The Archives nationales’ data preparation workflow
Émeline Levasseur and André Falut, data migration project managers, responsible for the digital preservation strategy

Département de l’administration des données – Department of Data Administration
Direction du numérique et de la conservation – Direction of Digital Support and Preservation
Archives nationales
Organization of the archives network (national scale)

- Appraisal
- Preparing transfer
- Transferring

OR

- Archives services of the ministries ("missions")
- Archives services of the national operators, the high courts

Interministerial Service of French Archives (SIAF)

Digital repository based on Vitam

Archives nationales

Preserving, giving access...

• Ideally, preservation should be planned as early as possible... but the National Archives are not in charge of appraising and preparing data
  • Except in two cases: private archives and data migration
Constatations

• The ingestion process to our Vitam-based digital archiving platform includes format identification (Siegfried) but so far, this is the only preservation action (no metadata extraction except in the case of mailboxes, no validation)

• As a result, in our digital archiving platform:
  • Most objects are identified and readable, but...
  • Some formats are incorrectly identified (due to the limitations of the identification tools, but also because of their implementation in Vitam)
  • Some files cannot be accessed (due to being damaged, or due to being encrypted, password-protected...
Our formats policy

• Meant to provide guidelines for evaluation to the missions (archivists in ministries)

• The principle is not to attribute ratings, but rather to indicate the degree to which the National Archives can guarantee the long-term preservation of a given format

• Most formats will be accepted or tolerated, only excluding formats corresponding to:
  • files with no inherent archival value (system files, temporary files, shortcuts...)
  • or files that present obstacles to preservation (zip files, password-protected files...)
  • As databases represent a large proportion of what the National Archives collect, they also have stricter criteria regarding their formats

• The goal is not to constrain collecting, but to adapt to its reality. Our formats policy must be articulated with our archiving policy, which has to comply with legal obligations

• Validation is not appropriate for our context
Preservation issues and goals

• In the context of a massive production of digital archives and a wide range of contents and formats, ensuring a certain level of quality at ingest without impeding collection

• At the very least, the files and data we ingest should be:
  • Identified
  • Readable and intelligible
  • Any damage to their contents should be extensively documented

• These goals must be taken into account before ingestion, and therefore during data preparation (ideally as soon as archives are received, at the very latest before SIP creation in ReSIP)...

• ... So we created a workflow to help data preparers do so easily, using a semi-automated process
AN~DROID

• Goal: allowing data preparers (mission archivists, archivists at the National Archives in certain cases) to conduct an easy and early preservation audit on a file hierarchy using a DROID report

• The process is simple and relies on widely known and used tools: DROID and Excel (macro)

• Two main steps: running DROID on a file hierarchy and exporting the results as a CSV, then using a macro on that export

• Limitations: requires obtaining the macro (which might not be transferrable by mail for security reasons) and installing DROID (not always possible, depending on security constraints as well)
Why DROID?

• Format identification issues can be symptoms of preservation issues
• But ReSIP and VITAM obscure these identification issues and therefore potential file quality problems
• A DROID report provides more accurate information, which along with other elements also featured in the report (size, names, extensions...) can be used for a first assessment of file quality
• The format identifications themselves can be compared to our formats policy
• The "AN~DROID" macro-enabled Excel sheet partly automates this analysis of the DROID report.
The end result is a report on:

- Empty files and folders
- Files with no archival value: system files, temporary files (with a ~$ prefix), shortcuts (PUID x-fmt/428)
- Compressed files (identified from PUIDs)
- Extension issues: no extension or extension mismatch
- Unidentified files, or files with multiple identifications (symptoms of readability issues)
- The status of the identified file formats in the National Archives' formats policy (accepted, tolerated, rejected, requiring further analysis, missing)
- The macro also highlights duplicates on the basis of their hash (if available in the export)
Demo!
As part of a wider reflection on data preparation best practices, we have been gathering the issues data preparers might encounter, ways to detect them when auditing a file hierarchy and potential fixes in a "Little Data Preparer’s Handbook", as a reminder.

The Excel AN~DROID sheet provides a semi-automated way to detect these issues in a single audit, but these checks can also be performed manually.

For data preparers who might not be able to run DROID, the Handbook provide alternative auditing tools.

It also indicates solutions for each problem the audit might reveal.

These solutions can be implemented using various tools (Windows explorer, ReSIP, Powershell commands...) and at different stages (before or while the files are processed in ReSIP).
Conclusion

• Our data preparation workflow attempts to answer a particularly complex collecting situation, within the specific context of the French public archives network

• But we are convinced that it could prove useful in other contexts with some adaptations, particularly since DROID is widely used internationally

• Some archivists (in mission or at the National Archives) have started using it and we received very positive feedback, but reflections and experiments are ongoing
Further prospects: SOFOCLE

• To go beyond what the DROID report might reveal, development of SOFOCLE, a Python executable aiming to check the readability and accessibility of files (for certain formats)

• Allows to perform these checks in cases where volumes make it impossible for archivists to manually open every potentially problematic file

• At this point, for internal use only

• Has proven useful, but has limitations (attempts to open files on the basis of their extension, flags issues using categories which might have different meanings depending on content types...)

• Raises theoretical questions we don’t have firm answers for regarding the definition of these categories (corrupted files, inaccessible files)
Thank you!

Émeline Levasseur
emeline.levasseur@culture.gouv.fr

André Falut
andre.falut@culture.gouv.fr

Site of Pierrefitte-sur-Seine
59, rue Guynemer
93380 Pierrefitte-sur-Seine