Huma-Num
la TGIR des humanités numériques

ANF Huma-Num

Qualité des données, comprendre les enjeux des formats de fichiers

Michel Jacobson IR* Huma-Num

8-décembre-2023

cnrs Aix*Marseille université CAMPUS CONDORCET Paris–Aubervilliers

La notion de format de fichier

Point terminologique

La norme ISO-14721:2012 OAIS (Open Archival Information System) distingue :

- l'information: connaissance que l'on peut échanger ;
- la donnée: représentation formalisée de la connaissance.

Pour manipuler une information, on lui donne une forme (la donnée)

Cette forme suit des règles qui permettent son codage et son décodage.

Un format de fichier constitue un ensemble de règles qui décrivent comment coder les informations sous forme de données et comment décoder celles-ci pour retrouver les informations. Ces règles sont généralement consignées dans des « spécifications » et sont mises en œuvre par des outils logiciels (éditeurs, convertisseurs...) pour permettre de manipuler l'information.

Point terminologique

- Les données DOIVENT être accompagnées de métadonnées afin de comprendre et réutiliser les informations qu'elles contiennent.

L'identification du format est une de ces métadonnées

- Mais les métadonnées sont aussi des informations, qu'il faut coder, formater et décrire…

- Les formats sont très liés au marché et plus globalement à l'ensemble de l'environnement numérique. Ils sont donc soumis à l'obsolescence plus ou moins rapidement.

Le choix des formats cibles

- Quels sont les enjeux/risques ?
- Comment faire le choix dans le cadre du projet ?
- Comment atteindre les formats choisis ?
- Comment vérifier que les données sont bien formatées ?

Quels sont les risques ?

Dépendance

ex. Un format fermé lisible que par un logiciel propriétaire d'un éditeur dont on ne connaît pas la feuille de route

Non maîtrise de la qualité des données

ex. Impossibilité de valider par un outil autre que celui de l'éditeur. Tout repose sur la confiance...

Perte de l'information

ex. Un format propriétaire dont l'éditeur ne maintient plus les outils d'édition/lecture et qui ne fonctionne plus avec la nouvelle version de MS-Windows

Comment évaluer les risques ?

1) Des critères basés sur les spécifications

- Explicites ou non
- Secrètes / Publiques / Standardisées / Normalisées
- Avec ou sans entrave légale (brevet, copyright…)
- Qualité des spécifications
  - Complexité
  - Auditabilité

2) Des critères liés au marché

- Existence d'un outil d'édition et d'alternatives
- Existence d'outils de conversion
- Existence d'outils de validation
- Mais aussi évaluation des risques sur les logiciels (nombre d'utilisateurs, dynamisme des communautés d'utilisateurs, dépendances logicielles, fréquence des mises-à-jour...)

« Guide méthodologique pour le choix des formats numériques pérennes dans un contexte adaptés aux données orales et audiovisuelles » https://francearchives.fr/fr/circulaire/DGP_SIAF_2010_010

Huma-Num

Spécifications

Absence de spécifications ou spécifications implicites

ex. SQL définit un langage d'interrogation pour base de données mais pas de format de représentation

Spécifications non publiées

ex. les anciens formats de Microsoft (doc, ppt...)

Spécifications standardisées/normalisées

ex. ISO OpenXML (pptx, docx…)

W3C XML

IETF CSV

Huma-Num

Complexité 1

Taille des spécifications

- CSV 8 pages
- XML 35 pages
- OpenXML + de 6000 pages

Complexité 2

Dépendances

ex. La recommandation « Extensible Markup Language (XML) »

- est liée à d'autres recommandations « Namespaces in XML », « XML Inclusions (XInclude) », « Associating Stylesheets with XML », « Associating Schemas with XML »

- basé sur un format texte mais ne spécifie pas ce qu'est un format texte

- ne spécifie pas l'encodage des caractères mais juste comment l'expliciter, + sa valeur par défaut (UTF8), + les encodages que tous les outils doivent permettre (UTF-8 and UTF-16), mais les outils peuvent en proposer d'autres

- ne spécifie pas les langues mais juste la norme à utiliser pour les exprimer

ex. Le format OpenDocumentFormat utilise une enveloppe ZIP contenant une arborescence de fichiers essentiellement formatés en XML

Complexité 3

Variabilité

ex. Le format WAV est un format conteneur qui permet de stocker des données sonores de différents encodages (PCM, ADPCM, µ-law…)

On peut avoir un fichier WAV avec un codex mp3. Ce qui n'est pas la même chose qu'un format mp3 dont le codec est forcément mp3

ex. Un format MP4 est un format conteneur qui peut contenir des pistes audio, des pistes vidéo, des sous-titre, des métadonnées, etc.

On peut avoir un fichier MP4 qui ne contient que de l'audio ou qui ne contient aucune piste audio ni vidéo

Silence

ex. Un format CSV ne dit rien de l'encodage des caractères qu'il utilise

Connaître l'organisation du conteneur ne veut pas dire maîtriser ses contenus

Initiatives d’évaluation des risques

Huma-Num

Preservation Risk matrix

NARA (Achives du gouvernement fédéral des États-Unis)

- ✔ https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix
- ✔ Grille d'évaluation pour déterminer le niveau de risque que représente l'utilisation d'un format
- ✔ Synthèse du risque en 3 niveaux : haut, moyen, bas
- ✔ Une cinquantaine de critères
  - Le format est-il propriétaire?
  - Les spécifications sont-elle publiées?
  - Existe-t-il des outil de validation du respect des spécifications par un fichier?
  - Les spécifications ont-elles été validées par un organisme de normalisation?
  - Le format est-il utilisé dans le "federal government"?
  - Le format est-il utilisé à l'extérieur du "federal government"?
  - Le format est-il activement maintenu et mis à jour par une organisation, une personne individual ou une communauté?
  - Existe-t-il plusieurs outils de visualisation?
  - [...]
  - Le format permet-il l'encapsulation d'information comme le watermarking?

Preservation Risk matrix

	A	B	C	D	E	F	G	H	I	J
1	NARA Guidance : Preferred	NARA Guidance: Acceptable	Numeric Risk Rating	Risk Level	NARA Format ID	Format Name	File Extension(s)	Category/Plan(s)	Is the format proprietary?	Does the format have a published open specification ?
11			-26,00	High Risk	NF00182	Executable file	exe	Software and Code	-1	-1
12			-25,00	High Risk	NF00702	BlackBerry Binary Executable	cod	Software and Code	-1	-1
17			-25,00	High Risk	NF00473	Nikon RAW (NRW)	nrw	Digital Still Image	-1	-1
18			-25,00	High Risk	NF00474	Olympus RAW	orf	Digital Still Image	-1	-1
19			-25,00	High Risk	NF00356	PCPaint Image	pic; clp	Digital Still Image	-1	-1
20			-25,00	High Risk	NF00475	Pentax RAW	pef; ptx	Digital Still Image	-1	-1
21			-25,00	High Risk	NF00476	Sigma RAW	x3f	Digital Still Image	-1	-1
22			-25,00	High Risk	NF00477	Sony RAW	srf	Digital Still Image	-1	-1
634	X		45,00	Low Risk	NF00650	Broadcast Wave (BWF) unspecified version	wav	Digital Audio	-1	2
635	X		45,00	Low Risk	NF00561	eXtensible Markup Language 1.1	xml	Web Records; Softwar	-1	2
636			45,00	Low Risk	NF00646	Portable Document Format/Archiving (PDF/A-4)	pdf	Presentation and Publ	2	2
637	X		45,00	Low Risk	NF00406	SIARD 1.0	siard	Databases	-1	2
647			47,00	Low Risk	NF00618	SIARD 2.0	siard	Databases	-1	2
648	X	X	48,00	Low Risk	NF00602	Portable Document Format/Archiving (PDF/A-2a) accessible	pdf	Presentation and Publ	2	2
649	X	X	48,00	Low Risk	NF00634	Portable Document Format/Archiving (PDF/A-2b) basic	pdf	Presentation and Publ	2	2
650			48,00	Low Risk	NF00642	Portable Document Format/Archiving (PDF/A-2u) unicode	pdf	Presentation and Publ	2	2
651			49,00	Low Risk	NF00619	SIARD 2.1.1	siard	Databases	-1	2

Num

Liste des formats archivables au CINES

- https://facile.cines.fr/

Liste des formats validables

⚠ Attention : le validateur de formats permet de valider certains formats qui ne sont pas pris en charge par la plateforme d'archivage du CINES.

Format	Nom	PRONOM PIUD	Type MIME	Commentaire	Archivable dans PAC
AAC AAC	Advanced Audio Codings	[fmt/199]		Format Mpeg-4 contenant uniquement un flux audio au format AAC.	✔
AIFF PCM	Audio Interchange File Format	[fmt/414]	[audio/x-aif, audio/x-aiff]	Format audio contenant uniquement un flux PCM.	✔
APNG	Animated Portable Network Graphics	[fmt/935]	[image/vnd.mozilla.apng, image/apng]	L'APNG est une extension du format PNG permettant de réaliser des animations graphiques.	✖
DAE UTF-8 1.4.1	Collada		[application/xml]	Format permettant de stocker des données géométriques sous forme de scènes (plusieurs objets combinés dans le même référentiel), et d'y ajouter des informations supplémentaires pour décrire la scène et les objets (matériaux, environement lumineux, animations, ...) ou pour ajouter des notions sémantiques (relations entre les objets, découpage d'un objet en plusieurs éléments fonctionnels, etc...).	✖
FLAC FLAC 1.2.1	Free Lossless Audio Codec	[fmt/279]	[audio/ogg, audio/x-flac]	Format audio compressé sans perte.	✔
GIF 87a	Graphics Interchange Format	[fmt/3]	[image/gif]	Format image pouvant contenir également des animations.	✔
GIF 89a	Graphics Interchange Format	[fmt/4]	[image/gif]	Format image pouvant contenir également des animations.	✔
GeoTIFF	Geographic Tagged Image File Format	[fmt/155]	[image/tiff]	Format dérivé du TIFF contenant des informations de géoréférencement et de géolocalisation.	✔
HDF5 1.0	Hierarchical Data Format	[fmt/286]		Format de données à caractère scientifique.	✖
HDF5 2.0	Hierarchical Data Format	[fmt/287]		Format de données à caractère scientifique.	✖

Gestion des fichiers

Des formats riches en FAIR

Pour qu'un fichier de données soit lisible et réutilisable

1) Son format doit être identifié explicitement et précisément
2) Son format doit présenter le moins possible de risque (multicritères)
3) Sa forme doit être vérifiée et valide (conforme aux spécifications des formats et codages utilisés)
4) La donnée dans ce format doit être dans un circuit de maintenance qui surveille l'obsolescence

Huma-Num

Quelques bonnes pratiques de gestion

nommage des fichiers

rangement des fichiers

éviter/repérer de doublons

Objectifs de gestion

Le nommage des fichiers et l'organisation des dossiers sont deux opérations importantes pour

- faciliter l'accès aux données (rôle des extensions, du plan de classement)
- faciliter le tri des documents
- faciliter la compréhension de l'ensemble des documents du projet
- éviter la perte et la duplication erronée de fichiers

Convention de nommage

Exemple 1 : Doranum* :

https://doranum.fr/stockage-archivage/comment-nommer-fichiers/

- Donner un nom bref et explicite
- Ne pas mettre d’espace ni de caractères spéciaux
- Indiquer les dates au bon format
- Placer l’élément important en premier
- Indiquer les versions des documents

*Doranum : plateforme de formation en ligne sur la gestion et le partage des données de la recherche selon les principes FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable), réalisée par l’Inist-CNRS et le GIS « Réseau Urfist » depuis 2015

Huma-Num

Convention de nommage

Exemple 2 : Alain Rivet, Marie-Laure Bachèlerie, Auriane Denis-Meyere et Delphine Tisserand « Traçabilité des activités de recherche et gestion des connaissances : Guide pratique de mise en place », 2018

https://qualite-en-recherche.cnrs.fr/wp-content/uploads/2021/08/guide_tracabilite_activites_recherche_gestion_connaissances.pdf

Les intitulés des fichiers :

- doivent être succincts et précis ;
- doivent être uniques ;
- peuvent être caractérisés a minima par une date, un sujet et un type de document ;
- ne doivent pas excéder 31 caractères (extension comprise).

Éviter :

- les accents (é, è ë, ê), cédille (ç), caractères spéciaux (, ; . : ! ? * »%...@ &) ;
- les mots vides : le, la, les, un, une, des, et, ou… ;
- les dénominations vagues, par exemple « divers », « autres », « à classer » ;
- les espaces que l'on peut remplacer par des « _ » (tiret 8).

Convention de nommage

Quelques monstres rencontrés

« vidéo de l'entretien de Monsieur Pierre Martin en 2002 à son domicile du 43 rue Paul Doumer.mp4 »

- Des accents : ne vont pas passer indemne à travers tous les systèmes
- Des espaces : sans précaution certains logiciels vont considérer qu'il y a autant de fichiers que de mots
- Mélange minuscules majuscules : risque de confusion
- Long : risque de troncation
- Contient des métadonnées : c'est un rôle pris en charge par d'autres outils que le simple nommage des fichiers
- Contient des informations à caractère personnel

Mais aussi :

✓ « 马孝珍采访音频.doc » → « é©¬å?ç??é??è®¿é?³é¢?.doc »

✓ « toto.bin »

✓ Des caractères retour chariot en fin de nom

✓ Des caractères BOM « byte order mark » en milieu de mot. Le BOM est un caractère invisible 'espace insécable sans chasse' U+FEFF ou U+FFFE qui indique l'ordre des octets pour composer les caractères multi-octets

Huma-Num

Quelques outils de traitement de fichiers de données

- ✓ Outils d’identification de format
- ✓ Outils de validation du formatage
- ✓ Outils de conversion de format

Identification des format des fichiers

Huma-Num

Comment identifier un format ?

- Quels indices pour « deviner » le format d'un fichier ?
  - Son extension (.jpg, .wav, .pdf...)
  - Un outil de lecture
  - Les métadonnées embarquées
  - Des outils spécialisés : DROID, FIDO, file,...

- Où l'indiquer ?
  - Dans le DMP ; dans les métadonnées

- Comment l'exprimer ?
  - En utilisant des identifiants de référentiels connus (type-mime, PRONOM)

- Pourquoi ?
  - Guider la bonne lecture des données

Huma-Num

Un exemple d’outil d’identification

« Digital Record Object Identification » DROID

- Logiciel d’identification de format de fichiers (Java, New BSD License)

- Lié au registre de formats de fichier PRONOM

- Éditeur : Les Archives nationales du Royaume-Uni

Huma-Num

Un vrac numérique


1archive-yann-bisiou_02-42097.jpg


3-IMG20211113121429.jpg


06-07_Accueil_Iresco.pdf


7savigny_50_01-0cadcdESSINJEUNE.jpg

14_01_zoom.jpg


14_02_zoom.jpg


20100501_FoireDuTrone_003.JPG


20100501_FoireDuTrone_018.JPG


20100501_FoireDuTrone_021.png


20210920_104035.jpg

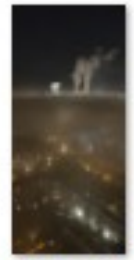

20210920_104035.pdf


20210920_104107.jpg


20210920_104107.pdf

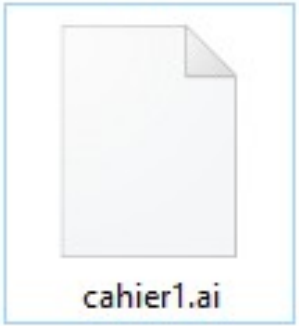

9327631989_9d4ee03ed1_o.jpg


1636737300376.jpg


1636737320330.jpg

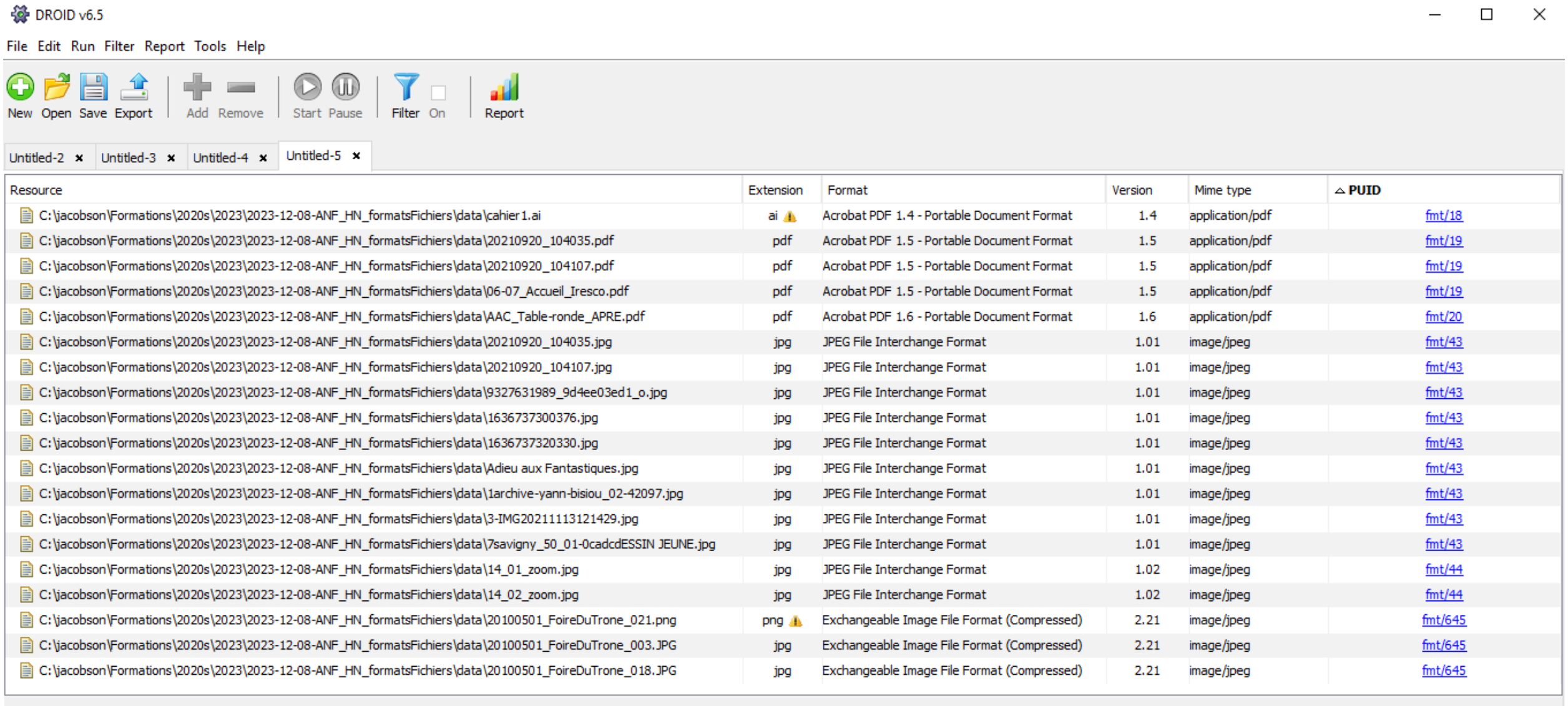

AAC_Table-ronde_APRE.pdf


Adieu aux Fantastiques.jpg


cahier1.ai

Huma-Num

Rapport DROID



Resource	Extension	Format	Version	Mime type	PUID
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\cahier1.ai	ai	Acrobat PDF 1.4 - Portable Document Format	1.4	application/pdf	fmt/18
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20210920_104035.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	application/pdf	fmt/19
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20210920_104107.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	application/pdf	fmt/19
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\06-07_Accueil_Iresco.pdf	pdf	Acrobat PDF 1.5 - Portable Document Format	1.5	application/pdf	fmt/19
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\AAC_Table-ronde_APRE.pdf	pdf	Acrobat PDF 1.6 - Portable Document Format	1.6	application/pdf	fmt/20
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20210920_104035.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20210920_104107.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\9327631989_9d4ee03ed1_o.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\1636737300376.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\1636737320330.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\Adieu aux Fantastiques.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\1archive-yann-bisiou_02-42097.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\3-IMG20211113121429.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\7savigny_50_01-0cadcdESSIN JEUNE.jpg	jpg	JPEG File Interchange Format	1.01	image/jpeg	fmt/43
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\14_01_zoom.jpg	jpg	JPEG File Interchange Format	1.02	image/jpeg	fmt/44
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\14_02_zoom.jpg	jpg	JPEG File Interchange Format	1.02	image/jpeg	fmt/44
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20100501_FoireDuTrone_021.png	png	Exchangeable Image File Format (Compressed)	2.21	image/jpeg	fmt/645
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20100501_FoireDuTrone_003.JPG	jpg	Exchangeable Image File Format (Compressed)	2.21	image/jpeg	fmt/645
C:\jacobson\Formations\2020s\2023\2023-12-08-ANF_HN_formatsFichiers\data\20100501_FoireDuTrone_018.JPG	jpg	Exchangeable Image File Format (Compressed)	2.21	image/jpeg	fmt/645

Huma-Num

Lecture du rapport DROID

- 5 extensions : (pdf, jpg, JPG, ai, png) dont 2 avec warnings
  - le fichier avec l'extension png est en fait un format jpeg
  - l'extension ai (Adobe Illustrator) est en fait un format pdf
- 2 types mime : application/pdf, image/jpeg
- 6 formats :
  - 3 formats PDF : « PDF version 1.4 »(fmt/18) , « PDF version1.5 » (fmt/19) et « PDF version 1.6 » (fmt/20)
  - 3 formats JPEG « JPEG File Interchange Format 1.01 » (fmt/), « JPEG File Interchange Format 1.02 » (fmt/) et « Exchangeable Image File Format (Compressed) 2.21 »

Retour d’expérience sur Nakala

Situation d’octobre 2023 sur un peu plus d’un million de fichiers

- 77 types mime
  - 68 % image/jpeg
  - 81 % image/*
  - 10 % application/pdf
- 186 formats différents au registre PRONOM

Huma-Num

Retour d’expérience sur Nakala

- 18 formats de type application/pdf

	A	B	C	D
1	fmt/276	Acrobat PDF 1.7 - Portable Document Format (1.7)	44953	36,16 %
2	fmt/20	Acrobat PDF 1.6 - Portable Document Format (1.6)	40023	32,20 %
3	fmt/95	Acrobat PDF/A - Portable Document Format (1a)	12997	10,46 %
4	fmt/18	Acrobat PDF 1.4 - Portable Document Format (1.4)	10294	8,28 %
5	fmt/17	Acrobat PDF 1.3 - Portable Document Format (1.3)	9370	7,54 %
6	fmt/19	Acrobat PDF 1.5 - Portable Document Format (1.5)	5449	4,38 %
7	fmt/477	Acrobat PDF/A - Portable Document Format (2b)	564	0,45 %
8	fmt/354	Acrobat PDF/A - Portable Document Format (1b)	389	0,31 %
9	fmt/16	Acrobat PDF 1.2 - Portable Document Format (1.2)	143	0,12 %
10	fmt/478	Acrobat PDF/A - Portable Document Format (2u)	41	0,03 %
11	fmt/157	Acrobat PDF/X - Portable Document Format - Exchange 1a:2001	31	0,02 %
12	fmt/488	Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4	15	0,01 %
13	fmt/476	Acrobat PDF/A - Portable Document Format (2a)	13	0,01 %
14	fmt/479	Acrobat PDF/A - Portable Document Format (3a)	9	0,01 %
15	fmt/480	Acrobat PDF/A - Portable Document Format (3b)	5	0,00 %
16	fmt/15	Acrobat PDF 1.1 - Portable Document Format (1.1)	2	0,00 %
17	fmt/158	Acrobat PDF/X - Portable Document Format - Exchange 3:2002	2	0,00 %
18	fmt/14	Acrobat PDF 1.0 - Portable Document Format (1.0)	1	0,00 %

Les registres de formats

Huma-Num

Les types MIME

En 1996, le standard de l'IETF, RFC-2046 « Multipurpose Internet Mail Extensions » définit des types de média dans le cadre des échanges mail (SMTP)

L'organisme d'enregistrement de ces types de média est IANA « Internet Assigned Numbers Authority »

- ✓ Les identifiants sont structurés en 2 parties
  - ✓ Types ou grandes familles : application, audio, example, font, image, message, model, multipart, text et video
  - ✓ Sous-types : par ex. audio/mpeg ; audio/aac ; image/png ; text/csv….
- ✓ Ils contribuent par exemple dans Nakala à orienter le choix d'une visionneuse adaptée

Le registre types MIME

Registries included below

- application
- audio
- font
- example
- image
- message
- model
- multipart
- text
- video

---

application

Available Formats

CSV

Name	Template	Reference
1d-interleaved-parityfec	application/1d-interleaved-parityfec	[RFC6015]
3gpdash-qoe-report+xml	application/3gpdash-qoe-report+xml	[_3GPP][Ozgur_Oyman]
3gppHal+json	application/3gppHal+json	[_3GPP][Ulrich_Wiehe]
3gppHalForms+json	application/3gppHalForms+json	[_3GPP][Ulrich_Wiehe]
3gpp-ims+xml	application/3gpp-ims+xml	[_3GPP][John_M_Meredith]
A2L	application/A2L	[ASAM][Thomas_Thomsen]
ace+cbor	application/ace+cbor	[RFC9200]
ace+json	application/ace+json	[RFC9431]

https://www.iana.org/assignments/media-types/media-types.xhtml

Le registre PRONOM

http://www.nationalarchives.gov.uk/pronom/

- ✓Éditeur : Les Archives nationales du Royaume-Uni (TNA)
- ✓Identifie les formats avec un PUID « PRONOM unique identifier
- ✓Décrit les formats de fichiers avec de nombreuses propriétés
- ✓Tout le monde peut proposer l'ajout de nouveaux formats par un formulaire. Par ex. suite aux travaux effectués avec le CINES sur la validation de la TEI, ajout de nouveaux formats fmt/1474, fmt/1475, fmt/1476 et fmt/1477
- ✓Un portail web pour accéder à la documentation structurée sur les formats
- ✓Un fichier de « signatures » de formats à utiliser dans le cadre de DROID pour l'identification
- ✓Liens directs par l'outil DROID au registre en ligne de PRONOM par le PUID

Le format « fmt/1 »

Details for: Broadcast WAVE 0 Generic

Save as... XML | CSV Print

Go to: Summary | Documentation › | Signatures › | Compression › | Character encoding › | Rights › | Reference files › Properties ›

Summary

Name	Broadcast WAVE
Version	0 Generic
Other names	BWAVE (0), BWF (0)
Identifiers	MIME: audio/x-wav PUID: fmt/1
Family	WAVE
Classification	Audio
Disclosure	Full
Description	Broadcast WAVE is a chunk-based audio format developed by the European Broadc based on the Microsoft WAVE format, which is in turn based on the generic Resourc Format (RIFF) specification developed by Microsoft and IBM. Structurally, a BWAVE fi a number of chunks, each comprising a four character code chunk identifier, the chu chunk data. It comprises a RIFF header with a WAVE data type identifier, followed b chunks. Every file must include a Broadcast Audio Extension chunk, containing meta exchange of information between broadcasters, a Format chunk, which describes t audio data, and a Data chunk, containing the audio data itself. BWAVE files which c encoded audio data must also include a Fact chunk, containing file-dependent infor audio data, and an MPEG Audio Extension chunk, containing extra information requ the MPEG encoding. A BWAVE identified as generic by DROID likely uses an encodin to PRONOM, or perhaps a structural difference, and users are encouraged to let th

Orientation	Binary
Byte order	Little-endian (Intel)
Related file formats	Has lower priority than Broadcast WAVE (0 PCM Encoding) Has lower priority than Broadcast WAVE (0 MPEG Encoding) Has lower priority than Broadcast WAVE (0 WAVEFORMATEXTENSIBLE Encoding) Has priority over Waveform Audio Has priority over Waveform Audio (PCMWAVEFORMAT) Has priority over Waveform Audio (WAVEFORMATEXTENSIBLE) Is previous version of Broadcast WAVE (1 Generic)
Technical Environment	
Released	01 Jan 1997
Supported until	01 Jul 2001
Format Risk	
Developed by	None.
Supported by	None.
Source	Digital Preservation Department / The National Archives
Source date	11 Mar 2005
Source description	
Last updated	19 Jul 2013
Note	

Validation du bon formatage des fichiers

Huma-Num

Comment tester la validité d'un fichier ?

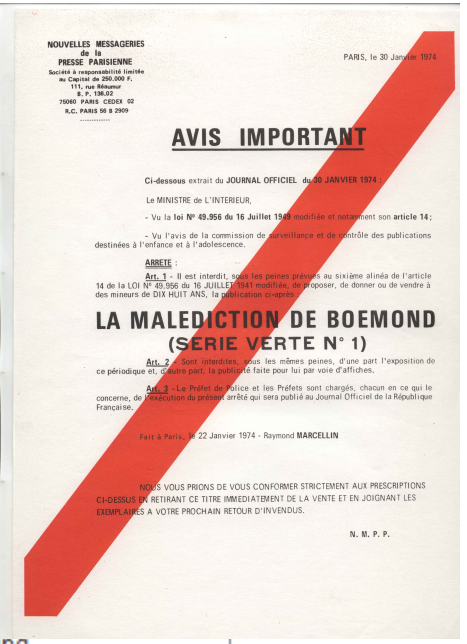
- La lecture avec son éditeur ne suffit pas
  - Peu fiable : les éditeurs sont souvent très tolérants aux erreurs afin de ne pas bloquer la lecture
  - Problème de « juge et partie »
- Il existe des outils dont c'est la finalité
  - Des outils spécialisés sur un format précis : Par ex. jpylyzer (Open Planets Fondation) pour le format JPEG2000
  - Des outils généralistes : Jhove (JSTOR and the Harvard Library)
  - Des intégrateurs d'outils : facile (CINES), FITS (Harvard Library)

Pourquoi valider ?

- ✓ Un fichier mal formé n'a pas toutes les garanties d'être correctement lu par tous les outils
- ✓ Il est hasardeux d'effectuer des conversions sur des fichiers mal formés : pas de garantie du résultat
- ✓ Les résultats aux tests de validation permet à des services d'archives de guider leur décision de prise en charge ou de niveau de service.
  - ✓ Prise en charge de la conservation de l'information
  - ✓ Prise en charge de l'intégrité de la donnée

Valider les données

Test du fichier « EF circulaire NMPP.jpg » avec Jhov


NOUVELLES MESSAGERIES de la PRESSE PARISIENNE

PARIS, le 30 Janvier 1974

AVIS IMPORTANT

Ci-dessous extrait du JOURNAL OFFICIEL du 30 JANVIER 1974 :

Le MINISTRE de L'INTERIEUR,

- Vu la loi N° 49.956 du 16 Juillet 1949 modifiée et notamment son article 14 ;

- Vu l'avis de la commission de surveillance et de contrôle des publications destinées à l'enfance et à l'adolescence.

ARRETE :

Art. 1 - Il est interdit, sous les peines prévues au sixième alinéa de l'article 14 de la LOI N° 49.956 du 16 JUILLET 1949 modifiée, de proposer, de donner ou de vendre à des mineurs de DIX HUIT ANS, la publication ci-après :

LA MALEDICTION DE BOEMOND
(SERIE VERTE N° 1)

Art. 2 - Sont interdites, sous les mêmes peines, d'une part l'exposition de ce périodique et, d'autre part, la publicité faite pour lui par voie d'affiches.

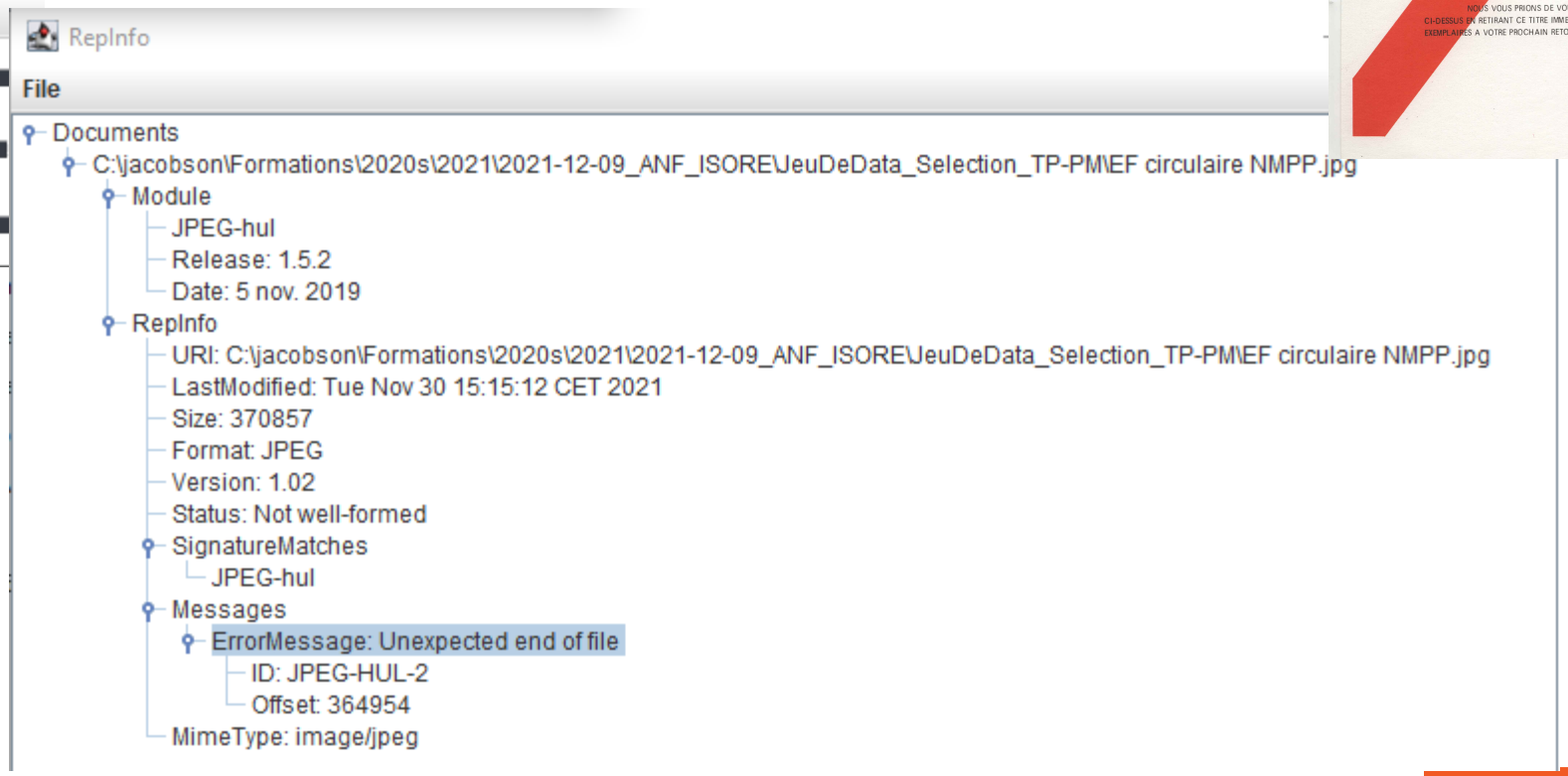
Art. 3 - Le Préfet de Police et les Préfets sont chargés, chacun en ce qui le concerne, de l'exécution du présent arrêté qui sera publié au Journal Officiel de la République Française.

Fait à Paris, le 22 Janvier 1974 - Raymond MARCELLIN

NOUS VOUS PRIONS DE VOUS CONFORMER STRICTEMENT AUX PRESCRIPTIONS CI-DESSUS EN RETIRANT CE TITRE IMMEDIATEMENT DE LA VENTE ET EN JOIGNANT LES EXEMPLAIRES A VOTRE PROCHAIN RETOUR D'INVENDUS.

N. M. P. P.

Format : JPEG 1.02
Statut : not well-formed
Message : Unexpected end of file à l'offset 364954 sur un fichier de 370857 octets



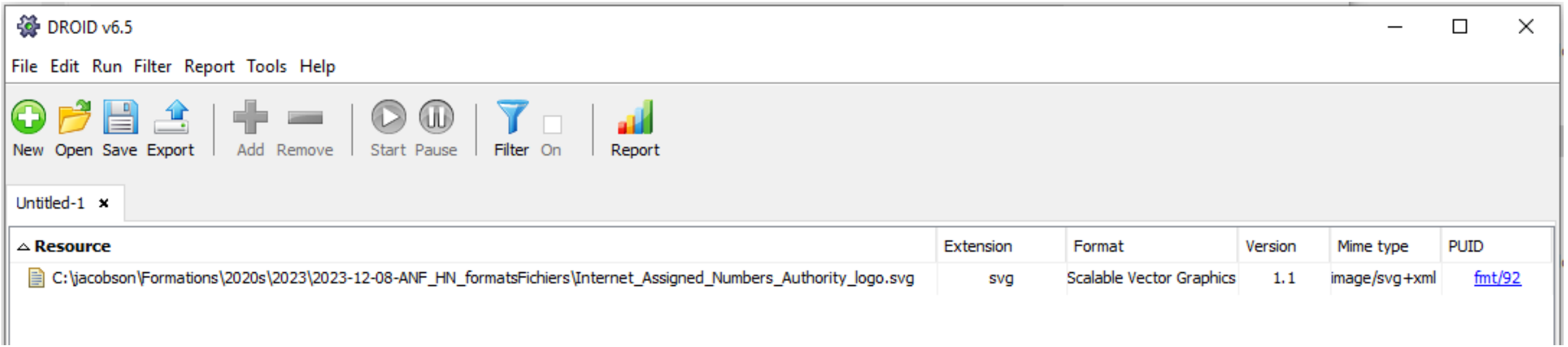
Huma-Num

Test sur un fichier iana.svg

La lecture du fichier avec LibreOffice

Test sur un fichier iana.svg

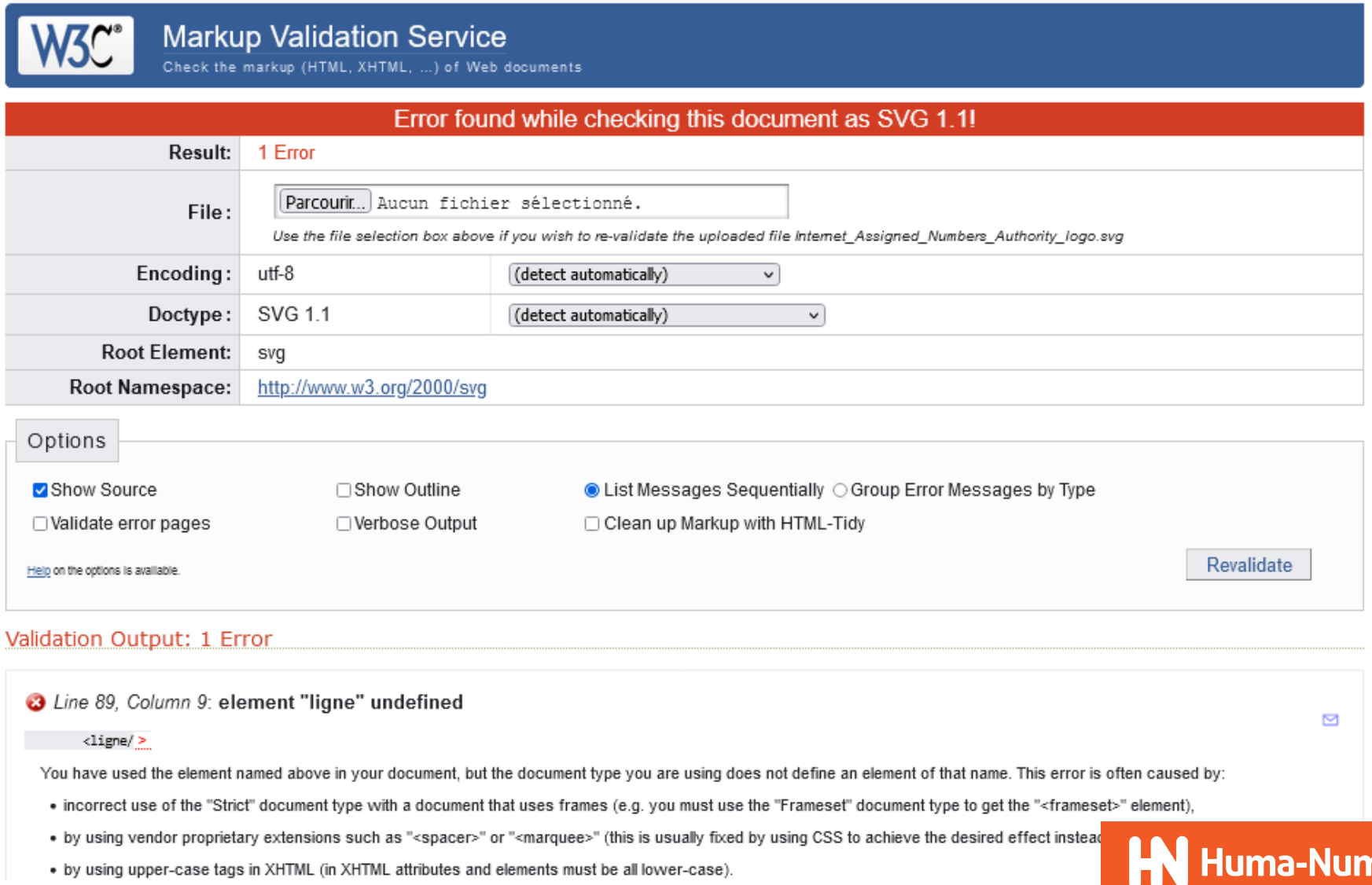
Identification du fichier avec DROID



Huma-Num

Test sur un fichier iana.svg

Validation du fichier avec Markup Validation Service



Stratégies mises en œuvre dans des services

Le service d’archives du CINES

CINES (Centre informatique national de l'enseignement supérieur)

- ✓ Un guide méthodologique pour choisir un format
- ✓ Une liste de formats acceptés
- ✓ Un outil de test : Facile

Huma-Num

La validation, c'est facile

Exemples de résultats

- Formats non pris en charge
  - jpeg-raw, md et mwv, pptx
- 1 fichier pdf non valide
  - Message : « Nombre d'objet ou de flux d'objet invalide »
- 1 format obsolete
  - PDF v. 1.3

Détails	Fichier	Format identifié	Bien formé	Valide	Archivable dans PAC	Commentaire
	nakala-faq.md		✖	✖	✖	
	Poster_ISIDORE_DH2018.pdf	PDF 1.3	✔	✔	✖	
	Human-numGBverticale.png	PNG 1.0	✔	✔	✔	
	equipe-HN-2021.csv	TXT UTF-8	✔	✔	✔	
	VIGNETTE_NAKALA_et_IIIF.png	PNG 1.2	✔	✔	✔	
	Poster_ISIDORE_DH2018.png	PNG 1.0	✔	✔	✔	
	logo-grand-nakala-rvb.png	PNG 1.2	✔	✔	✔	
	logo_rencontre_HN-2026.png	PNG 1.2	✔	✔	✔	
	Human-numGBverticale.jpg	JPEG RAW	✔	✔	✖	
	guide-module-isidore-suggestions.pdf	PDF 1.6	✔	✔	✔	
	ISIDORE_widget_2-0.png	PNG 1.2	✔	✔	✔	
	Human-numGBverticale.ai	PDF 1.5	✔	✔	✔	
	ANF2015-MOREL-C.pdf	PDF 1.5	✖	✖	✖	Corriger automatiquement votre fichier Consulter nos tutoriels pour corriger votre fichier
	lettre_info25-iINSHS_TribuneHumaNum.pdf	PDF 1.4	✔	✔	✔	
	ANF2015-Francart.zip		✖	✖	✖	
	ISIDORE Suggestions_low.wmv		✖	✖	✖	
	HUMA-NUM_PPTX_prez.pptx		✖	✖	✖	
	NAKALA_et_IIIF.mp4	MPEG-4 AVC/AAC LC	✔	✔	✔	

L’entrepôt Nakala

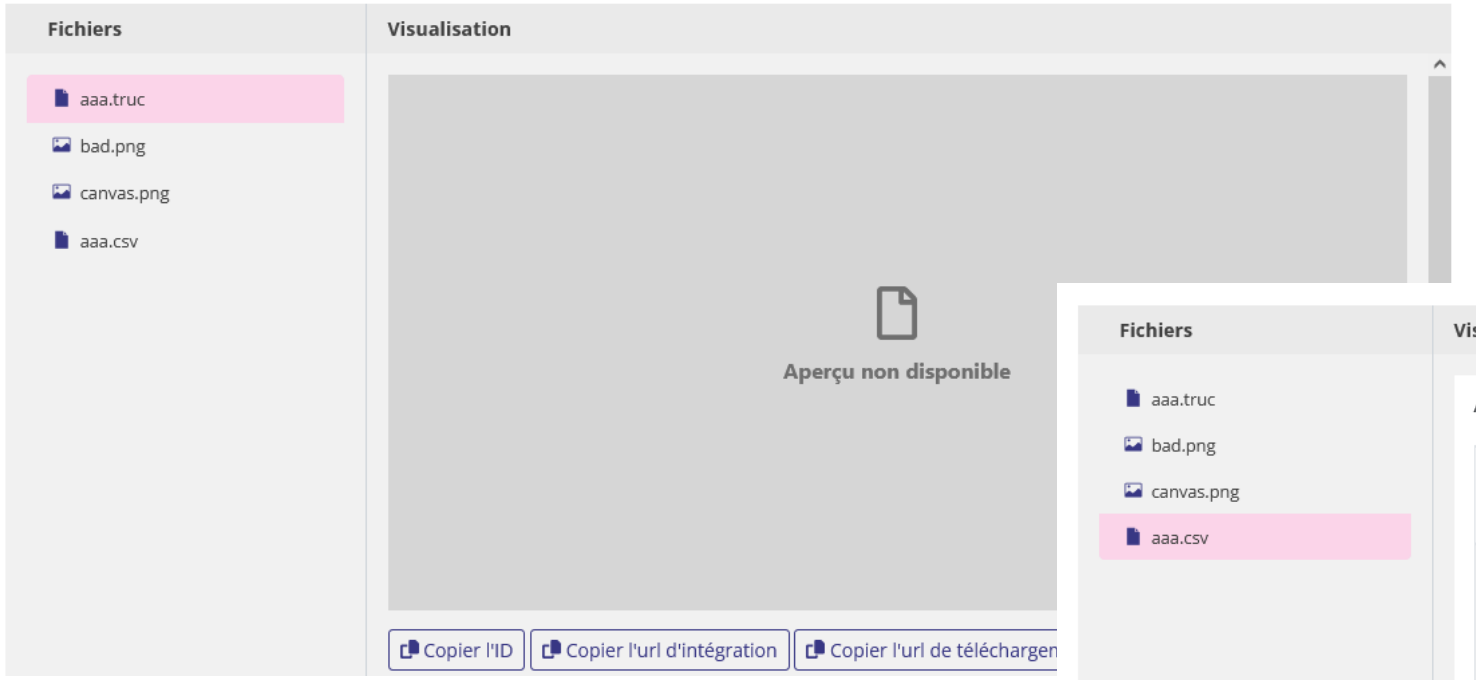
- ✓ Nakala accepte tous les fichiers (= sans contrôle)
  - ✓ Dans tous les cas, les fichiers déposés peuvent être téléchargés (au détail près des droits d’accès)
  - ✓ Certains formats peuvent être visualisés directement sur la page de présentation

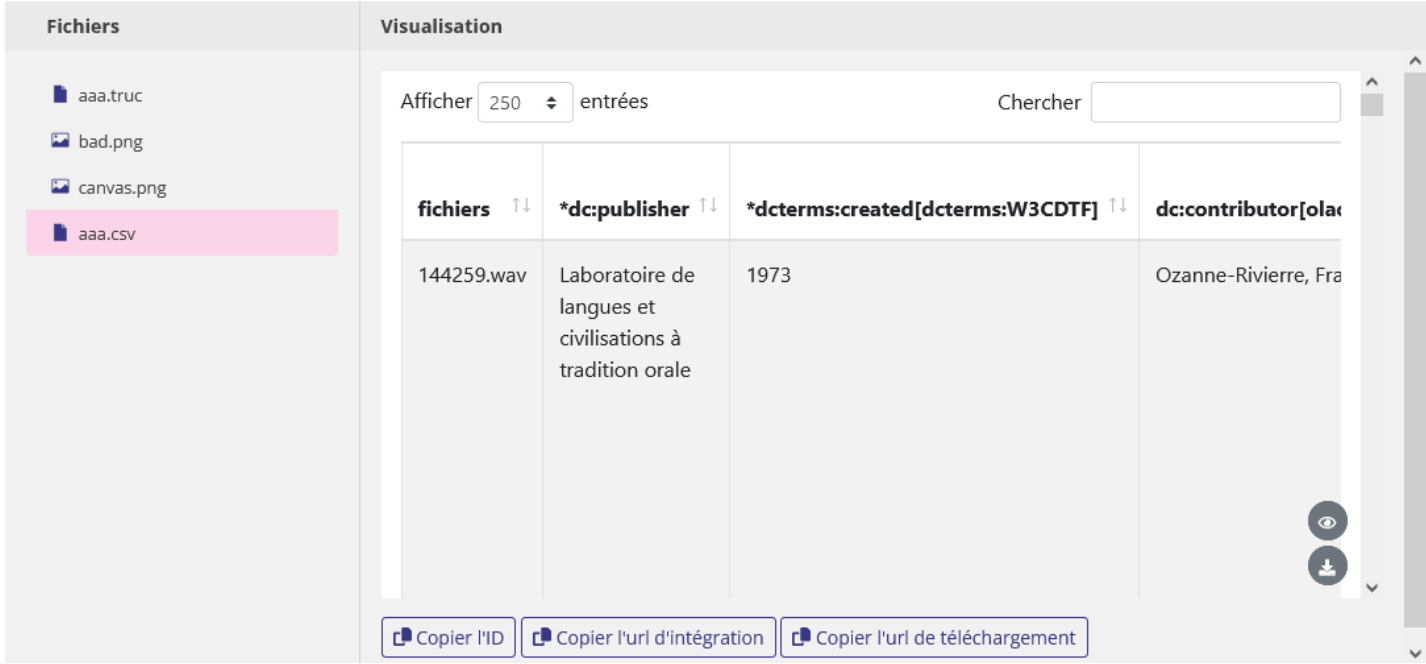
Les visionneuses proposées

- ✓ Fichier image : OpenSeadragon (tif, jpg, jp2, png, pdf)
- ✓ Fichier CSV : DataTables
- ✓ Fichier audio ou vidéo : Plyr
- ✓ Fichier PDF : PDF.js
- ✓ Fichier Markdown
- ✓ Archives : .zip, .rar, .phar, .tar, .tgz, .gz, .bz2
- ✓ Fichier de code : XML, HTML, JSON, etc.

Stratégie dans Nakala

✓ Extension non reconnue vs. Extension reconnue et prise en charge



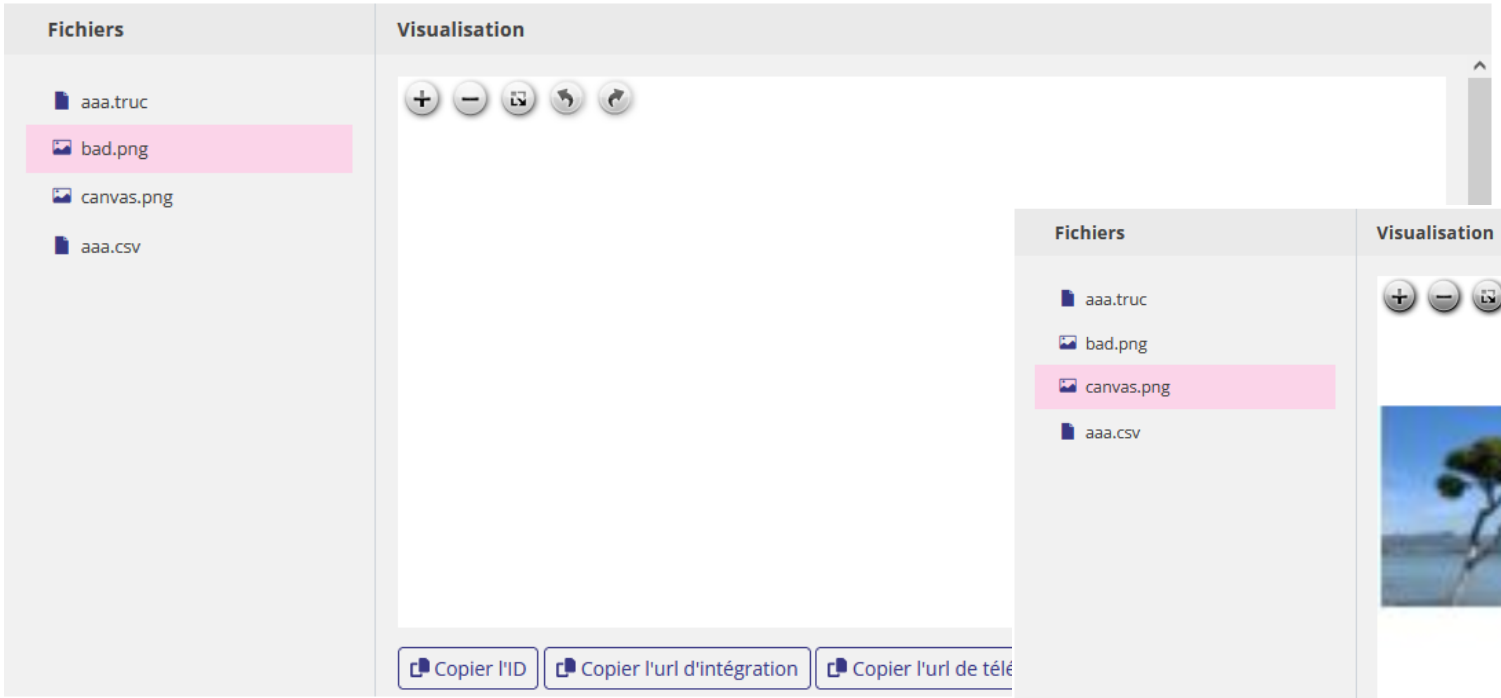


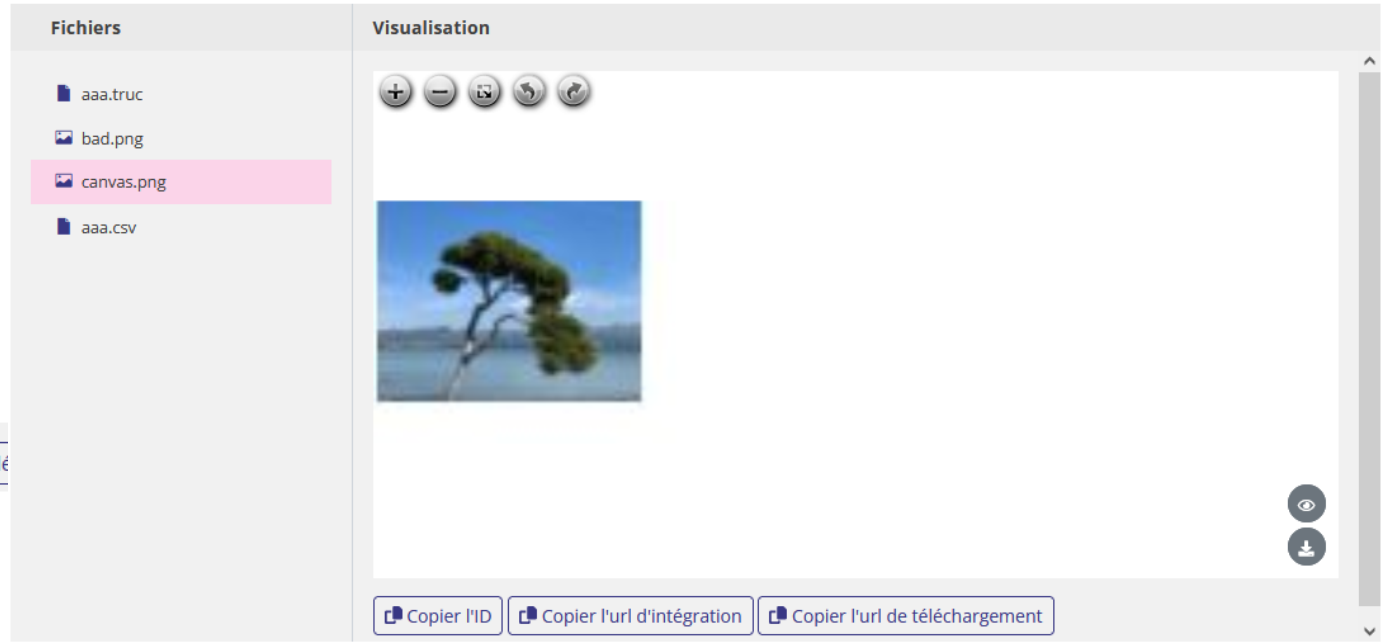
→ ne mettez pas des extensions exotiques, non-standards

Stratégie de Nakala

✓ Fichier bien formé vs. Fichier mal formé

Le fichier bad.png lu par LibreOffice





→ **vérifiez vos données avant de les publier**

Huma-Num

Stratégie de Nakala

- ✓ Enfin ce n'est pas parce qu'une visionneuse de Nakala affiche correctement votre fichier, que celui-ci est valide
- ✓ Si votre fichier est mal-formé bien que bien-représenté
  - ✓ En cas d'obsolescence vous serez amené à convertir votre fichier pour maintenir sa lisibilité, mais cette conversion sera risquée car votre outil de conversion risque de ne pas savoir lire le fichier ou ne sera pas comment corriger l'anomalie
  - ✓ En cas d'archivage le fichier sera testé et donc rejeté pour cause de malformation

→ **vérifiez vos données avant de les publier**

Huma-Num

Nakala (bonnes pratiques)

- ✓ Que faut-il déposer dans Nakala comme fichier ?
  - ✓ un format peu risqué (cf. matrice des risques de la NARA)
  - ✓ un format pris en charge par le CINES (cf. liste sur facile)
  - ✓ un fichier dont le formatage est valide (cf. facile)
  - ✓ un fichier bien nommé (extension standard)
  - ✓ le fichier le plus qualitatif que vous avez qui répond aux précédents critères (le protocole IIIF pour les images permet par exemple de générer des formats de consultation plus léger)
- ✓ Que faut-il décrire du format d'un fichier dans Nakala
  - ✓ Le système va extraire automatiquement des informations techniques de vos fichiers (l'extension, le type mime, le PUID (bientôt), la taille en octets. Il est donc inutile, risqué et inadapté de redéfinir ces informations dans les métadonnées DublinCore.

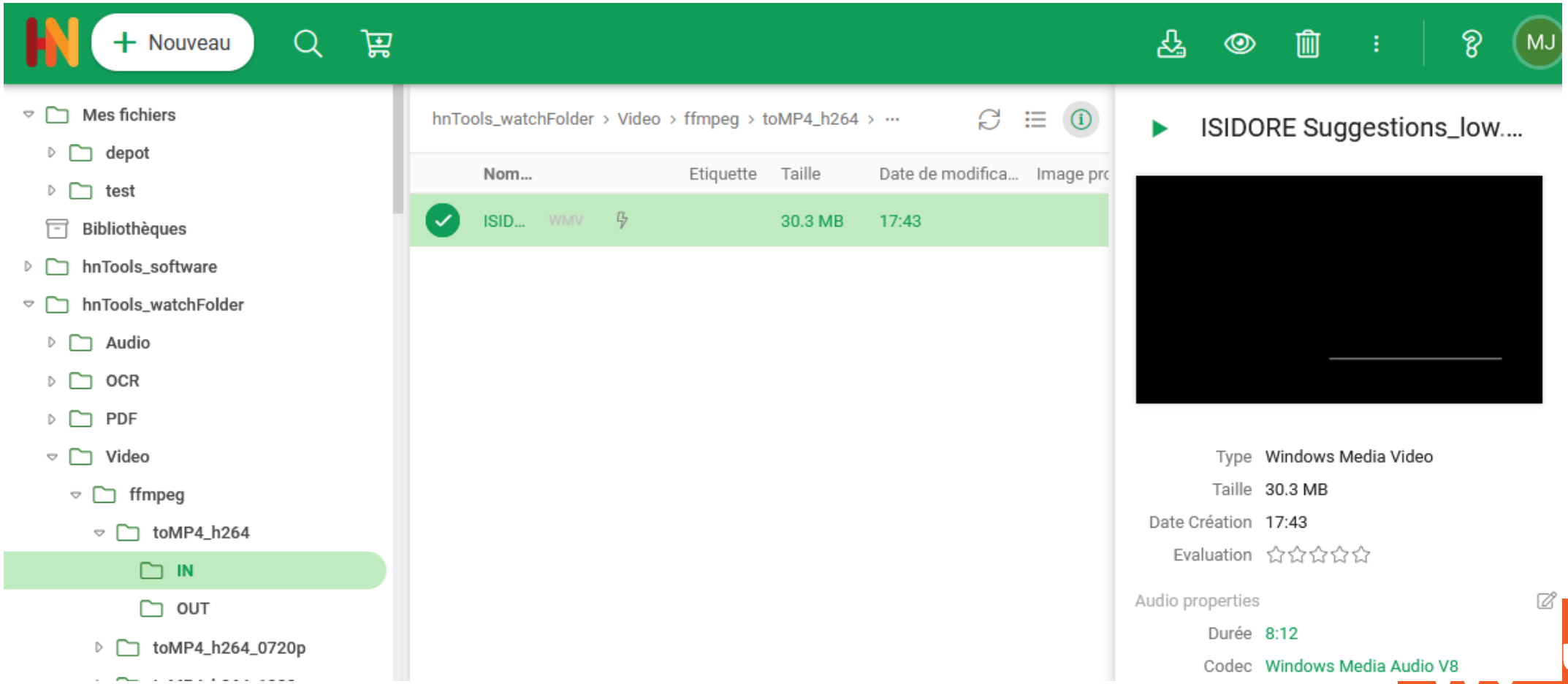
Huma-Num

Conversion de format de fichiers

un format maintenable

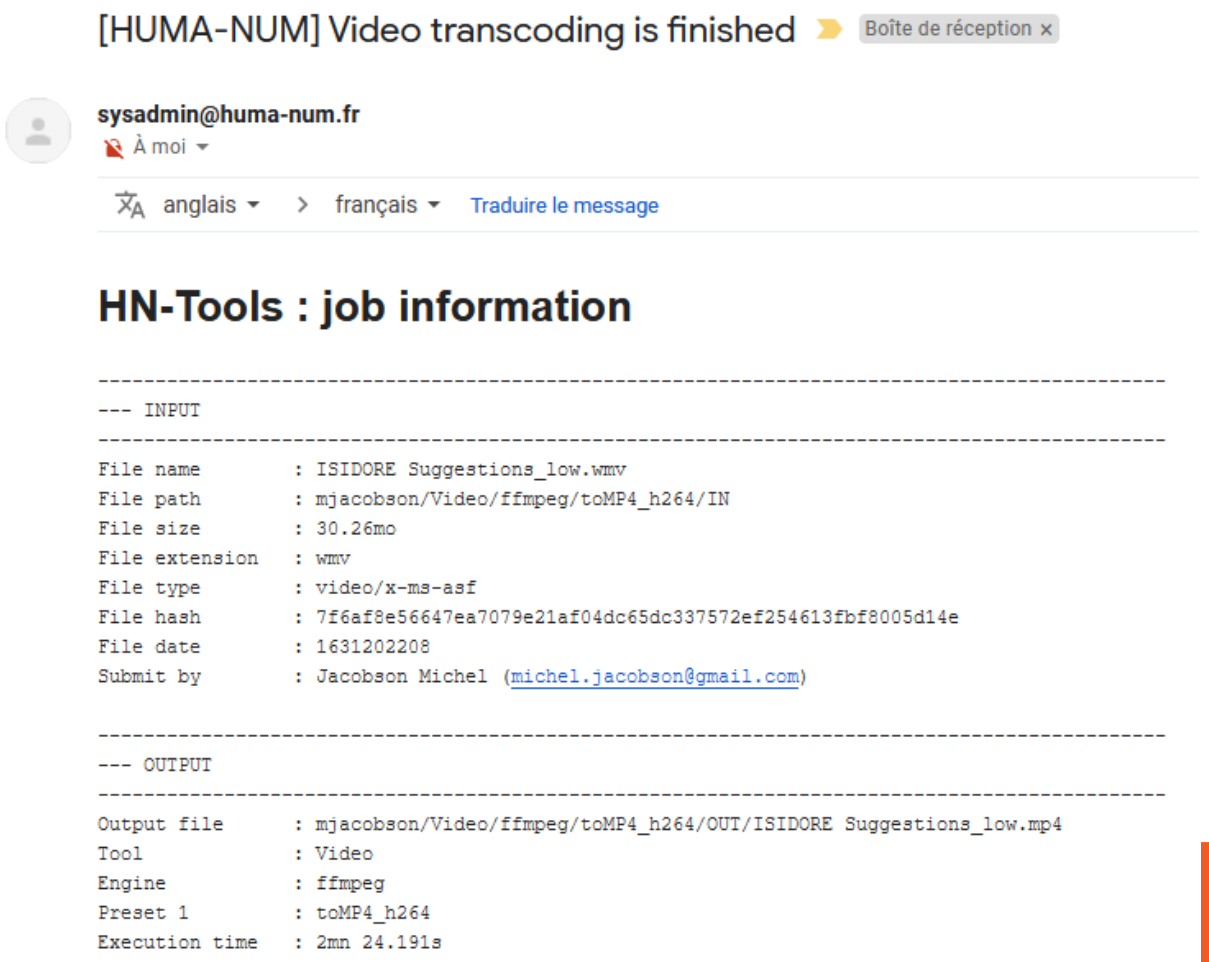
- Pour lutter contre l'obsolescence, pour maîtriser les risques, pour faciliter des usages
  - Trouver un format équivalent pour exprimer toute l'information et permettre de conserver les usages peut s’avérer difficile. Cela se fait parfois avec une perte d'information qu'il faut évaluer
  - Trouver des outils de conversion (il en existe plein) et les qualifier (résultat fidèle et valide)
    - Par ex. ffmpeg pour l'audio-visuel (utilisable à travers Sharedocs)
    - Par ex. ImageMagic pour les formats image
    - ...

Conversion de format pour normalisation WMV → MP4/H264



uma-Num

Conversion de format pour normalisation WMV → MP4/H264



[HUMA-NUM] Video transcoding is finished Boîte de réception ×

sysadmin@huma-num.fr
À moi

anglais › français Traduire le message

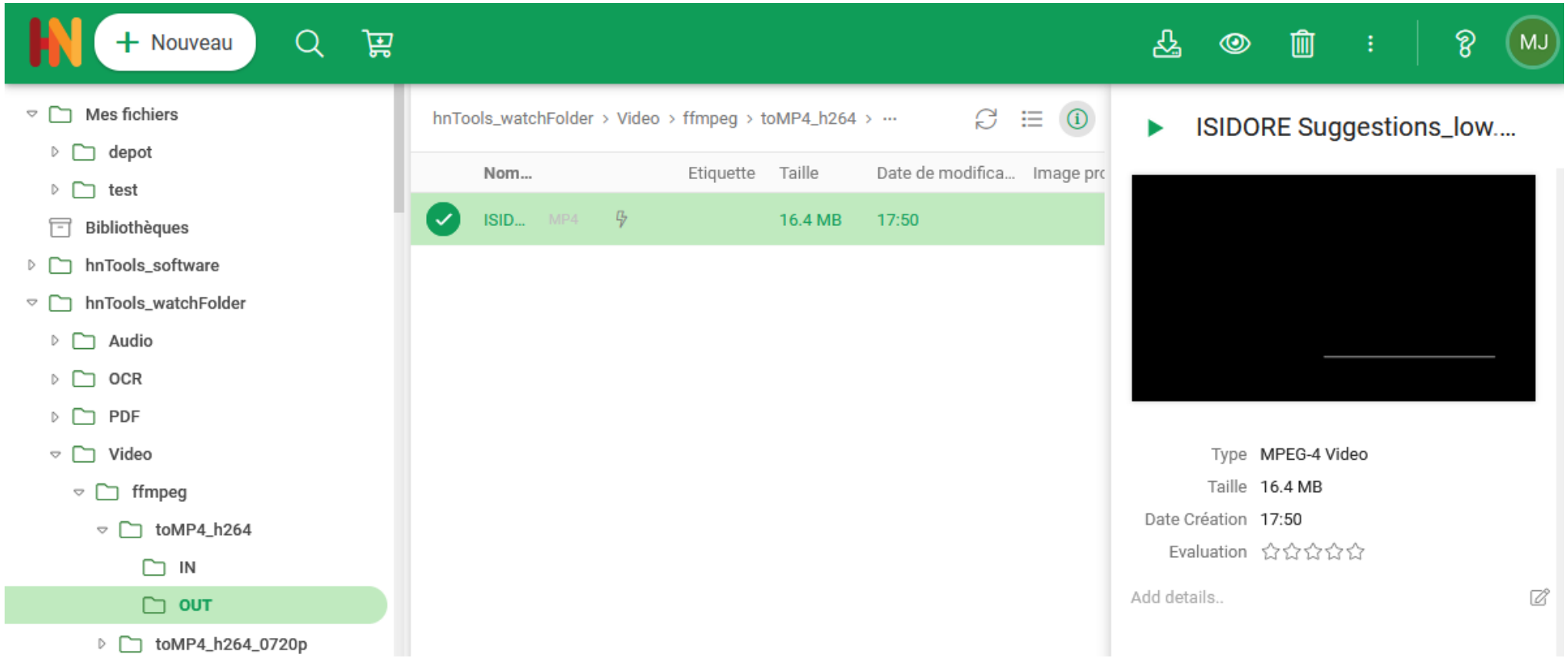
HN-Tools : job information

```
--------------------------------------------------------------------------------
--- INPUT
--------------------------------------------------------------------------------
File name        : ISIDORE Suggestions_low.wmv
File path        : mjacobson/Video/ffmpeg/toMP4_h264/IN
File size        : 30.26mo
File extension   : wmv
File type        : video/x-ms-asf
File hash        : 7f6af8e56647ea7079e21af04dc65dc337572ef254613fbf8005d14e
File date        : 1631202208
Submit by        : Jacobson Michel (michel.jacobson@gmail.com)

--------------------------------------------------------------------------------
--- OUTPUT
--------------------------------------------------------------------------------
Output file      : mjacobson/Video/ffmpeg/toMP4_h264/OUT/ISIDORE Suggestions_low.mp4
Tool             : Video
Engine           : ffmpeg
Preset 1         : toMP4_h264
Execution time   : 2mn 24.191s
```

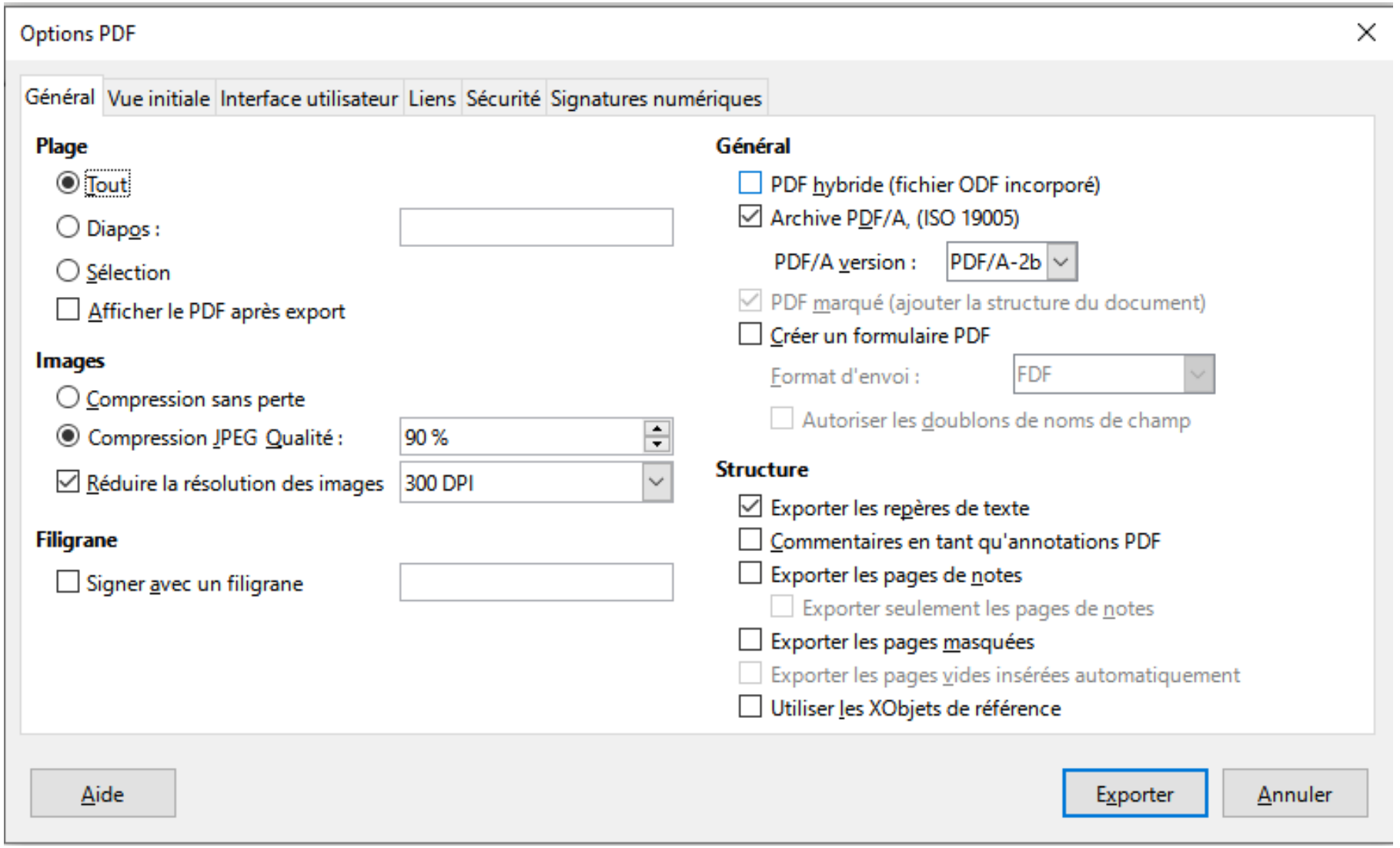
Huma-Num

Conversion de format pour normalisation WMV → MP4/H264



Huma-Num

LibreOffice



Conclusion

Le formatage des fichiers numériques n’est pas qu’un enjeu d’interopérabilité

C’est aussi un enjeu de pérennité, dans la mesure où c’est sur lui que se repose le codage de l’information sous forme de données et donc leur lisibilité