News Media Section

Lessons learned in 7 years of newspaper retro-digitization in Switzerland: How to learn lessons from mistakes, how to establish models and guidelines to avoid repeating them, how to share the lessons learned

Florian Steffen
Digitization Department, Swiss National Library, Bern, Switzerland.
florian.steffen@nb.admin.ch

Abstract:

Over the past 7 years, The Swiss National Library (NL) and its partners have digitized millions of historical newspaper pages. The main goal of these digitization projects has evolved: from making the papers accessible and searchable (full text) - at least in Swiss libraries - to providing online access all over the world, which should as far as possible be free of charge. Over the years we have been successful in our goals, but have also made mistakes along the way and we had to find ways to assure that these would not be repeated, but turned into lessons learned for future projects.
One important way to achieve this goal is to review our approach in view of the results achieved, document any errors (as well as successful approaches) and use them as the basis for the creation of guidelines. The NL and its partners have established guidelines for newspaper digitization, which aim to be simple to understand and apply and which cover all kinds and every phase of projects whether tasks are outsourced or everything is done in-house. These were based also on specialized literature, existing guidelines as well as experiences made by partners. As a next step, these guidelines have to be translated into workflows and basic models (contracts, inventories etc.) and implemented in our everyday work.
This is an iterative process and in every project, new knowledge is gained which can be used to improve existing guidelines, workflows and models. If one's goal is to improve accessibility and digitization quality at a national or regional level it is essential to share: the successes, the mistakes, the guidelines and the models. New techniques will also influence future practice.

Keywords: Newspaper, Library, Partnership, Access, Lessons Learned.

1 Introduction

In 2008, the NL started to digitize newspapers. In a public-private partnership project with the newspaper publisher Le Temps and the Bibliothèque de Genève, which was the first of its kind in Switzerland, the newspapers Le Journal de Genève, La Gazette de Lausanne and Le Nouveau Quotidien, representing 200 years of newspaper publication, were digitized.

Since then, many more titles have been digitized in a variety of projects. Over the past seven years, around 6.6 million pages have been made publicly available on the platform *Swiss Press Online*[1]. Full-text searching is available for millions of articles, giving a new lease of life to newspapers that would otherwise be forgotten. Given these positive results, it's worth looking at some of the errors made along the way. And there were quite a few over the years. There were stumbling blocks in all areas of newspaper digitization, some of relatively little consequence, others more serious. It is quite probable that organizations which had started digitization programs before the NL had a similar experience. And one can imagine that others starting newspaper digitization programs, now or in the next few years, will face the same problems. So, this is an incentive to provide some examples to talk about our experiences, both positive and negative. Our goal is to encourage newspaper digitization and help prevent costly, time-consuming and, above all, avoidable problems.
It is important to note the context in which the digitization activities described below take place. Resources for digitization projects are limited, when taking into consideration the complexity of projects and the high expectations of the public and the policy makers. The NL's digitization service has 7 staff members (5 FTE) charged with carrying out 15-20 projects simultaneously of differing scope, using different parts (and formats) of the collection. Newspaper projects are usually carried out within partnerships, occasionally with only one partner but more often with three or four, sometimes even up to six, not counting sponsors. The projects are frequently carried out under the pressure of deadlines such as a commemoration or an exhibition. So there is a lot of pressure to achieve results within a tight schedule.

2 What are the most common mistakes? Where do they arise?

During every digitization project at the NL, insights have been gained that can be included when planning subsequent projects. These may lead to changes in procedures, new definitions of technical parameters, adjustments to partnership models and communication structures, or to re-evaluation of partners. Decisions that were taken on a theoretical basis and which have proved unsuccessful or inapplicable in an operational environment may be re-visited. These include areas such as technical details (e.g. image format), procedures (e.g. the choice of in-house or outsourcing procedures), or conservation measures (to protect the originals). Additionally, technical and professional progress, changes in the institution's own staff or infrastructure and strategy change can lead to the reassessment of hitherto proven processes. In such cases, one cannot speak necessarily about making mistakes when the review of a project shows that other approaches might have had brought better results.

In this papers, we are only speaking of errors when the problems that occurred in projects were the result of ignoring or insufficiently considering evidence, when decisions were not

[1] http://www.swisspressarchives.ch

enforced consistently, when processes were not carried out correctly or existing tools were not installed correctly.

The main stages that each NL project follows are: preparation, digitization (scanning, quality control, data delivery), launch and marketing. This paper presents possible errors in the first two phases. The NL is currently still establishing better guidelines and developing tools for phases 3 and 4.

3. Lesson 1 : Preparation is everything

3.1. From the choice of title to the conclusion of cooperation agreements

The preparatory phase sets the course for the entire digitization project. Any errors or omissions here have an impact on the entire project. The preparation phase contains the following work packages:

- The exact definition and preparation of the collection to be digitized
- Definition of the digitization parameters
- Synopsis of the collection description and digitization parameters in specifications for the tender
- Clarification of the rights pertaining to the collection
- Partnership models and project organization

3.2. Insufficient knowledge of the physical collection

It is tempting to think that newspaper collections are similar enough that a detailed analysis of each title before digitization is unnecessary. This misconception can lead to significant problems. Newspapers differ in many ways, within the same title over the years, both in paper and print quality and are a challenge to digitize, as the NL has confirmed over the course of projects in which an analysis that was not detailed enough led to delays and increased costs.
At the beginning of this learning curve, in preparation for the call for tender of the first newspaper digitization projects, the NL's digitization service carried out a rough inventory of a newspaper collection, providing the following data for each call number:

- Title, subtitle
- Publication period
- Call number
- Number of physical volumes
- Format
- Number of linear meters

Based on the measured linear meters an estimate of the number of pages was created. It was calculated at 200 pages per cm. This collection data was part of the calls for tender. The digitization companies who received the calls had the opportunity to consult the collections on site, though they were not obliged to do so. In one example, the company that won the contract took this opportunity and examined a random sample of some of the volumes to be digitized. On this basis, the company made assumptions about the cost of the project and submitted a quote.

At the start of the digitization process in this project, neither the NL nor the digitization company had in reality a detailed knowledge of the collection: the proportion of coloured

pages in the collection was unknown; as was the fact that, for some years a morning and an evening edition was issued, which only slightly differed from each other; issues were sometimes wrongly dated, had missing or incorrectly numbered pages; the NL's collections was incomplete (individual issues and sometimes whole weeks were missing).
In the course of the project the most devastating problem turned out to be that the physical state of the collection, in particular the binding of the volumes, had not been evaluated correctly.
If a detailed examination of each volume had been carried out, it would have been clear that unless the volumes were disbound there would be a significant loss of information during digitization for around 70% of the collection. The newspaper was printed right across the page with very narrow margins. The bindings often left such deep curves on the left and right sides that OCR would be very inaccurate.

When this was recognized, much of the collection had already been digitized, and since the automatic curve correction had only produced marginally better results, the volumes had to be disbound and re-digitized. Here, of course, there were several problems combined: not only was the earlier collection evaluation too superficial, but so was the quality control during the digitization process, as well as the service provider's communication since they did not flag the problems they encountered but continued with their work.

In addition, the digitization company had miscalculated the workload required for identifying the pages numbers and assigning files IDs. In the tender, the NL had stipulated that each page be identified as follows: Title_year_month_day_page.format.

The date of the issue was therefore a part of the image identifier. In this particular case, the service provider had developed a tool that used OCR to identify the issue date automatically and assign identifiers to each image. Then we realized, as indicated above, that the issue numbers were sometimes wrongly indicated, so the dates in the image identifiers were incorrect. This required significant manual input to identify and change the incorrectly labelled images.

Further corrections, which were due to the lack of knowledge of the collection, resulted in additional costs and complex post-processing. The project was delayed by a total of just over a year.

3.3. Lessons learned

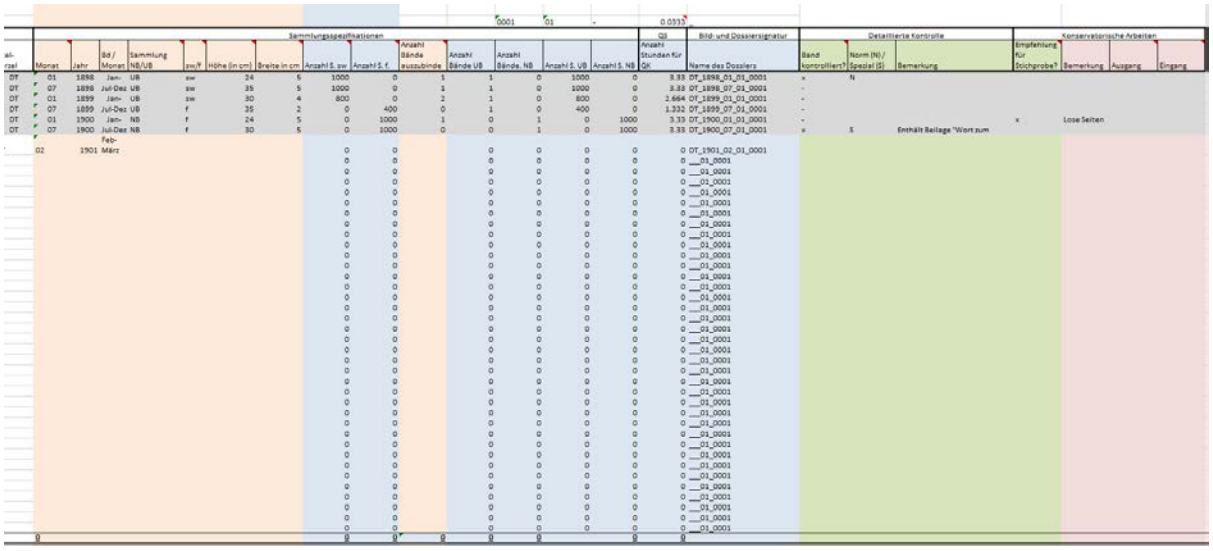
The issues described above underline how crucial detailed knowledge about the collections to be digitized is essential for good project planning[2]. This is especially (but not only) the case when a project is outsourced. External service providers require a reliable basis for their cost estimates, and the size of a collection in linear meters is not sufficient for the creation of binding offers. Ignorance about any attachments that may have been bound in with the volumes, or about gaps in publishing, the exact periodicity (which may have changed over time), or even about the scope and size of the collection can result in a big difference between the estimated and the actual project scope.

[2] See also *Planing Digitisations Projects – Best Practice Tips* (2014 – Point 3 *Assess the condition of your materials*): Online https://www.townswebarchiving.com/2014/09/planning-digitisation-projects-a-best-practice-tips/

Bibliographic metadata often provide only limited information about what the collection contains exactly. The physical collection must therefore be analyzed as closely as possible using a variety of parameters.

3.4. Solutions

Even under time pressure, a collection must be analyzed and described very thoroughly in order to prevent surprises during digitization projects. A model of an inventory has been developed that records different collection details up to the level of the physical volume, allowing the tenderer to determine which infrastructure is required and the level of difficulty involved when digitizing different volumes. In addition, the inventory shows exactly what preparatory work (especially on the part of the book binder) is required for a collection.



The inventory based on Excel is filled directly in the stacks. The collection is measured to estimate the number of pages. In addition, information about whether a volume should be scanned in grayscale or in colour is noted. Under certain circumstances, during the inventory, a sample size which defines the number of volumes to be examined in detail can be established. Within each sample volume all the pages are examined to see if they comply with the standard description or if there are any additional problems (such as morning and evening editions, Sunday papers, obvious errors in the issue number or pagination, etc.). In addition, recommendations can be made for volumes to be checked by the conservation department before and after digitization if required.

A very important piece of information, which is also indicated, is that if a volume needs to be disbound for optimal digitization. This guarantees an opening angle of 180 degrees so that each page can be completely scanned. Where this is not possible for conservation reasons, the volumes are treated separately.

This inventory accompanies the work throughout the process. After the collection of the basic data it is sent to the companies that perform the binding and the conservation and restoration examination. A copy of the basic inventory is sent to the digitization companies as part of the tender and then used to accompany the digitization process. After digitization, the comments of the digitization company are added to the inventory, and transferred to the company that

has carried out the conservation and restoration work. The conservation and restoration check and bookbinding work are once again documented in this inventory.

3.5. Conclusion - Outlook

This type of project preparation solves many, if not quite all the problems with which the NL was confronted within projects it has run in the last years. In any case, the re-digitization of a large part of the pages in one of the projects could have been avoided if the volumes that needed to be disbound had been identified during the preparation phase. All the other difficulties we encountered, especially with the quality of digitization, could be solved without re-digitization. To ensure that no uncertainties remain, it would be necessary to carry out a much more in-depth inventory looking at each page individually. Before spending huge human resources for such a complete inventory, it would certainly be worthwhile to evaluate carefully whether this effort does not exceed the cost of the subsequent correction of minor bugs.

4. Lesson 2 : Strengthening the contract during the test phase

4.1. From drawing up the specifications to accepting the product

Concerning the steps that are performed at this stage, there are no significant differences when the work is done in-house or externally. Since the NL outsources its newspaper digitization projects, the focus will be on that aspect.

The work packages in this phase are the following:

- Procurement
- Transport
- Test run / establishment of reference scans
- Quality control
- Acceptance of delivery
- Storage

4.2. Testing / Quality Control / Enforcement of digitization-schedule including partial deliveries

For outsourced projects, mistakes that occur in the digitization phase are the following: misunderstandings in the tender evaluation; hidden cost drivers or factors that can lead to delays; logistical weaknesses in the transport agreement and preparation; errors in scheduling and checking deadlines; lack of definition of digitization stages (deliveries in subsets); deficiencies in quality control; errors in data transfer (by the service provider to the contracting authority but also by the latter from the transport medium to the destination medium); and many more.

Here we will consider the consequences when unclear elements (which almost inevitably exist in a specification) are not identified and clarified in a test phase, and when quality shortcomings affect the whole collection because partial deliveries were not enforced. Experience has shown that uncertainties continue to exist even if a specification carefully and describes the expected results in detail.

Section 3 already explained that in some of our projects, the collections were described very superficially. This, in combination with a rather general description of the expected results

and sometimes significant deviations from these descriptions by the digitization providers, could result in significant delays and additional costs. In one specific project, after the digitization company had picked up the collection from the NL and prepared it in their studio, they immediately launched the operational digitization phase. A faulty collection description led to the defects described in section 3. Some elements of the tender were then misinterpreted by the digitization service provider or simply overlooked:

- Indicating whether pages should be scanned in colour or grey scale.
- Image format selection
- The main objective of the digitization process (quality enabling the best OCR possible)
- The requested post processing steps
- Delivery in at least 3 batches

The entire digitization process should have been carried out within 8 months. The delivery of the first lot was planned for three months after the start of digitization. The NL sent a reminder one month after this deadline (error: not checking project progress) and urged the supplier to deliver the first results. In an initial response, the supplier pointed out that, due to the wrongly dated issues in the collection, its image naming software had not delivered the desired results, and therefore the IDs had to be added later in the workflow. The supplier requested that only one delivery be made instead of the three previously agreed upon. Unfortunately, the NL agreed. Had the NL insisted on delivery in three batches, the project would have still been delayed, but it would have been possible to see that the image ID question was not the most serious problem in the scanned pages. It was only when the entire scan was completed and delivered to the NL, that the problem with legibility (and thus with OCR) of the too tightly bound pages was discovered. In addition the image format was incorrect as everything had been scanned in colour. Nor were all the pages post processed in the manner requested. As already indicated, the major problem here was the 70% bound pages as this required them to be disbound and digitized again. This re-digitization led to a 20% increase in costs and a shift in approximately six months compared to the original timetable. The cost increase was relatively moderate, as our contracts and specifications had stipulated that the digitized pages should be of a quality that allows "good OCR results". The new scans were carried out mainly at the expense of the digitization provider.

4.3. Lessons learned

In addition to the findings about documenting and preparing the collection, the NL also realized from this that even clearly formulated digitization parameters do not necessarily lead to the desired result. Experience has shown that there is always room for interpretation, misunderstanding or reasons for workflows to deviate from defined parameters. In addition, it was recognized that such differences must be identified at the earliest possible time and corrected in order to counteract significant cost increases and delays.

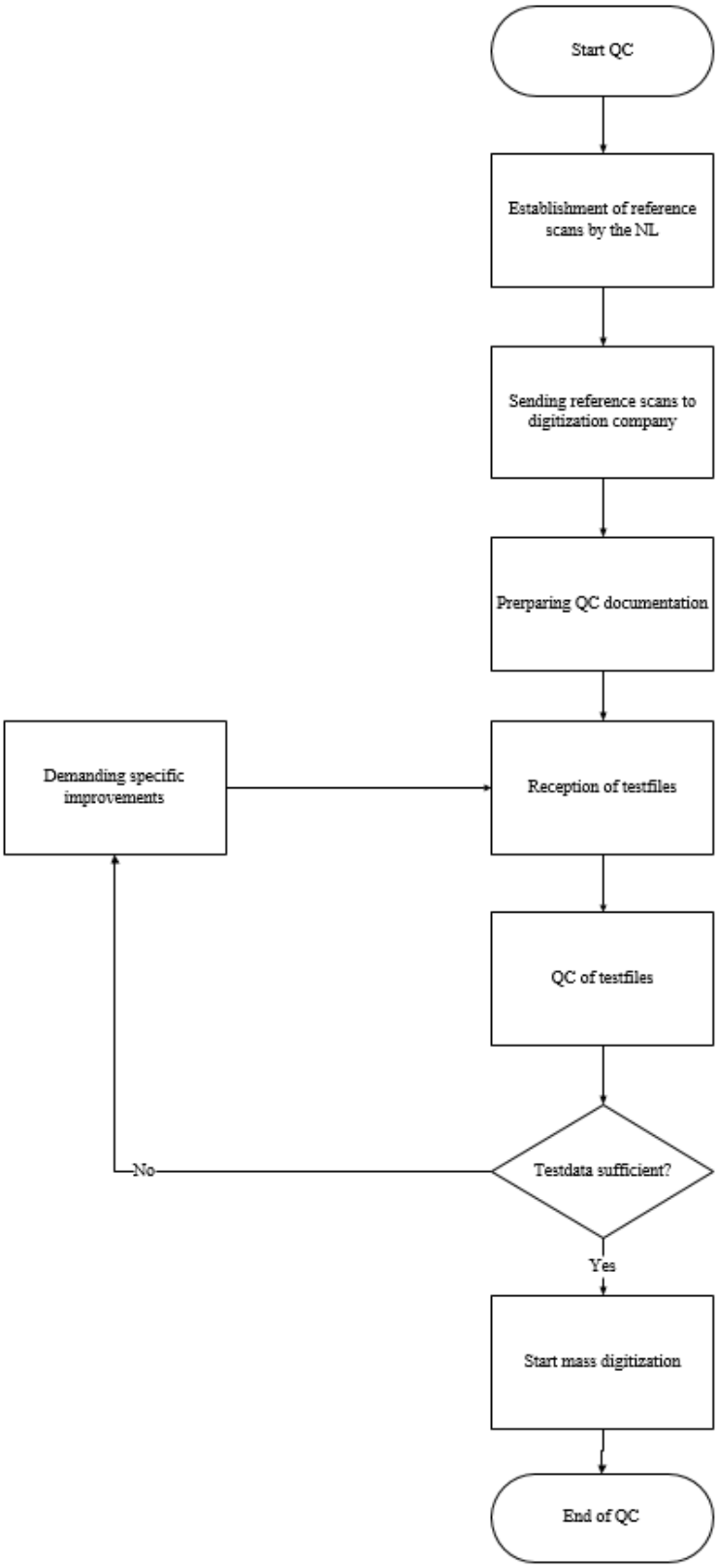
4.4. Solutions

Based on these findings, two measures have been introduced in the NL, which are described in detail in the tender and in the contract and which must be complied with throughout the project.

- The test phase with the objective to produce the requirements specification and reference scans, which form the basis of systematic quality control[3],
- Clearly defined partial deliveries, including delivery dates which are constantly monitored and immediately flagged if there are delays.

Ideally, the test procedure should begin with a discussion with the digitization service, enabling any ambiguity in the contract or specifications to be clarified. Within this discussion, the exact scope of the test delivery, including the publication years to be selected, are defined. The following process then starts:

[3] See also *Planing Digitisations Projects – Best Practice Tips* (2014 – Point 9 *Run a pilot project*): Online https://www.townswebarchiving.com/2014/09/planning-digitisation-projects-a-best-practice-tips/



The first step, the establishment of reference scans by the contracting authority is only carried out in the NL when working for the first time with a new service provider, or if a known service provider is going to digitize a specific type of collection for the first time. In the next steps, the subset defined for the test is processed by the service provider and delivered to the client, who checks it closely and indicates any requirements for improvement. The process is

repeated until the required quality is attained. When the required quality is not reached despite several test phases, the NL's contract stipulates that it is possible to stop work with the supplier at this stage. In most cases, the test set is then accepted and the resulting scans are defined as the reference scans. The scans from the operational phase are compared with the reference scans, giving both sides the most security.

The client can be sure that its requirements are precisely understood and will be implemented as desired; the digitization service can be assured that it will not need to repeat too many scans.
Nevertheless, the definition of deliveries and the monitoring of compliance with deadlines is essential. This measure ensures that the data can be checked in a timely manner. These two measures combined are perfectly suited to prevent large numbers of faulty scans being produced.

5. From lessons learned to guidelines

To ensure that errors will not be repeated, certain measures are taken. In its digitization projects the NL is constantly learning from its experiences, making any necessary changes and implementing them in its future projects[4]. It is important that projects, and especially lessons learned, be clearly documented. In a further step, the causes of the problems encountered need to be identified, and this information should be available to others. As an example, the causes of inaccurate OCR are not always clear. It may be the result of sub-optimum digitization or a bad choice of OCR software. When the cause (s) are identified they must be documented and the solutions noted in order to avoid repeating the mistakes. Outside opinions, guidelines and standards should be consulted and, if needed, interpreted in terms of institution standards, guidelines and policies. However, in many cases, such basic documents are not sufficient to avoid future similar problems. The reason for this is that this does not necessarily affect the operational workflows. So the people who are responsible for the effective monitoring and implementation of quality standards are not able to implement basic yet practical tools. The next step in this process is to build and test tools that ensure compliance with the defined quality standards in daily operations. Finally, it is important for an institution, which often works in partnership projects and which is not necessarily always the project leader to share knowledge and tools. Some practical examples follow.

5.1. Standards for newspaper digitization and digitization guidelines

The most important documents in the field of digitization for the NL are the standards for newspaper digitization[5] and the digitization guidelines[6].
The standards for newspaper digitization were drawn up in 2014 within a project of the Schweizerische Konferenz der Kantonsbibliotheken (SKKB) [Swiss Conference of Cantonal Libraries]. The NL was the project leader and had a strong influence on drafting of the standards. The draft was based on the experiences of the cantonal libraries involved in the area of newspaper digitization as well as third parties' standards, in particular the DFG Practical Guidelines on Digitisation of the German Research Society[7]. The standard is based

[4] See also the PDCA cycle: https://en.wikipedia.org/wiki/PDCA

[5] Standards Newspaper-digitization, 2013: online: https://www.digicoord.ch/images/0/08/Standards_Zeitungsdigitalisierung_DE_20140709.pdf

[6] Guidelines of digitization, 2014: online http://www.nb.admin.ch/themen/02074/02076/index.html?lang=de

[7] Deutsche Forschungsgesellschaft, 2013: „DFG Practical Guidelines Digitisation": online unter http://www.dfg.de/formulare/12_151/12_151_en.pdf

on the phases of digitization projects and proposals: "preparation of materials / Conservation examination", "Technical parameters of digital reproduction", "full text generation" and "quality control".

The digitization guidelines were drafted by in an NL internal working group and will be revised every 2 years. The guidelines determine the policies and activities of NL with regard to digitization. They answer the questions why, for whom, what, how is a collection digitized and which organizational units of NL are involved and to what extent. The "how" in particular is described in detail: here guidelines differ from the newspaper digitization standards, as the latter provide detail on the form of the collection. The guidelines cover not only technical aspects but also issues such as rights, partnerships, financing, access and promotion.

5.2. How to make sure that guidelines find their way to everyday work

As mentioned above, it is important that the standards and guidelines find their way into operational work. To this end, all steps that are carried out during a project are meticulously documented in workflows, in close cooperation with the services involved. In a next step, the necessary working instruments and tools are created. These are often modified and refined during the project. Each change is checked to see if it deviates from the above guidelines and standards. If such is the case, the change is evaluated to judge if it's valid or not. If it is not, the change is not accepted; if it is, the standards or guidelines are adjusted accordingly. It is important when developing tools that they are based on the defined workflows and always up to date. In the NL, the tools are currently mainly based on Excel spreadsheets, as the template for the inventory shows. A workflow tool is currently being evaluated.

5.3. How we share

As mentioned, it is very important for the NL to communicate its successes, but also the mistakes made and lessons learned, as well to share key documents and practical tools. In this way, the NL seeks to help other institutions, potential and current project partners and project leaders, who are digitizing Helvetica[8], to be faced with the same problems. For this reason, a platform - www.digicoord.ch - has been launched, in partnership with other institutions. It brings together information on current and planned digitization projects across the country, shares guidelines, models and experiences, and lists sites containing digitized collections. Guidelines and key documents are also available on the NL's own website[9].

Almost more important, however, is the exchange of experience between different memory institutions, openly communicating about omissions, mistakes and lessons learned. The NL staff are always available when it comes to planning and carrying out digitization projects even when not directly involved as an official partner.

6. Conclusion and perspectives

To summarize, it can be said that it is especially important to recognize one's mistakes, to reflect and to learn the necessary lessons from it. Afterwards, the causes must be identified and reference documents are drawn up for future failure prevention. It is very important when working with an external digitization service, to clearly define the exact requirements of the final product in a multistep process. This starts with the specifications of the collection and

[8] An Helvetica is a publication pertaining to Switzerland

[9] http://www.nb.admin.ch/themen/02074/02076/index.html?lang=en

the requirements arising from the projected uses of the digital copies. These are normally formulated in a specification and thus part of the call for tender. Then follows the tender submission by potential contractors, giving a first opportunity to remove possible ambiguities. The same applies to the negotiation of contracts and the preliminary discussions for receiving the final product. Tests to clarify the requirements and to obtain reference data are essential. Finally, especially for larger projects, deliveries need to be defined that allow a regular and reasonable quality control by the client and provide the opportunity to take corrective action where necessary. If the deadlines are then managed correctly, there is a good chance that the desired result will be achieved.

This process adds to the project preparation time, during which the collection is evaluated and described in as much detail as possible. This evaluation must be reasonably detailed, but should not overburden staff resources.

Finally, the NL has learned over the past 7 years of digitizing newspapers (and other formats) that errors are unavoidable. The most important thing is to learn from them and be open. Sometimes, technical developments also play tricks on the best and most careful planning, in which all avoidable errors were bypassed. This results in an important realization for the NL: only those who do not digitize, will never make any mistakes. In this sense, there is only one mistake to definitely avoid: being too afraid to make mistakes to begin digitization. We prefer to put a collection that was not perfectly digitized online, rather than not to be able to make the same collection accessible to many.

Acknowledgments

Thanks to Genevieve Clavel-Merrin for support, translation and proofreading