'The Image Is Not Enough'. Some Considerations on Digitization of Archival Documents

Bogdan-Florin Popovici

Keywords: *Digitization Projects*; *Analog Historical Records*; *National Archives*

In the last decade, digitization became one of the most popular activity in the archival world. I do not challenge if this is right or wrong, if all the processing backlogs or issues associated with paper records were solved, so we can move to another level. It is just an observation that almost all discussions about classical archives is departing from appraisal, arrangement, deacidification, etc. towards digitization. In part, this is due to the great advantages implied. Dissemination of archival holdings without reading room timetable or space limitations, the large-scale availability of digitizing hardware and software contributed greatly to the popularity of digitization projects. Hence, by using modern technologies, everybody, from individuals and small, but agile archives up to big National Archives institutions could make their holdings more visible to the world. The easiness of access sometimes created also the representation of a simple, straightforward process, and implicitly created the expectation for archives as being "at a click distance" from users.

I was involved in the last years in several projects of digitisation; I had the opportunity to see the plans, but also to draw the afterwards conclusions; to read the success receipts and learn from practical failures. And some of my observations I intend to share in this paper.

Definitions

In order to have a common ground I would like to start by saying that **digitisation**, in my understanding, is the process of converting the information from analogue/traditional instantiation into digital form. That is, digitisation does not imply, *per se*, making full text searchable textual from records; nor to associate it with metadata, i. e., descriptions or keywords. Digitisation means at minimum to have a digital picture of the record (in most cases, of the **appearance** of the record), but also to have an audio or video recording, on tape, transferred to its digital equivalent. It is true, however, that solely digitisation is rarely useful today; almost always it is accompanied by further processings that facilitate retrieval.

While digitisation can be performed on a variety of analogic documents, I would like to focus in this paper on **archival records**. By this term, I understand records (archival documents) that ceased their regular lifecycle and are considered for permanent preservation in a historical archive. Though further on almost all

examples will come from experience with paper archival records, the term itself is not carrier limited, but covers any record, on any type of carrier.

The record

The archival record has several characteristics that were outlined in professional literature for long time. These characteristics individualize it, making it different from a library document – a book or a journal, for instance. The archival record was generated in the course of a practical activity and it serves as evidence for it. The purpose of that activity was not to create that record, but the record is instrumental for attaining that purpose; this makes the record a by-product of the activity. For instance, in order to carry out an evacuation of persons during the war, a list of persons was created to serve the immediate administrative/military purposes, as a control list of persons. On long term, the contained information serves other different purposes, reflecting intentions of administration (to evacuate), profile of individuals (who is to be evacuated), history of administration (how the evacuation was performed), and so on. It would not serve any longer for listing persons to be evacuated and how many trucks they would need and how many supplies were necessary. Another important characteristic is that one record does not offer the full picture of an event, but rather a snapshot of it; one can say that a record is a picture of a given moment, but in order to see the whole movie of an event or activity it needs many pictures (records) in sequence. Using the previous example, historically speaking, the whole process of evacuation cannot be restored from a single list, but from the list and the other records that were created during that process.

Understanding these characteristics of (archival) records is very important in the process of digitisation, because it determines the limits of the records targeted. How relevant is to digitize partial fonds or excerpts from files? How useful for users is to bring out from various folders all the pictures, for instance, and have a wonderful collection of pictures online? Will they preserve their evidential value any longer? While there is an intuitive answer to this, I would say it is not always easy to answer. Any digitisation project should identify what is called "designated community", that is the audience of that project. If somebody is studying the history of photography, it is obvious that the fact that some digitized pictures came from courts records is almost completely irrelevant, because the information sought refers to something else than the administrative process from where those pictures emerged. But for somebody trying to restore the historical process that generated those records, the fact there are available only spots from the whole records production may be an issue.

Sometimes, the original documentary context was altered by ... additions. For instance, records from the 14th century were transcribed by researchers at the end of the 19th or during the 20th century, and those notes were preserved near original records. When digitising, a decision should be taken if those notes are considered historically as originals, if they are part of the record of not. Again, it

is a matter of envisaging possible audience: some users may be interested in the way in which records were transcribed, how other scholars, in maybe unpublished works, read and interpreted that record. On the other hand, it is obvious those records were not generated from the original process, and they are a different kind of evidence.

Determining what to digitize and to what extent may be a troubling issue for National Archives performing digitisation of their holdings. Such institutions do not have a specific designated audience; their mandate is to preserve and deliver information to any user, for any imaginable purpose. It is not for the institution to determine the purpose of research, but for the users. Therefore, the records should be digitized as to serve a broad range of purposes, and digital surrogate created should be as much as possible 'identical' to the original, transferring the same message, preserving **and** revealing the original documentary context.

The image

What makes digitisation so popular is the ability of creating similar documents, with enhanced accessibility. But talking about **similar** documents is not as simple as it may seem. Digitisation creates a copy, and in any copying process something is lost, and something is added. A digital copy of a 15th century record can be hardly considered a 15th century record... Date of creation is different, so obviously **it is a different document**. But the content is the same, so it is also obvious that **it is the same document**. For somebody who wants to study the watermark, the record **is not the same**, nor for the one who wants to recreate a carrier who may have the same texture as the original... Purposes for digitisation are the one dictating the process, and it is important for users to be acknowledged in what circumstances and for which purposes the digital copy was produced.

In digitisation, many archivists tend to focus on technical parameters, like resolution, bit depth, colours, and so on. While these parameters are undoubtedly important, these are often dictated by the purpose of the process, to what goal will serve the digital surrogates. If the process of digitisation aims to replace the original, there are other requirements than digital copy for dissemination. If the digital image will only serve for consultation onsite, there are other requirements than for dissemination online. In my opinion, a wise approach should take into consideration a broader range of usages, on the principle "copy once, use many times". Of course, this approach should be moderated; for instance, I can hardly believe there is a ground to digitize pictures at 4800 dpi, assuming somebody, some time, will want to generate a poster. The frequency of such uses, the size of such files, the hardware and software needed to process such big files are aspects that should be considered.

No matter how careful the decisional process would be, the evolution of hardware and software will make our current digital copy a part of history. Imagine digital copies made in 1990s and the one that can be produced today – it is easy to identify a difference in quality, faithfulness and so on. And this trend is very likely

to continue. That means, in all aspects, we need to be aware the digital copies we create should be named, cared and preserved considering that another version may appear in the future. Some may discard the previous copies, others may consider them as evidence of the appearance of records at a certain moment in time.

Dissemination conditions are also part from a project of digitisation. It would be a pity that such wonderful images of our wonderful and unique documents not to be shared with the world, so almost any project has a component of publication online. Today there are plenty of free or cheap possibilities to publish/share online the outputs of a digitisation project. But here too there are some aspects of concern. One of them refers to the right of dissemination of information, either under the sensitivity of information (that can vary in time!), or to the various copyright issue. Copyright may set conditions both for the institution (to publish copies) and for the users (to reuse the information published). Sometimes, the presence of a watermark on the image, indicating the holder of the original record is considered as excessive, under the claim of "public property". On the other hand, the free circulation of images, without acknowledging the holder, can lead to a low awareness about the holding institution, with possible effects on policies and funding. The issue is not easy to solve, but it is certain that it needs to be addressed.

The metadata

What Facebook or Instagram teaches people today is that it suffices taking a picture and put it online; that is all the effort needed in digitizing and displaying. While many tend to assimilate the process with digitizing archives, I think it is a bit more complex than that. And one of the most difficult points is where "old" information (description of records) joins with "new" information – the digital image.

In traditional environment, retrieval of record is based on finding aids, which are organized and presented in specific ways. In general, the finding aids reflect the physical organization of records in the repositories; sometimes, they create intellectual structures that map the physical arrangement. The structure of finding aids differs from tradition to tradition, from detailed abstract for each archival unit or even record up to general presentation of whole series or even fonds/collections. In the former case, the user must browse the finding aid, page after page, mostly if there is no other archival structure reflected (the case of simple chronological listing). In the latter, the user must browse the series, record after record, in order to find the information of interest. In both cases, the organization and presentation of records reflect mostly archivists' usage and needs, and they are not necessarily relevant for external users. And while in paper it makes some sense, publishing the records online sets the retrieval to a new level and claims new requirements, because of the possibilities offered by the technology.

For anyone using the internet today, Google approach is the main expectation for retrieval. That brings for many archival users the same expectation

– just ask a question in a box, hit Enter and the archival system should reveal the whole of information pertaining to a certain topic or person or place. But unfortunately, this is rarely doable, first of all because there are very few cases where the records of the past are transformed as to be content searchable. And those cases mostly come from dedicated Archives, and not general Archives, with huge holdings, spread across centuries, with records in a variety of formats, types, metadata, and so on. This is not a matter of technology, but a matter of resources: I am not yet aware of any National Archives who may afford an infrastructure where visual search engines look after landscapes and persons in pictures and video recordings, audio search recognize voices in audio recordings and identify persons, textual search engines bring to the user the whole list of records where a name was identified and artificial intelligence searches on main topics and related topics. And I presume in this scenario researching in the archives will be a very boring job...

Until this case will come true, the retrieval of digitized records will still rely on the metadata associated with the digital surrogates. In most cases, these metadata are in fact the traditional finding aids, which were simply converted into digital metadata. In this case, of course, tradition of archival processing will play a huge role. Those tradition that encouraged strict control over information from records (e. g., practices influenced by Russian traditions) have detailed information down to the file or record level; such information delivered online will offer a great help for archival users. The countries where the archival processing favorized descriptions for records aggregation of high level will need to rely more on browsing technique.

But traditional description of records implies some transformations in order to facilitate retrieval in automated systems. Both approaches mentioned above (browsing- or keyword-based retrieval) rely, in paper system, on the structure and context. In one case, structure of the fonds leads step by step from general to particular, narrowing the topics. For archives of a Mayor's office, for instance, **department of buildings** have a series about **building permits**, which, in a certain **year**, allows the identification of the file titled the **House no. 3**. In the other case, where the structure of archive was ignored and records are chronologically listed, the structure of the inventory may have relevance for retrieval. For instance, if we consider the same example above, there will be many files being related to building permits; the abstract may look like *Idem, House no 3*, since it is good enough for paper finding aids, because on one page it is easy to read the text above and understand the meaning. However, in automated retrieval system the meaning of both titles would be quite hidden. Querying for "building permit for House no. 3" will return no result, since the full meaning is dependent of the other descriptions ("building permits" return the series, not the file); querying for "house no 3" may be the best options, but with the assumed risk to have returned too many results. The point I liked to reveal is that simple transformation of finding aids may be too often not enough to respond to the needs

of retrieval in automated system. All along with records images, records descriptions need to be prepared for the new role and new functions. This adds a new layer of complexity the processes of digitization.

Closing lines

There is a big level of expectations about digitization of archives today. One may comment that it does not matter if your archival building is a ruin, or that you have kilometers of records unprocessed; if you are online with some records then you are in the trend. But, as we all know, a trend encompasses both best and worse practices.

Contrary to general belief, a good project of digitization is much more than getting the image of a record. It implies a lot of preparation work, in order to decide what to copy, how much to copy, for whom to copy. It implies a lot of anticipation about how that record image will be retrieved, where it will be stored, who will curate it, how the sustainability of the project will be accomplished. Metadata for the records digitized is also a challenging issue, because simple transformation of legacy finding aids in to digital ones is often not enough for a proper retrieval in automated systems.

Ignoring such facets may be a solution. But on long term, such approach would imply many efforts to patch the unanticipated situations and, sometimes, even to the resuming of the whole endeavor. For archivists, witnessing too often past mistakes of people, this will be an unwise approach, I believe...