Vers un recensement des formats conservés : échanges, réflexions, documents de travail

CELLULE NATIONALE DE VEILLE SUR LES FORMATS

Sous-groupe « connaissance des formats existants et émergeants et définition de critères d'obsolescence et de pérennisation »

Décembre 2020

Lorène BÉCHARD (Centre Informatique national de l'Enseignement Supérieur)

Bertrand CARON (Bibliothèque nationale de France)

Marion HUMBERT (Archives départementales de Moselle)

Anne-Flora JOLLY (Ministère des Armées)

Isabelle JOSSE (Ministère de l'Europe et des Affaires étrangères)

Emeline LEVASSEUR (Archives nationales)

Erwann RAMONDENC (Ministère de l'Europe et des Affaires étrangères)

Martine SIN BLIMA-BARRU (Archives nationales)

INTRODUCTION

Les premiers échanges au sein du sous-groupe ont abouti à l'ébauche de deux types de livrables :

- le rapport, par institution, de tous les formats présents dans leurs systèmes dès lors que le nombre de fichiers dépasse une proportion des fonds conservés qui reste à définir. En effet, ce « seuil critique » dépend de l'engagement de l'institution de conservation dans la préservation de certains formats prioritaires et de la stratégie adoptée (connaissance fine de la structuration de l'information, conversion, utilisation de conteneurs, etc.).
- le recensement et l'analyse des pratiques d'autres grandes institutions de conservation en matière de critères d'obsolescence et de pérennisation.

Elaborer un tableau de recensement des formats par institution à partir d'un modèle commun supposait de convenir en amont de la granularité à adopter. En effet, les pratiques d'identification et la prise d'informations sur les formats diffèrent d'un acteur à l'autre, par les outils (wikidata, base PRONOM, type MIME), la nature des informations collectées (compression, versions…) voire l'acception du terme « format ».

Même si les travaux de recensement ont eu lieu à l'aide d'une matrice établie en commun, les résultats obtenus possèdent chacun des caractéristiques propres. Les membres du sous-groupe ont estimé dans un premier temps que montrer les différentes approches serait plus intéressant pour refléter les différents niveaux de connaissance, politiques et processus propres à chacun, avant d'envisager une harmonisation des informations recueillies.

Après les premiers essais pour compléter le tableau et utiliser les résultats obtenus, il est apparu nécessaire de scinder le document en deux : un premier tableau uniquement destiné à établir la liste des formats conservés, avec leurs caractéristiques techniques, et un second correspondant aux neufs points identifiés par le sous-groupe comme les facteurs fondamentaux de la « préservabilité » d'un format.

Le document ci-dessous présente donc la démarche adoptée par chaque institution pour remplir les deux tableaux. Chaque document comporte un tableau synthétique des formats largement utilisés et connus de chaque institution qui comprend les précisions et les caractéristiques techniques des différentes versions des formats conservés, puis un second tableau qui tente de décrire les formats au regard des neuf critères de pérennité identifiés.

Dans un second temps, les membres du sous-groupe traiteront la « longue traîne » des nombreux formats moins maîtrisés, parfois uniquement identifiés et présent chacun en assez petit nombre dans chaque institution.

Dans un troisième temps, la BnF envisage également d'exploiter les informations issues de la collecte du web, qui permettraient de dresser un panorama de l'évolution des formats utilisés sur le web depuis les années 1990. L'objectif serait de répondre à la question suivante : quelle est notre politique par rapport à ce que l'on sait de l'utilisation massive des formats en ligne ?

TABLE DES MATIERES

Introduction 2

Table des matières 3

Présentation des colonnes du tableau 4

Archives nationales 5

Bibliothèque nationale de France 5

Ministère de l'Europe et des Affaires étrangères 6

Centre Informatique national de l'Enseignement Supérieur 6

PRESENTATION DES COLONNES DU TABLEAU

Domaines (prendre la catégorisation PRONOM)	Classification des formats proposée par la British Library et disponible à l'adresse suivante : https://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/pronom_4_info_model.pdf (page 25).
Type du format	L'intitulé courant du format utilisé par l'institution pour spécifier et/ou catégoriser le format étudié et qui présente le point d'entrée retenu par celle-ci pour son analyse. Le cas échéant, le numéro de version du format pourra être intégré à cette colonne.
Extension	Extension propre au format étudié (.doc, .pdf, etc.)
Type MIME	Identifiant unique du standard de description MIME s'il existe.
PUID	Identifiant unique de la base de données PRONOM élaborée par la British Library s'il existe.
Wikidata	Identifiant de la fiche format rédigée sur Wikidata si elle existe.
Caractéristiques.	Compression, modèle couleur, format des métadonnées etc. Le cas échéant, profil d'application, caractéristiques techniques du profil d'application
Outil particulier pour le manipuler au sein de l'institution	Equipement matériel et/ou logiciel utilisé pour ouvrir, consulter et visualiser des informations conservées dans le format au sein de l'institution.
Outil pour générer un fichier (*facultatif)*	Equipement matériel et/ou logiciel utilisé pour créer des fichiers dans le format au sein de l'institution
Nombre d'objets par institution	Nombre total de fichiers ou d'unités conservés dans un format donné.
Volume en pourcentage par institution	Pourcentage représenté par le nombre d'objets conservés dans un format donné par rapport à l'ensemble des objets conservés.
Contexte de production (*facultatif*)	Modalités matérielles, logicielles et organisationnelles de création des données.
Commentaire sur l'entrée	Choix opérés dans les intitulés et les types de formats, toute autre information jugée utile pour le récolement.

Les colonnes du second tableau reprennent les critères identifiés dans le document « Définir une politique formats : les neuf critères essentiels ».

ARCHIVES NATIONALES

Nous avons passé l'outil d'identification de format DROID sur l'ensemble des fichiers numériques conservés aux Archives nationales et archivés entre 1983 et 2017.

- Nous avons travaillé sur les formats de fichiers les plus représentatifs de nos fonds. Ces formats font partie des catégories de formats bureautiques, données structurées, image, son, audiovisuel et messagerie. Ainsi ce recensement ne comptabilise pas tous les formats rencontrés, seulement les plus importants.
- Sont inclues dans le périmètre de cette analyse, toutes les archives pérennisées sur bandes LTO, sauf les fichiers issus de la numérisation des questionnaires anonymisés du recensement de la population de 1999 qui n'ont pas été comptabilisées car ces données ne sont pas pérennisées. En les excluant, nous évitons de surreprésenter les fichiers tiff.
- En plus, des fichiers archivés sur bandes LTO, certains fonds en cours de traitement ont été inclus dans le recensement. Il s'agit des reportages photographiques de la présidence de Jacques Chirac qui représentent 381 227 fichiers, mais dont 18% n'ont pas été identifiés par DROID, dû aux limites de l'outil ; les archives audiovisuelles du procès en appel Ngenzi et Barahirwa (114 fichiers), et enfin les archives audiovisuelles du procès AZF (138 fichiers)
- Une ligne du tableau correspond à un intitulé de format, ce nom étant donné par PRONOM. Un intitulé de format peut couvrir plusieurs PUID.
- DROID n'étant pas l'outil recommandé pour l'identification des formats audiovisuels, les données « nombre d'objets » ne sont pas exhaustives pour cette catégorie de formats. En revanche, afin de préciser l'identification de ces fichiers, certains ont été analysés avec l'outil MediaConch et MediaInfo par échantillonnage afin de connaître les codecs vidéo et audio utilisés.

BIBLIOTHEQUE NATIONALE DE FRANCE

Le magasin numérique SPAR compte aujourd'hui plus de 431 000 000 fichiers. Sur ce nombre, 421 000 000 environ sont dits de format « connu » ou « maîtrisé », c'est-à-dire que la BnF considère avoir investi des moyens suffisants pour disposer

- d'une documentation composée au minimum de la ou des spécification(s) et de fichier exemples ;
- d'au moins un outil de caractérisation et un de validation ;
- d'outil(s) de visualisation et de restitution adaptés, ainsi que, le cas échéant, d'outils de production et de migration ;
- du maintien d'une compétence sur le long terme sur le format.

Contrairement aux archives, la BnF utilise pour l'identification des fichiers l'outil Unix File qui lui indique un type MIME. Si ce type MIME correspond à un format connu ou maîtrisé, on analyse le fichier et on compare les métadonnées techniques obtenues aux définitions des formats connus et maîtrisés. La BnF se repose donc d'un registre de formats qui lui est propre et non sur PRONOM,

celui des archives nationales du Royaume-Uni, ce qui implique qu'une différence de granularité apparaîtra nécessairement entre les recensements des archives et de la BnF.

On a choisi d'agréger plusieurs formats connus et maîtrisés par entrée du tableau. Les formats connus et maîtrisés sont généralement plus précis (par ex. : « TIFF BnF 24 bits sans compression »).

Le comptage des formats simplement « identifiés », c'est-à-dire dont on ne connaît que le type MIME, est l'étape suivante. Il fera apparaître une « longue traîne » de formats bien plus divers. Il serait également possible, dans un troisième temps, d'analyser le contenu des conteneurs ARC et WARC collectés par le Dépôt légal de l'Internet, plutôt dans l'optique d'identifier des grandes tendances par année de l'évolution de l'utilisation des formats de fichier sur le web.

MINISTERE DE L'EUROPE ET DES AFFAIRES ETRANGERES

L'approche du MEAE sur les travaux de récolement des formats de fichier a été la suivante :

- DROID a été utilisé en 2018 sur l'ensemble des archives collectées par les Archives diplomatiques pour identifier les formats les plus importants conservés dans les fonds. La version de PRONOM utilisée était la version 93 (2017).
- Pour chaque catégorie de format, dans un premier temps, les différents PUID de PRONOM des fichiers collectés ont été relevés.
- Pour la constitution du tableau de récolement, le choix n'a pas été fait de distinguer systématiquement les formats par PUID et des regroupements ont été tentés à partir du nom du format donné par PRONOM et de certaines caractéristiques qui ont pu être identifiées. Une distinction au type MIME semblait trop large.
- Le fait d'utiliser comme point de départ le nom du format a poussé à collecter par ailleurs, la documentation et les recommandations qui pouvaient être faites par les grandes institutions autour de ces formats ce qui nous a permis d'opérer certaines distinctions. Une étude est en cours au regard des 9 critères essentiels permettant de connaître les possibilités de conservation d'un format, ce qui nous permettra d'apporter nos premières conclusions ou perspectives.

CENTRE INFORMATIQUE NATIONAL DE L'ENSEIGNEMENT SUPERIEUR

Le tableau du CINES recense le format des documents archivés dans la plateforme d'archivage PAC à la date du 31 juillet 2020.

Ces documents correspondent aux données des établissements de l'enseignement supérieur et recherche qui ont confié la préservation de leurs archives au CINES.

On peut constater que la majorité des formats recensés sont de type Image, car les fonds conservés sont largement issus des programmes de numérisation des établissements.

Comment ces résultats ont-ils été obtenus ?

Lors de leur archivage, le format de chaque fichier est identifié au moyen de l'outil DROID et validé selon différents outils (Jhove, ImageMagick, etc.) en fonction du format identifié.

La liste des formats fournie dans le tableau est exhaustive car la politique du CINES est d'avoir une liste restreinte de formats acceptés pour archivage (accessible sur facile.cines.fr[1]).

Pour plus de lisibilité, le tableau de récolement a été complété en faisant des regroupements de PUID autour d'un même nom de format (par ex. : PDF, JPEG). Le détail des PUID Pronom concernés est indiqué dans la colonne E.

Les outils utilisés pour la validation de ces formats ont également été indiqués.

[1] Centre Informatique national de l'Enseignement Supérieur, « FACILE – Service de validation de formats » [en ligne], URL : https://facile.cines.fr/ (lien consulté le 14 décembre 2020).