La pérennisation de l'information numérique : une introduction

« Digital information lasts forever – or five years, whichever comes first. »

- Jeff Rothenberg

Bertrand Caron

expert de préservation numérique, département des Métadonnées, BnF

bertrand.caron @ bnf.fr

{BnF

Plan

- Prévenir l'altération des données numériques : l'intégrité et la sécurité de l'information
- Copier, transférer : une étape périlleuse
- Créer des unités autonomes et manipulables : l'empaquetage
- Choisir, connaitre, identifier et valider ses données : les formats de fichier
- Choisir une infrastructure : le stockage – l'offre Archivage numérique BnF
- Décrire et documenter les objets, les processus, les agents, les caractéristiques techniques… : les métadonnées
- (Nommer les objets préservés pour les retrouver et garantir leur accès sur le long terme : les identifiants)

BnF

Petit sondage pour commencer

- Quels types de contenus numériques à préserver ?
  - Image fixe
  - Textes / Documents
  - Son
  - Image animée
  - Autre ?
- Numérisé ou nativement numérique ? Ou numérique sur support, dématérialisé ?

BnF

DATA LOSS HORROR STORIES

Sharing my loss to protect your data

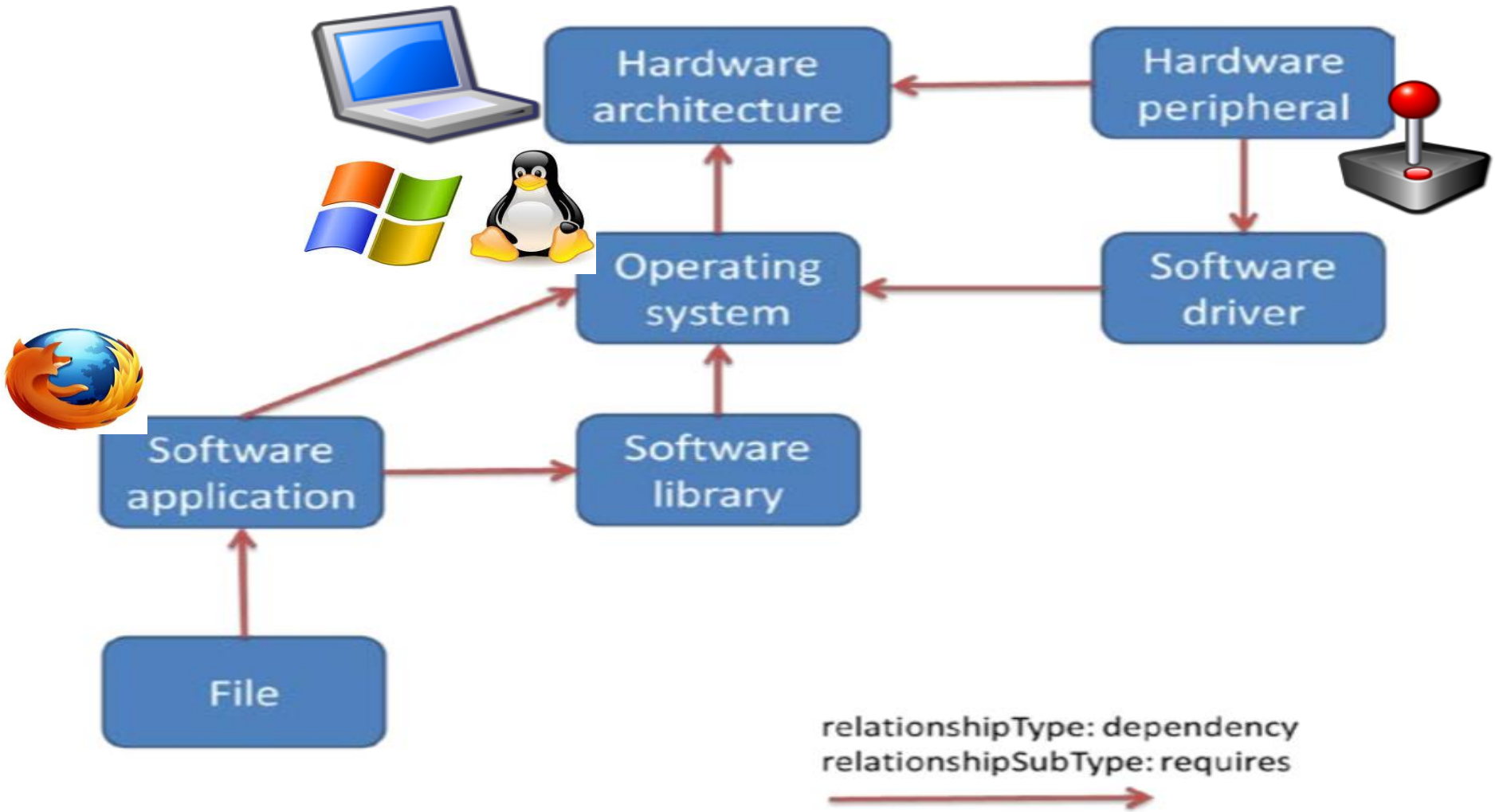
A story of unexpected data loss and how to do real preservation

Eduardo del Valle Pérez
eduard.delvalle@uib.cat

https://figshare.com/articles/presentation/Sharing_my_loss_to_protect_your_data_A_story_of_unexpected_data_loss_and_how_to_do_real_preservation/5415046/1

BnF

L'information numérique, un changement de paradigme : médiation et dépendances



BnF

De la donnée à l'information

01010000 01100001 01110011 01110011
01100101 01110010 00100000 01100100
01100101 00100000 01101100 01100001
00100000 01100100 01101111 01101110
01101110 11000011 10101001 01100101
00100000 01100010 01101001 01101110
01100001 01101001 01110010 01100101
00100000 11000011 10100000 00100000
01101100 00100111 01101001 01101110
01100110 01101111 01110010 01101101
01100001 01110100 01101001 01101111
01101110 00100000 01100011 01101111
01101101 01110000 01110010 11000011
10101001 01101000 01100101 01101110
01110011 01101001 01100010 01101100
01100101 00100000 01110000 01100001
01110010 00100000 01101100 00100111
01101000 01110101 01101101 01100001
01101001 01101110 00101110

UTF-8

« Passer de la donnée binaire à l'information compréhensible par l'humain. »

BnF

L’information numérique, un changement de paradigme : uniformisation du support

Image par Annie Spratt de Pixabay

Ce sont des données géographiques.

lausanne.gpx

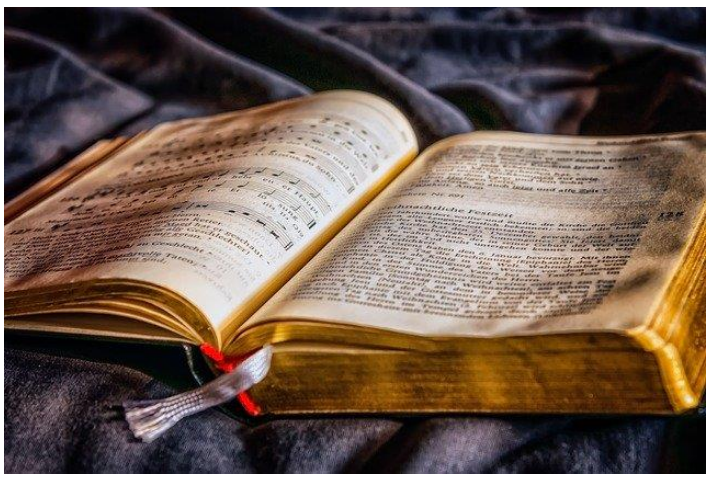


Image par Peter H de Pixabay

Ce sont des données textuelles.

miserables.epub

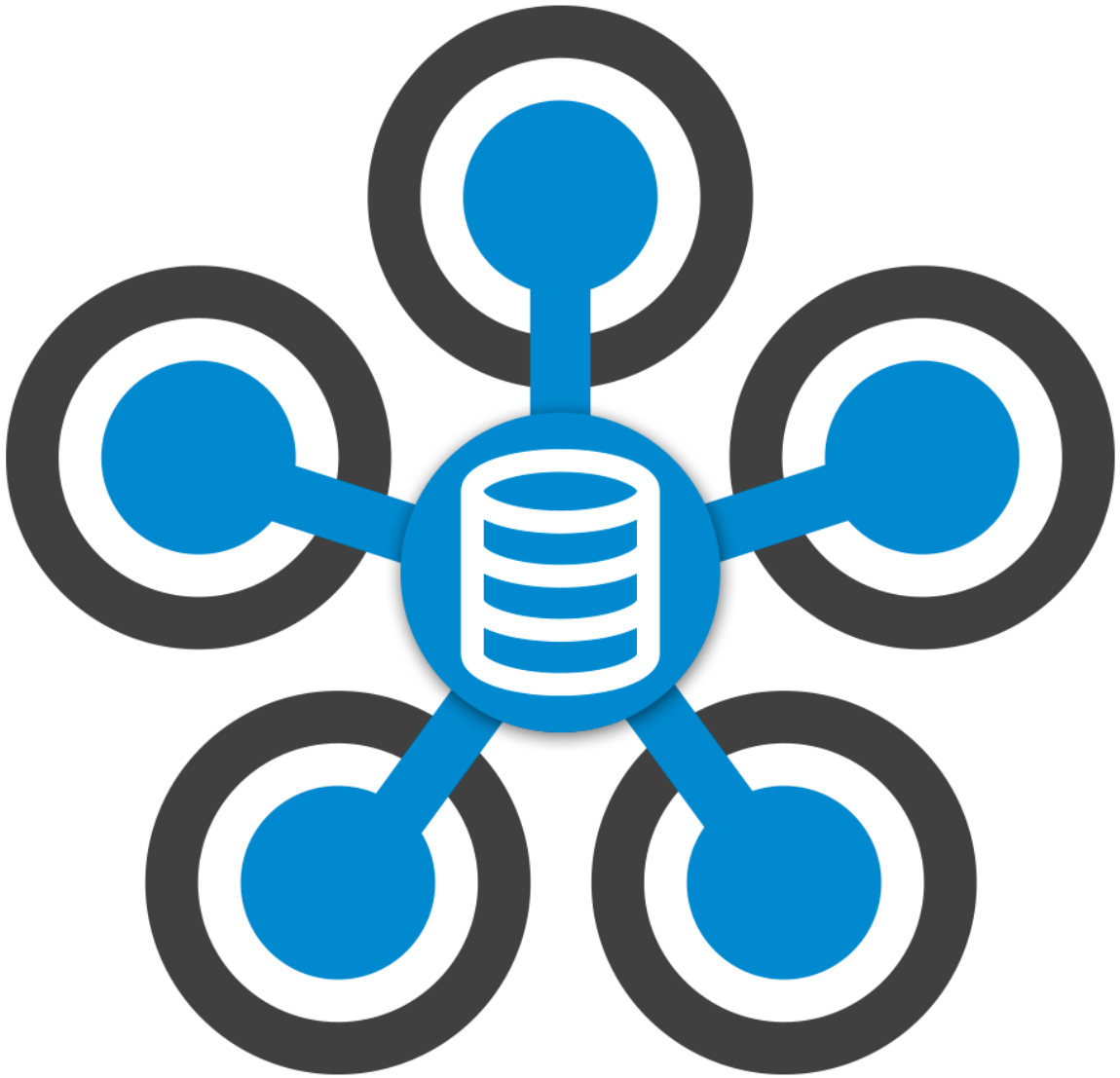
Image par Rudy and Peter Skitterians de Pixabay

Ce sont des données sonores.

kraisleriana.mp3

{BnF

L’information numérique, un changement de paradigme : ubiquité

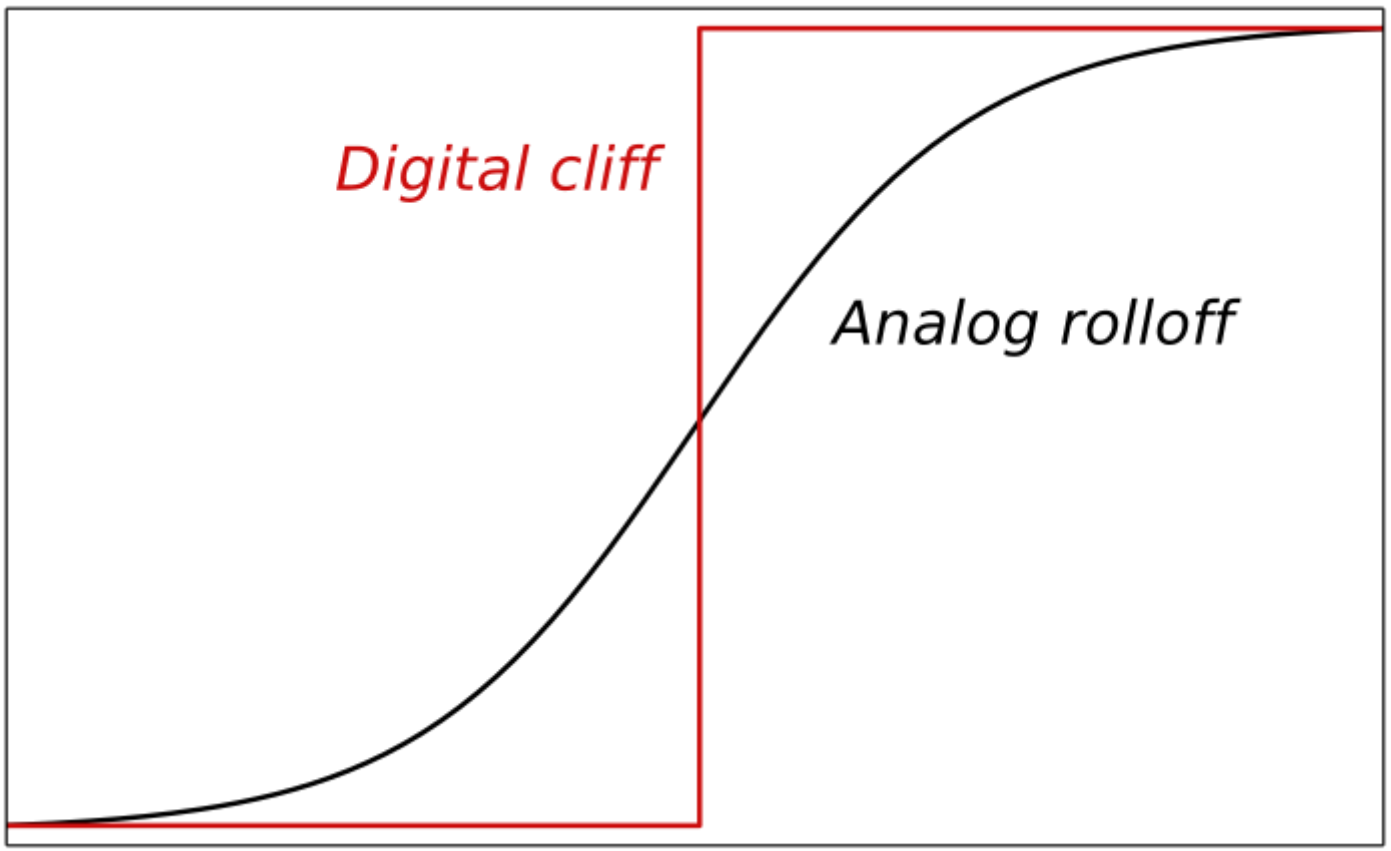


{BnF

L’information numérique, un changement de paradigme : dynamisme

BnF

L’information numérique, un changement de paradigme : irréversibilité des altérations



BnF

En cause : bon nombre d'idées préconçues

Sur l'air de « Le numérique, c'est magique. »

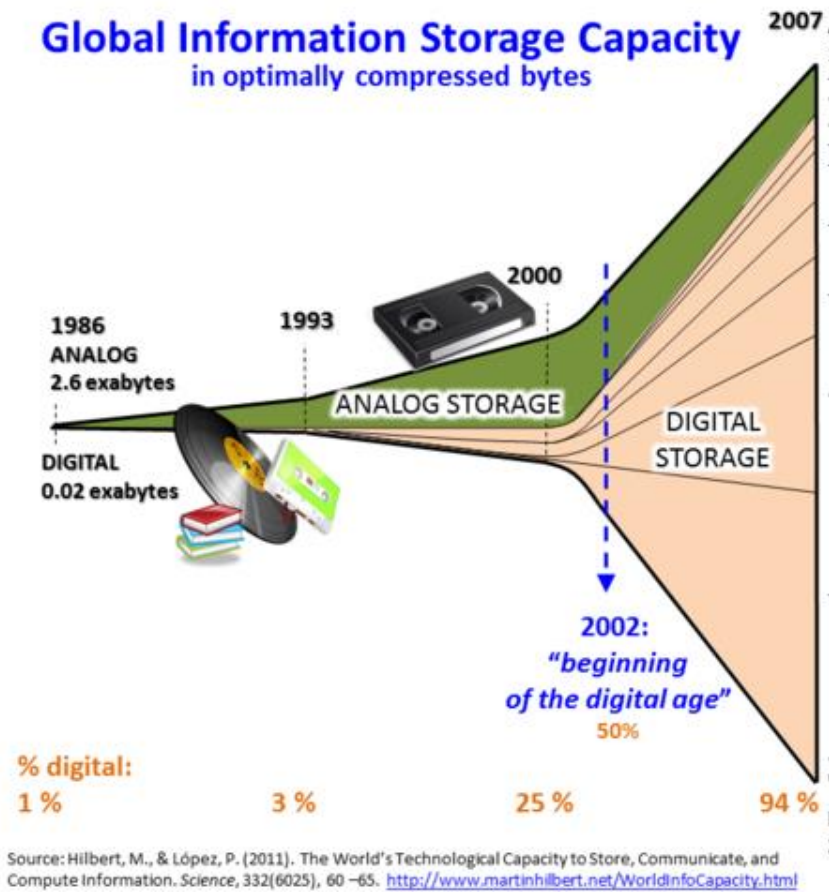
- « C'est en ligne – donc c'est préservé »
- « C'est numérisé – donc c'est préservé. »
- « C'est sauvegardé – donc c'est préservé. »
- « Quelqu'un doit bien s'en occuper. »
- « Le stockage, c'est dématérialisé, ça ne coûte rien. »
- « De toute façon, c'est l'affaire des informaticiens. »
- « Une fois stockée, l'information numérique n'a plus vocation à changer. »

…à confronter aux réalités suivantes :

- Le principal risque pesant sur l'information numérique, c'est l'abandon à des professions dont la préservation n'est pas la spécialité. Les activités des bibliothécaires (sélectionner, décrire, préserver) s'appliquent aussi aux contenus numérisés ou nativement numériques.
- L'engagement sur le long terme a un coût.
- La préservation numérique est une activité récurrente.
- Numériser équivaut à produire un second objet qui requiert des efforts de gestion et de préservation conséquents.
- Un outil de diffusion n'est pas un outil de préservation.
- Un outil de stockage n'est pas un outil de préservation.

L'information numérique, un changement de paradigme : quantité

- La quantité de données numériques stockées dans le monde doublerait tous les deux ans.
- 2020 : la masse de données stockées atteint 40 zettaoctets (1 Zo = 1 000 000 000 To). On produit 50 000 Go / seconde (en 1982, on en produisait 100 par jour).
- Dépôt légal du son : passage de 8000 CD audio déposés par an à 250 000 produits phonographiques produits dans le même temps.
- Des données dynamiques car actualisées en permanence.



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

BnF

Un outil d'auto-évaluation : les « niveaux de préservation numérique » du NDSA

https://hal-bnf.archives-ouvertes.fr/hal-02551807

- Cinq critères simples, avec une évaluation de 1 à 4.
  - Stockage et localisation géographique
  - Intégrité des données
  - Sécurité de l'information
  - Métadonnées
  - Formats de fichiers
- Les niveaux sont cumulatifs.
- Les moyens et les risques sont à évaluer, **il n'est pas forcément souhaitable de viser le niveau 4 partout.**

BnF

Le manuel de la préservation numérique

- Un document de référence, désormais traduit en français : le *Digital Preservation Handbook*, 2nd Edition, **https://www.dpconline.org/handbook**, Digital Preservation Coalition © 2015.
- Synthétique mais comportant de nombreux liens vers des ressources plus précises pour aller plus loin, il aborde autant les aspects organisationnels que techniques.

BnF

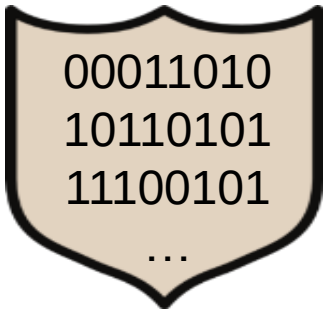
Prévenir l'altération des données numériques

L'INTEGRITE ET LA SECURITE DE L'INFORMATION

BnF

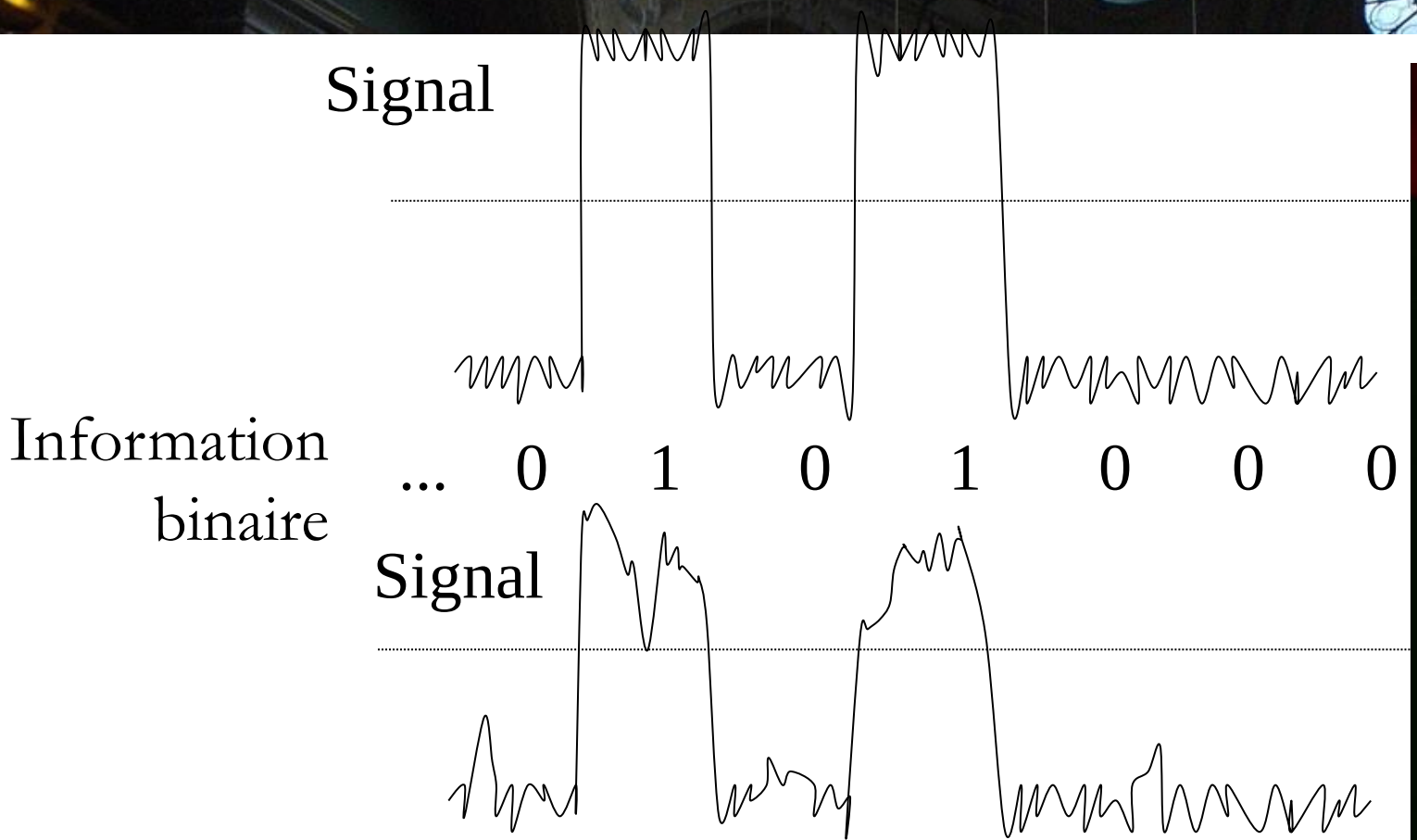
L'intégrité des données selon les niveaux de préservation NDSA

Risques : dégradation physique ou chimique des supports, usure, erreur de copie, interruption du transfert… => perte d'information, inaccessibilité du fichier ou du support complet.



- Niveau 1 (Protéger) : vérifier l'intégrité au versement si l'information d'intégrité a été fournie, la créer si elle ne l'a pas été.
- Niveau 2 (Connaître) : vérifier systématiquement l'intégrité au versement, travailler en « read-only » sur les contenus originaux, réaliser une analyse antivirus sur les contenus à risque.
- Niveau 3 (Surveiller) : vérifier l'intégrité à intervalles régulier (« auditer »), conserver les résultats de ces tests, réaliser une analyse antivirus sur tous les contenus.
- Niveau 4 (Réparer) : vérifier l'intégrité en fonction des opérations réalisées, capacité à remplacer les données corrompues.

L’altération du signal



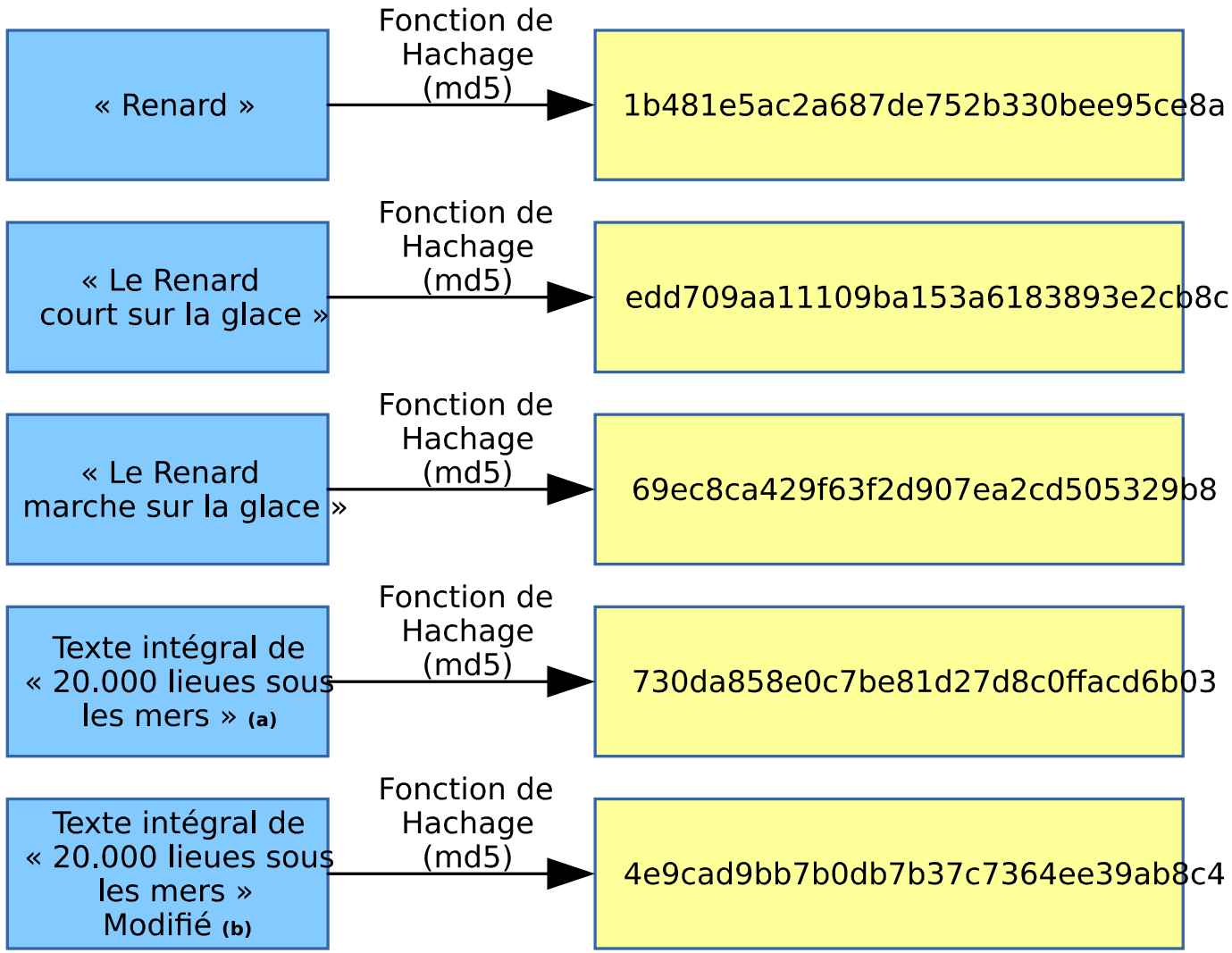
A terme, la dégradation de la modulation passera un seuil fatidique = perte d'un ou plusieurs signes

BnF

L’empreinte numérique

Principe (simplifié) : utiliser un algorithme analysant les données binaires (0 et 1) du fichier et renvoyant une chaîne de caractères.

Si le fichier se trouve altéré, le résultat du calcul d’empreinte changera, ce qui permettra de repérer l’altération.



Quel algorithme choisir ?

- Dans le domaine de la cryptographie et des signatures numériques où ils sont également utilisés, ces algorithmes sont régulièrement considérés comme obsolètes à mesure de l'augmentation des capacités de calcul des machines, qui arrivent désormais à produire des données différentes ayant la même empreinte.
- Néanmoins, dans le domaine de la préservation numérique où l'on cherche à prévenir des altérations accidentelles, il n'est pas nécessaire d'utiliser les algorithmes les plus puissants (qui demandent en outre plus de temps machine).
- MD5, SHA-1 et SHA-256 sont les plus répandus.

Des outils pour générer et vérifier l'intégrité de vos données

- QuickHash : https://www.quickhash-gui.org/ libre.
- Bagger (Library of Congress) : https://github.com/LibraryOfCongress/bagger (un guide sur Bagger pour les donateurs a été fait par les Gloucestershire Archives: https://www.gloucestershire.gov.uk/media/2084506/guidelines_bagger-v-1-draft-cc.pdf) libre.
- File Analyzer (National Archives and Records Administration) : https://github.com/usnationalarchives/File-Analyzer libre.
- Fixity (AV Preserve), libre mais non maintenu (voir https://blog.weareavp.com/fixity-pro-release-2020) et sa version payante (47,90 $ / an) : https://www.weareavp.com/products/fixity/

Tous ces outils sont multi-plateformes et disposent d'une interface simple d'utilisation.

BnF

Que faire en cas d'altération ?

- A moins de disposer d'une connaissance très poussée du format du fichier, il est impossible de « restaurer » des données corrompues.
- La solution : restaurer une copie intègre parmi les autres copies dont vous disposez (voir la partie Stockage plus loin).

Transférer des données : collecter ou recevoir

- De multiples méthodes :
  - Disques durs fournis par l'institution ou le fournisseur (chiffrés si le contenu est sensible) ;
  - Serveur sFTP ;
  - Service de transfert en ligne : préférer un outil officiel (ex. : France Transfert, https://francetransfert.numerique.gouv.fr/upload) ou celui maintenu par votre institution à d'autres services privés (wetransfer, OneDrive, DropBox, etc..
- Eviter les envois par messagerie électronique (sécurisation inexistante), sauf en dernier recours et si les données ne sont pas sensibles.

BnF

Focus : les dates d'un fichier numérique

- Lorsque la date d'un contenu n'est pas connue, ni apposée sur le contenu, les métadonnées internes sont utiles.
- Les dates des métadonnées internes spécifiques à un type de contenu (ID3, Vorbis, etc. pour l'audio, EXIF, IPTC et/ou XMP pour l'image, XMP pour les documents, etc.).
- En revanche, attention aux dates de création, dernière modification et dernier accès qui sont liées au **système de fichiers** et non au fichier lui-même : la plupart du temps, **un transfert les modifiera**.
- Si on souhaite préserver ces dates pour s'en servir ou garantir l'authenticité du contenu, il faut des **procédés de transfert spécifiques** (en particulier : utiliser des outils de copie avancés – Robocopy pour Windows, rsync pour Linux, etc. –, zipper les fichiers avant de les envoyer).

Créer des unités de gestion manipulables et autonomes

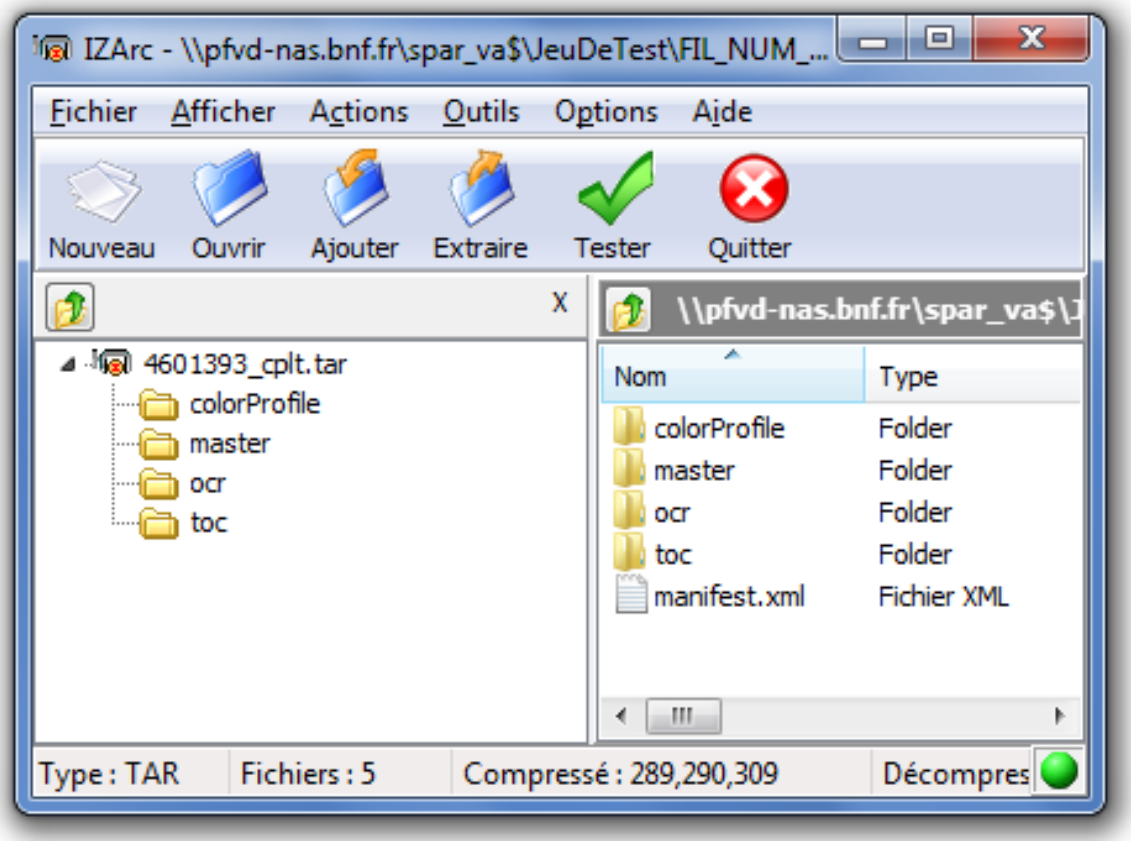
L'EMPAQUETAGE

BnF

Définir vos règles d'empaquetage

- Quel niveau de granularité (par ex. périodique ou fascicule ?)
- Quelle arborescence des dossiers ?
- Quel nommage des dossiers et fichiers ?
  - Rappel : pas d'espaces, uniquement des caractères alphanumériques, pas de diacritiques, éventuellement des tirets (« - », « _ »), minuscules ou majuscules exclusivement.
  - Une fiche en anglais sur la manière de bien nommer – et de bien documenter ses conventions –: https://authors.library.caltech.edu/103626/
- Quels formats attendre dans chaque dossier ?
- Quelle forme pour les paquets (dossiers, fichiers zippés, « bags ») ?
- Où se trouvent les métadonnées ?
- Quelle information nécessaire à la compréhension et à l'usage de mes contenus dois-je embarquer dans mes paquets ?

Un exemple concret d'empaquetage




```
<fileSec>
    <fileGrp USE="master" ID="GRP.1">
        <file CHECKSUMTYPE="MD5" CHECKSUM="580256aaea9079083786b867
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="master/T0
        </file>
        <file CHECKSUMTYPE="MD5" CHECKSUM="17990c6ef8e35cbdb73ebced
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="master/T0
        </file>
        <file CHECKSUMTYPE="MD5" CHECKSUM="24755472cab75fefac3350ce8
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="master/T0
        </file>
        <file CHECKSUMTYPE="MD5" CHECKSUM="c113c95aceb9e58effb92e807
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="master/T0
        </file>
        <file CHECKSUMTYPE="MD5" CHECKSUM="cdb27460859437fc402f43d1ff
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="master/T0
        </file>
        <file CHECKSUMTYPE="MD5" CHECKSUM="38bf8dd398e791772126e3df9
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="master/T0
        </file>
'fileGrp>
'ileGrp USE="colorProfile" ID="GRP.2">
        <file CHECKSUMTYPE="MD5" CHECKSUM="5b717845bd8b33d5d5c33e5f3
            <FLocat xlink:type="simple" LOCTYPE="URL" xlink:href="colorProfil
        </file>
</fileGrp>
```

{BnF

Un format d'empaquetage : BagIt

BagIt, c'est

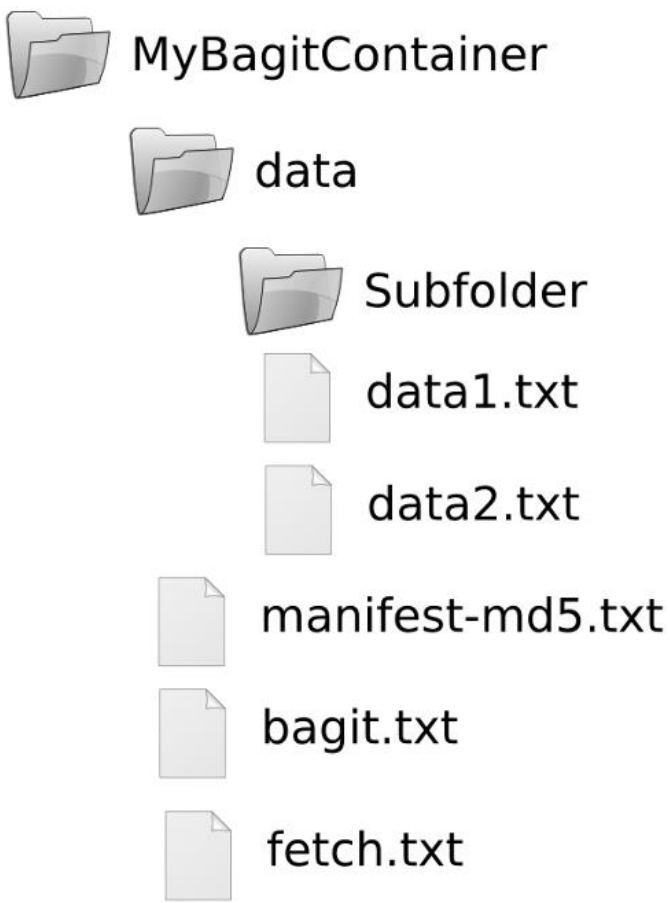
- La spécification IETF (une norme de l'Internet) RFC 8493, voir https://tools.ietf.org/html/rfc8493
- Un moyen standard d'empaqueter vos données, de structurer le contenu du paquet, d'associer des métadonnées.

Le « bag » produit est un fichier conteneur (ZIP ou TAR) qui embarque les données et leurs métadonnées dans des « manifestes » (inventaire des fichiers de données et métadonnées avec leur empreinte numérique).

Mais c'est surtout une série d'outils pour créer et gérer des paquets de données dans ce format (voir https://en.wikipedia.org/wiki/BagIt)

BnF

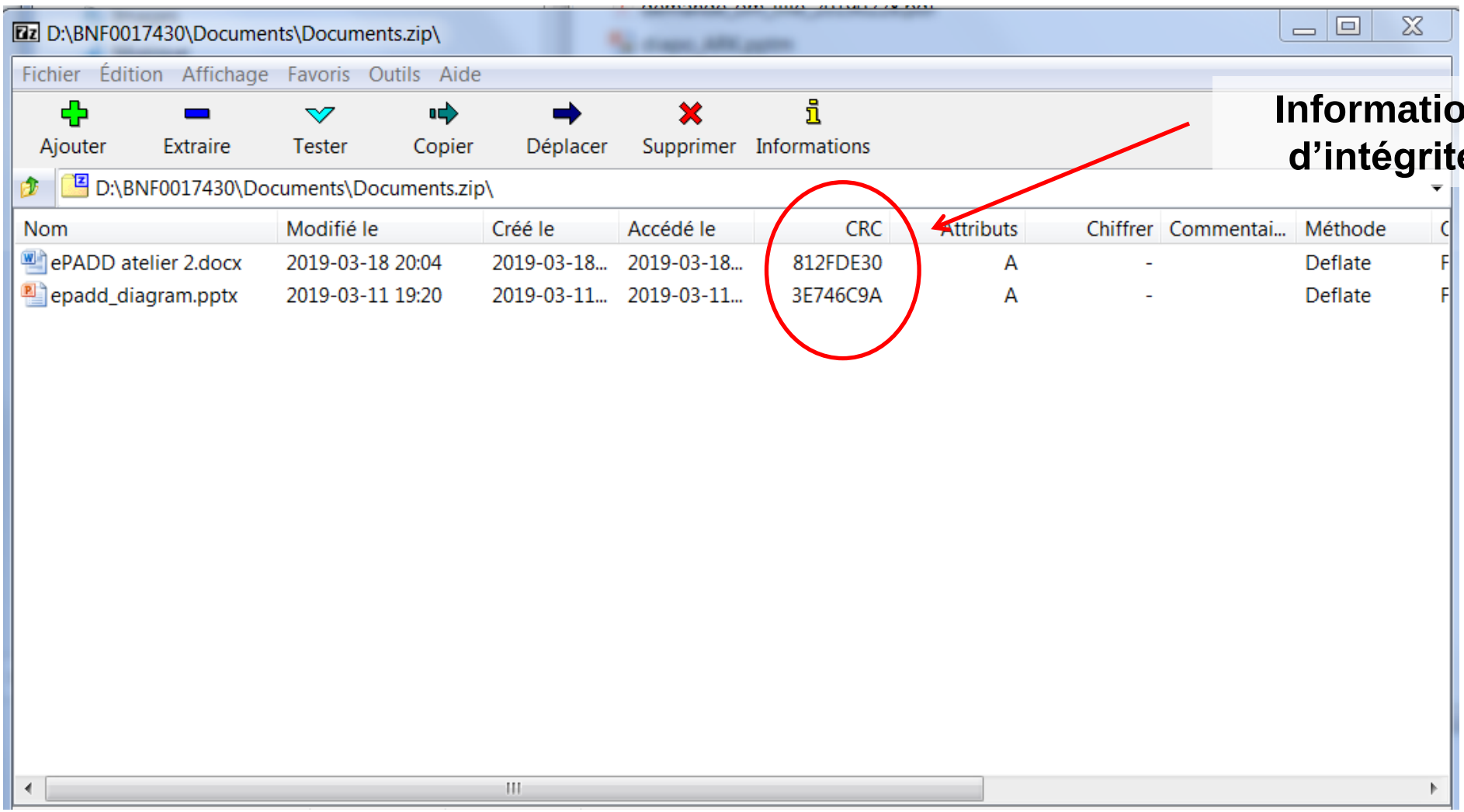
Un format d'empaquetage : BagIt



- data : un dossier où placer dossiers et fichiers selon l'arborescence voulue.
- manifest-md5.txt : l'inventaire des fichiers avec une empreinte numérique MD5
- bagit.txt : l'ensemble des métadonnées associées au « bag ».

BnF

À défaut : créer un fichier ZIP



Information d'intégrité

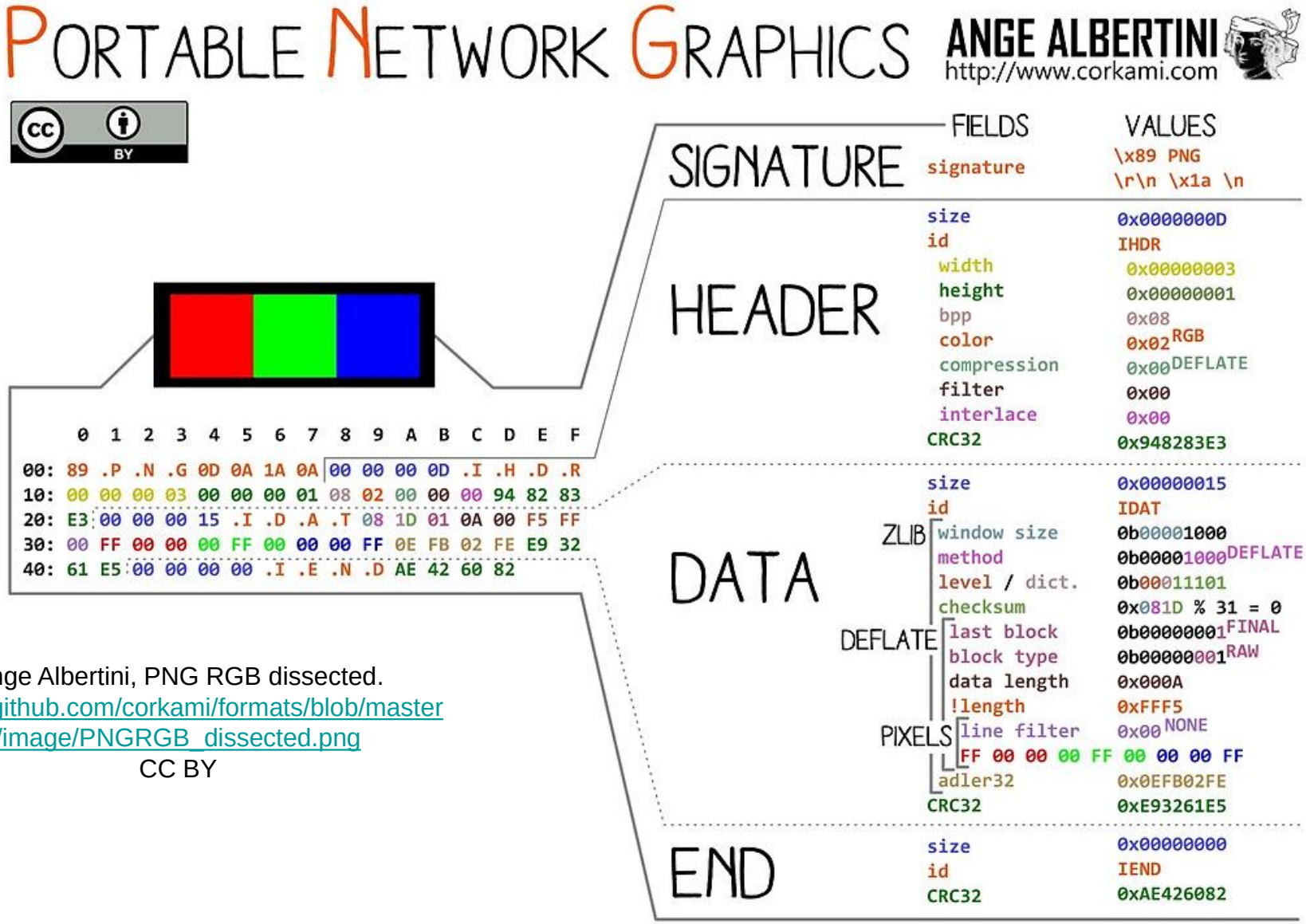
BnF

Choisir, connaitre, identifier et valider ses données

LES FORMATS DE FICHIER

BnF

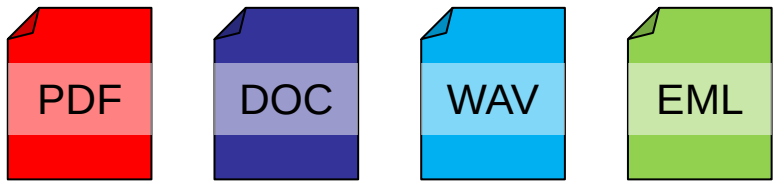
Un fichier, qu'est-ce que c'est ?



Ange Albertini, PNG RGB dissected.
https://github.com/corkami/formats/blob/master/image/PNGRGB_dissected.png
CC BY

BnF

Les formats de fichier selon les niveaux de préservation NDSA



Risques : perte de la compétence côté personnel ou usager, incapacité à maintenir les outils associés => impossibilité d'accéder au contenu des fichiers.

- Niveau 1 (Protéger) : fournir une liste de formats recommandés aux producteurs des données.
- Niveau 2 (Connaître) : avoir un inventaire à jour de tous les formats utilisés dans son entrepôt.
- Niveau 3 (Surveiller) : réaliser une veille technologique sur les problèmes d'obsolescence.
- Niveau 4 (Réparer) : réaliser des opérations de préservation (migration, émulation) si besoin.

BnF

Évaluer les caractéristiques d'un format de fichier dans une perspective de pérennisation

- Il n'existe pas de format pérenne !
- En revanche, il y a des formats pour lesquels **limiter les risques de perte d'information** sera moins coûteux en moyens humains et financiers que pour d'autres.
- Selon les **besoins** et les **moyens**, les choix de formats ne seront pas nécessairement les mêmes d'une organisation à l'autre.
- Pour s'en convaincre, consulter la comparaison des politiques de formats de plusieurs institutions de conservation dans le monde : Open Preservation Foundation, *International Comparison of Recommended File Formats*, 4 avril 2022. https://openpreservation.org/news/new-community-resource-international-comparison-of-recommended-file-formats/.

BnF

Les 12 critères objectifs (BnF) d'évaluation du niveau de « préservabilité » d'un format

Communauté d'utilisateurs / Sociabilité	Documentation	Liberté d'utilisation	Indépendance / autonomie
Résilience / robustesse / tolérance à l'erreur	Compacité	Disponibilité d'outils de traitement	Contenu additionnel embarqué
Mécanismes de protection	Simplicité	Stabilité / évolutivité	Transparence

BnF

A compléter par deux critères subjectifs

- Expressivité
- Maîtrise

BnF

Quelques sources d'information en français

- Le Référentiel général d'interopérabilité (RGI) publié par la DGME (Direction générale pour la modernisation de l'État), révisé en 2016 : https://references.modernisation.gouv.fr/sites/default/files/Referentiel_General_Interoperabilite_V2.pdf
- Groupe Formats de la BnF, *Formats de données pour la préservation à long terme : la politique de la BnF*. [Rapport Technique] Bibliothèque nationale de France (Paris). 2021. ⟨hal-03374030⟩
- BnF : des référentiels (TIFF, JPEG 2000, EPUB) : https://www.bnf.fr/fr/les-referentiels-de-numerisation-de-la-bnf et un document de recommandations et des fiches par format en cours d'élaboration
- Les formats de fichiers recommandés par le CECO (Centre de coordination pour l'archivage à long terme de documents électroniques, Suisse) : https://kost-ceco.ch/cms/kad_image_fr.html

BnF

Les trois fonctions des outils d'analyse

- Identifier
  - « image/jpeg »
- Valider
  - « Bien formé et valide »
- Caractériser
  - Profondeur couleur = « 8,8,8 »
  - Résolution = « 400 dpi »
  - Profil couleur = « sRGB »

A quel format ai-je affaire ?

Le fichier est-il bien conforme aux spécifications du format ?

Quelles sont ses caractéristiques ?

Est-ce que j'accepte ce fichier ?

BnF

Identification : connaître le format d'un fichier

- Par son extension ?
- Par un outil
  - En masse, avec production d'un rapport : DROID (https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/)
  - Unitairement, pour obtenir un type MIME : TrID (https://www.mark0.net/soft-trid-e.html) et sa version en ligne : https://www.mark0.net/onlinetrid.html.
- Comment nommer les formats ?
  - Type MIME ;
  - PUID (identifiant dans la base PRONOM).

BnF

Caractérisation et validation de certains formats : JHOVE

TIFF, JPEG, GIF, JPEG 2000, TXT, XML, HTML, WAVE et PDF

(http://jhove.openpreservation.org/)

Interface, mais peu développée – il est fait pour être utilisé en ligne de commande.

Quelques autres outils complémentaires

- Caractérisation des fichiers audiovisuels : MediaInfo (http://mediaarea.net/fr/MediaInfo). Existe aussi en version en ligne : MediaInfo Online (https://mediaarea.net/MediaInfoOnline)
- Validation des fichiers PDF et PDF/A : veraPDF (http://verapdf.org/)
- Identification et correspondance avec la politique formats du CINES : outil en ligne Facile (http://facile.cines.fr).

BnF

Un exemple parmi d'autres de problèmes liés au format

La méconnaissance d'un format peut générer de réels problèmes d'authenticité :

Thomas Ledoux, Alix Bruys, Bertrand Caron, Yannick Grandcolas, « Le Blues du JPEG : Une histoire haute en couleur » sur OPF Blogs, 9 novembre 2019. Accessible sur https://openpreservation.org/blogs/le-blues-du-jpeg/ (consulté le 6 avril 2021).

BnF

Les formats de fichier : BnF Archivage numérique *vs.* CINES

- BnF : les formats acceptés sont négociés avec le partenaire, ainsi que l'acceptation ou non de fichiers invalides. Du niveau de maîtrise de la BnF dépendra le niveau de service.

  Liste des formats connus ou maîtrisés dans SPAR : https://www.bnf.fr/sites/default/files/2018-11/spar_formats_techniques_fcais.pdf

- CINES : les formats acceptés sont une liste fermée mais évolutive définie sur https://facile.cines.fr/. Les fichiers doivent être valides ; le validateur Facile permet de contrôler avant versement leur validité.

BnF

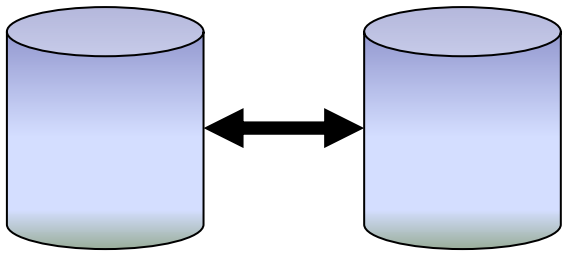
Choisir une infrastructure

LE STOCKAGE

BnF

Le stockage et la localisation géographique selon les niveaux de préservation NDSA

Risques : catastrophe naturelle, incident interne…



- Niveau 1 (Protéger) : deux copies distinctes, migration des données sur des supports hétérogènes vers un système unique
- Niveau 2 (Connaître) : au moins trois copies, dont une sur site distant, documentation du système de stockage et des supports
- Niveau 3 (Surveiller) : au moins une copie sur un site présentant des risques différents, surveillance de l'obsolescence des supports
- Niveau 4 (Réparer) : au moins trois copies sur sites présentant des risques différents, plans d'urgence et de reprise d'actitvité.

BnF

La durabilité des supports

On ne recherche plus la durabilité à tout prix On préférera contrôler l'état des supports et les renouveler régulièrement ou lorsque trop d'erreurs sont repérées.

La mixité des supports (mais non l'hétérogénéité !) est recommandée.

Image par Gerd Altmann de Pixabay

BnF

Les critères de choix d'un service de stockage

- Combien de copies ?
- Quelles localisations ?
- Comment les espaces sont-ils sécurisés ?
- Quels coûts : à la volumétrie ou à l'accès ?
- Quel délai d'accès ?
- Quel mode de versement ?
- Quelle taille maximale ?

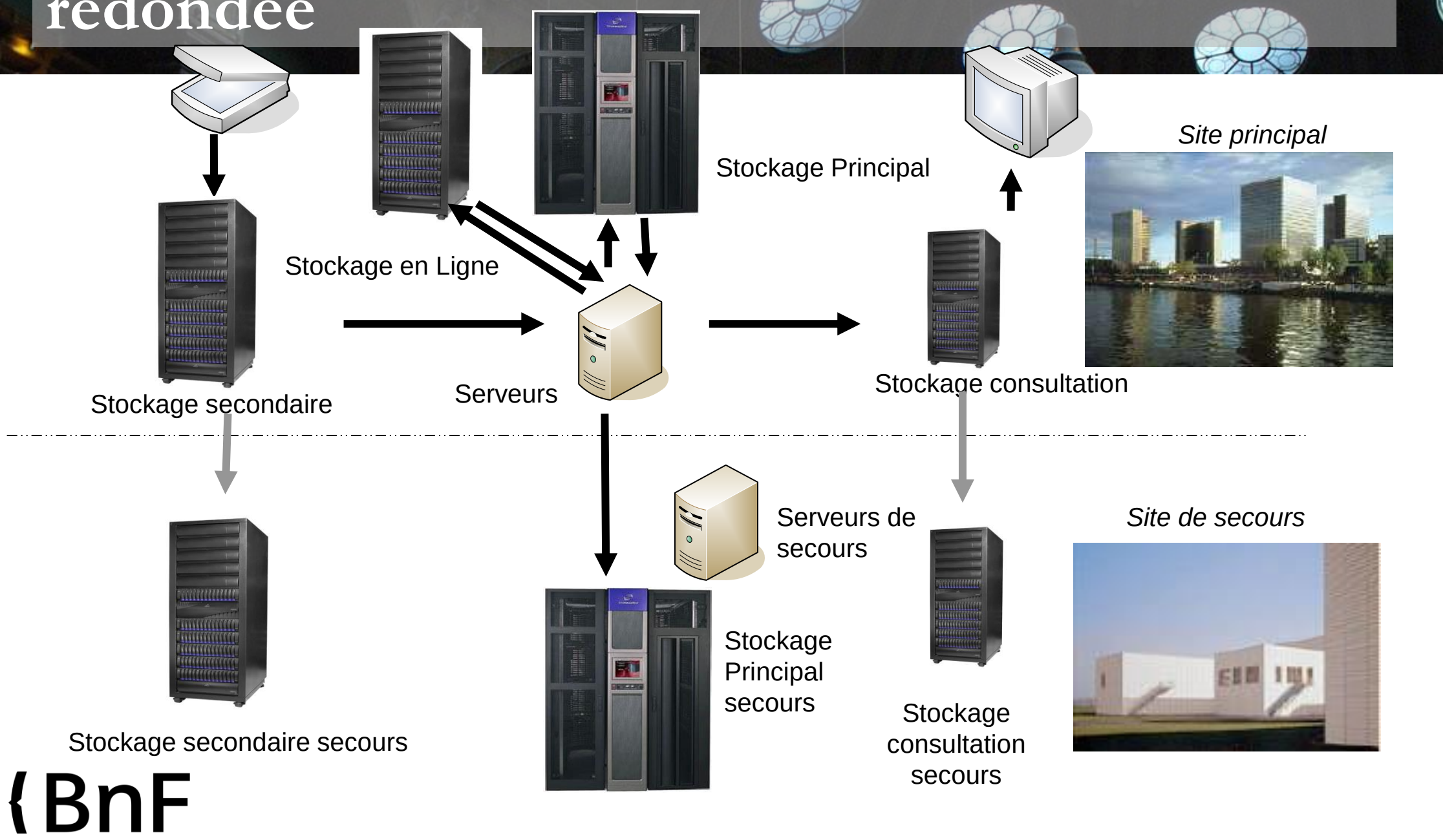
BnF

Le tiers-archivage, *by* la BnF

L'OFFRE BnF Archivage numérique

BnF

À la BnF : une infrastructure évolutive et redondée



BnF

Les tarifs

- Les tarifs votés en conseil d'administration le 21 novembre 2012 ont permis de lancer l'offre officiellement.
- Ils ne sont pas ceux d'un service commercial mais une participation aux coûts.
- Ils reposent sur un Accord de qualité de service (AQS) déterminé en préambule avec le futur tiers-archivé et intégré à une convention
- Les tarifs standards sont basés sur :
  - La volumétrie de stockage (de 1 à 50 To)
  - Le mode de téléchargement des données à archiver (intégration sur-mesure ou par l'extranet Coopération à partir de mi-2013)
  - Le support de stockage
    - Bandes : support peu coûteux mais lent
    - Disques : support plus cher, mais disponibilité plus rapide des données
  - La durée du contrat (3, 5 ou 8 ans).

BnF

Tarifs pour un contrat de 5 ans

BnF Archivage numérique Tarif général annuel, au téra octet, en € HT, basé sur **5 ans**		
Volume des données à archiver sur 2 bandes :	**extranet**	**sur-mesure**
pour 1 To	2 545 €	3 990 €
de 2 à 5 To	1 960 €	2 680 €
de 6 à 9 To	1 740 €	2 160 €
de 10 à 29 To	1 590 €	1 840 €
de 30 à 49 To	1 420 €	1 530 €

Options

Archivage complémentaire sur 1 disque : + 315€

Archivage complémentaire sur 2 disques : + 630€

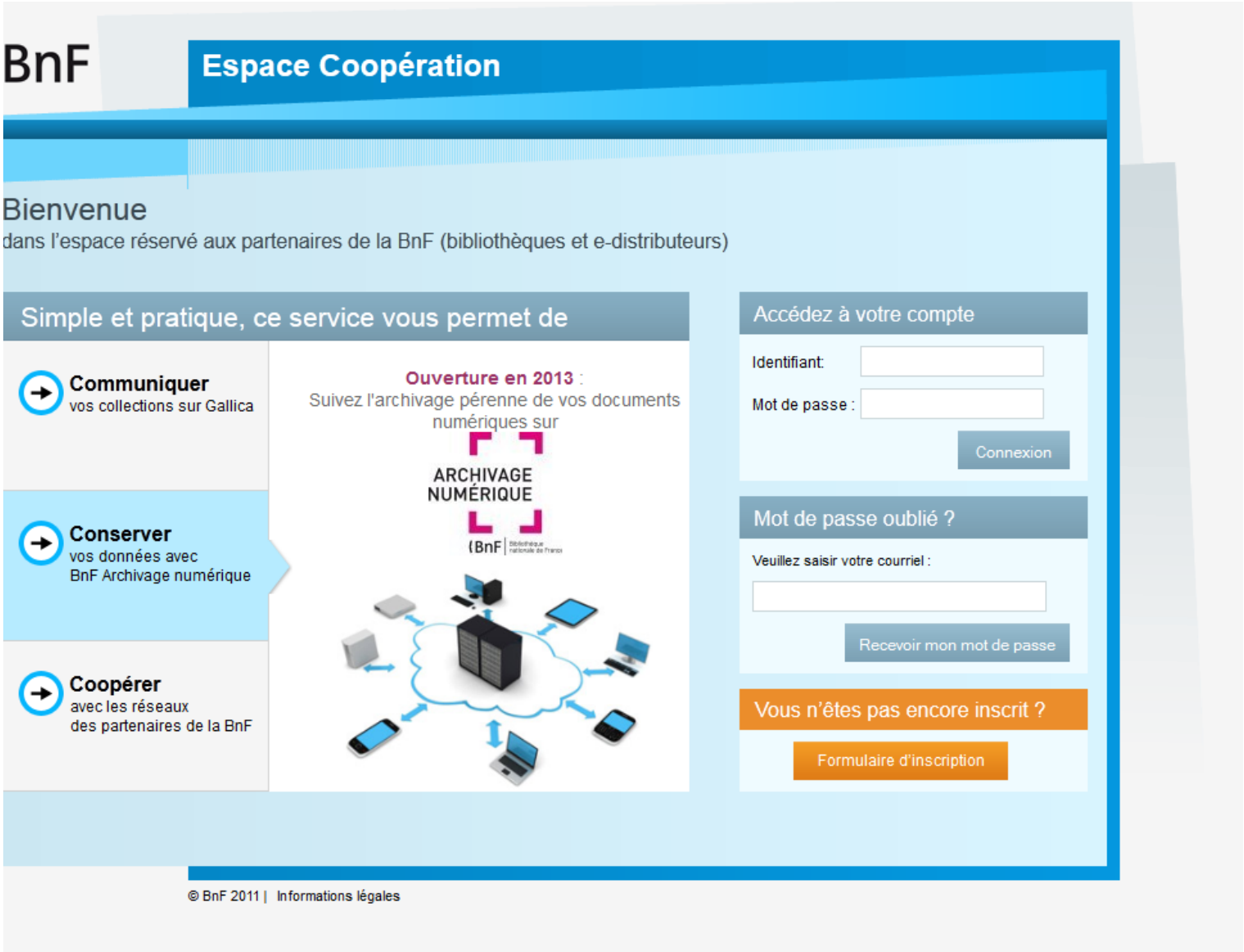
BnF

BnF Archivage numérique *vs.* Gallica Marque Blanche

	Gallica Marque Blanche	BnF Archivage numérique
Politique documentaire	Complémentaire BnF	Libre
Données	Copropriété BnF	Propriété unique de votre institution et restituées intégralement à la fin du contrat
Trajectoire de pérennisation	Les données doivent être compatibles avec les standards BnF et sont migrées selon la politique BnF	Les données sont conservées dans le format choisi par l'institution ; sur proposition de la BnF, elles peuvent bénéficier de la politique BnF
Métadonnées	Migrées dans le formalisme BnF	Conservées dans le format d'origine
Accès	Refacturé	Compris dans le prix, fréquence et volume négociés
Coût	Selon le nombre de documents	Selon la volumétrie (+ autres critères additionnels)

BnF

L'Espace Coopération pour gérer sa collection



BnF

Pour en savoir plus :

Bibliothèque nationale de France

Direction des Services et des réseaux

Sébastien Pétratos

01 53 79 45 31

adresse générique :

archivage-numerique@bnf.fr

Page d'information :

http://www.bnf.fr/fr/professionnels/numerisation/a.bnf_archivage_numerique.html

BnF

Décrire et documenter les objets, les processus, les agents, les caractéristiques techniques…

LES METADONNEES

BnF

Les métadonnées : de quoi parle-t-on ici ?

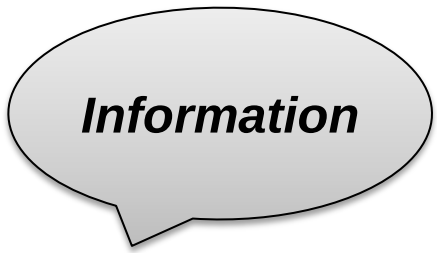
- Du sous-ensemble des informations décrivant un objet numérique nécessaire à sa préservation à long terme.
- Elles sont indispensables pour retrouver le contenu, dès lors que la collection numérique est importante, pour gérer une collection numérique, pour prouver son authenticité.
- On les classe généralement selon leur usage : décrire le contenu lui-même (métadonnées descriptives), les fichiers (métadonnées techniques), l'historique de production du contenu ou de ses fichiers (métadonnées de provenance), l'organisation et les relations des fichiers entre eux pour représenter le contenu (métadonnées de structure)

Les métadonnées : quelle structure ?

- Elles gagnent énormément à être **structurées**, ce qui permet de les interroger par des protocoles adaptés et de contrôler leur validité.
- Outre une utilisation individuelle, un outil de gestion de contenu permettra de réaliser des interrogations pour les exploiter efficacement selon des protocoles divers (base de données XML : XQUERY, base de données relationnelle : SQL, entrepôt de triples : SPARQL, etc.).
- Il est largement préférable qu'elles soient réunies dans un fichier unique contenu dans le Paquet d'informations (selon une structure texte, XML, JSON, CSV, XLSX, etc).. Néanmoins, certaines informations peuvent être transmises autrement : nom de fichier ou métadonnées internes (intégrées au fichier dans un bloc de données dédié, voire dans le bloc de données principal – une vidéo où la date et le lieu sont indiquées par une voix, par exemple – et selon un formalisme spécifique au format du fichier)

Les métadonnées selon les niveaux de préservation NDSA

Risques : importance des métadonnées sous-estimée => incapacité à retrouver un contenu, à identifier la version pertinente, incapacité globale à intervenir sur les contenus…



- Niveau 1 (Protéger) : produire un inventaire des contenus et de leur localisation de stockage, en assurer la préservation.
- Niveau 2 (Connaître) : conserver des métadonnées de gestion et de provenance.
- Niveau 3 (Surveiller) : conserver des métadonnées descriptives et techniques dans un format standard.
- Niveau 4 (Réparer) : conserver des métadonnées de préservation dans un format standard.

Quelles métadonnées minimales pour la numérisation et la préservation ?

Quelques éléments de métadonnées à conserver :

- Identification : **identifiants** (ISBN, ISSN, ISRC, etc.) ; autres si besoin.
- Descriptives, de niveau document, article ou page (pagination, type de page)
- Techniques : poids, format, **empreinte numérique**. Pour les images : taille (définition), profondeur couleur, résolution de capture. Pour les sons : fréquence d'échantillonnage, codec, débit.
- De provenance : **événements de production** (numérisation, océrisation, insertion de métadonnées internes, contrôle qualité, archivage…), date, résultat, agent.
- De structure : **liste des fichiers, usage des fichiers**, ordre de lecture, rapports entre eux (dérivation par ex.).

Une source d'inspiration : le référentiel d'enrichissement des métadonnées METS (https://www.bnf.fr/sites/default/files/2018-11/ref_num_metadonnees_mets.pdf)

BnF

Quelques formats de métadonnées parmi les plus utilisés en bibliothèque

Type de métadonnées	Normes et standards
Formats de métadonnées descriptives	Dublin Core, MODS, EAD, etc.
Format de métadonnées de pérennisation	PREMIS
Formats de métadonnées de droits	ODRL, metsRights, PREMIS, MPEG-21, etc.
Formats de métadonnées techniques	MIX, audioMD, videoMD, documentMD, etc.
Formats de métadonnées de structure	METS, PREMIS, SEDA

L'intérêt de ces standards et normes n'est pas nécessairement de permettre une implémentation complète mais également de constituer des manières de penser le sujet, le décrire et identifier les éléments d'information fondamentaux à documenter.

BnF

Dois-je utiliser un standard / produire du METS pour travailler avec la BnF ?

Réponse courte : pas nécessairement.

- Utiliser un formalisme XML vous permet d'utiliser tous les puissants outils et technologies associés à ce mode de structuration des données…
- …mais si vous n'êtes pas équipés (éditeur XML, compétence en interne), mieux vaut un jeu de métadonnées en CSV / XLSX bien spécifié entre vous et votre prestataire que des métadonnées METS en XML que vous ne maîtrisez pas.
- Aucun mode de partenariat avec la BnF (moissonnage, intégration, Marque Blanche, tiers-archivage) n'est conditionné par l'existence de métadonnées METS.
- Certains outils, dont NumaHOP (https://www.numahop.fr/) et Archifiltre (https://archifiltre.fabrique.social.gouv.fr/) proposent une sortie METS des métadonnées dont ils disposent.

BnF

Métadonnées requises au versement : BnF Archivage numérique *vs.* CINES

	CINES	BnF Archivage numérique
Métadonnées d'identification de niveau document	Format maison sip.xml au format XML intégrant les 15 champs Dublin Core + identifiant producteur	Format libre sous forme CSV (Excel), XML ou XML-METS, converti par la BnF en Dublin Core
Métadonnées métier de niveau document	Fourniture d'un fichier de métadonnées métier riches recommandée. Fichier stocké comme un fichier de données	Possibilité de fournir un fichier de métadonnées métier riches
Métadonnées de niveau collection	Optionnelles, à fournir avec un paquet de données à verser avant les paquets de documents.	Optionnelles, à fournir avec les métadonnées de chaque document
Métadonnées de niveau projet	Format maison ppdi.xml décrivant les droits, processus, le service versant, le fonds	Format maison AQS (Accords de qualité de service) définissant les engagements mutuels
Métadonnées d'intégrité	Obligatoires par fichier, à fournir dans le fichier sip.xml	Obligatoires par fichier, à fournir dans un fichier .md5
Métadonnées techniques	Certaines obligatoires (format, encodage, compression) à fournir dans le fichier sip.xml	Optionnelles, calculées au versement

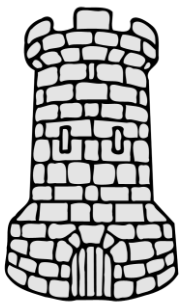
Contrôler les métadonnées

- Dans le contexte d'un marché de numérisation, ces livrables sont trop rarement contrôlés.
- On recommande un contrôle systématique ou par échantillonnage – au moins pour vérifier la validité des fichiers XML, voire la conformité à des règles « maison ».
- Divers langages pour contrôler des métadonnées au format XML : DTD, schéma XML, RelaxNG, schematron.


Pour quelques éléments de plus…
Sécurité de l'information,
chiffrement, etc.
BnF

La sécurité de l'information selon les niveaux de préservation NDSA

Risques : intrusion et dégradation volontaire ou effacement accidentel



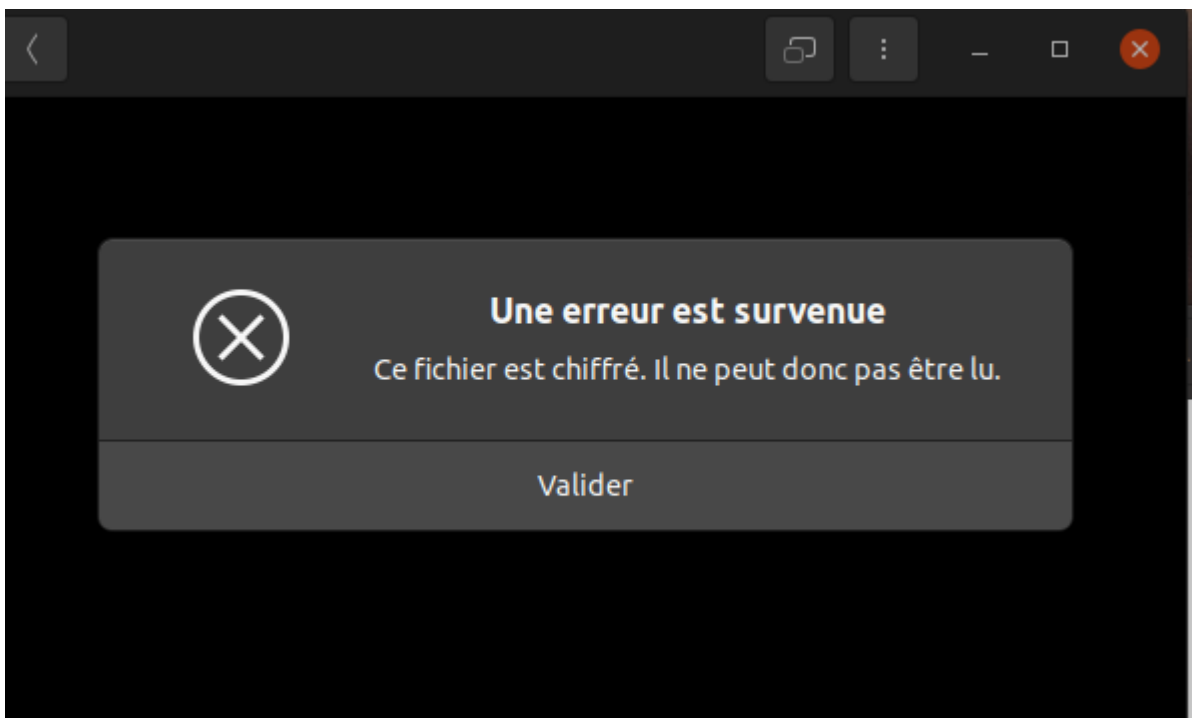
- Niveau 1 (Protéger) : identifier qui a les droits de lecture, modification, déplacement et suppression et limiter à des agents formés.
- Niveau 2 (Connaître) : documenter les restrictions d'accès
- Niveau 3 (Surveiller) : produire un historique des opérations (ou « audit trail ») de préservation, modification, suppression…
- Niveau 4 (Réparer) : contrôler l'historique des opérations, s'assurer qu'aucune personne n'a tous les droits sur toutes les copies.

BnF

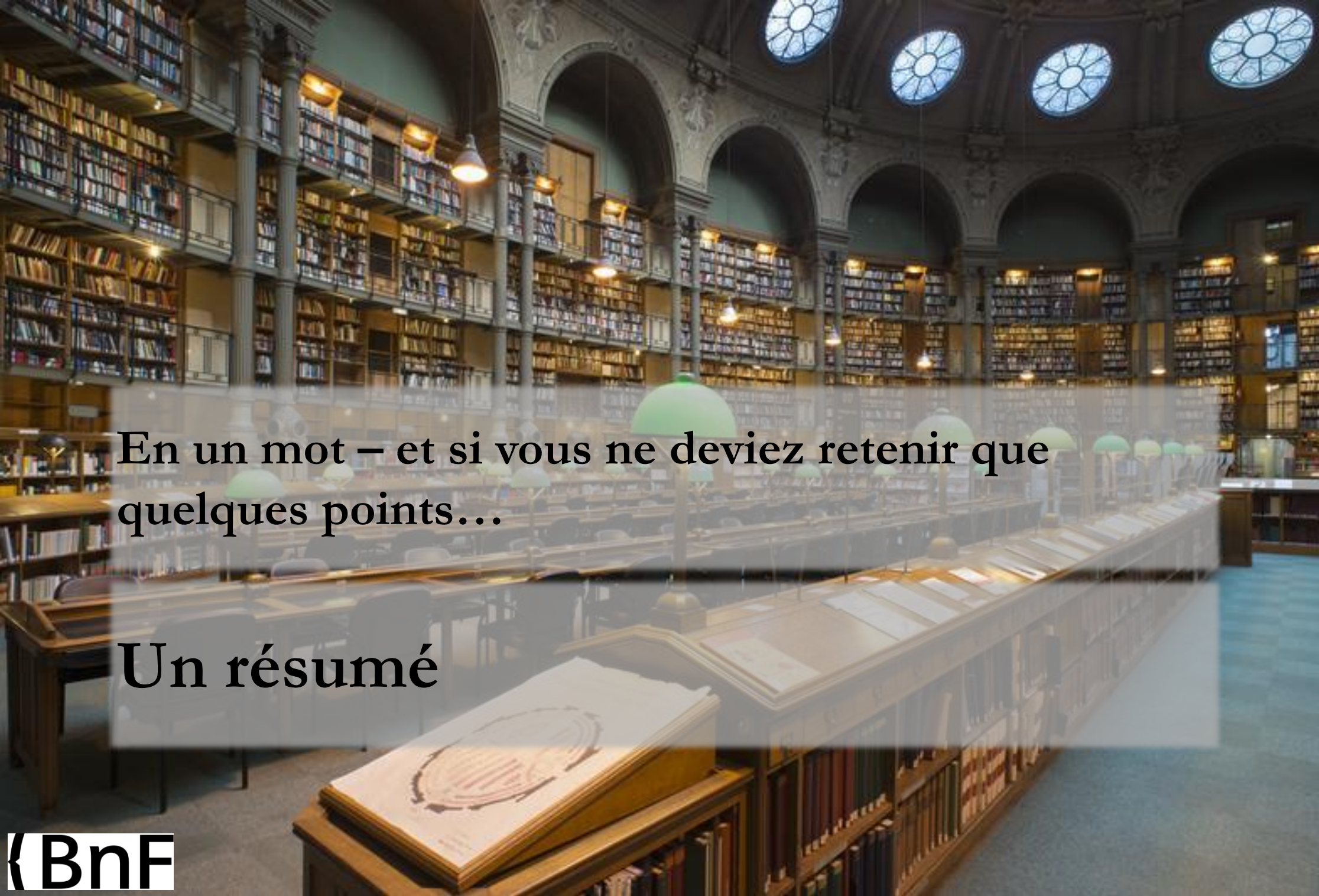
Inaccessibilité technique : les risques du chiffrement

- Tout fichier peut être chiffré (mais les spécifications de certains formats intègrent une méthode privilégiée de chiffrement)
- Tous les DRM (mesures techniques de protection) se fondent sur le chiffrement.

Exemple de fichier M4P (fichier Quicktime diffusé par iTunes avec DRM) ouvert avec l'application Videos/Totem (OS Ubuntu 20,04)



BnF



En un mot – et si vous ne deviez retenir que quelques points…

Un résumé

En un mot – et si vous ne deviez retenir que quelques points…

- Demandez une empreinte numérique, contrôlez-la (de préférence régulièrement).
- Stockez sur au moins deux copies distantes, trois étant le mieux. Formalisez et maîtrisez vos méthodes de sauvegarde et de synchronisation. *Pour les petites institutions, ne négligez pas les services de stockage cloud.*
- Demandez ou recommandez un ou plusieurs formats de fichier que vous maîtrisez, utilisez un outil de contrôle adapté.
- Faites des paquets : définissez des règles de nommage et d'arborescence des dossiers et des fichiers, et des modalités d'empaquetage (format, localisation des métadonnées) *et choisissez un format d'échange adapté – BagIt et/ou à défaut ZIP.*

BnF

En un mot – et si vous ne deviez retenir que quelques points…

- Obtenez la maîtrise de vos métadonnées de gestion (notamment de provenance et techniques) et exploitez-les. Vos contenus doivent survivre à votre outil de diffusion, ne laissez pas à ce dernier le monopole de ces informations.
- Utilisez un nommage pratique et solide de vos fichiers (et contrôlez son application, voir https://bibliotheques.wordpress.com/2019/01/07/un-logiciel-de-controle-de-noms-de-fichiers-parce-que-ca-peut-servir-parfois/), mais attribuez des identifiants opaques à vos documents.
- Faites très attention aux conversions (même l'enregistrement dans un autre format de fichier que celui d'origine est risqué !), évitez-les à moins d'une absolue nécessité ou d'une assistance par des spécialistes.

BnF


Merci de votre attention !
bertrand.caron@bnf.fr
BnF