires work with cataloguing. E.g. Bavarian state library lists monographs, musicological journals, sheet music, music manuscripts, sound carriers and libretti [5]. More technical issue is how should the post-processing of the digitisation identify and utilize the musical notes, which OMR-quality level would be acceptable? Do stakeholders want OMR? How results from OMR could be utilized in the presentation system? In addition, there are multiple formats, protocols and tools to choose from [6] where selection requires evaluation. For example, the International Image Interoperability Framework (IIIF) for illustrations could be useful for the music sheets images, too. The annotation support of the IIIF standard could enable crowdsourced fixes to the most complicated OMR cases, like we have previously already seen in the text recognition research [7]. Most effort will be put to the configuration of digitisation process to be able to process different kinds of sheet music documents. For manual annotation or quality control we could only put a few days of work. Therefore, in the enrichment phase existing open-source tools could be invaluable when they could be plugged-in to the digitisation process with minimum changes.

III. Conclusions

The digitised collections of a library create a new kind of environment in which existing data are used in different ways [8]. This creates new challenges with regard to the copyrights, use, re-use and availability of the materials. The additional enrichment of the materials could also be done together with the researchers. In that case the library would act as content provider but also utilize the research in the presentation system. For example, in Denmark, sheet music is in the top-5 search words [9], so improving access could hopefully increase awareness of that collection. Any OMR technology chosen should fit to the existing digital chain, and provide standardized format for preservation and further use.

In the long run, in addition to the digital chain, a new enrichment chain can be created to include additional automatic enrichments, which aim to ease the researcher use of the materials. Sheet music could act as one more material type, which could invite more users to the collections.

References

[1] H. Arpiainen, "Digiss jo yli 13 miljoonaa sivua aineistoja," Mar 2018, (in Finnish). [Online]. Available: https://tinyurl.com/digi1813

[2] N. L. of Finland, "The digitisation policy of the national library of finland," 2010. [Online]. Available: http://urn.fi/URN:NBN:fi-fe201401151119

[3] TensorFlow, "Tensorflow for poets." [Online]. Available: https://tinyurl.com/digicl

[4] *Punalippu: tylistyttjen ja -poikien lehti*, no. 2324, p. 19, Jan 1929.

[5] J. Diet, "Digitization and presentation of music documents in the bavarian state library," *Fontes Artis Musicae*, vol. 61, no. 3, p. 275283, Jul 2014.

[6] R. Freedman, R. Viglianti, and A. Crandell, "The collaborative musical text," *Music Reference Services Quarterly*, vol. 20, no. 34, p. 151167, Oct 2017.

[7] K. Kettunen and T. Pääkkönen, "Measuring lexical quality of a historical finnish newspaper collection analysis of garbled ocr data with basic language technology tools and means," May 2016, p. 956961. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2016/index.html

[8] KB.se, "Digitaliserade svenska dagstidningar 1758-1926 (sida 1) av kb digitaliserade dagstidningar, andra tjnster," 2017, (in Swedish). [Online]. Available: http://feedback.tidningar.kb.se/viewtopic.php?id=89

[9] A. . Jensen, "Experiencing materials and knowledge: The danish national librarys music collection," *Fontes2011*, vol. 58, no. 3, p. 253259, Jul 2011.