First published in: Bulletin du CIRA, Lausanne: Centre international de recherches sur l'anarchisme, 2016, 72, pp. 10-27 (in French and English). Contact: info@cira.ch

Best practice for digitizing documents

This article is based on an intervention by the CIRA-Lausanne during a meeting of the FICEDL[1] in April 2016 in Bologna.

Introduction

This contribution offers general thoughts on the main issues which come up in the planning stages of any (more or less) serious digitization project. Based on our own modest and mostly empiric experience, it centers on methodological proposals, rather than on technical recommendations. Thus, our aim is to allow FICEDL members, as well as other activist libraries or archive centers, to acquire, regardless of their computer skills, some useful notions in order to understand exactly what is at stake when choosing guidelines to help plan such a project. Preliminary remark: we are dealing here with digitized documents, and not with the production and management of born-digital documents, which raise their own set of issues.

Digitization: a new quest for the Holy Grail?

We regularly observe at the CIRA that the word “digitization” is a generic term which encompasses various meanings, including when used by fellow librarians and archivists. It is a buzzword used to show we are up to date on the latest developments, and the mere fact of uttering that word seems imbued with some kind of magical function, which will supposedly solve all the problems we are faced with. However, if we were to ask around what digitization practically entails, we can be sure to get dozens of different answers.

With almost every new visit at the CIRA, we are inevitably asked whether we plan to “digitize all this”. Countless enthusiastic comrades scan or photograph everything at hand, often without much technical knowledge, without any real method or long-term plan. It is not rare for users to hand us a flash drive containing thousands of pictures taken from our collections or from elsewhere, without coherent file names and in many different formats. These files are hardly usable if they are not organized and comprise no references. Friendly libraries ask for advice, or rely on comrades who “know about computers” and are convinced that they hold the one key to the problem (we cannot stress enough that in order to elaborate serious projects, we need not only computer know-how, but also some notions in library and archival sciences).

We are ourselves assailed by doubt every time we need to make a technical decision. For digitizing our posters collection, we have had to start afresh several times because we had used different machines, formats, and methods depending on who had offered to help. And, retrospectively, we do not think the method we finally fell on is even the best one.

Digitizing certainly offers numerous advantages:

- better preservation of the originals (especially interesting for delicate or large-size collections like periodicals and posters);
- easy reproduction which allows to share and communalise documents, while increasing the

1 < http://ficedl.info/ >

probability of a copy being preserved in case of the disappearance of the physical documents;
- and, if we put more resources towards it, distant access and full-text search within a corpus.

However, there is a high risk of:

- spending a lot of energy for a result which is not always satisfying;
- scanning collections which have already been digitized by others;
- not being able to make the files accessible or long-lasting, etc.

In short, it is not enough to invest in some hardware and find volunteers to push a button and turn pages.

Need for historical distance

To this day, no institution in the world has the necessary experience to look back and judge whether this or that technical choice will be appropriate in the long run. With increasing dependence on technology and built-in obsolescence, the reach of the digital projects of our small-scale and often precarious activist structures must be considered with modesty. To offer just one example of this, the PDF/A international standard (supposed to offer an archival format more sustainable than ordinary PDF) is already on its third version in the ten years of its short existence.[2]
Fundamentally, it is important to keep in mind that, until now, paper, if preserved in good conditions, has a theoretical lifespan that is far superior to that of any digital data.

Digitization against preservation?

If we take a look around at the libraries of the FICEDL, we can see that most are housed in precarious buildings which are not suitable for the long-term preservation of documents. Their main function is often to be used as spaces for local activist gatherings rather than as spaces for heritage preservation. This explains their enthusiasm to digitize in order to "save" collections and make them accessible. This should not replace the need to first ensure the preventive conservation, description and localization of their physical collections[3]. This work will not be useless, since one of the basic rules of digitization says that there should be first of all sufficiently-detailed catalogs and inventories in order to be able to link digital documents to them.
It is also necessary to keep the original physical documents, especially when collections are unique or rare. We can never be certain we didn't skip a couple of pages, didn't leave a crumpled page folded, or even that we won't simply lose the digital files. Digitization will not allow us to gain some room on shelves.

There is digitization and then there is digitization...

Depending on what our objectives are, technical choices differ (and so does the volume of the stored data![4]). If we aim at storing real preservation files, usage commands to keep as close as possible to the original documents[5] and to choose high-quality formats and resolutions (often uncompressed TIFF or

2 The PDF/A format includes in the file metadata the fonts used within the document in order to restore it as accurately as possible if these fonts are no longer available. It is however complicated to check whether files are really set according to norm, since different validation tools exist and they do not always send back the same result.

3 For advice on preservation, see: Conservation préventive, à l'usage des bibliothèques militantes, *Bulletin du CIRA* n° 66, 2010, pp. 11-19 (online: < www.cira.ch/bulletins/cira-bull-066.pdf >).

4 The size of a document only a few pages long can then vary from a few dozens Ko to several hundreds Mo!

5 For iconographic documents, we can for example include a color scale.

PNG), whereas low-resolution files (for example JPEG) are enough for consultation, online access or inventory purposes.
In professional projects, we generally mix both, the former being used as "master files" from which lighter consultation files are generated (in the hope of being able to get back to the original, non- or lightly-processed files if need be). Some activist projects which make books, periodicals, or fanzines available online are satisfied using only low-resolution files[6], which demand less storage space and allow fast online access, but do not allow documents to be reprinted identically.
If we are to spend hours on end digitizing documents in any case, we can ask ourselves whether it is not worth it always to produce high-resolution files, even if it means delegating their conservation to other, better-equipped libraries, for example within the FICEDL, and only storing on our own computers the files suitable for common usage.

Text vs. image

We still find too many text documents only accessible as simple images (like pictures taken with a camera, or PDF files generated without using OCR software[7]). The files are then of no use if we want to find out quickly whether a thick volume, or a periodical with hundreds of issues, mentions this event or that person. People then have to scroll through the screens one by one. This process sometimes takes longer than with paper, especially for larger-size periodicals which require people to zoom in to read. For potential research, it leads to the same situation as when we used to microfilm our collections.
When we digitize text documents, we must therefore always use some OCR software when we generate consultation PDF files, while keeping in mind that despite some progress, no software on the market can reach an absolute recognition rate[8] and that the result is not always displayed correctly.

Opting for documented formats

As for the choice of formats, especially those used for master files, it is a good idea, whenever possible, to opt for standardized formats, or else it might create problems such as only being able to open the files with software from this or that company, or even not being able to open them at all when the company no longer provides updates, goes bankrupt or is bought up.
Some formats like TIFF are not formally international standards, but are however considered reliable because they are documented.[9]

To compress or not to compress?

The need to save on storage space can make us compress some data. Generally speaking, file

6 For example, see the DIY Bookscanning project presented at this same FICEDL conference in Bologna: < www.grafton9.net >.
7 OCR : optical character recognition.
8 Older fonts, particular characters (ex. Gothic alphabet), a damaged collection or hand-written notes can make the recognition rate fall drastically. Also some desktop OCR software are particularly useless. That is one more reason to keep preservation files, since they can be used once we have access to better OCR software, without having to scan everything all over again.
9 TIFF is often recommended for preservation files. We must however carefully check the settings on our machines, since it is not rare for some compression to be used by default (since TIFF is an extensible format, hardware companies can include a lot of things which are not detectable by untrained eyes). When we bought a new copy machine at the CIRA, we realized only after hundreds of scans that the TIFF files we had obtained used JPEG compression, which we had wished to avoid!

compression (for example JPEG) leads to irretrievable data loss[10]. However, some modes of compression are reversible without any data loss, they allow to save some disk space while storing high-quality files[11]. Still, unnecessary format changes should be avoided, since with every transformation (compression or decompression), there is a risk to lose some data.

Original size

A large number of pictures can be found on the web without any indication as to whether the original document was a postcard or a large-size poster. This missing information is especially embarrassing for libraries or archives which are supposed to be careful around those issues.
If we use flatbed scanners, this information is by default stored in the metadata (so many pixels = so many centimeters), as long as we do not resize the digital "originals". On the other hand, this information is not stored when using ordinary cameras[12]. An alternative can then be to associate the files with information from a preexisting catalog. Despite the disadvantages of measuring each and every placard by hand, it allows us to store the actual dimension of the document, as it can vary from the dimensions of the file when the frame is voluntarily wider in order to show the material condition of the original (folds, tears, etc.). This is the method used for the posters at the CIRA and on the *Placard*[13] website.

Set priorities

Since it is impossible to scan all the documents in a library or archive center, we must make choices and set priorities. We generally start with digitizing the most popular, rare, or politically, historically, or intellectually relevant documents first. Sometimes, choosing a collection of local interest can also allow us to obtain some external funding.

Digitize wholesale

At the CIRA, we regularly digitize just one issue of a periodical, an excerpt from a book, or one illustration in order to meet the needs of our distant-access readers, but these scans are generally not kept, because they can hardly be managed systematically.
On the other hand, if we are concerned with digitizing with the aim of preserving documents, we must stress the importance to scan entire sets of documents. What could be more frustrating than to find a digitized periodical, only to find out that the pages or issues we are looking for are missing? What is more time-consuming than to try and assemble fragments scanned by various people, with different

10 This means that, although it is technically possible, there is no sense in converting a JPEG file into a TIFF, which will then only be pointlessly heavy.

11 An LZW compression is sometimes used on conservation TIFF files. The jpeg2000 format also theoretically allows compressions without loss, but it is currently rarely used and most software are unable to read/write it correctly. We must also note that exporting as a PDF generates a compression, generally of the JPEG type. When dealing with older collections without colors or pictures, it can be interesting to produce conservation files in greyscale, then to transform them into two-tone, and to use the PDF fax compression. This allows to make extremely light consultation files, with a quality that is at least equal to a JPEG compressed PDF.

12 Some professional photo set-ups however allow for automatic distance calculation. In a more DIY fashion, we can imagine a fixed set-up in which we know that the camera is at a set distance from the document (as long as the documents do not vary in size and thickness). If we take this into account, using a camera can still be useful for large-scale documents, since scanners for documents larger than an A3 format are extremely expensive.

13 < http://placard.ficedl.info/ >

naming and format standards?
If we do not possess the entire collection, we should try to collaborate with other libraries.

Use the network

The first thing to do is to check whether said collection has not been digitized elsewhere, which is not an easy thing to check (people may have digitized documents without the information being online). Informing people in the FICEDL network, checking the online platforms of institutional libraries (for example, in France, the vast online library *Gallica*, which contains some anarchist publications). Depending on contexts and political principles, relations with national libraries can be difficult. However, what use is there in scanning a collection which already exists in a digital format, if it has been done according to standard?
When we digitized the main collection of the CIRA, *Le Réveil anarchiste / Il Risveglio anarchico*, the Geneva library lent us the issues which were in too bad a state in the CIRA collections and a file-sharing agreement was made with the Swiss national library, which contributed to the cost of outsourcing the digitization process. The CIRA however keeps its own copies of the files and its own access platform. Only afterward did we discover an almost complete collection of the periodical at the bottom of some storage space in a union library in Geneva, but unfortunately it didn't appear on any catalog.

Communicate about your projects

In a similar fashion, it is important to announce the projects we are planning, which are underway, or finished on collective platforms. The Berlin comrades manage the *Lidiap* database for specifically-anarchist periodicals[14]. In some countries, some institutional national databases also exist[15].

Outsourcing?

Our small structures often have no other choice but to digitize documents ourselves for financial reasons, but this also offers other advantages, like greater autonomy, and a greater safety for the collections which remain on-site. However, it requires a large enough work space, to be equipped with appropriate machines (which are expensive for everything larger than an A3 format) and to master technical parameters (lighting, anti-reflection, how to use the machines, end formats...).
Using an external company allows us to delegate the issue such as hardware and the management of technical issues to professionals. It must be noted that cost rapidly increases when documents have to be digitized by hand[16], which is generally the case for delicate, bound, or badly-preserved special collections.

Draw a detailed plan

Whether you digitize documents in-house or outsource the project, you must define, as precisely as possible, the technical standards you choose: what kind of machine, black/white or color, resolution, metadata, file naming system and format, and, eventually, filters, OCR, etc. Then you need to check

14 Lidiap (List of digitized anarchist periodicals) : < http://www.bibliothekderfreien.de/lidiap/ >.
15 For example, in Switzerland, the platform < www.digicoord.ch/ >. These also indicate some best practices in French and German.
16 As opposed to piles of non-stapled A4 sheets of paper which can be put into an automated loading tray.

that the result is satisfying (quality check). In order to save on cost and time, it is a good idea to use median settings which can be used for the whole collection. Tests can be necessary before starting on the project for good.
Writing a detailed plan is also very useful in order to know how part of a collection has been digitized for later reference.

Document access

Once digitization is done, the issue remains of how to manage the files and make them accessible.

- Should they simply be stored on a hard drive with a file-naming system which allows us to navigate them?
- Should they be linked to catalog entries (if those exist)?
- Should a complete digital library be set up with some suitable software and stable electronic addresses which allows for reliable citations in our research[17] ?
- Do we offer solely an on-site access or do we aim at making files available online?
- What tools do we provide for consultation (readers, full-text search engines?)
- Are we eventually able to articulate them with digital-born collections which we might have?

There are many possible answers to each of these questions.

For on-site access, we recommend to install at least a small-scale indexation system[18], which allows to search rapidly through the whole digitized collection (as long as it has been treated with some OCR, see above).

For online consultation, we need the necessary material and bandwidth if we wish to avoid using "free" online apps in the hands of companies which care little for user confidentiality.

Should we bother about rights?

Even though the issue of intellectual property may seem absurd for anarchist publications[19], some issues may nevertheless arise. We must distinguish between intellectual property rights or copyright per se and issues around the protection of a person's image.[20]

The principle used when considering whether or not to put something online is one of risk assessment. Within the FICEDL, those are low concerning copyright issues, as long as sources are cited and that there is no intended monetary gain. We can however imagine more delicate issues on whether to put something online, for example the case of a contemporary poster accusing an activist of rape (rightfully or wrongfully, who are we to know?). Should we take position as a library? What is more important: individual right to privacy and to erase online data about oneself, or collective access to historical sources?

Back-ups and migration

17 Nothing is more raging than to click on a reference link only to find a massage saying that "this page no longer exists".
18 Such as Recoll on Ubuntu, or FileSearchy for Windows.
19 By publishing their books through commercial publishers, some authors however relinquish their rights (this could be the subject of a debate in itself). In other cases, the copyright holders might be the authors' heirs.
20 At the CIRA, we have for example been contacted regarding a poster in support of a strike in the 1980s denouncing an employer in Spain. The plaintiff thought the image of his late brother was unfairly used.

We now address a core issue around *long-term* digital preservation. Some conceptual models exist (OAIS[21]), but, at our scale as small activist libraries, setting up such structures has so far been utopian. We can however make sure to make some regular back-ups (which should be checked), with at least one back-up stored off-site (in case of fire, flooding, theft or ill-intent, back-ups are useless if they are stored next to the main computer). In this case, collaboration within the FICEDL could take the form of reciprocal storage of back-ups.

One of the problems with activist library practices is that files can sometimes be stored on the personal computers of some comrades who help out, or own specialized software. This is probably the worst situation for transparent, collective, and long-term management. All the work regarding the library should be possible on the structures' own computers, or, if that is not the case, be systematically transferred as quickly as possible.

Also, nothing is eternal, especially when computers are concerned. For back-ups, ideally we should have a check-sum system in order not to duplicate corrupted data for years unknowingly. It is likely that few members of the FICEDL currently use this kind of safety system.

It is also recommend to replace hard-drives at least every 5 years and to check periodically whether files are still readable with the evolution of software, operating systems, hardware and reading devices, and, if need be, to operate the necessary migrations.[22] All of this represents a large and rather technical aspect of our work if we wish not to lose in the longer term what we put so much energy towards "saving". It is not a mystery if for professionals, long-term digital preservation costs more than the storage of paper documents.

On the importance of documentation

While technology evolves (quickly), activist groups and work methods also change (more slowly!). Therefore, it is important to document the choices made in order to avoid losing the trace of those decisions, as it is part of the history of the collections. This must make it possible to keep some sort of coherence for later projects, and later to understand the different "technological strata" which will inevitably form (it is also useful in order not to centralize knowledge within the brains of one or two indispensable people).

Conclusion

As librarians in charge of a historical memory to preserve and transmit, we must take time to think properly before blindly launching digitizing projects spurred by enthusiasm or what is fashionable at the time. Since people and energies at our disposal are rare and precious, we must use them wisely and aim at the more sustainable solutions possible given what we currently know.

Translated from the French by Corinne Chambers.

21 OAIS (Open archival information system): it is a reference model for long-term digital archives, defined in the ISO 14721 norm.

22 Recent studies show that most data loss is due to back-ups, migrations or software updates.