

Bogdan Florin Popovici¹

“WORKING WITH PETA”. SOME LESSONS FROM A LARGE DIGITIZATION PROJECT

Abstract

Purpose: *The paper aims to present some lessons learnt from a large digitisation project, for the use of audience of IIAS Autumn School.*

Approach: *the presentation is made step by step, following the workflow of a digitisation project.*

Results: *The paper emphasizes key points and decision to be taken, as well as impact of them on the results of the project.*

Conclusions: *The size of the project may add new layers of complexity in a digitisation project. The impact of some decision is bigger, so the project needs to be carried out with a good planification and considering the delays and unpredicted issues that may appear.*

Keywords: *digitisation, digital archives, National Archives Romania*

“LAVORARE NELL’ORDINE DEI PETABYTE”. COSA CI INSEGNA UN GRANDE PROGETTO DI DIGITALIZZAZIONE

Astratto

Finalità: *Il presente contributo intende illustrare, a beneficio dei partecipanti alla Scuola archivistica d’autunno dell’IIAS, alcuni insegnamenti tratti da un progetto di digitalizzazione di ampia portata.*

Metodo: *La presentazione dei contenuti avviene per step, seguendo il flusso di lavoro di un progetto di digitalizzazione.*

Risultati: *Il presente contributo mette in evidenza gli aspetti chiave dell’esperienza e le decisioni via via assunte, ivi incluso il loro impatto sui risultati del progetto.*

¹ Bogdan-Florin Popovici, Ph. D, archivist, National Archives, Romania; assist.prof., Alma Mater, Slovenia, email: bogdanpopovici@gmail.com.

Conclusioni: *Un progetto di digitalizzazione, se di una certa entità, può acquisire inediti strati di complessità. Alcune decisioni pesano più di altre, perciò la fase di implementazione progettuale implica non solo una buona pianificazione, ma anche mettere in conto ritardi ed eventuali imprevisti.*

Parole chiave: *Digitalizzazione, archivi digitali, Archivi nazionali di Romania.*

“DELO S PETO”. NEKAJ LEKCIJ IZ VELIKEGA PROJEKTA DIGITALIZACIJE

Izvleček

Namen: *Prispevek želi predstaviti nekaj izkušenj, pridobljenih pri velikem projektu digitalizacije, z namenom nadaljnje predstavitve slušateljem Jesenske šole IIAS.*

Pristop: *predstavitev je narejena korak za korakom, in sledi projektu digitalizacije.*

Rezultati: *Prispevek poudarja ključne točke in odločitve, ki jih je treba sprejeti, ter njihov vpliv na rezultate projekta.*

Sklepi: *velikost projekta lahko projektu digitalizacije doda nove ravni kompleksnosti. Vpliv neke odločitve je večji, zato je treba projekt izvesti z dobrim načrtovanjem in upoštevanjem zamud in nepredvidenih težav, ki se lahko pojavijo.*

Ključne besede: *digitalizacija, digitalni arhivi, nacionalni arhiv Romunije*

1. INTRODUCTION

The present paper aims to share some of the lessons learned and challenges that we faced during a large digitization project undertaken by the National Archives of Romania. What we discovered, and often in the hard way, is that the big scale project has many shared points with smaller ones, but, in the same time, because of the magnitude of work, some new challenges appear and decisions need to envisage more aspects.

From the beginning, the word “large” needs to be clarified. Considering big digitization companies’ experience, dealing with 10 million pages scanned may look trivial. For our institution though, as for those involved in the project, it was not. Our assumption with this paper is that other organisations not centred on digitization as core activity—but interested in performing digitisation actions in house—may be interested in our experience, both as perspective and as solutions. In this regard, this presentation during the Autumn school of the International Institute for Archival Science may prove to be useful.

2. GENERAL APPROACH

Even before the starting of the project several decisions needed to be taken and the best way to proceed had to be examined. Initially, the project aimed to digitize 9.5 million pages (then 10 million), in a limited period (less than 2 years).

A). OUTSOURCED OR IN-HOUSE?

It was obvious that the target could only be attained by outsourcing, but the intention of the project was to also acquire expertise for the internal staff, on the whole process of digitisation and indexing. Therefore, one first decision was to split the task, and outsource the digitisation of 8.5 million pages, while the rest should be realised in-house.

B). WHERE TO DIGITIZE?

The records that planned to be digitized were historical records, and the need to protect them was evident. In this regard, it was decided that the records should not be digitized externally, but within the offices of the National Archives, considering the better implemented and manageable security measures. Moreover, since those records were very much required in reading rooms, the decision

adopted was not to centralise the records in one National Archival office, but to ask the provider to visit each individual territorial office. The issues associated with these decisions were revealed during implementation. The external provider had difficulties in (re)packing/(re)installing the hardware in every territorial division, basically covering the whole country. That hardware required high power supply, that was not always available in the offices, not to mention the spatial needs for the hardware and the activity of the provider's team, which had to be accommodated in some smaller buildings. Also, the workflow required internet with a large bandwidth, that was not always available.

C). WHO TO DIGITISE?

Apart of the provider's work, as I mentioned above, one of the goals of the project was to increase the competences of the National Archives staff in records digitisation and indexing, and in curating digital outputs. Possible options included the training of a core team or a more "democratic" approach, training one team for every territorial unit of the National Archives. Our previous experiences, with regional centres for microfilming, were not really successful, therefore we opted for the training of staff from each division. Several workshop and training sessions were organised, with presentation of the workflow, procedures on each step of the flow and practical usages of hardware and software.

D). HOW TO DIGITISE?

Since training would have been nothing without proper equipment, the project and its implementation solutions adopted required equipment for all teams; since the budget was limited, we needed to decide what type of professional capturing device can be purchased. The alternatives were scanners or cameras. After a market scrutiny and considering the usage of equipment after the termination of the project, the decision was to acquire professional cameras, as being cheaper, more versatile for various purposes in the Archives, and with sufficient quality in capturing images (300 dpi, optical, with small degree of distortion and excellent fidelity of colours).

3. TECHNICAL PREPARATIONS FOR IN-HOUSE DIGITISATION

As in any digitisation process, the flow has several distinct phases: capture the image, preparation of derivatives, quality check, and storing/preservation of the outputs. For each of these steps some decisions should be taken.

A first very important analysis was made in what concerned the purpose of digitization: to generate a replacement copy or only an access copy. A decision in this regard would have had impact on the number of copies, the file format, and the workflow itself. We decided to use the digital images also as replacement copy (security digitisation), which implied a more complex workflow and higher quality for the outputs.

A second step was to imagine and describe in a procedure the steps of the workflow. This procedure needed a simulation of activity, going through all steps and anticipating the information needs of the trainee. Though the process was carried out carefully, several issues appeared. The procedure was written by staff who had a certain experience, with both hardware and software to be used, and have a clear perspective on the starting point and arriving point, and the way to lead the flow between them. This is why some steps were missing in the procedure, since they seemed “implicit” for authors—while the working teams disagreed later, in practice. Also, due to some external factors, the procedure steps designed were altered. The procedure envisaged a flow where the image is captured, and the quality check would be performed during indexing. However, due to delays in the process of purchasing cameras, these steps could not be followed. Such situations required adaptation of the procedure and a flexibility of the workflow.

One very tricky issue concerned with the performance indicators. As in any project, time was a very rigid constraint and we needed control keys for the amount of work to be performed. Simulations have been made and an average production of captured images and indexes per day was set. In practice although, various unexpected factors generated delays: medical leave for the staff, drainage of the camera’s batteries after 4 hours of intensive work, malfunction of the lights system, slow performance of some computers, bandwidth of the network, the re-capture of defective images etc. The obvious lesson was to keep a buffer of time as to accommodate these delays and to avoid missing the target.

4. CAPTURING IMAGES

The most challenging analysis for the process of capturing the image was to decide upon the format needed. We considered the fact the images that were taken would have a double role, both as possible safety copy (paper original replace-

ment) and as access copy. That would imply the choice of a long-term format. Moreover, for the access copy, the constraints of the online portal were considered: a file of over 70-80 MB would be difficult to be accessed online and, due to technical specifications, also a large batch of small files would generate a delay in page loading. Due to the large amount of files involved, we needed to focus on the files with the best ratio size-quality.

In our analysis we have tested 5 formats: PDF/A and simple PDF, TIFF, JPG and JP2000. The cameras could produce JPG, NEF and TIFF files. NEF files have some prerequisites in being rendered, so it is rather complicated to use it. JPG is a lossy format. TIFF files are rather big—so, several dilemmas. In the end, considering that JPG is a lossy format, and it is not considered robust enough due to its compression (Allegrezza, 2021), we inclined to use TIFF format as master, with a LZW lossless compression.

The constraints of the portal implied that presenting copies as batches of TIFF or JPG files were excluded. JP2 is not well rendered by some browsers, so, it was also disqualified. PDF family remains as choice. Test of files showed that PDF/A files are bigger than simple PDF. Considering that access copy should focus on smaller size more than of resilience in time, PDF was eventually the format adopted.

Based on these initial choices, we made same tests of the necessary storage space and possibility to get as small files as possible. The results of the tests were not quite intuitive.

Table 1: Comparison table of formats and sizes

	No. test files	Master size (MB)	Access copy files (MB)	Total no. files	Total storage master (1 instance, TB)	Total storage access copy (1 instance, TB)	Total storage (1 instance) TB
a	b	c	d	e	$f=(c*e)/12$	$g=(d*e)/12$	$h=f+g$
TIFF	12	970	20	1.500.000	115,63	2,38	118,02
JP2	12	581	581	1.500.000	69,26	69,26	138,52
JPG/100	12	291	291	1.500.000	34,69	34,69	69,38
JPG/80	12	47	47	1.500.000	5,60	5,60	11,21

As it was expectable, the batch of 12 files in TIFF format was several times bigger than JPG files with 80% quality. The surprise was when we generated access copies: the PDF engine assembled the TIFF files in much smaller files than the set of JPG files. While all the other formats kept the size of access copy rather similar

in size with the master, in TIFF case the compression was almost 50 times and generated the smallest access files.

Considering the total planned number of images (1.5 million), a total of approx. 116 TB of data were necessary to store 1 instance of the master file in TIFF, comparing with only 5.6 TB if used JPG/80. But having the total size of files produced, JP2000+PDF were the most “expensive” as stored space, TIFF+PDF were ranked second, and JPG files were the smallest, but with rather big access files. The final decision was to stay of TIFF files (LZW compressed), 200 dpi and “classic” PDF format for access, combined from TIFF files.

For the capture process an important dimension of planning was the time. If images were transferred to computer directly, the process was not instantaneous (press the button and record the image on the hard disk), so it delayed the time for each picture and, also, contributed to the exhaustion of the camera battery. The decision was to capture the files on the camera card, which offered the speed and the possibility to transfer images to the computer in an unattended mode, streamlining the process and saving time.

5. PRODUCTION OF COPIES

Once downloaded in the computer, the flow of generating the final output consisted of several steps. Image was capture in TIFF files, 300 DPI. Then it was converted to TIFF LZW 200 dpi, which became the production master (but also preserved as archival master). These files were named according to a set of rules of syntax. All the files pertaining to the same archival unit was then assembled into PDF files. If the PDF file would go over 80 MB, then the PDF was split in volumes.

The hardest challenges in this phase were the one derived from the calculation power of the computers we had available. Converting **large** amount of **big** files are computing intensive, and using regular PCs is not the optimal matching. The best solution in such cases would be performant dedicated processing server(s). In our case, since the activity was distributed, it would imply a rather complicated architecture and an extra-budget that was not available. We stood then on regular PCs, but this was translated in longer times for processing.

6. QUALITY CONTROL

The process of checking the output included several decisions of planning and procedure.

A). WHO CHECKS?

There were 2 options — the quality check may be performed by the operator itself for their own products or have a different person check. The latter seems a better option, since it is in the human nature to be indulgent with own work. However, due to staff unavailability, different working pace etc., it was not possible to follow this approach in all cases.

B). WHEN TO CHECK?

The options are that the check may be performed after each capture, at the end of a daily batch or at the end of the whole capturing process. The first option would be a reasonable option only if the operator used a larger screen and not the camera display. But even in such cases, the multiple parameters to be analysed and the pressure of daily pace may lead to lack of focus. The latter option was unacceptable, because not all material was taken out from the repository at once; hence, in case of an error, the archival material should be taken out from the repository again. As a result, we focused in general on the second approach, which implied that a batch of material was imaged (for instance, half day of work), and then it was checked. That led the possibility to have the material in proximity for a re-capture, in case of low quality.

C). HOW TO CHECK?

It comes as a fact that the more critical a digital copy is, the more extensive the quality check should be. The plan was to have a full check by examining every image during the indexation process, but the context did not allow for this. As such, images were examined randomly, in samples, from the whole batch, in rare cases the time allowing for a full check. However, even the sampling was difficult. TIFF files were rather big, therefore the pace of loading the image was rather slow if the computers were not performant enough. And we did not have in all cases powerful computers. This situation led to delays and in some cases, to the reduction of the sample batch. Fortunately, at least for some parameters of the checking process, we managed to automate the process, as I shall present bellow.

D). WHAT TO CHECK?


The quality control checked the quality of images following several usual parameters (like clarity, colour fidelity, correct framing etc.), the completeness of capture of the archival material, the naming syntax, and its consistency. But one very important aspect was the correct technical codification of files.

For the TIFF files, we were aware of a project from some years ago, Preforma (Preforma, 2023), which had a tool for checking the conformity of TIFF files, DPF manager. We analysed several TIFF images with it, and we received some errors.


Files Check | **Reports** | Statistics | Periodical Checks | Conformance Check


Reports / Results / **HTML**


I	283	YResolution	300/1
	284	PlanarConfiguration	Chunky
	296	ResolutionUnit	2
	305	Software	LIBFORMAT (c) Pierre-e Gougelet
	306	DateTime	2022:12:06 10:34:22
	315	Artist	????????????????????????????????

 **Metadata analysis**

Description

 No metadata incoherencies found

 **Conformance checker**

 **Baseline TIFF 6.0**

Type	ID	Location	Description
✖	IDFE-0004	tag 315 Artist	Only one NUL is allowed between ASCII strings
✖	IFDR-0002	IFD1	Invalid Compression for RGB image
✖	TAG-259-0003	IFD1	The Compression tag must have a valid value

Figure 1: Results from DPF manager of a visually defective TIFF

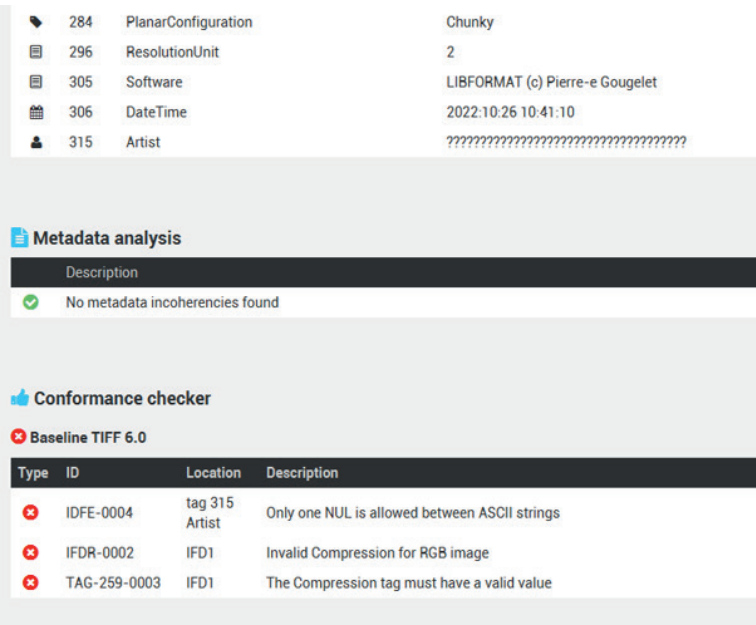


Figure 2: Results from DPF manager of a visually correct TIFF

The first remark was that we were not able to interpret the errors returned: how critical they were, what was their source etc. The tests showed the error were present in the first instance of image; basically, this is how they were captured by camera. We could not fix the errors since they seem to be generated by the camera software. We could not abandon the cameras, as they were the available devices. In the end, what we did was to check the files with several image viewers, in order to see if the files are readable; they were. Then, we accepted the images as such. The second remark was that this the software gave us the same errors and only the same errors even though the images looked like this:

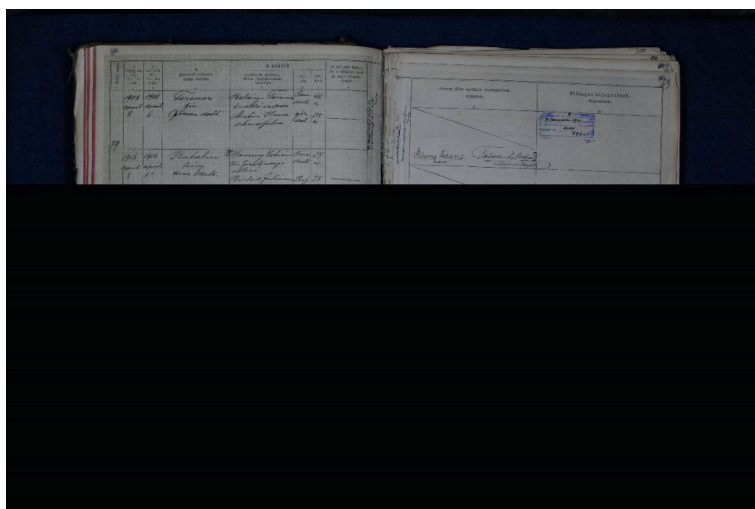


Figure 3: The TIFF file visually defective



Figure 4: The TIFF file visually correct

So, the tool we used it was rather a conformance test on some items in the structure of TIFF files but could not detect erroneous coded TIFFs. And we did not manage to find a tool that could check for the quality of a TIFF in terms of visualisation and to work on big amount of files.

As it was stated above, the camera produced TIFF files without any compression. We appreciated that preserving TIFF files uncompressed with 300 DPI would be excessive, so the procedure stated a downsize and a lossless compression—200 dpi and LZW TIFF. This approach managed to solve several things. First, it

helped reduce the size of a file with 30-40% (which counted a lot, both in terms of total space occupied on storage, but also in the ability to view TIFF files on less performant machines) and solved the previous issues, that is, if a file was corrupted in visualization, the converter software could not “load” it neither. This way, we managed to automatically identify the random corrupted files among other hundreds of thousands of good ones and ask for re-capture when necessary. The process of generating the access copy or the transfer of files over network were not protected from such errors neither. As aforementioned, TIFFs were assembled in PDFs and then all were moved to the centralized storage area. During the process, sometimes hidden errors occurred. These errors were sometimes hard to determine by direct visualisation, because the errors in displayed images were very small (see figure 5), and because of the huge number of files.

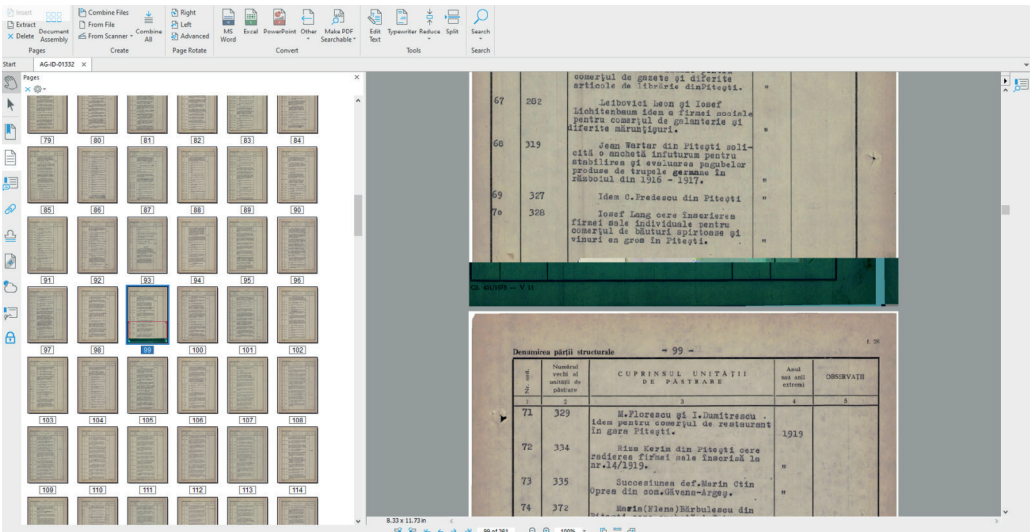


Figure 5: Error in PDF file

We tried to automate the process, by using some tools of checking for corrupted PDF files. While the file looked like in (Figure 6)), two checking software indicated no error (Figure 7 and 8), while another one gave a hint there is a problem with the file (Figure 9).

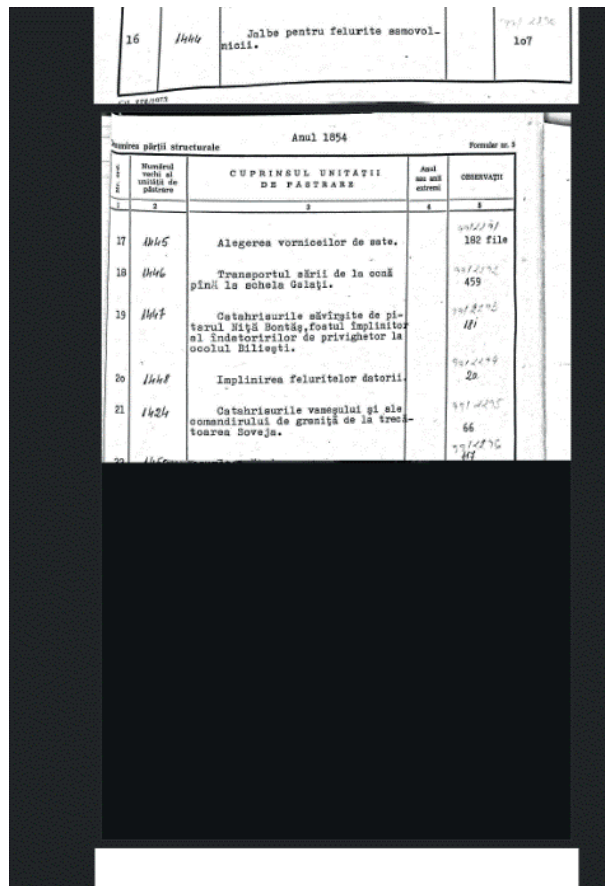


Figure 6: Visualisation of the file

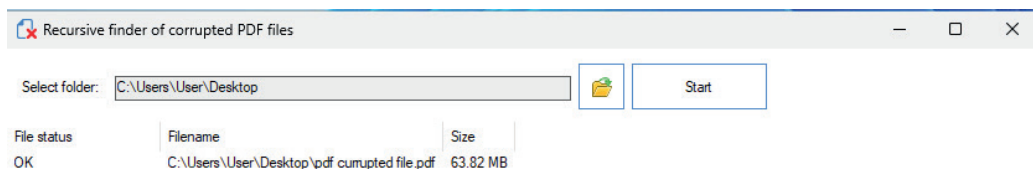


Figure 7: Test1 of defective PDF file

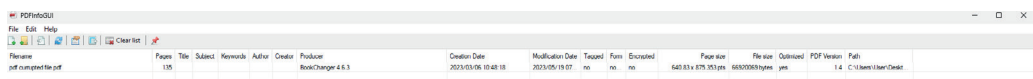


Figure 8: Test2 of defective PDF file

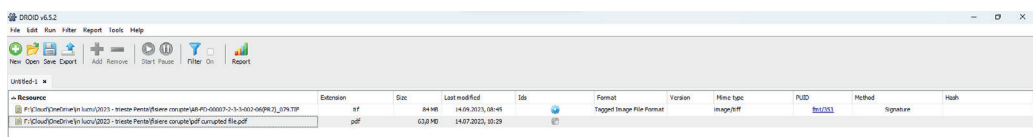


Figure 9: Test3 of defective PDF file

Such situation led us to the conclusion that, mainly after producing and checking the files, all should also have calculated the checksum and in any future examination of files the comparison between original checksum and the current checksum should be performed. And, of course, it is the general conclusion that it is a vital need to have software able to perform batch checking for various corruption aspects on technological level of bitstreams.

E). HOW TO STORE?

Storing of the results of a large digitisation project can also be very tricky. Letting aside various storage infrastructure [see, for a point of view (Popovici, 2022)], when we had as a result several hundreds of terabytes of data some challenges are revealed: sizing the storing capacity, distribution of storage, bandwidth to transfer all data to the storage, organisation of files on storage etc.

The transfer of data was really a big challenge, and sometimes it proved to be quicker to do the transfer by physically connecting small storage devices than to rely on the network. Checking for the checksum of files to detect possible errors after the transfer also required a lot of time. Respecting the rule for geographically distributed storage and separation of storage between master and access copies asked for extra storage room and remote storage units, fitted for the resulted amount of files. And even though the price/storage unit is cheaper than ever, acquiring enterprise-level storage for hosting hundreds of terabytes was not that cheap...

7. CONCLUSIONS

In spite the fact we have had experience with various smaller projects, we faced some unexpected issues during this project. The main lesson is that “big data” changes dramatically the perspective in terms of time of implementation, complexity, or budget. A small error if it is multiplied by thousands, may have a huge impact. An error in training that is replicated by tens of colleagues in practice, led to extensive corrections that may take days to fix. A miscalculation of necessary storage space may complicate the budget incredibly. A too tight schedule for tasks may be impacted by slow computers and if you do not have a time buffer, this small detail can make you miss the targets.

It is obvious that large digitization process requires generous budget, skilled staff, and professional tools. The difficulties we were encountered were facilitated maybe because we aimed too high, and we wanted to produce a big amount of outputs, but this was assumed in terms of goals. That led to bigger efforts and a lot of stress, but, in the end, these will be forgotten, and the results will be the one to stay.

One of the greatest accomplishments was the fact we managed to teach (and learn...) all these lessons with many colleagues, extending the competences in digitisation. And even though one project was hard, we know how to do it better in the future.

In the end, I would like to quote my colleague, Martin Stürzlinger (Vienna), who used to say years ago that digitization is not the end of the problems, but only the beginning. This is true. Starting from the issue of storage capacity and going towards the processes of digital preservation of large amount of data, these are difficult tasks to perform and a direct result of digitisation. Once we used to say that we do not have enough digitised material. Now, we have a problem of having much of it. As such, ending a digitization project is not a relief, but rather a door opens to new tasks.

REFERENCES

- Allegrezza, S. (2021). The Italian Guidelines on Creation, Management and Preservation of Digital Records: a Quantitative Methodology for File Formats Evaluation. Lecture at 15 th International Autumn Archival School (online), 1. 12. 2021. International Institute for the Archival Science Trieste (Italy) - Maribor (Slovenia).
- Popovici, B. F. (2022). Some Considerations On The Archival Storage In Digital Preservation of Records. *Moderna arhivistika*, 5(2), 462-466.
- Preforma. (20. 11. 2023). Preforma - Smart solutions for digital preservation. Available at <https://www.digitalmeetsculture.net/article/preforma-smart-solutions-for-digital-preservation> (accessed 5. 3. 2023).

SUMMARY

The paper aims to present to the IIAS Autumn School audience some lessons learnt from ta large digitisation project. The paper goes to steps of the project, asking some key questions and presenting the rationale behind decisions adopt-

ed. Planning had to decide if the records will be digitised in-house or outsourced, place to collect the material for digitisation, who would be in charge for digitisation, how to digitise. The project split the material in two, part of it being outsourced, but in the offices of National Archives, part of it being processed internally, by the staff. In this regard, hardware has been acquired and staff was trained.

The capturing process was generating TIFF files, 200 dpi LZW compressed, with access copies in PDF. Various facets on this decision are presented. A significant portion of the text is devoted to the quality control, showing difficulties in batch identification of corrupted files. In the end, some considerations on storing these materials are presented.

The conclusions highlight that “big data” changes dramatically the perspective in terms of time of implementation, complexity, or budget. Large digitization process requires generous budget, skilled staff, and professional tools. The final remark state that digitization is not the end of the problems, but only the beginnings ending a digitization project is not a relief, but rather a door opens to new tasks.

Typology: 1.04 Professional Article