https://doi.org/10.2352/issn.2168-3204.2024.21.1.10

How to (mass-)digitize newspapers in Switzerland – The Swiss National Library approach - renewed

Martina Hoffmann, Swiss National Library, Bern, Switzerland

Abstract

The Swiss National Library (SNL) operates a variety of different digitization projects for different kinds of materials. The biggest part of the pie contains newspapers. Newspapers are of high interest to the public and to researchers in the Digital Humanities field. The effort must be made to put as much as possible newspapers online. The SNL as per her strategy is to take a leading role in this digitization effort. This paper will describe how the newspapers are digitized, what the SNL's role is and how the pipeline is structured from original to online presentation.

Background

Switzerland is comprised of twenty-six cantons, each with their own state government, which each have their own cantonal library. The Swiss national library is the countries governmental library.
It has the duty to collect, preserve and convey Swiss cultural heritage called Helvetica. To reach people with the collection it has a reading room, which can be visited during opening hours and with a library card people can borrow books to read. Like every other library the Swiss National library holds in its underground depots a vast number of items that cannot be displayed or borrowed at once. Also, in times where research is more and more taking place on digital resources like in the field of Digital Humanities, the library has to fulfill a need in in the digital presentation too. Other factors like the past lockdowns during the pandemic have shown that the need for digital data is on the rise and we all live in a global environment where customers can come from any place in the world and are unable to visit the SNL in person to research.

The Digitization Service

The Digitization service in the SNL is a small team of people who are responsible for the digitization of everything apart from audio and A/V material which means all cultural heritage on paper like books, newspapers, journals but also handwritten manuscripts, prints, billboards, and all other items that are part of our collection. It also researches the future digitization of 3D objects and its possibilities to incorporate this possibly in the library's digitization services. The SNL has its own service for A/V material which is in Lugano where the experts for audiovisual materials take care of those items. For classic reproduction and print publishing materials for marketing purposes the SNL has a separate service making the digitization service the sole responsible service for mass digitization efforts and standardization of digitization and preservation of digitized materials.

The digitization service of the SNL works on 40 to 50 different projects at any given time and does this with external partners and vendors as well as internal partners and internal in-house digitization. Internally projects with unique items that cannot leave the premises are carried out while other projects can be done with external service suppliers. The conservation service has an important role in that decision and can exclude items from being sent out to external suppliers.

Newspaper Digitization

Switzerland has a huge number of newspapers being published every day and because Switzerland has four official languages, the newspapers also appear in those four languages. This makes the landscape of Swiss newspapers a well of research opportunities.
Patrons all over the world use Swiss newspapers to research their own genealogy or topics of interest. Most newspapers are presented on the platform e-newspaperarchives.ch [1]. Not all cantons are represented on this platform and some cantons have their own solutions to present newspapers to the public. Nevertheless, the platform is steadily growing and has ongoing projects also in cantons that are at this moment not yet represented on the platform.

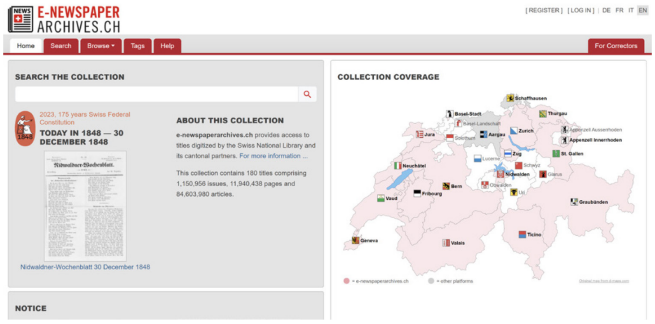

Illustration 1: e-newspaperarchives.ch landing page

At this point the platform contains 180 different titles with 11.9 million pages. The platform is maintained by the SNL with external service providers. The upload of final data is a responsibility of the digitization service of the SNL.

Partnerships

All newspaper digitization is carried out as partnership projects between the SNL, the cantonal library of the canton the newspaper is published in and the newspaper publisher. Each partner has its own roles and duties within such a project. Those projects are so-called PPPs: Private-Public-Partnerships.

For example, the cantonal (or university) library is responsible for the planning of the project which includes contracting and finances as well as promoting. The publisher is a most crucial partner in all those projects where the newspaper is not in public domain. Without permission of the publisher, we are not able to publish any data online. We could of course digitize it and store it somewhere but the amount of work to be done and having such a huge amount of material in our depots already means we must prioritize what the resources will be spent on.

Like any other large volume institution, the SNL has only limited number of staff and finances available each year and therefore the policy for digitization of newspapers is clearly that those, where we are able to present them online, are prioritized.
Once all the legal challenges are cleared the preparation work must begin.

Policy and changes

The SNL has been digitizing newspapers with its partners for several years before making a significant change in policy around 2018. Before that newspaper digitization was intended for online presentation only and the focus of the digitization was on the text presentation. Long term preservation factors for digital images did not play a role in the newspaper digitization projects for the SNL. Cantonal partners were responsible for the storage of the data and had to maintain their own protocols for that.

Classic digitization means transferring the information from an analog source to a digital source. In digitization of cultural heritage like newspaper it was often thought that only the text is a valuable information in a newspaper and therefore the requirements regarding this digitization were not that high. If one could read the text, it was fine. The Optical Character Recognition (OCR) had to be able to read the characters.
This resulted in a huge number of projects and title to be presented online with images that were altered to enhance the textual content without the use of technical specifications as we do know them now in digitization of cultural heritage. The images were highly manipulated to make more characters visible and readable in a time where OCR-software was not as far advanced as it is today. Because of the focus on text, grayscale images were preferred, and the crop as made to show text only.

Nowadays it is clear that researchers like to have more than just text on a manipulated page (cut to the text, enhanced grayscale image, small and ready for web viewing). They are keen on provenance, and this is way more difficult to prove without the look of the original when dealing with only digitized data.

The SNL had to rethink its strategy for newspaper digitization – as mentioned earlier – and decided to look for the implementation of best practices.

In 2017 the new ISO Standard 19264-1 (newest version now 2021-06) [2] emerged and the SNL was searching for a way to make her digitization output long term preservation ready.
After carefully reviewing the options with international best practices the SNL decided to change the workflows of image creation to ISO-Standard 19264-1 for all its digitization including newspaper digitization. This resulted in a huge effort to change the in-house workflows as well as the specifications for all projects to provide high end images of newspapers.

The SNL has adopted the principle of “Do it once, do it good” and has made improvements to her workflows over the whole digitization service. Having the most expertise in house the SNL is responsible for providing the specifications for digitization processes. This also effected the Swiss service suppliers as well as the partners in the newspaper digitization.

Switzerland has not adopted international best practices before in a systematic way and having to deal with new specifications and workflows were a challenge for all involved parties. The SNL however is convinced that the quality of the digitization is a major factor in the presentation of online materials and having long-term preservation ready images created in a standardized and controlled manner makes preservation on a large scale possible. The challenges with digital preservation are not to be underestimated and given the large volume of data that is being created automation has to be key for the future. The SNL also wants to provide the best possible data to the public and researchers in order to make the information accessible but also to prevent the originals from being pulled from their storages too often. Therefore, the “Do it once, do it good” principle is a starting point for all project ideas and carried out projects. The SNL was aware of the fact that this presents a change as well as an impact to the partners and service suppliers. With several measures the SNL seeks to help partners and service providers to make the necessary changes in their workflows as well.

To make the newspaper digitization ready for long term preservation more information about the collection or title has to be gathered than before. For high end digitization an assessment of the original is also necessary in order to create the best possible digital images in a standardized way. Therefore, the SNL had not only to change the specifications for the creation of digital images but also the level of information about the originals had to be increased. The SNL implemented an issue level inventory for newspaper digitization that allows to have control of the transport of originals to and from the service suppliers as well as assist in the quality control in several steps.

The way of the newspaper

Cantonal libraries hold a lot of newspapers in their facilities. As they are the regionals first supplier for the public, they often have very good stock of the specific newspaper of interest. As mentioned before, they also have contacts with the publishers of those

newspapers and are the first in line to clear the right to digitize and publish online – if the newspaper is not in the public domain. The SNL aides with the steps of rights clearance. For mass digitization projects those originals are being sent to external service providers which will digitize the newspapers creating digital images.

The SNL in her new workflows and attempts to make the most of the valuable resources available is requiring an inventory of each issue of a planned newspaper project. This inventory has to be made by looking at the originals and marking the date, page number, other remarks and things that can prevent digitization (like missing pages, torn pages, missing issues, text loss due to binding, etc.). When cantonal partners discover that parts of their collection are not suitable for digitization or missing, the inventory serves as a request to the SNL to investigate her stock and possible match the collection of the partner with missing issues or exchange issues with better condition. The SNL makes of her original the same inventory and will provide a final inventory of all original items when the delivery to the service provider starts. In newspaper digitization we are lucky that more institutions have copies of the same newspapers at hand, so it makes sense in a digitization effort like this to search for the most complete set of originals up front. Doing this after the fact when the collection is already presented online and in the long-term preservation storage is a huge effort, time consuming and at high staff expense.

This takes a lot of time and effort, but the inventories will then serve multiple purposes. First, one needs to know what will be sent out and what should be expected back digitally. The inventory serves the external service suppliers as a document for checking their stock as well as the expected number of digital files and, they can look up file naming in an automated way as well as other remarks that concern the originals (like torn pages where it serves as prove that the damage was there before handling by the operator for example).

Another use of the inventory comes into play at the quality control stage in the SNL where all the digital data will be delivered for quality checks. The SNL is using the inventories to check file naming and remarks as well. At the end of a project the SNL should hold a certain number of digital files that correspond to the number of pages in the inventory.

The inventory will also accompany the collection in the long-term preservation storage where it will serve as information for future inquiries about the state of the collection at the point of digitization for example and it will also be used to easily access information about certain issues by our staff in case of user requests.
The inventory – although it takes a lot of time to create – therefore serves so many purposes that it was deemed worthwhile to make for each newspaper collection that is being digitized from 2021 onwards.

Another big part that has changed is that the requirements for digitization have changed for all internal and external digitization for the SNL. The SNL has adopted the ISO-Standard 19264-1 as its quality standard for digitization.
The requirements for newspaper digitization changed from a web ready text-based image to a long-term preservation master that adheres to ISO Standard 19264-1 and shows the complete original as it would be seen in a reading room: in color, with a background around the original, etc. Swiss service suppliers were not ready for this kind of digitization in the early days of the adoption and the SNL has put in a lot of effort and continues to do so to help service suppliers change their production to a controllable standardized workflow adhering to ISO-Standard 19264-1. The SNL finds it tremendously important to educate and advocate for the ISO-Standard 19264-1 and is committed to helping improve the quality of digitization. The SNL has taken several steps to support external service suppliers to change their workflows to adhere to the requirements.

One of the common arguments was, of course, that it was never done before but international examples have been used to show that change is possible and could possibly benefit the service suppliers as the process does not require enhancements on digital images and the inventories can be used to automate internal processes and QA as well. Another argument of course is the cost of targets and software. The SNL has made the decision to always provide the service suppliers with targets in sizes matching with their equipment used for newspaper digitization. The SNL has decided to use UTT (Universal test target) [3] mounted and measured in their newspaper digitization projects and recognizes that those have a certain monetary impact on an operation at a service supplier. Therefore, the SNL has bought and keeps a stock of various UTT targets and ships them to each service supplier for each project. If a service provider resides outside of Switzerland, the SNL is prepared to ship the targets also abroad.
Together with explanations of the workflows and changes to service suppliers as requested by them, the SNL also seeks to work with them to ensure that their production can adhere to the specifications of a newspaper digitization project. This includes testing their machines and results on UTT Targets upfront of projects as well as their output images on file naming conventions or header information. The SNL carries out pilot and test batches with service suppliers that seek to be part of the newspaper digitization process with the SNL. The SNL is also regularly scheduled to visit service suppliers at their facilities to give advice about the operations/test results or to exchange on new processes and automation ideas. Part of the visits are also about the storage facilities for the originals to keep the originals safe at any time.

The SNL also provides and keeps a stock of external 4TB hard drives to be filled with the production and sent to the SNL. This way the SNL ensures a fast and steady data delivery with service providers. Even if the internet has data transfer protocols the sheer amount of data that has to be transferred is enormous and therefore slow. If like in the SNL there are also security protocols in place which prevent traffic from various protocols, data transfer becomes a challenge and therefore the SNL has decided to use external hard drives as means of transportation. Once inside the building and secured the data will be transferred internally to servers for further use. Especially within Switzerland the postal service is extremely well organized and fast beating transfer of large amounts of data by internet by a lot.

Giving service suppliers the means to enhance their production does not mean that the responsibility of producing quality and also doing the internal quality control can be transferred to the SNL. The SNL expects her service suppliers to have suitable internal processes for this end.

The SNL expects the deliveries of pilot batches of projects or production batches to be adhering to all specifications of a project. ISO Standard processes with UTT targets being only one part of those specifications. The SNL specifies also file naming, header tags, crop, and other things in her project specifications. All items will be quality controlled by the SNL herself in and a final report will be sent to the service supplier as well as the cantonal partner involved. If the batches are accepted, the process will go further to the segmentation phase. If a batch is rejected, rework and repairs have to be done by the service supplier.

Only after successful delivery of the images, the originals will be returned to the cantonal library resp. the SNL or other partners.
After the digitization phase the images are long-term preservation master files that will resemble the original in the best possible way.

Having created only archival images that are not searchable on the internet, the SNL now proceeds with a further step in the newspaper digitization where external service suppliers do perform OCR and segmentation work on the images. In this step the digital images will be sent to service suppliers that will create a METS/ALTO structure of each page of the newspaper issue and perform OCR on it to make full text search possible.

The segmentation phase will provide presentation packages that can be used to be presented at the e-newspaperarchives.ch platform.

As with the earlier steps in those projects the SNL works closely with the cantonal libraries who officially have the role as project leader and contracting authority.
The SNL again writes for each project the technical requirements and specifications and takes care of sending the images created in the former phase to the service suppliers for segmentation. The SNL also does the quality assurance on this process step before a title will be made publicly available through the platform.

Leaning on METS/ALTO standards that are internationally used for structures of textual information and a METS-schema suitable for newspapers the SNL creates several derivatives for presentation.
Each issue will get a JP2000 file per page as well as a PDF per page including the OCR. One ALTO XML file per page and one METS XML per issue as well as one multipage PDF per issue including the OCR.

After successful delivery the complete newspaper title will be uploaded to the presentation platform and the long-term preservation masters will be directed to the long-term repository of the SNL. The cantonal partners don't store the data anymore in their own repositories as the SNL will keep the data. This also makes a huge impact for cantonal libraries especially for smaller institutions that either don't have their own repositories or that have less budget to keep all this data.

Challenges

The change to this workflow came not easy and not fast. It has taken several years in some cases to change old habits and old ways of creating or storing data.
The SNL has always done a lot of partnership projects and changing the requirements incorporated the need of a lot of explanation why this was a necessary and right decision.
The SNL is now proud of the results of the first digitization projects that produced images according to this new workflow and the adaption of partners and service suppliers alike.

While on the image producing side the main challenges where that the equipment had to perform to quality levels never asked before as well as having an internal quality assurance program set up – which of course was an investment for the service suppliers - the cantonal partners needed to see the differences in quality of information in a new way of digitization. Both topics were addressed by the SNL in several presentations to each group. The service suppliers were offered workshops and help with feedback rounds in order to discover the flaws in their production workflow. While this might seem like an undoable task, the SNL has found it profoundly useful to establish good working relationships with the service suppliers and enhancing the overall quality of the production.

For the segmentation service suppliers, the input images made a bigger impact as they were used to the same output files. However, the images that were now provided resemble the original even if it was cut, torn, unbound, wavey or showing another page. The images were not enhanced for OCR recognition anymore and no contrast was altered. The images showed the originals in color as if they would be in the reading room. It took several rounds of conversations to get also the same look out of the derivatives.

The cantonal partners had also a very impactful change to make in this workflow. While the SNL takes over the costs for storage, it is necessary to provide more data with a long-term repository than merely for presentation purposes. Also, the SNL and her partners would like to upscale the newspaper digitization which means the whole workflow has to be standardized. That also means that the whole process chain has to be more controllable and that starts with transports of originals and even before that with the selection of originals. The work for inventories is a tremendous effort undertaken but the benefits no doubt outweigh the work. Cantonal partners have now changed to making issue level inventories and therefore also different time frames for digitization projects. This step was a major improvement for the rest of the process chain because the inventories now give information to all the succeeding steps and aid in the location tracking, quality control and automation of processes at the service suppliers as well as the SNL.

Lessons learned

Changing existing programs takes time and effort as well as lots of talking.

That might come as no surprise but creating a completely new workflow with new specifications for a country wide digitization effort multiplies the challenges. While most of the time one creates and changes programs for one institution, the SNL newspaper digitization has over 50 partners to consider. Given the four official languages of Switzerland adds to the challenge if most of the technical requirements that are referred to are in English. It added another layer of complexity but the Swiss translation skills are fantastic.

Also, Swiss cantons have all their own governmental state rules and ideas. Some may be able to contract out a project anywhere, some can only contract local service suppliers. Some may have other rules for projects than others. Furthermore, the cantons are completely different in all possible ways and also their relation to the federal government is not always the same. Given the complexity of the landscape and the impact of the changes, it has been a huge success to have been able to talk to all partners and service suppliers in order to explain the ideas behind the change and what would be the benefit for each party involved.

The SNL did travel around the country giving workshops to institutions, service suppliers as well as making online video meetings and workshops available for interested parties. Having done this in the past pandemical period between end of 2020 to 2022 added to the challenge.

Organization is a big factor in undertaking such an effort. Making time for each party and listening to their arguments made a huge impact on our buy-in from the partners.
It is also wise to have good examples from other institutions or countries where similar products are being presented online or similar technical standards are being used.

Last but not least: Providing aid throughout the process like the support with own UTT Targets, providing feedback continuously to service suppliers and making yourself available to questions before product delivery has helped our process to smoothen a lot and improve the overall quality and time to process large amounts of newspaper pages.

Conclusion

The Swiss National Library together with its partners has changed its quality level for newspaper digitization from web presentation only to long-term preservation ready. This is a continuous process as more and more partners are starting new projects and more service providers are looking to change their workflows as well. The SNL will continue to support her partners in those efforts and is convinced that with this huge effort the quality of digitization of Swiss newspapers improves a lot for researchers and customers. The quantity of digitized newspapers will continue to grow in the coming years, making more information accessible to the public. The quality of cultural heritage digitization in Switzerland has improved by this effort alongside as more and more people continued to learn about standardization and provenance as well as how to present newspapers with more information alongside the text and OCR. With this the SNL carries out one of her important tasks of making its collection available.

References

[1] www.e-newspaperarchives.ch
[2] https://www.iso.org/standard/79172.html
[3] https://www.image-engineering.de/products/charts/all/579-te262

Author Biography

Martina Hoffmann is the head of digitization services at the Swiss National Library and works internationally as independent consultant in Cultural Heritage digitization and QA workflows. She was Senior Production Manager digitization at the National Library for the archival section of Metamorfoze and operational manager QA of digitized products in the National Archives in the Netherlands. She designs and operates workflows from original to long term preservation for many different cultural heritage materials.