”Un Manuscrit Naturellement” Rescuing a library buried in digital sand

Nicolas Larrousse, Christophe Jacobs, Michel Jacobson, Gilles Kagan, Joël Marchand, Cyril Masset

“Un Manuscrit Naturellement”
Rescuing a library buried in digital sand

Nicolas Larrousse (nicolas.larrousse@huma-num.fr), CNRS (Centre National de la Recherche Sientifique), France

Christophe Jacobs (christophe@limonadeandco.fr), Agence Limonade & Co

Michel Jacobson (michel.jacobson@huma-num.fr), CNRS (Centre National de la Recherche Sientifique), France

Gilles Kagan (kagan@cnrs-orleans.fr), CNRS (Centre National de la Recherche Sientifique), France

Joel Marchand (Joel.Marchand@huma-num.fr), CNRS (Centre National de la Recherche Sientifique), France

Cyril Masset (cyril.masset@cnrs-orleans.fr), CNRS (Centre National de la Recherche Sientifique), France

1. Manuscripts in a digital necropolis

This story really began during the Middle Ages, with the creation of manuscripts by copyist monks[1].

In 1930, Félix Grat a French archivist paleographer, experimented with a sophisticated camera to create microfilms of manuscripts in order to make them widely available and facilitate their study. This resulted in the creation of IRHT[2], an institute devoted to fundamental research mostly on medieval manuscripts.

In 1979, an agreement was signed with the Ministry of Culture and IRHT to digitize all the manuscripts stored in French public libraries. This corpus is among the largest digitized medieval sources: the work is still in progress!

The “original digital” copy was stored on Huma-Num’s[3] infrastructure: it was made of files for the manuscript pages encoded in TIFF format representing a huge volume of data, around 40 TBs distributed in 2 million files.

The fantasy of digital immortality is widely shared, but in reality, digital resources are highly fragile. Even if we are able to store them safely in a readable format, they prove to be totally unusable if we don’t provide related information to understand their content and their organization.

In short, over many years, we have built a very safe digital necropolis progressively covered by layers of digital sand rather than a clean organized library.

2. The context

As the original material of this set of data is clearly part of French cultural heritage, it is important to conserve both the manuscripts themselves, some of which are no longer physically available for consultation due to their poor condition, and the scientific work already done on them. It was becoming a matter of urgency to take action as the memory of the project began progressively to disappear mostly because of the numerous changes in human resources.

We have an institution in France, the CINES[4], dedicated to the long-term preservation of research data. But the cost of preservation in our case was too high. In the meantime, the French National Library had successfully converted part of its resources from TIFF format into JPEG2000, thereby reducing the data to one-third of the original size, and prices at CINES had fallen dramatically.

It now became financially reasonable to consider the preservation with the CINES using the JPEG2000 format.

We therefore decided to begin the preservation project.

3. Dealing with the data abundance and imperfection

The first need was to sort and organize the huge number of files: deleting files is a serious decision to take. We made a copy of the corpus on the new storage system, and began to clean the data. We kept track of every single step in order to be able to go back over all the changes.

Eventually, we succeeded in getting rid of one million files, mainly technical files and redundant images. The file tree also required some transformations as it was based on library names, and of course some of them had changed over this long stretch of time. We also needed to solve traditional encoding problems and special characters in both file and directory names.

The analysis of the technical metadata also showed that some files were missing. A further check showed that some other files were empty or corrupted. All these files were manually checked, and some were regenerated from a copy.

The next step was to transform all the files into JPG2000 format.

The goal was to ensure that the transformation was technically correct but also that the image was still human-readable and of good quality. The result of this test workflow on a small sample of pictures showed that different TIFF encodings caused many errors and that it was again necessary to carry out a manual check on some files.

It took us no less than two years just to do this part of the work.

4. Documenting data

Then, it was time to retrieve the corresponding metadata from various technical and scientific databases. During this process, we discovered that for some manuscripts, the identification number was not correct: it was again necessary to re-adjust the file tree. To encode metadata, we chose a mix of different standards encapsulated in METS format to describe all the metadata available: TEI for scientific metadata and XMP for technical stuff.

Lastly, we had to abide by French law on archives, which complicated matters even further.

At last, we were able to create nice packages compliant with the needs of the CINES platform based on the OAIS[5] model.

5. What we learned

To achieve this project, it was necessary to assemble a team of people with very different skills and backgrounds: it was not easy, to say the least, to make this team operate smoothly! We had on board Huma-Num's system administrator and also people from IRHT, the database experts to take care of all the relevant metadata and the manuscript photographer who also happened to be the living memory of the project, and last but not least some archivists.

It's impossible do this kind of work if you don't have access to a proper infrastructure: we had around 80 TBs to deal with, and we also needed computing power to proceed with the format migration. You can't use the same approach with millions of files as you do with a standard corpus.

The person who was the "memory of the project" was the key whenever decisions needed to be taken: this is due to the complexity of dealing with material created over a long period of time (nearly 40 years) by different persons.

Bibliography

1. **Baude, O. and Joffres, A. and Larrousse, N. and Pouyllau, S.** (2017) Huma-Num, DH Conference 2017, *Une infrastructure française pour les Sciences Humaines et Sociales. Stratégie, organisation et fonctionnement* Available at <https://dh2017.adho.org/abstracts/242/242.pdf>
2. **CINES** - Centre Informatique National de l'Enseignement Supérieur (2019), *A digital archiving solutions for long term preservation*, Available at <https://www.cines.fr/en/long-term-preservation>
3. **Jacobson M. and Larrousse N. and Massol M.** (2014), ICA/SUV Conference, *La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales),* Available at < https://halshs.archives-ouvertes.fr/halshs-01025106>
4. **Loebbecke C. and Thaller M.** (2005), ECIS conference 2005, *Preserving Europe's Cultural Heritage in the Digital World* Available at <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1000&context=ecis2005>
5. **Mounier P.** (2018), *Les humanités numériques, Une histoire critique,* Editions de la Maison des sciences de l'homme, Available at <https://books.openedition.org/editionsmsh/12006>

[1] Un manuscrit naturellement (Foreword from The Name of the Rose / Umberto Eco 1980)
[2] Institut de Recherche et d'Histoire des Textes. See https://www.irht.cnrs.fr/?q=en
[3] Huma-Num is the French national infrastructure for humanities which provides mostly digital services. See http://www.huma-num.fr/about-us
[4] Centre Informatique National de l'Enseignement Supérieur. See http://www.cines.fr
[5] Open Archival Information System. See https://fr.wikipedia.org/wiki/Open_Archival_Information_System