It Takes a Whole Village to Define a Preservation Strategy

Bertrand Caron, Alix Bruys, Thomas Ledoux, Jordan de La Houssaye

IT TAKES A WHOLE VILLAGE TO DEFINE A PRESERVATION STRATEGY

Formalizing Policies on Data Formats Normalization at the National Library of France

Alix Bruys
Direction of Collections
Bibliothèque nationale de France
Paris, France
alix.bruys@bnf.fr

Bertrand Caron
Department of Metadata
Bibliothèque nationale de France
Paris, France
bertrand.caron@bnf.fr

Thomas Ledoux
Department of Information Technology
Bibliothèque nationale de France
Paris, France
thomas.ledoux@bnf.fr

Jordan de La Houssaye
Department of Information Technology
Bibliothèque nationale de France
Paris, France
jordan.de-la-houssaye@bnf.fr

Abstract – After publishing its policy on data formats for digital preservation, the National library of France (BnF) had to formalize its method to deal with collected data that did not meet its requirements. This paper describes several significant examples that led BnF from preconceptions to pragmatic decisions upon normalization and preservation strategies for content that could not be ingested as is. Collective intelligence was highly required; this paper is also intended as an attempt to identify which conditions made it possible to emerge between experts, collection managers and process managers.

Described cases tackle issues with PDFs with protection, 48 bits images, PSD files, PDF transformation to JPEG and Final Cut Pro projects. These cases helped define empirically a method, still a work in progress, briefly presented in the last part of the paper.

Keywords – normalization, data formats, preservation strategy, collaboration.

Conference Topics – collaboration; exchange.

I. PREVIOUSLY, ON THE BNF FORMATS WORKING GROUP...

Enters the whole working group, guards standing at the door

The National library of France (BnF) started collecting born-digital content at scale six years ago: donated and acquired texts and still images since 2016[1], ebooks and sound obtained by legal deposit since 2019. Since then, it strives to take the full measure of the differences between digitized and born-digital documents in terms of Quality Assurance (QA), preservation and dissemination.

This is why, since 2018, BnF has reactivated its activity of studying data and metadata formats for the preservation of digital information. As described in an OPF blog post [1], the dedicated working group, named "Groupe Formats de données et de métadonnées pour la préservation numérique (quickly abbreviated "Groupe Formats", in English "Formats Working Group") faced in 2017 a need for continued monitoring of data formats in the context of increasing flows of born-digital content.

The working group is composed of around thirty members working in specialized departments (Engravings and Photography, Performing Arts, Audiovisual, Maps, Music) and in support departments (Information Technology, Preservation, Metadata, Cooperation, Images and Digital Services, Institutional Archives). To gather this team, knowledge of specific content types was requested from different BnF organizational units, expertise was identified in some individuals, and participation from collection departments was demanded.

[1] Dates correspond to the ingestion of the first Information Package in the digital preservation repository.

iPres 2022: The 18th International Conference on Digital Preservation, Glasgow, Scotland.

DOI: 10.1145/nnnnnnn.nnnnnnn

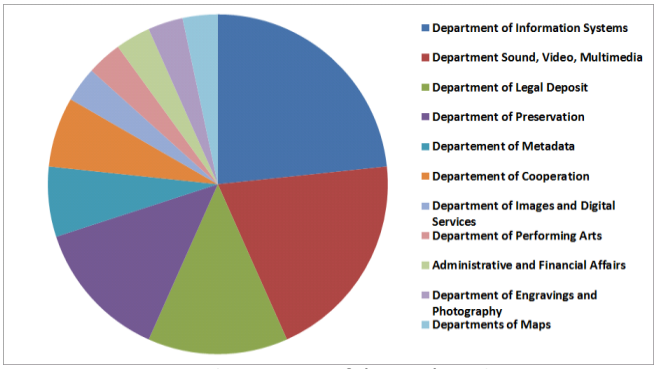

Figure 1. Composition of the Working Group

From 2020 on, the working group's mission has been to publish a revised, justified and accepted formats policy. It was officially released in October 2021, after a five-month review period [2].

This reference document determines and justifies the choices made by BnF in terms of data formats: which data formats it accepts, which properties it extracts from the data, by which tools and how it compares such properties to its requirements. It also started scratching the surface of a difficult question: how does BnF act when it gets data in a format (or with properties) that does not correspond to its standards? This paper reports BnF's efforts to structure its practices and policies one step further.

In this situation, preliminary negotiation with the Producer is preferred, but whenever this is not possible, one among four options must be chosen:

1. Simply refusing the accession of the content;
2. Requesting a new transfer from the Producer;
3. Accepting the content as is and changing the QA, preservation and dissemination environment to take the new data format into account;
4. Transforming the content in order for it to comply with BnF's requirements.

Each of the next five sections will present a real world use case, how it enriched BnF's policy, and/or how this policy in turn informed the Formats working group in order to address the problems at hand. Because these use cases were far and wide across the range of BnF's activities, the paper intends to show that the diversity of the working group was not only useful, but necessary.

The last section of this paper describes the methodology that emerged in this process.

Note: each section is mischievously introducing actors of the preservation operations in the scenery, identifying them by their first name.

II. Refusing, in The Name of The Formats Policy

Featuring Olivier (collection manager), Alix (process manager), Thomas, Jordan & Bertrand (preservation experts).

In this first use case, Olivier (a collection manager from the Maps and Plans Department) wanted to acquire a simple cartographic document, in the form of 11 PDF files constituting the different parts of an atlas. In the end, he had to give up the acquisition of this resource, despite its value for the BnF collections.

These files were acquired in May 2021, in a context where we couldn't negotiate neither the format nor the rights associated with this set of files. This will rapidly prove important to consider.

At BnF, when documents enter our collections, we try to confront as soon as possible the properties of the files received with the BnF standards. This comparison is first handled by a visual assessment of the documents which exposed no problem. Then an internally developed tool called "Frontin", which retrieves characterization metadata (extracted by Apache Tika and JHOVE, as far as PDF is concerned) and issues an alert in case of properties different from those expected. In this case, Frontin first called Tika, at the time in its 1.12 version (slightly out of date at this time), and reported the following error:

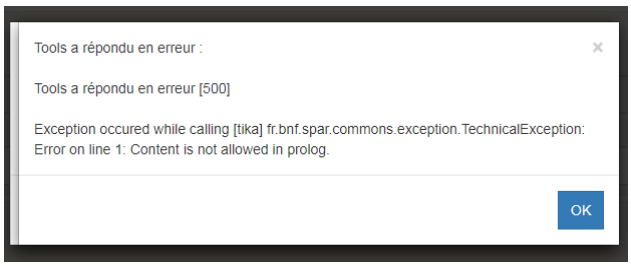

Figure 2. Error reported by Tika

Alix, the Digital Donations and Acquisitions process manager then sought to refine the advice rendered by Frontin, parsing the files with JHOVE (version 1.12.1), which brought up a "Compression method is invalid or unknown to JHOVE" error. The working group was asked to refine the diagnosis which led Jordan, a preservation expert, to speculate about the presence of TPMs (Technological Protection Measures) as the cause of these errors. It was indeed the case. However, BnF's policy is to accept documents in PDF format as long as they do not contain TPMs (see [3] or [4]). Indeed, TPMs add a layer of complexity to the content that is not tractable in the near future. They jeopardize the accessibility of the content and impede the use of the migration strategy in order to preserve the content.

Subsequently, Thomas (one of the preservation experts) recommended that BnF consider not transforming the file. Indeed, the hypothesis of removing TPMs did not seem clearly authorized, at least as opposed to the case of legal deposit where the absence of TPMs is legally required. The other possibility, which would have consisted in "printing" the file in an image format, would have resulted in the loss of significant properties by going from vector information to a raster image. The re-delivery of the

file without TPM by its producer was therefore to be preferred, but this proved impossible due to a lack of respondents. The final decision was therefore to abandon the processing of the document, by eliminating all other possible options in case of data that did not comply with the BnF's format policy.

Note that, as both analysis tools failed to return an explicit error message, the BnF digital preservation "village" joined on this occasion its efforts to the international community, as Bertrand submitted an issue to Apache Tika developers[2] and supported a similar issue in JHOVE[3]. In the case of Tika, TPMs were already better recognized by a new version we had not yet implemented. It turned out however that our issue allowed Tika's developers to correct a bug concerning Open Document formats.

This use case reveals several fundamental aspects of the implementation of a format policy, common to all digital entries. It confirms the importance of making a reliable and understandable diagnosis when files arrive. This diagnosis is facilitated by the use of up-to-date and explicit analysis tools that combine identification, characterization and validation tools and synthetize it in a simplified form. But the diagnosis is only complete after an analysis by a human. It is primarily the responsibility of the process manager, whose role is to ensure that the data entering the process is suitable, natively or after normalization, for access by BnF readers in a permanent manner.

In more complex cases, the process manager solicits and connects different expertises. This use case also makes it possible to evoke the involvement of preservation experts at a very early stage: solicited by the process manager, they refine the diagnosis, evaluate the feasibility of data normalization and accompany the collection manager in the decision regarding the fate of the files received.

III. Changing The Environment Instead of The Content?

Featuring Rime (collection manager), Thomas (digital production coordinator), Yannick (product owner), Anne (image signal specialist),

A new challenge was faced when Rime, the collection manager, wanted to process a set of photographs from Brigitte Pougeoise's collection, acquired in 2014. This collection was a mix of ordinary JPEGs as well as TIFFs coded in 48 bits (3 color channels coded with a depth of 16 bits). Our current policy, based on what we can process and what we can give access to, is limited to the more common 8-bit depth. Two approaches were considered: either we expanded our policy or we transformed the content.

Knowing that contemporary practices lean toward better resolution in capturing images, Thomas, the coordinator, explained that the acceptance of these files entailed a revision of our policy, even though this would mean an important evolution of the whole digital environment. Not only the parameters of the assessment tools should be adapted, but more profoundly the whole chain of ingestion and access should be modified in order to take full account of the accuracy of the image (for us, this is a change as important as going from TIFF to JPEG2000 to process images). Yannick, the product owner, was the one who could measure when such modifications could be carried out and what would be the consequences of this choice.

However, after convening the working group, Anne, the image signal specialist, was able to detect that the use of 16-bit depth was merely an artifact of the post-processing of the images by the photographer. This fact was correlated with the camera model (as described in the images metadata) which would not have been able to encode images in 16-bit depth, as well as the analysis of the color histogram which shows that not all of the space available for encoding had been used. A careful transformation to a more typical 8-bit depth image was then deemed possible by following the usual decision workflow for designing such a transformation[4], as described in Section V. In taking the decision, the group was helped by the notion of "preservation intent"[5]. The 16-bit depth was not used to capture a richer image, nor was it intended to express a richer image. It was only used in the post-processing of the image, never to be shown. Therefore, should we try to preserve this particular property of the image, our preservation intent would not align to the artist's intent.

Even though it is not this case that will make us change our processing environment, we are fully aware of the rapid evolution of digital practices, thanks to experts of this domain such as Anne. It is not up to us to avoid it but to be able to take it into account at the right moment and to invest in new formats when they become mainstream. This means that the experts should remain fully vigilant and connected to their communities. The technology watch activity, as described "Preservation Planning" entity of the OAIS [6], is a permanent activity which must enable us to regularly update our policy and leave us sufficient time

[2] "Return a more informative error when trying to parse encrypted ODT", issue 3331 on Apache Tika, available at <https://issues.apache.org/jira/browse/TIKA-3331> (accessed on March 4th).

[3] "Report a more informative error message for encrypted PDFs", issue 640 on JHOVE, available at <https://github.com/openpreserve/jhove/issues/640> (accessed on March 4th).

[4] It should be noted that the actual procedure is not yet decided at the time of writing this article.

[5] This notion is being developed by the digital preservation community for several years. See in particular [5].

to make the necessary changes to our processing environment.

Indeed, there is a balance to be found between restricting ourselves to the available formats we already know about and accepting all the particularities that can be thrown at us. It's not just about having a trustworthy environment that does not distort reality; it's also about sustainability where we couldn't cope with the countless forms of creation.

Again, such a decision is only made possible through teamwork where various expertise can be brought together to evaluate the cost of the developments, the content itself and the preservation intention of the creator and of the institution.

IV. Transforming Content: Multiple Choices for Complex Contexts[6]

Featuring Sandrine (collection manager), Chloé (collection manager), Rime (collection manager) Bertrand (preservation expert), Anne (image signal specialist).

Three different collections, received as donations by BnF, had in common the fact of having a strong component of digital images intended for consultation in Gallica[7], the BnF digital library, mostly in formats mastered by BnF (PDF, TIFF and JPEG). These three collections also included some files in PSD format, which is the proprietary format created and used by Adobe for its Photoshop suite. Because this format is proprietary and undocumented, our first intent was to consider a migration for these files. However, because these collections differ in the nature of their content, these PSD files had to be treated differently. Here are the main characteristics of these collections:

- The Philippe Apeloig collection documents the creation of posters by the graphic designer Philippe Apeloig for the book festival in Aix-en-Provence between 1997 and 2015. The collection is hybrid (printed and digital materials) and contains about 300 digital sketches and about fifteen source files of the final printed poster; 3 PSD files are represented among the digital sketches.
- The Amos Gitai collection gathers archives of the film *Rabin, the Last Day* including nearly 2000 photographs of the shooting; 3 PSD files are present among them.
- The Michèle Laurent collection is composed of a hundred photographs of the actor Philippe Caubère's performances, including some digitizations of book covers; 7 PSD files are present among the scanned images.

After eliminating the other options (request a redelivery, exclude the contents), a study was initiated to define the preservation strategy for these PSD files, gathering Sandrine, Chloé and Rime, the collection managers concerned, Bertrand, the preservation expert and Anne, the image specialist. To begin, Anne shared her knowledge of the PSD format and the expected uses of the software that produces it. She also revealed the use that had been made of it in the three use cases, according to the properties of the different files after opening them with Photoshop. Subsequently, the working group used the "in-house" method, described in the policy document[8], consisting in analyzing each use case according to a grid of criteria, structured by three questions:

- Is it necessary to transform the received data?
- If so, in which format?
- Should the source files be retained?

The choice of transforming the data, instead of accepting them as they are, was quickly made for three reasons. First, BnF did not wish to invest in the preservation of a proprietary format. Second, we didn't have the evidence of an intentional technical choice from the data producers. Third, it was necessary to integrate these PSD files into image batches with other formats.

Once this decision was made, the choice of a target format required further investigation, using three criteria relevant to these use cases, taken from the grid defined in the policy document. These three criteria were as follows:

- Format category: identification of a preferred format for the type of content concerned, if applicable.
- Consistency within the information package or the collection: identification of the formats present in the information package or the collection, to be preferred in case of multiple preferred formats.
- Preservation of significant properties or functionalities: definition of a preservation intention, i.e., the set of informational properties and usage modalities of a digital object to be preserved over the long term for a community of users.

In the case of the Apeloig collection, Sandrine, the collection manager, wanted to offer Gallica users the possibility of consulting the information content of the sketch as part of a batch presenting the successive explorations of the graphic designer. To meet this intention (to show "flattened" image content in Gallica), JPEG was chosen as the target format, even though Anne, the image signal specialist, recommended TIFF as the best option for capturing the maximum amount of information contained in the PSD. In the case of the Gitai collection, JPEG was also chosen, but for slightly

[6] See the BnF blog post [7].

[7] Available via https://gallica.bnf.fr.

[8] See [2], p. 19.

different reasons: on the one hand, because Rime wished to privilege access to the visual content like Sandrine, but on the other hand, because Thomas and Bertrand had noted the presence of JPEG files with identical naming, suggesting that JPEGs were the source of PSDs. The proximity of nearly 2000 other photographs in JPEG format also weighed in the decision. For the Laurent collection, there was no doubt that the images were the result of a scanning process. Bertrand therefore advocated the formats retained in the BnF format policy, namely uncompressed TIFF or JPEG 2000. TIFF was finally chosen, because of the exclusive presence of this format in the rest of the collection.

The study also included whether or not to keep the PSD files after they were transformed into the target format. For the Apeloig collection, Chloe, collection manager, wanted to keep all traces of the designer's creative process, including layers and editing history. For the Gitai collection, on the other hand, the files contain layers but are not activated, which makes the PSD format less relevant to these files. For the Laurent collection, the files contained no trace of modifications, which made the PSD even less relevant. Nevertheless, the source files were kept, because they belonged to research-level collections, but also for more pragmatic reasons of prudence and low cost (due to the small number of files involved).

Through three similar and simultaneous use cases, we have experimented with the fact that the choice of a target format is not the result of a miracle recipe. In particular we learned that one cannot simply choose a destination format for a migration based on the source format.

In the field of still images, the BnF's format policy had retained preferred formats for images resulting from digitization or for edited digital photographs, but had not yet pronounced itself, for lack of cases, on images in their production stage.

In the end, these cases did not lead us to change our policy: the presence of PSD files in these collections was too anecdotal, and sometimes not even significant. These cases have taught us how to manage the exception in the search for homogeneity of information packages.

V. Transforming Content: Which Method & Tools to Use?

Featuring Sandrine & Bérenger (collection managers), Alix (process manager), Thomas & Bertrand (preservation expert), Anne & Patrick (image signal specialists).

Another case arose with the aforementioned Apeloig collection. The digital assets were of two different kinds:

- Final version of the poster, ready to be printed, in PDF;
- For each poster, several sketches successively made. These files were in different formats: PDF, TIFF and JFIF.

The sketches of the same poster, gathered in the same Information Package, had to be normalized; indeed, BnF policy requires that files with the same use in the same Information Package be in the same format. Sandrine's (the Apeloig collection manager) intention was that only the final version would be reprinted for an exhibition. She considered that the interest of the sketches was limited to documenting the creative process. The informational preservation approach[9] allowed for a transformation to image format, while retaining the original PDF files.

Thomas noted that the PDFs of the sketches contained some superimposed elements (text, sometimes transparent graphic elements). In order to transform the PDF sketches into JFIF images it was therefore necessary to opt for a rasterization solution instead of a simple image extraction.

A short list of object properties has been determined by Bertrand[10] in order to judge the result of a transformation:

- Definition (width and height of the image in pixels);
- Weight (in bytes);
- Resolution (number of pixels per size unit)
- Dimensions (size in centimeters / inches, depending on the definition and resolution);
- Visual quality (estimated visually by the image signal specialist).

These criteria were completed by some others regarding the software tools[11]:

- Availability of the tool (free or not, deployment on BnF standard workstations, price);
- Implementation mode (CLI / GUI);
- Possible automation of the tool.

The correct treatment of certain components of the object was also considered:

- Color profile management;
- Presence of internal metadata.

To determine which method and tool would be most effective, the group compared the proposals of several of its members. These proposals came from Thomas and Bertrand, preservation experts, from Anne, an

[9] For a definition of informational vs. artifactual preservation approaches, see [8], p. 15 sqq..

[10] This list is a subset of the properties proposed by [9].

[11] The distinction between criteria for evaluating transformation consequences and criteria for transformation process is inspired by [10].

image signal specialist, but also from Bérenger, an audiovisual collection manager. The following tools were evaluated:

- PDFCreator[12], a tool deployed on all BnF workstations and with a GUI;
- XnView[13], also available on all BnF workstations, with or without the help of Adobe Reader;
- pdftoppm[14], a tool found thanks to Johann van der Knijff's excellent list of PDF processing tools [11], which by coincidence was published at the time of the study;
- PDFBox[15], already used in BnF processes to generate thumbnails from PDFs for digital books;
- Photoshop[16], a tool favored by image signal specialists.

Resulting files were examined by Anne and Patrick, our image signal specialists.

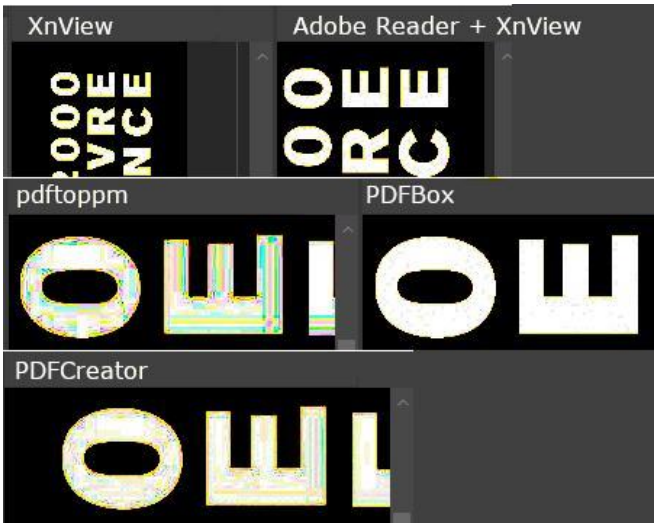

Figure 3. Visual comparison of the transformation to JPEG image.

For each tool, the method was recorded:

- Step-by-step instructions, possibly with screenshots, for GUI-driven tools;
- Command line for CLI-driven tools.

One important result of such a process is a publishable, justifiable and reproducible, though always questionable, policy. In case such a situation occurs, BnF determined that "born-digital" PDF such as those in the Philippe Apeloig collection will be processed by PDFBox, a tool capable of rasterization, while a PDF resulting from a digitization, containing only one image per page, will be processed by an extraction tool such as Apache Tika[17].

PDFBox was then added to our preconditioning tool, Frontin, to handle automated transformations; moreover, shortly after, and following the appearance of a new use case, Thomas studied the automatic distinction between these two types of PDF [12].

This process also showed that comparing results in a working group plenary session had pedagogical virtues. The diversity of the results obtained demonstrates that not all conversions are equal. Moreover, it proves once again that two objects of different nature can be recorded in the same format, and that the strategy adopted will depend on the object nature.

Some organizational issues emerge: the choice of a method cannot omit the "human resources" dimension: depending on whether one chooses a tool with a GUI or only a command line, the personnel capable of implementing the transformation is not the same. This consideration is all the more important as the transformation of born-digital content is time-consuming, for signal specialists as well as for collection managers, who currently tend to consider that these operations are not, or not exclusively, of their responsibility.

VI. When There Is No Identified Target Preservation Format Yet: Creating Dissemination Surrogates

Featuring Jean-Yves (audiovisual expert), Rime (collection manager), Bertrand (metadata specialist).

The ultimate challenge arises when we receive material that is not only not currently accepted, but whose formats are either proprietary or require specific hardware. One such recent example comes with the film daily rushes[18] from the FCP (Final Cut Pro) program. This software is one of the classic tool for filmmakers but it is completely tied to the Apple platform and has already undergone one breaking change with version X which chooses an XML-based representation and force the use of a third-party utility to migrate to the new version[19].

In the donation we have daily rushes in FCP7 as well as FCP-X format. Neither of these can be read in an ordinary workstation in the library and the management of such files requires specific competences. Moreover,

[12] Adobe PDFCreator, PDF converter, https://www.pdfforge.org/pdfcreator.

[13] XnView, free software to view, edit and resize images, https://www.xnview.com/.

[14] Poppler pdftoppm, PDF converter to image files, https://www.mankier.com/1/pdftoppm.

[15] Apache PDFBox, open-source library for handling PDFs, https://pdfbox.apache.org/.

[16] Adobe Photoshop, raster graphics editor, https://www.adobe.com/fr/products/photoshop.

[17] Apache Tika - a content analysis toolkit, https://tika.apache.org/.

[18] Daily rushes are the raw, unedited footage shot during the making of a motion picture (definition taken from Wikipedia).

[19] Refer to https://support.apple.com/en-us/HT208054 and https://en.wikipedia.org/wiki/Final_Cut_Pro_X

the edit decision list[20] contained in the central file makes direct references to other files (the raw audio or video parts) with absolute paths. The first manipulation that requires the use of the software and a well-equipped hardware is to recreate the links with the new installation. In order to try to figure out how we can manage this material, we first look for an expert (fortunately, there are knowledgeable people in the audiovisual department) and wait for a compatible hardware workstation. Having both of them provides us the ability to better understand the material (delimited all the files involved in a FCP project) and try to figure out the main piece of information.

Even though it was clear from the beginning that we would have to store the information as is and provide a basic bitstream preservation, we also intend to provide in an easy manner to our users some sort of substitute. Indeed, we don't view our preservation system as a dark archive but more like a repository of information that needs to be accessible as far as the legal restrictions permit us. In this case, because of the kind of material, a direct access through our digital library, Gallica or its version accessible only in its precinct, Gallica Intra Muros, is not envisioned but we intend to provide enough information so that the researchers know if the material is of interest to them.

In the case of film daily rushes, we are willing to provide a list of the material involved in the making (images, sound recordings, video footage) as well as the images of the timeline. Those advanced descriptions of the original material will be used as a surrogate for the original material. It allows us to give access to certain information in a simple way and, if necessary, to accept justified requests for communication that would involve the installation of specific equipment and the associated logistics.

For practical reasons dictated by our preservation system, we intend to ingest the original material and their surrogate in two different Information packages, probably at two very different times. From the preservation point of view, this is the first time we intend to ingest both an original and the result of a migration in two different packages. Usually the two representations are archived together and the relationship between what constitutes an original and a master is stated in the package. Moreover the migration itself can be described in the provenance metadata. This allows us to apply a strict policy for the master version (target of the migration) and a less strict one for the original (source). Here, we will need to ingest the FCP project as a master, even though we have no control on its format whatsoever. This implies lowering the bar of entry so much for this case that any kind of data could enter our systems afterwards, which we do not want to happen.

Therefore, once the decision of acceptance has been made, the original material is stored in a specific location and documented so that the intention for migration is clearly stated and the reason and needs formalized as much as possible. A complete documentation of our level of knowledge is written and the risk associated with a possible loss of control is stated: proprietary format, hardware specificities, legal issues... A PREMIS Event [13] of type `migrationIntended`, informs about it:

```
<premis:event>
  ...
  <premis:eventType>
  migrationIntended
  </premis:eventType>
  ...
  <premis:eventOutcomeInformation>
    <premis:eventOutcome>
    type=transformationWithBackup,
    sourceUse=master,sourceFormat=fcp
    </premis:eventOutcome>
  </premis:eventOutcomeInformation>
  <premis:linkingAgentIdentifier>
    <premis:linkingAgentIdentifierType>
    documentCode
    </premis:linkingAgentIdentifierType>
    <premis:linkingAgentIdentifierValue>
    BnF-ADM-2021-012345
    </premis:linkingAgentIdentifierValue>
    ...
  </premis:linkingAgentIdentifier>
</premis:event>
```

In our preservation system, we will implement a rule of a new kind stating that these files are allowed, but only if a `migrationIntended` event is attached. Therefore the existence of this case in our collections will be exposed.

In parallel, a surrogate is built that provides as much information as possible using only managed formats: it can be screenshots, a video of the representation of the material, part of it. This surrogate is directly linked to the original. Again, if the original is preserved at the bitstream level, the surrogate is meant to be enriched as we gain more information about the original material or find new ways to provide access to it.

```
<premis:event>
  ...
  <premis:eventType>
  migrationProcessed
  </premis:eventType>
  ...
  <premis:eventOutcomeInformation>
    <premis:eventOutcome>
    type=transformationWithBackup,
    sourceUse=master,
    sourceFormat=fcp,
    targetFormat=jpeg,
```

[20] An edit decision list contains an ordered sequence of audiovisual material used in a film editing project.

```
    satisfactionLevel=poor
  </premis:eventOutcome>
  </premis:eventOutcomeInformation>
  <premis:linkingAgentIdentifier>
    <premis:linkingAgentIdentifierType>
    documentCode
    </premis:linkingAgentIdentifierType>
    <premis:linkingAgentIdentifierValue>
    BnF-ADM-2021-012345
    </premis:linkingAgentIdentifierValue>
    <premis:linkingAgentRole>
    performer
    </premis:linkingAgentRole>
  </premis:linkingAgentIdentifier>
  <premis:linkingObjectIdentifier>
    <premis:linkingObjectIdentifierType>
    ark
    </premis:linkingObjectIdentifierType>
    <premis:linkingObjectIdentifierValue>
    ark:/12148/m0n4rk
    </premis:linkingObjectIdentifierValue>
    <premis:linkingObjectRole>
    source
    </premis:linkingObjectRole>
  </premis:linkingObjectIdentifier>
</premis:event>
```

As you may understand, even for complex or unreachable material, the preservation process starts at the beginning by capitalizing on the information available to us and seeking skills either inside or outside the library. Even if our grasp is weak, we do not intend to bury the material but on the contrary to make it visible by cataloging it, preserving it and providing access to a direct surrogate for it. In this way, we hope to be able to monitor it and possibly find innovative means of access. The mere fact that we record all this material could be an incentive to seek sponsorship or to consider a research program on it.

VII. Which Regular Process Emerged from These Experiments?

Featuring Benjamin (functional analyst), Anne-Lise (collection manager)

A. Vocabulary

Benjamin: "In the triage and appraisal application we are currently developing, what should we call the operations that change the bitstream of objects we want to accession, prior to ingestion?"

Working together between people of different backgrounds implies agreeing on a common terminology. Thus the working group had to recommend a term corresponding to a "preservation operation carried out before ingesting into the preservation system and resulting in the modification of the bitstream". The candidates were the terms "migration", "conversion", "transformation" and "normalization".

The term "transformation" was preferred in the dialogue between different BnF entities. It corresponded indeed to a term defined by OAIS and was generic enough to be understood by all. In international writings, the term "normalization" is also used, according to the generally adopted meaning.

On the other hand, were rejected:

- "Conversion", which was too restrictive because it suggested a change in container format, whereas the operation could affect the signal alone (a change in color model from CMYK to RGB, for example);
- "Migration", which evoked a migration of the system or the supports for the computer specialists;
- A variant of the previous one, "format migration", because it is not the format which is affected but the content.

B. Roles and missions

Although the decisions taken on the occasion of the various cases cited above are always questionable, they are indisputably better than those that the members of a single BnF department could have taken. But what are the profiles and skills of the agents involved in these decisions?

Four main profiles stand out today among the members of the Formats group, from the perspective of analyzing and processing natively digital objects before they enter the preservation system:

- The **collection manager** knows and understands the institution's documentary policy, the context of content creation, and maintains contact with the creator; they selects the content to be acquired by BnF, defines the intention of preservation, makes an informed decision on the acceptability of the content (appraisal) and on its technical and bibliographic treatment with the help of diagnostic tool(s), and justifies and documents these decisions, in agreement with the process manager. *Note: the collection managers were originally seen as relays for the working group's recommendations in their departments; it turned out that no decision could be taken without them!*
- The **process manager** is responsible for the overall operation of the circuit, from the controls carried out by the QA services to the dissemination; they leads a community composed of the profiles mentioned above, ensures the coherence of the decisions taken by the collection managers, makes sure that the collections deposited are accessible, formalizes and expresses the needs of these communities to the preservation experts.
- The **preservation expert** has skills on data and metadata formats (especially internal), on analysis tools, on the functioning of the preservation system; they analyzes the feedback from the diagnosis tools, they makes sure they are updated, they eventually makes them evolve, they defines the controls to be put in place in the preservation

system or upstream, and helps documenting the transformation methods.
- The **signal specialist** has skills in editing and transforming signals - in OAIS vocabulary [6], this generally corresponds to "Content Information" -, on the uses and practices of the creators of these contents; they carries out complex transformations, evaluates the methods and results of a transformation, and helps documenting the transformation methods.

A gap in this organization remains: there is no profile that takes care of simple transformations. BnF "digital stacks managers'" role is currently limited to the preservation system perimeter; as normalization takes place before ingestion, they are not engaged in this process yet.

Note that the organizational logic presented above is empirical and derived from the use cases described in the article. Eventually, a more thorough analysis of the missions and the skills required to carry them out, as well as the integration of these elements into job descriptions should be carried out. We could then rely on multiple works from the digital preservation community such as the DigCurV initiative [14].

C. Modeling a Regular Normalization Process

These cases forced BnF to reflect on the decision-making processes and the means of documenting them, in order to show how the documentary choices condition the technical decisions.

The normalization process was therefore defined as follow:

1) **Diagnose**. The diagnosis stage consists of determining whether the content as received by BnF can be deposited in the form of the file currently in its possession. It consists of comparing the properties of a file using analysis tools (characterization) with those of the preferred and accepted formats by BnF for a given context and with the rules for constituting the package.
2) **Decide**. If the file is not in one of the formats acceptable to a given channel, decide what to do with the contents. It is necessary to make a choice between:
   - Rejection of the file, and therefore of its content (as described in section II);
   - Identification of another form of the digital representation or request for a new delivery after transformation by the Producer;
   - Acceptance of the file as it is (this option implies adapting the ingestion, preservation and access environments (as described in section III);
   - Transformation carried out by BnF to meet its own requirements (as described in sections IV and V).
3) **Study**. If the last option was chosen, determine whether an existing preservation strategy applies; if not, define a suitable transformation method: software tool, parameterization, implementation method.
4) **Perform**. Implement decisions taken in the previous step.
5) **Control**. Verify that the file produced complies with BnF's deposit and preservation requirements, and that the significant properties and functionalities of the content have been preserved during the transformation.
6) **Document**. Keep track of the transformation operation and, if a new study was needed, define BnF's policy in the form of a preservation strategy.

D. Documentation

Anne-Lise: "But how do we keep track of these decisions? We chose the other option one year ago… How can we improve consistency?"

Having noticed conflicting decisions for which the reasons were unclear, collection managers emphasized the need to document the transformations. The documentation process is linked to the transformation process described above, in the following way:

1) **Diagnosis and decision stages**: upon receipt of a homogeneous set of contents that do not comply with BnF's format policy, a **diagnosis and decision form** is created, documenting the nature of the contents, their production history, their use by the Producer, the collection manager's preservation intention, the identification of their format, the analysis of the set according to the criteria grid in the policy document,[21] and the appraisal decision. The form is filled out by the collection manager assisted by the process manager and possibly by preservation experts.
2) **Study stage**: If the decision concludes that the content needs to be transformed, the **list of transformations** is consulted to determine if one of them fits the case. In addition to recording the source format, the target format and the tool used, this local, non-automated "preservation action registry"[22] emphasizes the justification for using such a transformation, its objectives and the above-mentioned criteria that were decisive in choosing the transformation.
3) If in the previous stage no existing transformation is applicable, the **study stage** results are recorded in a **report** listing the criteria for evaluating the transformation process and produced data (as

[21] See [2], p. 19.

[22] In reference to the PAR international initiative [15], whose ambition is to register preservation actions across different repositories.

described in part V). The document contains a detailed description of the implementation of each solution, the choice of a method and its justification. The **list of transformations** is also updated to include the new transformation.

4) **Documentation stage**: if the implementation method is manual, a **tutorial document** to reproduce it is produced in order to guide step by step the agent who will perform it in the future.
5) **Documentation stage**: in the METS manifest accompanying each Information Package, a **comment** describing the transformation operation is added to keep track of it and inform the reader.
6) **Documentation stage**: If the transformation appears to be sufficiently mastered and broadly applicable, it is considered a validated policy and will appear in the next version of the **policy document** [2].

EPILOGUE

In the last years, the 'Formats' working group appears to have gained maturity in both technical and organizational domains. It has become clear that on preservation strategy issues no-one can take a decision alone, the right decision being the one that is both driven by librarians and informed and implemented by technicians.

Discussions happening in this working group made clear that expertise is not about developing a comprehensive knowledge on a specific domain, but rather about gathering insights from agents all around the institution and building a consensus by bringing together different points of view.

As it was recently recalled by William Killbride, "if you're doing digital preservation alone you're not doing it right" [16]!

Exeunt all softly

ACKNOWLEDGMENT

The authors would like to make a warm round of applause for the whole BnF "village". Particular gratitude goes to all members of the 'Formats Working Group' who are daily contributing to make this team a welcoming and inclusive place to learn and improve knowledge on digital preservation.

Authors owe special thanks to Chloé Perrot, Yannick Grandcolas, Rime Touil and Anne Paounov, who made particular contributions to the work described in this paper.

REFERENCES

[1] B. Caron, "A Love Letter to Formats," *Open Preservation Foundation blogs*, November 2021. Available at <https://openpreservation.org/blogs/a-love-letter-to-formats/> (accessed on February 25th 2022).

[2] *Formats de données pour la préservation à long terme : la politique de la BnF*, Bibliothèque nationale de France, October 2021. Available at <https://hal-bnf.archives-ouvertes.fr/hal-03374030> (accessed on February 25th 2022).

[3] S. Hein & T. Steinke. *DRM and digital preservation: A use case at the German National Library, in Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014*, Melbourne, Australia, October 6-10, 2014 Available at <https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf> (accessed on June 1st 2022).

[4] S. Derrot, J.-P. Moreux, C. Oury & S. Reecht. *Preservation of ebooks: from digitized to born-digital, in Proceedings of the 11th International Conference on Digital Preservation, iPRES 2014*, Melbourne, Australia, October 6-10, 2014 Available at <https://ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf> (accessed on June 1st 2022).

[5] C. Web, D. Pearson, & P. Koerben, "Oh, you wanted us to preserve that?!" Statements of Preservation Intent for the National Library of Australia's Digital Collections. *D-Lib Magazine*, 2013, 19(1/2). Available at <https://doi.org/10.1045/january2013-webb> (accessed on June 23rd).

[6] CCSDS, *Reference Model for an Open Archival Information System (OAIS)*, 2012. Available at <https://public.ccsds.org/Pubs/650x0m2.pdf> (accessed on February 25th 2022).

[7] A. Bruys, B. Caron, Y. Grandcolas, T. Ledoux & A. Paounov, "If we want things to stay as they are, things will have to change," *Open Preservation Foundation blogs*, November 2021. Available at <https://openpreservation.org/blogs/if-we-want-things-to-stay-as-they-are-things-will-have-to-change/> (accessed on February 25th 2022).

[8] T. Owens, *The Theory and Craft of Digital Preservation*, Baltimore: John Hopkins University Press, 2018.

[9] L. Montague, A. Brown, G. Knight, S. Grace, *InSPECT Significant Properties Testing Report: Raster Images*, September 2018. Available at <https://figshare.com/articles/journal_contribution/InSPECT_Significant_Properties_Testing_Report_Raster_Images/7137803/1> (accessed on March 1st 2022).

[10] F. Luan, M. Nygård, G. Sindre et al., "Using a multi-criteria decision making approach to evaluate format migration solutions", in *Proceedings of the International Conference on Management of Emergent Digital EcoSystems - MEDES '11*, presented at the International Conference, San Francisco, California, ACM Press, 2011. Available at <http://dl.acm.org/citation.cfm?doid=2077489.2077498>. (accessed on February 22nd 2022, p. 48.

[11] J. van der Knijff, "PDF processing and analysis with open-source tools", *Bitsgalore*, September 6th 2021. Available at: <https://www.bitsgalore.org/2021/09/06/pdf-processing-and-analysis-with-open-source-tools>. (accessed September 6th 2021).

[12] T. Ledoux, "Scanned vs native PDFs, how to differentiate them?", *OPF Blogs*, February 11th 2022. Available at: <https://openpreservation.org/blogs/scanned-vs-native-pdfs-how-to-differentiate-them/?q=1>. (accessed on March 3rd 2022).

[13] B. Caron, A. Di Iorio, C. Blair, L. Bountouri, R. Guenther *et al.*: *PREMIS 3 OWL Ontology: Engaging sets of linked data, in Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018*, Boston, MA, USA, September 24-28, 2018. Available at <*https://hdl.handle.net/11353/10.923631*>

[14] DigCurV Curriculum Framework, a European Union funded project, 2013. Available at <https://digcurv.gla.ac.uk/> (accessed on March 7th 2022).

[15] Preservation Actions Registry. Available at <https://parcore.org/> (accessed on March 7th 2022).

[16] W. Killbride, "Why I iPres", *iPRES 2022 blogs*, [2021]. Available at <https://ipres2022.scot/blog/> (accessed on March 8th 2022).