Bibliothèque nationale de France

Formats de données pour la préservation à long terme : la politique de la BnF

Version initiale pour appel à commentaires

Bibliothèque nationale de France

direction des Services et des réseaux
département de la Conservation
service Numérisation

version 2 du 9 avril 2021
émetteur : Etienne CAVALIÉ
affaire suivie par : Bertrand CARON
référence : BnF-ADM-2018-021638-02

Table des mises à jour du document

Version	Auteur	Date	Objet de la mise à jour
0.1	Groupe Formats de la BnF	12/04/2021	Publication de la première version du document pour appel à commentaire.

TABLE DES MATIERES

1. Introduction

1.1. Objet

Ce document décrit la politique globale de la BnF vis-à-vis des formats de données dans la perspective de la préservation à long terme de l'information.

Il détaille en particulier :

- les critères que la BnF a adoptés pour évaluer les formats de données dans cette perspective (section 2.4) ;
- les méthodes d'analyse que la BnF met en œuvre pour mesurer la conformité des données à sa politique (section 2.5) ;
- la liste des formats qui ont sa préférence (section 3) ;
- une description de la plupart de ces formats (section 4) ;
- les stratégies de préservation mises en œuvre par la BnF, en particulier lorsque les données ne sont pas dans un format préféré ou accepté (section 2.6).

1.2. Gestion et évolution

Ce document est publié en avril 2021 dans une version 0.1 pour appel à commentaires.

Il est géré par le groupe de travail BnF « Formats de données et de métadonnées pour la pérennisation de l'information ». Il reflète l'état de la connaissance et du savoir-faire de la BnF en matière de formats de données pour la préservation à une date donnée. Dans la mesure où l'activité du groupe de travail est continue, ce document a vocation à être périodiquement révisé pour refléter l'évolution de ses connaissances et compétences. Les changements suivants sont attendus.

- De nouveaux types de contenu seront traités.
- De nouveaux formats seront ajoutés, voire retranchés si nécessaire.

1.3. Documents applicables et de référence

Certaines restrictions peuvent être appliquées selon la filière d'entrée des collections. Lorsqu'ils seront publiés, les documents incluant des recommandations sur les formats spécifiques à une filière seront référencés dans cette section.

Intitulé	Description	Référence
Norme OAIS	*Norme CCSDS 650.0-M-2 (F) « Modèle de référence pour un système ouvert d'archivage d'information (OAIS) », version française*	*https://public.ccsds.org/Pubs/650x0m2%28F%29.pdf*
Référentiel OCR – version 2	*Référentiel décrivant le profil d'application BnF du format XML-ALTO*	*https://www.bnf.fr/sites/default/files/2018-11/ref_num_ocr_v2.pdf*
Référentiel d'enrichissement des textes	*Référentiel décrivant le profil d'application BnF du format METS enregistrant le résultat d'opérations de reconnaissance optique de la mise en page*	*https://www.bnf.fr/sites/default/files/2018-11/ref_num_enr_texte.pdf*
Référentiel ePub 3	*Référentiel présentant le profil d'application BnF pour le format EPUB 3 de numérisation*	*https://www.bnf.fr/sites/default/files/2018-11/referentiels_num_epub3.pdf*

Référentiel DAISY	*Référentiel présentant le profil d'application BnF pour les fichiers DTBook 2005-3 de numérisation*	*https://www.bnf.fr/sites/default/files/2018-11/ref_num_daisy.pdf*
Référentiel de numérisation des documents opaques	*Référentiel décrivant le profil d'application BnF pour les fichiers TIFF de numérisation*	*https://www.bnf.fr/sites/default/files/2018-11/ref_num_doc_opaques_v2.pdf*
Référentiel de format de fichier image, version 2	*Référentiel décrivant le profil d'application BnF pour le format JPEG 2000 de numérisation*	*https://www.bnf.fr/sites/default/files/2018-11/ref_num_fichier_image_v2.pdf*
Référentiel de traitement des tables	*Référentiel décrivant le format tdmNum de la BnF*	*https://www.bnf.fr/sites/default/files/2018-11/referentiels_num_tdm.pdf*

1.4. Glossaire

Les termes listés ci-dessous sont entendus dans leur acception communément admise au sein de la communauté internationale de la préservation numérique.

Certains d'entre eux proviennent de la norme OAIS citée en section 1.3 « Documents applicables et de référence *» ci-dessus. La forme de leur définition diffère parfois légèrement de celle donnée dans la norme. Comme dans cette dernière, les termes OAIS sont identifiés dans le corps du document par une majuscule initiale (ex. : « le Paquet d'informations »). Une connaissance des notions de base (section 2) et du modèle fonctionnel de la norme OAIS (section 4.1) est conseillée mais non indispensable pour la compréhension du présent document.*

Libellé	Terme anglais correspondant, le cas échéant	Définition
Analyse (du fichier)	/	Ensemble des opérations de préservation (identification, caractérisation, validation) visant à extraire et/ou calculer des métadonnées à partir d'un fichier.
Caractérisation (du fichier)	Characterisation	Opération de préservation consistant à analyser le contenu d'un fichier afin d'obtenir, par simple lecture ou par calcul, des informations sur le contenu, la forme et/ou l'historique du fichier. Elle englobe l'opération d'extraction de métadonnées.

Codec	Codec	Dispositif matériel ou logiciel permettant de mettre en œuvre l'encodage ou le décodage d'un flux de données numérique. Autrement dit un codec prend en entrée des Données encodées dans un Format A et donne en sortie des Données encodées dans un Format B. Appliqué dans un sens on parle d'encodage ; dans l'autre on parle de décodage. Si une opération d'encodage suivie d'une opération de décodage sur des Données X reproduit exactement ces Données X, on parle d'un codec sans perte ; sinon on parle d'un codec avec pertes. Par commodité, on utilisera parfois le terme de « codec » (par opposition à celui de « format conteneur ») pour désigner les règles d'encodage d'un flux de données à l'intérieur d'un fichier.
Communauté d'utilisateurs cible	Designated community	Terme provenant de la norme OAIS et désignant un ou plusieurs groupe(s) d'utilisateurs identifiés par l'institution de conservation pour lesquels celle-ci effectue son activité de préservation.
Données	Data	Terme défini par la norme OAIS comme « une représentation réinterprétable formalisée de l'information, adaptée à la communication, à l'interprétation ou au traitement. Exemple : une séquence de bits, un tableau de nombres, les caractères d'une page, un enregistrement de paroles ou un échantillon de roche lunaire. »
Extraction de métadonnées	Metadata extraction	Opération de préservation consistant à extraire, par lecture du fichier, des informations sur le contenu, la forme et/ou l'historique du fichier.
Fonctionnalité signifiante	/	Terme provenant de la norme OAIS et désignant une modalité d'utilisation d'un objet numérique que l'institution de conservation souhaite préserver sur le long terme, au fil de migrations successives afin de satisfaire les besoins de sa Communauté d'utilisateurs cible. Il ne s'agit pas d'une caractéristique intrinsèque des données mais d'une politique, tacite ou explicite, de l'institution.
Fonds	Fonds	Ensemble de documents de toute nature constitué de façon organique par un producteur dans l'exercice de ses activités et en fonction de ses attributions[1].

[1] Cette définition est tirée du *Dictionnaire de terminologie archivistique*, Direction des archives de France, 2002, disponible en ligne sur <https://francearchives.fr/file/4f717e37a1befe4b17f58633cbc6bcf54f8199b4/dictionnaire-de-terminologie-archivistique.pdf> (consulté le 22 mars 2021).

Format conteneur	Container format	Format de données adapté au stockage et/ou au transfert des Données dans un ordinateur. Un format conteneur est généralement un moyen de combiner plusieurs flux de Données dans un seul flux. Par exemple on peut combiner une piste audio avec une piste vidéo dans un conteneur mp4. Il peut arriver que l'on stocke un flux de données sans l'embarquer dans un format conteneur.
Format de données	Data format	Ensemble des règles d'encodage d'une information sous forme de données binaires interprétables par un ordinateur.
Identification (du format)	Identification	Opération de préservation représentant l'étape préliminaire de l'analyse de fichier. Elle consiste à déterminer le format d'un fichier. On peut considérer que c'est une forme sommaire de caractérisation.
Information	Information	Terme défini par l'OAIS comme « toute connaissance pouvant être échangée. Lors de l'échange, elle est représentée par des données. Exemple : une séquence de bits (les données) accompagnée d'une description permettant d'interpréter cette séquence de bits comme des nombres représentant des mesures de températures en degrés Celsius. » L'Information est ce que l'on cherche à pérenniser.
Intention de préservation	Preservation intent	Ensemble des fonctionnalités et propriétés d'un contenu numérique que l'institution de conservation considère, tacitement ou explicitement, comme « signifiantes » et s'engage à préserver sur le long terme.
Métadonnées de caractérisation	/	Métadonnées extraites ou calculées par un outil logiciel à partir d'un fichier. Il peut s'agir de métadonnées descriptives, techniques et/ou de provenance.
Métadonnées de provenance	Provenance metadata	Métadonnées documentant l'historique de production, de transfert et de préservation d'un fichier.
Métadonnées descriptives	Descriptive metadata	Métadonnées documentant le contenu informationnel d'un fichier.
Métadonnées techniques	Technical metadata	Métadonnées documentant la forme et les caractéristiques techniques d'un fichier
Mesures techniques de protection	Digital Rights Management	Dispositifs visant à contrôler l'usage de contenus numériques.
Migration	Migration	On entend par « migration » toute opération de préservation destinée à modifier le format de données numériques pour assurer la pérennisation de la totalité, ou au moins d'une portion conséquente, de l'information qu'elles véhiculent, c'est-à-dire les propriétés signifiantes de l'Information.

Paquet d'informations	Information Package	Terme provenant de la norme OAIS et désignant un conteneur abstrait réunissant les contenus numériques faisant l'objet de la préservation et l'ensemble des ressources nécessaires à leur compréhension, réutilisation et restitution sur le long terme.
Préservation du train d'octets	Bit-level preservation	Activité consistant pour une institution de conservation à garantir l'intégrité des données, c'est-à-dire la suite fidèle de 0 et de 1. On parle aussi de « stockage sécurisé ».
Préservation sémantique	Functional / logical preservation	Par opposition à la simple préservation du train d'octets, activité consistant pour une institution de conservation à assurer sur le long terme l'accessibilité, l'utilisabilité et la compréhensibilité de l'information que les données véhiculent. La préservation sémantique inclut la préservation du train d'octets. Le terme « préservation (numérique) » renvoie à cette activité.
Producteur	Producer	Terme provenant de la norme OAIS et désignant tout agent qui fournit des informations à préserver à l'institution de conservation.
Profil d'application (d'un format)	Application profile	Ensemble des règles spécifiques à une institution déterminant les contraintes additionnelles à la spécification qu'elle applique aux fichiers d'un format donné.
Propriété signifiante	Significant property	Terme correspondant à la notion OAIS de « Propriété d'information de transformation » et désignant une caractéristique du contenu ou de la forme d'un objet numérique que l'institution de conservation souhaite préserver sur le long terme, au fil de migrations successives, afin de satisfaire les besoins de sa Communauté d'utilisateurs cible. Il ne s'agit pas d'une caractéristique intrinsèque des données mais d'une politique de l'institution.
Référentiel de numérisation	/	Terme spécifique à la BnF désignant un document détaillant les attentes de la BnF concernant les livrables d'opérations de numérisation.
Spécification (d'un format)	Specification	Documentation complète décrivant la structure d'un format.
Stratégie (ou trajectoire) de pérennisation	Preservation strategy	Méthode envisagée par l'institution de conservation pour préserver l'information contenue dans un fichier d'un format donné.
Système ouvert d'archivage d'information	Open Archival Information System	Ce terme désigne ici la norme OAIS citée en section 1.3 « Documents applicables et de référence ».
Validation (de fichier)	Validation	Opération de préservation consistant à évaluer la conformité d'un fichier au regard de sa spécification.

1.5. Abréviations

Sigle ou abréviation	Développement
BnF	Bibliothèque nationale de France
DRM	*Digital Rights Management*
OAIS	*Open Archival Information System*

2. Principes directeurs de la politique

2.1. Contexte

Dans le cadre de ses missions de collecte, conservation, enrichissement et communication du patrimoine documentaire national, la Bibliothèque nationale de France (BnF) a développé des techniques appropriées de conservation curative ou préventive. De même qu'elle surveille l'état de ses collections physiques, elle réalise une veille technologique sur les formats d'encodage des fichiers. La donnée numérique étant stockée sur support sous une forme binaire – une série de « 0 » et de « 1 » –, la préservation numérique consiste à garantir à la fois l'intégrité des données (préservation du train d'octets) et l'accessibilité, la compréhensibilité et l'utilisabilité des données pour une Communauté d'utilisateurs cible (préservation « sémantique »).La connaissance du format d'encodage est donc une condition nécessaire si ce n'est suffisante de la transformation des données en information accessible, compréhensible et utilisable par les usagers.

Au-delà des quelques dizaines de formats les plus courants, la variété de formats utilisés par les producteurs de données est considérable, particulièrement dans les communautés scientifiques spécialisées. Une bonne partie d'entre eux est susceptible de se retrouver dans les collections numériques des institutions de conservation. La diversité des formats augmente d'autant les risques pesant sur les capacités de telles institutions à restituer correctement leurs collections. Parmi ces risques multiples, on peut citer :

- l'incapacité à identifier le format et donc l'outil de restitution adapté ;
- la disparition de logiciels de lecture ou leur indisponibilité en raison de leur coût, de la politique de l'entreprise qui les développe ou de leur mode d'achat ;
- l'utilisation d'outils inadaptés ou mal employés donnant lieu à une restitution non fidèle, ou à la perte d'information involontaire à la suite d'une migration mal maîtrisée.

Afin de réduire ces risques et l'investissement de l'institution (achat et maintenance de matériel, de logiciels, de normes, développement et maintien de compétences, etc.), il est donc souhaitable, autant que faire se peut, de se limiter à une liste de formats maîtrisés.

- En tant qu'institution commanditaire, la BnF veille à choisir avec soin les formats dans lesquels elle demande la production de données.
- Lorsque la BnF acquiert des données de producteurs, elle négocie la remise de celles-ci dans un format maîtrisé.
- Lorsque la BnF hérite de données dans un format qu'elle ne peut choisir, elle envisage en cas de risque avéré portant sur elles de les convertir dans un format maîtrisé.

2.2. Objet du document et lectorat attendu

L'objectif de la BnF dans le présent document est donc de formaliser les principes de politique de l'institution concernant la forme des données numériques dont elle a la charge. Ces principes s'accompagnent de la liste des formats de données pour lesquelles elle s'engage à maintenir la compétence et les outils nécessaires à leur restitution, compréhension et exploitation, et à faire bénéficier ses partenaires de son expertise. Le présent document est donc une composante essentielle de la politique de préservation numérique de la BnF.

Ce document présente également un intérêt pour :

- les bibliothèques conservant des données numériques qui souhaiteraient développer une politique formats ou comparer la leur à celle de la BnF ;
- les donateurs potentiels soucieux de fournir leurs créations sous une forme maîtrisable par la BnF ;
- et plus généralement, tout producteur de données intéressé par leur préservation, qu'il soit susceptible ou non de les confier à la BnF.

Les fiches Format, en section 4, s'adressent à un public plus spécialisé. Elles doivent permettre aux personnels chargés de la production et de la préservation numériques de déterminer l'adéquation des formats décrits aux besoins et aux moyens de leur institution. En outre, elles décrivent les paramètres du profil d'application BnF du format qui semblent essentiels à celle-ci dans une perspective de préservation à long terme afin que les institutions ne disposant pas de personnel compétent puissent connaître et éventuellement adopter les choix de la BnF.

Toute question portant sur ce document pourra être adressée à formats.num@bnf.fr.

2.3. Une politique, des pratiques

Le présent document vise à formaliser la politique de la BnF concernant les formats de données pour la préservation, c'est-à-dire l'ensemble des choix qu'elle a faits en connaissance de cause et après des études ou l'accumulation d'une expérience de longue durée. Pour autant, les pratiques de la BnF dans ce domaine ne se limitent pas à celles décrites ici. Ainsi, on trouvera en section 2.6.1.2 ci-dessous quelques exemples de pratiques appliquées à des formats qui ne figurent pas dans les listes de ce document. Néanmoins, ces pratiques ont vocation à être progressivement consolidées et à intégrer la politique générale lorsque la BnF jugera qu'elles sont suffisamment étayées.

Certains autres cas particuliers peuvent amener la BnF à déroger, temporairement ou durablement, à la politique décrite dans le présent document :

- pour les données menacées de perte dans un délai qui ne permet pas à la BnF de mener une instruction suffisamment approfondie, et en attente de cette dernière ;
- pour les données qui n'appartiennent pas à la BnF, en particulier lorsqu'elle fournit un service de préservation pour un tiers[2] et que ce dernier dispose d'une maîtrise supérieure à celle de la BnF sur les contenus ;
- pour les données qui ne sont pas l'objet principal de la préservation (données « master »)[3], c'est-à-dire les dérivés, éléments d'accompagnement et documentation diverse.

Dans certains de ces cas, la stratégie adoptée par la BnF peut se limiter à garantir l'intégrité des données (« préservation du train d'octets »).

2.4. Critères de pérennité pour le choix d'un format de données

La capacité des institutions de conservation de restituer fidèlement le patrimoine numérique et d'en conserver l'utilisabilité repose sur un certain nombre de facteurs dépendants du format d'encodage dans lequel les données sont enregistrées. Dans le but d'évaluer les formats les mieux adaptés à la conservation à long terme, la BnF a élaboré la liste de critères suivante, à la suite d'autres institutions de conservation à travers le monde[4]. **Cette liste n'est pas classée par ordre d'importance.**

On distinguera douze critères objectifs et deux critères subjectifs. Ces deux derniers critères sont fondamentaux dans l'évaluation de la capacité de l'organisation à conserver ses données numériques et rappellent que la pérennité n'est pas une caractéristique intrinsèque d'un format mais découle de l'engagement de l'institution qui les utilise à disposer d'outils de lecture, d'analyse, de traitement, et de migration, ainsi que des compétences nécessaires à leur utilisation.

Afin d'élaborer sa propre politique sur les formats pour la préservation numérique, chaque organisation est invitée à pondérer l'évaluation des critères objectifs à l'aune de ses priorités. Développer une politique sur les formats consiste à expliciter et justifier le compromis entre les objectifs définis par les critères mentionnés ci-dessous. Ainsi, si une organisation dispose déjà d'une forte compétence en interne et d'un outillage adéquat, mais que son budget destiné à l'augmentation de ses capacités de stockage est limité, elle valorisera sans doute la compacité (CPO-COM) du format et réduira l'importance de la complexité (CPO-SIM).

Ce présent document résulte donc de la démarche spécifique de la BnF, incluant l'évaluation pondérée des formats selon les deux critères subjectifs. Il présente les formats qui demandent un investissement raisonnable dans une optique de préservation à long terme, et reflète l'état actuel des connaissances, des politiques, des pratiques et des capacités de la BnF. Il n'a donc vocation ni à l'exhaustivité ni à l'universalité.

---

[2] Voir sur le site institutionnel de la BnF la page « Prestation d'archivage numérique », disponible sur <https://www.bnf.fr/fr/prestation-archivage-numerique>, (consulté le 31 mars 2021).

[3] , Ces données sont nommées « Objets-données de contenu » par l'OAIS.

[4] On en trouvera des exemples en section 5.2 du présent document.

2.4.1. Critères objectifs

Afin d'y faire référence plus aisément, ces critères sont identifiés par une série de lettres : « CPO » (pour « Critère de pérennité objectif ») et les trois premières lettres du terme le plus significatif.

Identifiant et intitulé(s)	Définition	Justification
CPO-SOC. Communauté d'utilisateurs / Sociabilité	Le format est-il largement utilisé dans sa communauté cible ? Par le grand public ? Par les institutions de conservation ?	L'utilisation d'un format au sein de sa communauté est un indice de son adaptation aux besoins spécifiques de cette communauté. Un format également utilisé au-delà des institutions de conservation fournit des garanties supplémentaires, car les moyens de telles institutions pour maintenir un format sont limités comparés à ceux des industries culturelles. Ce critère est lié à celui des outils disponibles : plus la communauté d'utilisateurs est conséquente, plus elle est susceptible d'avoir développé ou fait développer des outils adaptés.
CPO-DOC. Documentation	Les spécifications du format sont-elles publiées ? Si oui, sont-elles maintenues par un organisme de normalisation reconnu ? Quel est leur coût ?	Si les spécifications du format sont librement accessibles, il est possible à tout un chacun d'en comprendre la structure et, si le format est également libre, de développer des outils qui le prennent en charge. La documentation peut être partielle : ainsi certains industriels publient-ils des documents décrivant seulement la structure générale de leur format. Un processus de normalisation garantit que l'on dispose de spécifications décrivant l'intégralité des caractéristiques significatives d'un format. Les formats maintenus par des organismes de normalisation nationaux (AFNOR) ou internationaux (ISO, W3C, IETF, etc.) offrent de meilleures garanties de maintenance et de disponibilité des spécifications, mais peuvent se révéler payants. On parle de standards *de facto* lorsqu'une spécification produite par une organisation est majoritairement adoptée dans une communauté bien qu'elle n'ait pas fait l'objet d'un processus de normalisation officiel.
CPO-LIB. Liberté d'utilisation	Existe-t-il des obstacles juridiques à l'utilisation du format ?	Si un format peut être totalement ouvert (documenté et utilisable par quiconque), il arrive également que des limitations d'usage pèsent sur des formats documentés, notamment en raison de brevets accordant des droits de propriété industrielle déposés au profit d'une organisation donnée. Ces brevets peuvent limiter ou interdire le développement d'outils prenant en charge le format.

CPO-AUT. Indépendance / autonomie	L'utilisation du format requiert-elle d'autres formats, encodages, environnement logiciels ou matériels ?	La consultation et l'utilisation d'un fichier numérique sont systématiquement dépendantes d'un environnement technique. Outre la dépendance à un environnement logiciel qui peut être propriétaire, abordée dans le critère « Liberté d'utilisation », l'utilisation de certains formats est tributaire d'environnements matériels, de bibliothèques logicielles, ou d'éléments habituellement non embarqués dans le fichier (par exemple, la dépendance de la plupart des PDF aux polices installées sur le poste de l'utilisateur).
CPO-ROB. Robustesse	Le format dispose-t-il de mécanismes pour repérer, ignorer voire corriger des parties altérées du signal ?	Ce critère évalue la résistance des fichiers de ce format à l'altération. Cette altération peut provenir d'une dégradation du support ou d'une erreur du matériel de lecture, mais elle est plus souvent encore le fait d'un transfert interrompu, notamment en raison d'une défaillance du réseau ou de la connectique. De ce fait, les formats de fichiers destinés à être échangés sur le réseau par le biais du *streaming* sont souvent conçus pour être robustes. La robustesse inclut les notions de résilience et de résistance à l'erreur. Elle dépend de la structure du format. Elle peut être renforcée par la présence d'empreintes numériques caractérisant chacune des zones d'un fichier, ce qui permet à un outil de validation d'identifier précisément la zone corrompue On notera que certaines méthodes de compression, particulièrement celles s'appliquant à l'ensemble du fichier et non à chacune de ses parties, peuvent avoir un effet négatif sur la robustesse des données. En raison de la réduction de la redondance que ces méthodes impliquent, une altération pourra affecter simultanément plusieurs zones du fichier.
CPO-COM. Compacité	Le format exprime-t-il une quantité d'information conséquente dans un espace contraint ? Si cette compacité est liée à une méthode de compression, celle-ci est-elle réversible (sans perte d'information) ou non ?	Un des risques majeurs pesant sur la pérennité des données est le risque budgétaire. Si les budgets que l'utilisateur peut allouer à l'achat d'espace de stockage sont limités, le critère de compacité peut devenir décisif. La compacité peut être liée à la structure d'encodage des données ou, le cas échéant, à la méthode de compression. Une compression sans perte, également dite « réversible », permet généralement une réduction significative du poids du fichier tout en garantissant la possibilité, grâce à la même méthode, de décompresser le fichier et d'obtenir une copie exacte, au bit près, du fichier source.

CPO-OUT. Disponibilité d'outils de traitement	Existe-t-il des outils de restitution, de validation, d'analyse, de migration ? L'organisme de maintenance du format en développe-t-il officiellement ?	La disponibilité d'outils de traitement est liée à la documentation (plus un format est documenté, plus les chances sont élevées que des outils de traitement l'exploitent avec précision), à la sociabilité (plus sa communauté d'utilisateurs est étendue, plus elle a de chances d'avoir développé des outils de traitement nombreux et efficaces). On prêtera une attention particulière au degré de prise en charge native par les navigateurs web comme l'indice de l'adoption du format et la garantie supplémentaire de la diffusion d'outils de lecture auprès du grand public.
CPO-ADD. Contenu additionnel embarqué	Le format permet-il d'embarquer des flux complémentaires nécessaires à l'utilisation, l'identification et la gestion du fichier (métadonnées, documentation, visuels associés, etc.) ?	Outre le contenu principal du fichier, le format peut être conçu pour permettre d'embarquer du contenu additionnel nécessaire à son utilisation. En l'absence de telles dispositions, l'utilisateur sera amené à transmettre séparément ce contenu, avec les risques de perte que cela implique. Le contenu additionnel peut être constitué de métadonnées permettant l'identification précise du contenu du fichier, des agents ayant contribué à sa création, des droits associés, etc. Il peut également s'agir de flux spécifiques comme un visuel pour un fichier audio MP3, des sources pour un document PDF/A-3 résultant d'une migration, etc.
CPO-PRO. Mécanismes de protection	Le format dispose-t-il de mécanismes de protection de son contenu ?	Ce critère est ambivalent. Certains mécanismes de protection visant à interdire l'accès ou l'utilisation d'une ou plusieurs des fonctions du fichier, tels que les DRM, peuvent empêcher l'utilisateur de mener à bien des opérations à but de conservation. Ceux à l'inverse qui n'ont pas pour but de limiter l'accès ou l'utilisation, tels les signatures électroniques embarquées comme XMLDSig pour XML, sont un atout pour garantir l'intégrité et l'authenticité du fichier.
CPO-SIM. Simplicité	Le format a-t-il une structure simple ou complexe ?	Maintenir une compétence et des outils sur un format complexe demandera nécessairement un investissement plus lourd que sur un format simple. Une méthode de compression ajoute un niveau de complexité supplémentaire. Selon les méthodes, ce niveau de complexité peut être conséquent (ex. : JPEG 2000) ou plus limité (ex. : MP3).

CPO-STA. Stabilité / évolutivité	Le format connaît-il une évolution soutenue et des versions qui se succèdent à une fréquence élevée ?	Suivre l'évolution d'un format fréquemment mis à jour peut s'avérer complexe et coûteux en investissement ; à l'inverse, un format qui ne connaît plus d'évolutions depuis plusieurs années supposera un effort moindre d'adaptation à son évolution. Plus généralement, ce critère interroge le stade de développement du format : est-il dans sa phase d'expansion initiale ou a-t-il atteint sa maturité, voire est-il toujours maintenu ?
CPO-TRA. Transparence	Le format est-il aisément lisible et compréhensible ou sa structure est-elle opaque ?	En l'absence d'outils spécifiques, un format relativement transparent pourra être plus facilement interprété et compris par un humain à l'aide d'outils génériques tels que des éditeurs de texte, XML ou hexadécimaux. La mise en œuvre d'une compression limite généralement le niveau de transparence d'un format.

2.4.2. Critères subjectifs

Ces critères reposent sur une évaluation directement liée aux besoins et aux moyens de l'utilisateur réalisant l'évaluation.

Afin d'y faire référence plus aisément, ces critères sont identifiés par une série de lettres : « CPS » (pour « Critère de pérennité subjectif ») et les trois premières lettres du terme le plus significatif.

Identifiant et intitulé	Définition	Justification
CPS-EXP. Expressivité	Le format permet-il d'encoder toute l'information que le producteur souhaite exprimer ?	Lorsqu'on évalue un format candidat pour une création ou une migration, on considérera sa fidélité à l'intention originelle : toute l'information exprimée par le créateur ou contenue dans le fichier source peut-elle être adéquatement transformée dans les structures de données du fichier cible ? Une expressivité excessive au regard du besoin peut se révéler contre-productive à long terme car elle révèle généralement l'inadéquation du format, trop riche pour l'emploi que l'utilisateur souhaite en faire.
CPS-MAI. Maîtrise	L'utilisateur dispose-t-il déjà des compétences et des outils nécessaires à une exploitation précise du format ?	L'enregistrement de fichiers dans un format de données différent de celui d'origine équivaut à une migration. Si celle-ci n'est pas réalisée en connaissance de cause, elle peut avoir un impact négatif sur la fidélité de la copie à son original. Dans le cas de créateurs confiant leurs données à une institution de conservation, il est important que le format soit maîtrisé par les deux parties.

2.5. Analyse des fichiers

L'activité de préservation numérique repose sur la connaissance fine des contenus versés dans le système de préservation et de leurs représentations numériques sous forme de fichiers. Afin d'acquérir cette connaissance, le processus d'analyse préalable au versement dans le système de préservation met successivement en œuvre plusieurs opérations à but de préservation : l'**identification** du format du fichier, la **caractérisation** du fichier, c'est-à-dire sa lecture en vue de produire l'information nécessaire à sa compréhension et sa réutilisation (ce que l'OAIS définit

comme l'Information de représentation et l'Information de pérennisation), et enfin la **validation** du fichier au regard de sa spécification.

Afin de réaliser ces opérations d'analyse, la BnF emploie des outils de validation du fichier au regard de son format et de caractérisation du fichier. On indiquera donc, au début de chaque sous-section concernant un type de contenu dans la partie 3, le format de métadonnées de caractérisation, c'est-à-dire la forme dans laquelle la BnF souhaite conserver le rapport de l'outil d'analyse, ainsi que, dans chaque fiche Format en section 4, un ou plusieurs outils permettant de valider et de caractériser les fichiers de ce format.

Outre la connaissance des caractéristiques de chaque fichier[5], le processus d'analyse a pour objectif immédiat de décider si le fichier est acceptable dans la filière[6]. La BnF définit des politiques de versement, de préservation et d'accès spécifiques à chaque filière. Ainsi, selon sa politique de numérisation de conservation actuelle, elle imposera le format JPEG 2000 BnF 24 bits. La validité du fichier et sa conformité à un profil d'application donné seront, selon la filière, considérées comme des exigences contraignantes ou non. Autrement dit, certaines filières (telle la numérisation de conservation) rejetteront les fichiers invalides tandis que d'autres (telle l'acquisition et le don de documents numériques) pourront les accepter. Dans ce dernier cas, néanmoins, la BnF portera une attention particulière aux fichiers invalides et leur appliquera une politique de préservation spécifique.

Afin que le processus d'analyse puisse être mené à bien, on portera attention à l'absence de **chiffrement**, à la base de toutes les mesures techniques de protection (DRM) de contenu[7]. Il arrive souvent que le chiffrement fasse échouer les processus d'analyse, l'outil ne pouvant accéder au contenu[8].

Le processus d'analyse des fichiers mis en œuvre par la BnF présente des particularités qui conditionnent les politiques que l'institution appliquera à ses données. Il est décrit dans les sections suivantes.

2.5.1. Identification du format[9]

L'identification est une opération de préservation qui initie le processus d'analyse du fichier. Elle vise à donner une première indication, plus ou moins sommaire, sur son format. En pratique, elle procède à une analyse afin de déterminer à quelle entrée d'un registre de formats le fichier correspond. Le degré de précision de l'information résultante dépendra du niveau de détail des entrées du registre de formats. L'identification s'appuie généralement sur la détection de motifs signifiants dans le code du fichier, appelés « nombre magique » ou « signature », mais il arrive qu'elle doive se limiter à l'analyse de l'extension du nom du fichier[10].

À la BnF, cette première étape consiste à obtenir le type MIME[11] du fichier. Par exemple, un fichier communément appelé « JPEG » sera identifié comme « `image/jpeg` », La première partie de ce terme donne une

---

[5] Le rapport d'analyse des fichiers, lorsque la BnF a pu en produire un, est enregistré au format XML dans les métadonnées du Paquet d'informations et est indexé pour pouvoir être par la suite interrogé par les experts de préservation.

[6] Dans ce document, on identifie par le terme « filière » des ensembles documentaires groupés par leur contenu, leur mode d'entrée (dépôt légal, don, acquisition, etc.), leurs conditions juridiques d'utilisation, leur méthode d'obtention (collecte automatique, dépôt à l'unité, etc.) et leurs caractéristiques techniques.

[7] Le chiffrement est applicable à tout fichier, mais il est plus couramment employé sur certains contenus que sur d'autres ; en outre, la manière de l'intégrer dans un format est parfois prévue par la spécification elle-même.

[8] L'opération d'analyse n'est pas la seule à être potentiellement affectée par le chiffrement : toute opération de préservation (copie, migration, etc.) peut également être compromise.

[9] Pour un panorama général sur les problématiques liées à l'identification des formats de fichiers, on se reportera au document *Identification des formats de fichiers*, Programme VITAM, version 2.0, février 2020, disponible à l'adresse http://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-identification-format-v2.0.pdf (consulté le 4 septembre 2020).

[10] C'est ainsi que fonctionnent les systèmes d'exploitation Windows, qui fondent l'identification du format sur la seule extension du nom de fichier.

[11] La liste officielle des types MIME est maintenue par l'organisation IANA (Internet Assigned Numbers Authority, https://www.iana.org/) sous la forme du standard RFC 6838 (https://www.iana.org/go/rfc6838). On notera que nombre de types MIME effectivement utilisés ne sont pas officiellement enregistrés par l'IANA car cet enregistrement suppose la soumission d'une RFC (« Request for comments », standard du web maintenus par l'IETF, *Internet Engineering Task Force*) officielle. Ces types MIME officieux sont théoriquement identifiés par un caractère « x » et un tiret précédant la seconde partie du type (par exemple, le format de livre numérique MOBI n'a pas fait l'objet d'une RFC officielle et est donc identifié par le type MIME « application/x-mobipocket-ebook »).

indication du type de contenu (image, texte, audio, etc.), la seconde du format des données qu'il contient[12]. Cette opération est mise en œuvre par l'outil logiciel « Unix file ».

Dans certaines institutions de conservation, notamment les archives, l'étape d'identification fournit davantage d'information car elle se fonde sur le registre de formats PRONOM. Celui-ci est une base de connaissance collaborative sur les formats maintenue par les archives nationales du Royaume-Uni et enrichie par la communauté internationale de la préservation numérique. Ce registre compte 1899 entrées à la date du 15 février 2021, à un niveau de détail généralement plus fin que le type MIME. Par exemple, l'identification du format d'un fichier dit « JPEG » à l'aide du logiciel DROID pourra identifier qu'il correspond plus précisément à l'entrée JPEG File Interchange Format (JFIF), version 1.02 dans la base PRONOM.

Dans le cas où l'outil ne peut déterminer le type MIME du fichier, il sera considéré comme un simple train d'octets et le type MIME retourné sera « `application/octet-stream` ».

2.5.2. Caractérisation du fichier[13]

La caractérisation d'un fichier, dans son sens adopté par la communauté de la préservation numérique, est une opération de préservation consistant à analyser le contenu d'un fichier afin d'obtenir, par simple lecture ou par calcul, des informations sur le contenu, la forme et/ou l'historique du fichier.

La BnF ne se contente en effet pas du type MIME comme Information de représentation ; elle complète cette première étape d'analyse par la mise en œuvre d'un outil de caractérisation associé au type MIME ; par exemple, l'outil associé au type MIME « `image/jpeg` » est le logiciel JHOVE. La caractérisation permet d'extraire des métadonnées internes au fichier ou de les calculer par l'analyse de la structure du fichier. La caractérisation produit majoritairement des métadonnées considérées comme « techniques », mais, dans la mesure où elle englobe l'opération d'extraction de métadonnées internes, le processus peut également renvoyer des métadonnées descriptives et de provenance.

La BnF définit de préférence au niveau du type de contenu, mais parfois à plus bas niveau, la forme dans laquelle elle enregistre le rapport de caractérisation. Elle privilégie, là encore, des formats de métadonnées de caractérisation standardisés. Par exemple, la caractérisation de tous les fichiers de type de contenu « image » donnera lieu à un rapport conforme au format de métadonnées Metadata for Images in XML (MIX), maintenu par la *Library of Congress*. De ce fait, le format de métadonnées de caractérisation choisi par la BnF sera indiqué dans la section correspondante de chacune des fiches Format.

Dans le cas du fichier JFIF (JPEG), l'outil JHOVE produit un rapport de forme suivante :

```
<mix:mix xmlns:mix="http://www.loc.gov/mix/v10">
  <mix:BasicDigitalObjectInformation>
    <mix:byteOrder>little_endian</mix:byteOrder>
    <mix:Compression>
      <mix:compressionScheme>4</mix:compressionScheme>
    </mix:Compression>
  </mix:BasicDigitalObjectInformation>
  <mix:BasicImageInformation>
    <mix:BasicImageCharacteristics>
      <mix:imageWidth>1348</mix:imageWidth>
      <mix:imageHeight>2473</mix:imageHeight>
      <mix:PhotometricInterpretation>
        <mix:colorSpace>0</mix:colorSpace>
      </mix:PhotometricInterpretation>
    </mix:BasicImageCharacteristics>
  </mix:BasicImageInformation>
  <mix:ImageCaptureMetadata>
    <mix:GeneralCaptureInformation>
      <mix:dateTimeCreated>2009-04-05T22:29:56</mix:dateTimeCreated>
```

---

[12] On omet ici une troisième partie, optionnelle, qui indique des « paramètres ». Ainsi, le type MIME « audio/vnd-wave; codec=1 » identifie un fichier WAVE dont le flux audio n'est pas compressé.

[13] Pour un panorama général sur les problématiques liées à la caractérisation des formats de fichiers, on se reportera au document *Extraction des métadonnées techniques*, Programme VITAM, version 2.0, février 2020, disponible à l'adresse http://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-extraction-MD-v2.0.pdf (consulté le 4 septembre 2020).

```
        <mix:imageProducer>Biblioth%E8que nationale de France - cote :
          N5516370</mix:imageProducer>
      </mix:GeneralCaptureInformation>
      <mix:ScannerCapture>
        <mix:ScannerModel>
          <mix:scannerModelName>d1</mix:scannerModelName>
        </mix:ScannerModel>
        <mix:ScanningSystemSoftware>
          <mix:scanningSoftwareName>Spi.Factory</mix:scanningSoftwareName>
        </mix:ScanningSystemSoftware>
      </mix:ScannerCapture>
      <mix:orientation>1</mix:orientation>
    </mix:ImageCaptureMetadata>
    <mix:ImageAssessmentMetadata>
      <mix:SpatialMetrics>
        <mix:samplingFrequencyUnit>2</mix:samplingFrequencyUnit>
        <mix:xSamplingFrequency>
          <mix:numerator>629145600</mix:numerator>
          <mix:denominator>2097152</mix:denominator>
        </mix:xSamplingFrequency>
        <mix:ySamplingFrequency>
          <mix:numerator>629145600</mix:numerator>
          <mix:denominator>2097152</mix:denominator>
        </mix:ySamplingFrequency>
      </mix:SpatialMetrics>
      <mix:ImageColorEncoding>
        <mix:bitsPerSample>
          <mix:bitsPerSampleValue>1</mix:bitsPerSampleValue>
          <mix:bitsPerSampleUnit>integer</mix:bitsPerSampleUnit>
        </mix:bitsPerSample>
        <mix:samplesPerPixel>1</mix:samplesPerPixel>
      </mix:ImageColorEncoding>
    </mix:ImageAssessmentMetadata>
</mix:mix>
```

On constate la coexistence, dans cette sortie, de métadonnées techniques (l'espace couleur, `//mix:colorSpace`, vaut « 0 », correspondant aux images binarisées ou en niveau de gris) et de métadonnées de provenance (le nom du logiciel de numérisation : « Spi.Factory »).

L'adoption d'un format de métadonnées de caractérisation commun permet la comparaison des caractéristiques de tous les fichiers véhiculant un même type de contenu, quel que soit leur format. Elle ouvre également la voie à la comparaison de ces caractéristiques, avant et après une migration, afin de contrôler que les propriétés considérées comme signifiantes pour l'institution ont été conservées.

2.5.3. Validation du fichier[14]

La validation consiste à évaluer la conformité d'un fichier au regard de la spécification de son format. Ainsi, dans le cas d'un fichier TIFF, validé à la BnF à l'aide de l'outil JHOVE[15], la validation vérifiera dans le rapport d'analyse que le statut a pour valeur « Well-formed and valid », ici dans la sortie XML de JHOVE :

```
<jhove xmlns="http://schema.openpreservation.org/ois/xml/ns/jhove">
  <date>2021-03-01T11:56:20+01:00</date>
  <repInfo uri="/data/spar-rea/ptools01/tmp/rep2100950133715637140tmp">
    <reportingModule release="1.9.1" date="2019-04-17">TIFF-hul</reportingModule>
```

[14] Pour un panorama général sur les problématiques liées à la validation des formats de fichiers, on se reportera au document *Validation du format des fichiers*, Programme VITAM, version 2.0, février 2020, disponible à l'adresse http://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/20200131_NP_Vitam_preservation-validation-format-v2.0.pdf (consulté le 4 septembre 2020).

[15] Open Preservation Foundation, JHOVE, http://jhove.openpreservation.org/ (consulté le 04 mars 2021).

<lastModified>2021-03-01T11:56:20+01:00</lastModified>
<size>33313830</size>
<format>TIFF</format>
<version>5.0</version>
<status>Well-Formed and valid</status>

La validation peut être réalisée à l'aide du même outil que celui employé pour la caractérisation ou faire appel à un outil distinct, notamment dans le cas où l'outil de caractérisation ne fournit pas de retour, ou pas de retour jugé suffisamment fiable, sur la validité du fichier au regard des spécifications de son format[16].

2.5.4. Contrôle de conformité à un profil d'application

Dans le cas de formats bien maîtrisés par la BnF, en particulier lorsqu'il s'agit de fichiers produits selon ses référentiels, la politique de la BnF va au-delà de la simple conformité aux spécifications du format et définit un profil d'application du format que le fichier doit suivre.

Afin de vérifier que ce profil d'application est bien respecté, un contrôle compare les métadonnées de caractérisation et les propriétés attendues par le profil d'application. Ainsi, l'image TIFF dont le rapport de caractérisation est présenté plus haut sera déclarée conforme au format « TIFF BnF noir et blanc » notamment parce que les éléments suivants présentent la valeur attendue :

- `//mix:samplesPerPixel` (échantillons par pixel) a pour valeur « 1 », correspondant à des images en niveau de gris ou binarisées ;
- `//mix:bitsPerSample` (bits par échantillon) a pour valeur « 1 », correspondant à des images binarisées, où la valeur de chaque pixel est codée sur un seul bit ;
- `//mix:compressionScheme` (mode de compression) a pour valeur « 4 », correspondant au mode de compression attendu pour ces images : CCITT groupe 4.

2.6. Stratégies de préservation

Maintenir l'information numérique accessible, compréhensible et utilisable sur le long terme suppose une activité régulière et la mise en place d'actions de préservation pour prévenir des risques budgétaires ou organisationnels autant que techniques. Le choix de formats maîtrisés et adaptés à la préservation ainsi que la limitation du nombre de formats représentés au sein de l'institution de conservation lui permet néanmoins de maintenir le coût de cette activité à un niveau acceptable. Traditionnellement, on identifie deux grandes catégories de stratégies de préservation visant à limiter les risques liés au format des données :

- l'adaptation du contenu à l'environnement : il s'agit des stratégies de migration ;
- l'adaptation de l'environnement au contenu, qui comprend notamment les stratégies d'émulation.

La stratégie de préservation ne dépend pas uniquement du format mais également du type de contenu des données, de leur usage et de l'intention de préservation de l'institution de conservation. Ce présent document évoquera les stratégies de préservation adoptées par la BnF selon les deux modalités suivantes.

- Lorsqu'une migration n'est pas envisagée à moyen terme, on indiquera les éléments significatifs de l'environnement de restitution choisi par la BnF dans la section « Outils » de chaque fiche Format.
- Lorsqu'une migration est envisagée à court ou moyen terme, on indiquera la stratégie adoptée par la BnF (format source, format cible et procédure de migration) dans la section 3. et non dans les fiches Formats en section 4. En effet, de telles stratégies seront appliquées non à tous les fichiers d'un format donné mais prendront en compte outre le format lui-même, et entre autres critères détaillés ci-dessous, le type de contenu et le contexte d'utilisation.

Toute migration présente un risque de perte du contenu informationnel des données[17]. De ce fait, l'étude d'une méthode de migration adaptée vise à s'assurer qu'elle n'a engendré aucune perte ou à limiter cette perte aux propriétés que l'institution de conservation ne juge pas signifiantes. La BnF peut néanmoins être contrainte d'opérer une migration dont elle sait par avance que les données résultantes ne contiendront pas l'intégralité de l'information véhiculée par les données originales, en particulier dans le cas de formats propriétaires. La BnF envisage alors la conservation des données originales en complément de leur version migrée dans un format préféré. Un cas de ce type est décrit dans la section 2.6.1.2.5 ci-dessous.

---

[16] Ainsi, l'outil JHOVE est un outil de caractérisation et de validation pour de nombreux formats d'images courants. A l'inverse, l'outil MediaInfo ne propose que la caractérisation des fichiers audiovisuels. La BnF juge donc nécessaire de mettre en œuvre un second outil de validation pour certains fichiers audio ou vidéo.

[17] On considère comme « migration » toute transformation des données, y compris un simple enregistrement dans un format différent de celui dans lequel les données ont été originellement créées.

La BnF portera une attention particulière à la méthode de migration, et en particulier à l'outil, son paramétrage et la compétence des agents qui l'opèrent. Dans le cas de données dans un format propriétaire, on privilégiera généralement le logiciel qui les a produites pour effectuer la migration. On part en effet du principe que le logiciel propriétaire de création bénéficie de la meilleure connaissance du format source et sera donc capable de conserver la majeure partie du contenu informationnel.

2.6.1. Opérations préalables au versement

Avant de prendre la responsabilité de la conservation de données qu'elle n'a pas fait produire, la BnF tente d'engager une discussion avec le Producteur afin d'identifier un ou plusieurs formats satisfaisant les deux parties au regard des critères de pérennité édictés plus haut. En particulier, elle s'assure que le ou les formats choisis sont maîtrisés par le Producteur et par elle-même.

Si une telle négociation avec le Producteur n'a pu avoir lieu et que la BnF a obtenu des données dans un format qui ne fait pas partie de ses formats préférés ou acceptés, elle envisagera une migration. Celle-ci peut affecter le format du conteneur ou celui du flux des données. Il peut s'agir d'opérations ciblées : pour une image, une migration du modèle couleur CMJN destiné à l'impression vers RVB, destiné à l'affichage, par exemple. Comme pour toute autre opération de préservation, les migrations effectuées entre le transfert des données et leur versement seront tracées et enregistrées dans les métadonnées de provenance du Paquet d'informations.

2.6.1.1. **Critères de migration au versement**

Pour chaque fonds, la BnF décide du format cible, de la méthode de migration et de la conservation des fichiers d'origine, en prenant en compte les critères suivants, selon une pondération spécifique à chaque fonds. Les cases notées « N/A » (non applicable) signalent que la BnF considère le critère comme n'influant pas sur la réponse qu'elle apporte à la question, ou qu'elle choisit de ne pas le prendre en compte. **Cette liste n'est pas classée par ordre d'importance.**

Afin d'y faire référence plus aisément, ces critères sont identifiés par une série de lettres : « CM » (pour « Critère de migration ») et les trois premières lettres du terme le plus significatif.

Identifiant et intitulé	Réalisation ou non d'une opération de migration ?	Quel format cible ?	Conservation ou non de l'original ?
CM-CAT. Catégorie du format	Si le format du fichier livré par le Producteur n'est pas dans un format préféré pour son type de contenu et son usage, on envisagera une opération de migration.	On choisira pour le format cible un des formats préférés identifiés pour le type de contenu véhiculé par le fichier livré.	N/A
CM-CHO. Choix artistique ou technique délibéré	On prendra en compte le caractère délibéré ou non de la livraison des contenus dans un ou plusieurs formats donnés pour déterminer la pertinence d'une opération de migration, notamment dans le cas d'un fonds de contenus de même nature mais dans des formats hétérogènes.	Les choix artistiques ou techniques peuvent également influencer le choix du format cible, notamment si le Producteur a spécifié une méthode de migration vers un format donné ou a réalisé lui-même une représentation numérique alternative dans un format préféré.	On tendra à conserver le fichier original si le choix de son format est délibéré.

CM-JUR. Contraintes juridiques	On doit s'assurer qu'on est en droit d'opérer des migrations (notamment dans le cadre des dons et acquisitions), afin de ne pas contrevenir au droit de la propriété intellectuelle[18].	N/A	Dans certains cas, les contraintes juridiques peuvent obliger à conserver la forme originelle du contenu[19].
CM-AUT. Autonomie / dépendances du fichier	Si le fichier livré n'est pas autonome mais fait partie d'un ensemble requis pour constituer un contenu, on peut envisager une opération globale, par exemple en embarquant les données dans un fichier conteneur.	Le format cible d'un contenu cohérent sera envisagé globalement et non à l'échelle d'un de ses fichiers composants.	N/A
CM-COH. Cohérence au sein du Paquet d'informations ou du fonds	Afin de simplifier la préservation et la réutilisation, il est préférable que les fichiers destinés au même Paquet d'informations et ayant le même usage soient conservés dans un format homogène.	Entre deux formats cibles candidats, on préférera celui qui est représenté au sein du fonds.	N/A
CM-SOU. Soutenabilité	La volumétrie et la complexité des données sources influent sur le choix d'opérer une migration ou non.	N/A[20]	Conserver le fichier original, dans le cas d'un fichier très volumineux ou complexe et non maîtrisé, peut représenter une charge importante pour l'institution de conservation.

[18] L'exception « Bibliothèques » du code de la propriété intellectuelle (Art. L.122-5, alinéa 8, disponible sur https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000037388886/) permet à la BnF de réaliser des opérations de copie et de migration à des fins de conservation, dès lors que les documents ont été divulgués, sans autorisation préalable des ayants-droit. Cette exception concerne les documents soumis au dépôt légal et ceux entrant par acquisition ou par don s'ils ont été divulgués. En revanche, tout document non divulgué entrant dans les collections par don ou par acquisition doit faire l'objet d'une cession de droits de la part du ou des titulaires des droits de propriété intellectuelle afin d'autoriser la BnF à réaliser des opérations de copie et de migration à des fins de préservation. Dans la pratique, les dons et acquisitions font l'objet d'une cession de droits même lorsqu'ils portent sur des documents divulgués, afin de clarifier les opérations autorisées.

[19] Selon sa rareté, le document peut être considéré comme étant un document numérique original (si le donateur fournit la seule copie existante ou du moins, la seule dont il dispose). Celui-ci relevant par ailleurs du code général de la propriété des personnes publiques, il ne peut de fait être détruit. La BnF a alors pour mission de le conserver dans son format d'origine. En outre, certains contrats conclus avec les donateurs pourraient conditionner le don du fonds à la conservation et/ou à la diffusion du format d'origine.

[20] Le format cible ayant vocation à être un format déjà connu par la BnF, le critère de soutenabilité correspond à son évaluation au regard des critères de pérennité, entre autres la compacité (CPO-COM) et la simplicité (CPO-SIM).

CM-EQU. Existence d'un équivalent dans un autre format	L'existence d'une autre représentation numérique dans un format préféré influe sur le choix de réaliser une migration ou non[21].	N/A[22]	N/A
CM-SIG. Conservation de propriétés ou de fonctionnalités signifiantes	N/A	La définition par l'institution de conservation de son intention de préservation, c'est-à-dire de l'ensemble des propriétés informationnelles et des modalités d'utilisation d'un objet numérique qu'elle souhaite préserver sur le long terme pour sa Communauté d'utilisateurs, conditionnera le format cible, la méthode de migration et la conservation ou non de l'original[23].	
CM-REV. Réversibilité de l'opération[24]	N/A	Entre deux formats cibles candidats, on préférera celui vers lequel la migration est réversible.	Si l'opération est réversible (ajout de métadonnées, compression sans perte, etc.), la conservation de l'original pourra être considérée comme non nécessaire.

2.6.1.2. **Exemples d'application des critères de migration**

Ces exemples sont inspirés de cas d'usage réels rencontrés par la BnF mais ont pour but essentiel d'illustrer le tableau des critères de migration ; ils ont été simplifiés et ne reflètent donc pas nécessairement toute la complexité de la situation. Les choix mentionnés reflètent en outre des **pratiques** (voir section 2.3 ci-dessus) qui n'ont donc pas forcément vocation à être systématisées.

2.6.1.2.1. Migration de photographies numériques au format PDF en images JPEG

La BnF obtient par don un lot de quinze photographies numériques dans des formats différents (PDF et JFIF JPEG). Le PDF n'étant pas un format préféré pour les photographies numériques (CM-CAT), elle envisage une migration. Elle s'assure d'abord que le format PDF ne résulte pas d'un choix délibéré du Producteur (CM-CHO). Elle constate par ailleurs que le fichier PDF a pour source une image JPEG de la même série que les autres photographies au format JFIF, et que l'image contenue dans le PDF est compressée selon l'algorithme DCTDecode, également utilisé pour la compression JPEG. Dans la mesure où elle souhaite constituer un unique Paquet d'informations à partir de ce lot pour des raisons documentaires, elle préfère homogénéiser le format des fichiers destinés à ce Paquet (CM-COH) et migrer les fichiers PDF vers le format JFIF (JPEG).

2.6.1.2.2. Migration vers JPEG et conservation du PSD

La BnF obtient par don un lot de trois fichiers PSD (format d'enregistrement du logiciel Photoshop). Ce format ne fait pas partie des formats préférés de la BnF, notamment parce qu'il est propriétaire. Elle envisage une migration vers JFIF (JPEG). Elle constate néanmoins que le fichier PSD comprend deux calques, dont l'un est la version de la photographie avant recadrage. Considérant que la présence de cette seconde image est une propriété signifiante documentant la provenance de la photographie recadrée, et que la migration vers le format JFIF (JPEG), envisagé comme format cible, fera disparaître cette propriété (CM-SIG), elle décide de réaliser cette

---

[21] La solution la plus simple est évidemment de ne prendre en charge que la représentation numérique dans un format préféré par la BnF.

[22] Néanmoins, l'existence dans le fonds d'une version convertie depuis le même format source que le fichier étudié peut amener à préférer le format cible dans la mesure où il a déjà été choisi par le Producteur comme format de conversion.

[23] Dans le cas où la migration est réversible (CM-REV), la perte de fonctionnalités et propriétés signifiantes peut ne pas conduire à la conservation de l'original puisqu'il peut être restauré dans sa version strictement équivalente par une migration inverse.

[24] Attention à certaines conversions qui peuvent être réversibles pour le flux principal de données, mais non réversibles pour les flux secondaires (métadonnées, vignettes, etc.).

migration tout en conservant le fichier PSD original. Elle emploie pour cela le logiciel Photoshop en sollicitant les compétences de son experte en imagerie numérique.

2.6.1.2.3. Migration d'AIFF vers MP4

Au sein d'un ensemble conséquent de fichiers, la BnF reçoit une série de quatre fichiers AIFF (équivalent du format WAVE développé pour les systèmes Apple pour l'encodage du son non compressé). Après étude, il s'avère que ces quatre fichiers sont les pistes sonores synchrones d'un contenu vidéo (CM-AUT) dont la partie image est enregistrée dans un fichier MXF ; le contenu des fichiers AIFF et du fichier MXF est donc migré dans un unique fichier conteneur au format MP4.

2.6.1.2.4. Migration de WAVE vers FLAC

La BnF étudie l'opportunité d'une migration de ses fichiers sonores au format WAVE non compressé vers le format FLAC pour des questions de soutenabilité, le format FLAC étant deux à trois fois plus compact que le format WAVE non compressé (CM-SOU). Elle prend en compte, parmi d'autres critères, la réversibilité intégrale (flux de données audio et métadonnées) de la migration du format WAVE non compressé vers FLAC (CM-REV), ce qui lui permet d'envisager une migration sans conservation de l'original.

2.6.1.2.5. Migration de FCP vers un format image

Le format FCP est un format utilisé par le logiciel propriétaire Final Cut Pro et développé par Apple pour réaliser le montage de films. Un don d'archives par un cinéaste a mis la BnF en possession des montages successifs d'un même film, sous la forme de « bibliothèques » ou « projets » FCP. Ces fichiers, déjà très volumineux, n'embarquent pas tous les médias utilisés mais y font référence par le biais d'un chemin relatif. La BnF envisage :

- d'opérer une migration partielle n'épuisant pas le contenu informationnel des projets FCP afin de mettre à disposition des chercheurs un aperçu du contenu de ces projets, notamment la capture écran de la *timeline* du projet et l'export du montage final sous forme d'un fichier MP4 ;
- de conserver le projet FCP, dans la mesure où la migration mentionnée ne permet pas de conserver les fonctionnalités et propriétés signifiantes[25] (CM-SIG), malgré le poids conséquent de ces fichiers (CM-SOU) ;
- de conserver les médias sources référencés par les projets comme des contenus unitaires, et de leur appliquer une migration spécifique, mais également d'en conserver un état original empaqueté avec le projet lui-même (CM-AUT).

2.6.1.2.6. Traitement de contenus disponibles au format DOCX et PDF

Dans de nombreux fonds donnés à la BnF, on trouve des contenus sous deux formats : l'un au format de production DOCX, format utilisé par le traitement de texte Microsoft Word à partir de sa version 2007, l'autre dans une version de publication au format PDF. La BnF choisit dans ce cas de conserver le contenu sous les deux formats, considérant que la version DOCX est le fichier original, et la version PDF son équivalent dérivé dans un format préservable à long terme.

2.6.2. Opérations postérieures au versement

Sur le long terme, il arrive également que la BnF doive effectuer sur les objets numériques dont elle a la charge des migrations. Si une telle opération est prévue à moyen terme par la BnF, elle est indiquée pour un format, un usage et un type de contenus donnés, avec les outils pressentis pour la réaliser. Cette section du document a vocation à être enrichie à mesure que la BnF développe l'expérience de telles opérations.

Comme les migrations préalables au versement, celles postérieures au versement sont tracées et enregistrées dans les métadonnées de provenance du Paquet d'informations.

---

[25] Si la BnF ne considère pas la possibilité de reprendre et modifier le montage comme une « fonctionnalité signifiante » à maintenir, elle a conscience que les projets FCP conservent de nombreuses informations sur le processus de création qui seraient perdues dans une migration vers un autre format. La BnF a également conscience que les limitations liées à ce format propriétaire l'empêcheront sans doute à terme de restituer correctement ces projets, et que les dérivés, si insuffisants soient-ils, deviendront alors la seule représentation du montage.

3. Liste des formats préférés et acceptés

Ce document propose de classer les formats par type de contenu et non par type de format, un format pouvant être adapté à un type de contenu et non à un autre. Ainsi le PDF est-il idéal pour l'encodage du texte dont on souhaite préserver la mise en page, mais non à la numérisation patrimoniale (même s'il a pu être utilisé dans ce cadre). La première version de ce document détaille les préférences BnF pour cinq types de contenus (image fixe, texte, son, image animée, contenu web), mais la liste est amenée à se développer à mesure que la BnF mène des études sur les formats. Les catégories prévues à l'avenir sont : jeux de données structurées, données cartographiques, objet (3D), support dématérialisé (image disque), notation musicale, correspondance. À l'intérieur de chaque section, on a défini des sous-sections correspondant à un usage ou à un mode de production des données conditionnant les propriétés signifiantes à préserver.

Pour chaque type de contenu, la BnF distinguera quatre catégories de formats, dont la définition découle de ses pratiques et de l'état de ses connaissances. Du niveau de maîtrise le plus élevé au plus bas, donc par niveau de risque croissant et par niveau de service décroissant, on trouvera donc quatre catégories de formats.

1. **Formats préférés** : formats ou profils d'application d'un format qui présentent du point de vue de la BnF les meilleures garanties de préservation pour un type de contenu donné. La BnF indique ainsi aux producteurs de données que la création de contenus dans ces formats limitera l'investissement nécessaire au maintien de leur accessibilité, utilisabilité et compréhensibilité à long terme pour les usagers de la BnF.
2. **Formats acceptés** : formats présentant de moindres garanties de pérennité, soit qu'ils soient moins favorables à la qualité du contenu, soit qu'ils demandent une surveillance accrue. La BnF ne recommande pas leur utilisation pour la création de données, mais celles héritées dans de tels formats seront considérées comme acceptables. Elles pourront néanmoins faire l'objet d'opérations de préservation lors de leur prise en charge par la BnF, allant jusqu'à une migration de format.
3. **Formats à l'étude** : formats que les experts de la BnF étudient ou ont prévu d'étudier à court ou moyen terme en vue d'une éventuelle adoption.
4. **Formats reconnus par la communauté** : formats non étudiés par la BnF, mais qui sont utilisés par la communauté de la préservation numérique, et qui à ce titre présentent un certain nombre de garanties de préservation. On indiquera ainsi les formats recommandés par d'autres institutions patrimoniales reconnues (une liste de documents de politique similaires est disponible en section 5.2 ci-dessous).

À l'intérieur des sections du document, un éventuel classement par préférence est indiqué par une numérotation 1), 2), 3) ; si à l'inverse la liste n'est pas classée par préférence, les formats sont introduits par un tiret (–).

Les données dans un format ne figurant pas dans ce document ne seront pas nécessairement refusées par la BnF et leur absence ne signifie pas qu'ils sont jugés impropres à la préservation de l'information**.** Dans le cas où un producteur disposerait de données dans d'autres formats que ceux mentionnés dans le présent document, et souhaiterait les confier à la BnF, **il est invité à entrer en contact avec la BnF avant de se lancer dans une migration non maîtrisée** afin d'engager un dialogue sur le format préférable pour un transfert, sous réserve de l'acceptation du don par la BnF. En effet, une migration opérée avec un outil inconnu ou un enregistrement sous un format différent que celui dans lequel le fichier a été originellement créé présente un risque élevé de perte d'information.

3.1. Images fixes

On entend, par « image fixe », une représentation visuelle en deux dimensions, opaque ou translucide et sans mouvement, par opposition à l'image animée. Sa représentation numérique prend la forme d'un ou plusieurs fichiers en couleur ou en niveau de gris. La couleur est exprimée le plus couramment selon les modes RVB[26] ou CMJN[27]. Le premier correspond principalement aux domaines de la production d'images et de la visualisation sur écran. Le second est communément dédié à l'impression en quadrichromie. L'information de couleur peut être encodée avec plus ou moins de précision : sur 8 ou 16 bits par couche pour les cas les plus fréquents. Les images sans information de couleur peuvent être décrites en noir et blanc (binaire) ou en niveaux de gris (on parle également de « nuances de gris »).

La BnF préfère les images en couleur décrites en RVB avec une profondeur de codage de 8 bits par couche (soit 24 bits) et les images sans information de couleur décrites en niveaux de gris avec une profondeur de 8 bits par couche. Les images décrites différemment sont converties lors de leur versement dans le système de préservation.

La BnF impose que les fichiers images soient décrits dans un espace colorimétrique standard. Elle utilise l'espace Adobe 98 mais d'autres peuvent être choisis selon l'usage (profil plus limité pour la diffusion, par ex. sRGB, ou plus large pour la conservation, par ex. ProPhoto).

La BnF identifie quatre sous-ensembles d'images fixes pour lesquelles elle définit une politique distincte, décrite ci-dessous : images issues de la numérisation, photographies nativement numériques, documents graphiques destinés à l'impression et autres créations graphiques.

3.1.1. Métadonnées techniques de caractérisation produites par la BnF

Pour toutes les images fixes, le format des métadonnées de caractérisation enregistré par la BnF est le format *Metadata for Images in XML (MIX)*[28]. Les métadonnées techniques suivantes sont encodées selon ce formalisme :

- format du fichier conteneur ;
- format du flux image / type de compression ;
- profondeur couleur par échantillon ;
- nombre d'échantillons ;
- taille (hauteur, largeur) ;
- résolution de capture (seulement pour les images issues de la numérisation),
- espace colorimétrique ;
- profil ICC ou autre mode de spécification couleur.

3.1.2. Images issues de la numérisation

Cette section concerne les images numériques issues de numériseurs linéaires (scanners) ou matriciels (appareils photo) visant à reproduire en deux dimensions des documents ou des œuvres en deux dimensions. Sont par exemples concernés les documents opaques tels que journaux, livres, cartes, etc., et transparents tels que supports argentiques photographiques ou micrographiques.

Formats préférés

- JPEG 2000 conforme au référentiel BnF
- TIFF v. 6 monopage sans compression
  - Trajectoire de préservation envisagée : la BnF étudie la migration rétrospective de ses images TIFF couleur vers le JPEG 2000 à l'aide de l'outil *kakadu*.

Formats acceptés

- TIFF noir et blanc compression CCITT groupe 4
- JPEG 2000 compression sans perte
- JFIF (JPEG)

Formats reconnus par la communauté

- PDF multipage embarquant des images JPEG ou JPEG 2000

---

[27] Voir l'article « Quadrichromie » de l'encyclopédie Wikipédia, accessible sur <https://fr.wikipedia.org/wiki/Quadrichromie> (consulté le 24 mars 2021).

[28] *Library of Congress*, *NISO Metadata for Images in XML Schema*, site web officiel, dernière modification le 8 septembre 2020, disponible sur http://www.loc.gov/standards/mix/ (consulté le 10 mars 2021).

- PNG
- DJVU
- TIFF 48 bits (16 bits par couche)

3.1.3. Photographies nativement numériques

On entend par « photographies numériques » des prises de vue visant à représenter des scènes avec profondeur de champ, par opposition à la numérisation de documents patrimoniaux. Cette section concerne par conséquent les images issues de dispositifs matriciels (appareils photo), visant à enregistrer en deux dimensions des scènes en trois dimensions. Par exemple : photographies de paysages, de personnes, d'objets, etc.

Formats préférés

- TIFF v. 6 monopage sans compression
- EXIF JPEG
- JFIF (JPEG)

Formats acceptés

- TIFF v. 6 compressé (sauf compression JPEG)

Formats reconnus par la communauté

- JPEG XT (dont JPEG HDR) (ISO/IEC 18477)

3.1.4. Documents graphiques destinés à l'impression

Il s'agit ici de documents nativement numériques destinés à un système d'impression, quand on souhaite conserver le fichier ayant servi à l'impression en remplacement ou en supplément d'un tirage sur support physique.

Formats préférés

- PDF/X
- TIFF v. 6 monopage sans compression
- JFIF (JPEG)

3.1.5. Autres créations graphiques nativement numériques

Ce domaine est actuellement à l'étude à la BnF.

3.2. Contenus textuels

Pour l'ensemble des contenus textuels, la BnF préfère l'usage systématique de l'encodage UTF-8, et non ISO Latin-1 par exemple. Aujourd'hui très majoritaire sur Internet, UTF-8 prend en charge l'intégralité du répertoire Unicode. Conséquemment, il permet de gérer des textes dans la quasi-totalité des systèmes d'écritures et alphabets du monde.

3.2.1. Texte brut ou semi-structuré

Cette section concerne le texte brut, c'est-à-dire la simple représentation de caractères sans information de structure, d'apparence ou de présentation, et le texte semi-structuré grâce à un langage de balisage. Les formats principalement utilisés, à la BnF, pour la numérisation de documents textuels. La plupart des formats ainsi utilisés présentent des dépendances envers les fichiers images (voire envers certains autres fichiers texte).

3.2.1.1. Métadonnées techniques de caractérisation

Pour le texte brut ou semi-structuré, le format des métadonnées de caractérisation enregistré par la BnF est le format textMD[29]. Les métadonnées techniques suivantes sont encodées selon ce formalisme :

- boutisme ;
- taille d'octet ;
- taille d'un caractère en octets ;
- codage des caractères ;
- codage des retours chariot ;
- syntaxe de balisage (SGML, XML, GML, etc.) ;

[29] *Library of Congress*, *Technical Metadata for Text*, site web officiel, dernière modification le 10 mars 2021, disponible sur http://www.loc.gov/standards/textMD/ (consulté le 10 mars 2021).

– langage de balisage (par exemple, nom ou URI du schéma XML, dans le cas d'un fichier XML)

3.2.1.2. Structuration physique (segmentation et reconnaissance des caractères)

Format préféré

– XML-ALTO

Formats reconnus par la communauté

– hOCR
– Plain text-TXT, ASCII ou UTF-8
– PAGE XML

3.2.1.3. Structuration intellectuelle (reconnaissance de la mise en page, table des matières…)

Format préféré

– XML-METS

Format accepté

– XML-tdmNum

Formats reconnus par la communauté

– XML-TEI avec fichier ODD correspondant
– TEI lite

3.2.2. Documents

On réunit dans cette catégorie les contenus textuels nativement numériques dont la finalité est la publication, donc pour lesquels la mise en page et l'apparence formelle sont des propriétés signifiantes. Les formats présentés ci-dessous sont donc généralement des conteneurs capables d'embarquer d'autres types de contenu[30] (images, voire contenu audiovisuel ou interactif).

La BnF préfère acquérir le contenu textuel sous sa forme définitive et publiée. Dans le cas où on lui soumettrait un contenu textuel à une étape antérieure à la publication, par exemple afin de documenter la genèse d'une œuvre écrite, elle conservera, en plus de(s) fichier(s) dans un des formats de production ci-dessous, une version convertie dans un format de publication afin de conserver sa mise en page.

3.2.2.1. Métadonnées de caractérisation produites par la BnF

Pour les documents, le format des métadonnées de caractérisation enregistré par la BnF est le format XMP[31]. Les métadonnées suivantes sont encodées selon ce formalisme :

– titre ;
– auteur ;
– mots-clés ;
– nombre de pages ;
– outil de création.

3.2.2.2. Étape de publication

La BnF négocie avec les Producteurs afin que les documents collectés ou reçus par elle au titre du dépôt légal ne comportent aucune mesure technique de protection (DRM) susceptible d'entraver les opérations de préservation qu'elle mettra en œuvre.

Formats préférés

1. EPUB 3
2. PDF/A-1 ou PDF/A-2, toutes variantes

---

[30] On notera de ce fait que les types MIME des formats de texte brut ou semi-structuré présentent un type « text », (ex. : « `text/xml` ») alors que ceux des formats de documents présentent un type « application » (ex. : « `application/pdf` »).

[31] *Library of Congress*, *Technical Metadata for Text*, site web officiel, dernière modification le 10 mars 2021, disponible sur <http://www.loc.gov/standards/textMD/> (consulté le 10 mars 2021).

3. PDF

Formats acceptés

- EPUB 2
- DAISY Digital Talking Book avec texte integral

Format à l'étude

- (X)HTML + CSS associé
- DAISY Digital Talking Book avec texte audio

3.2.2.3. Étape de production

Formats acceptés

- OpenDocument Text (ODT)
- Office Open XML - Document (DOCX)
  - Trajectoire de préservation envisagée : les fichiers Office Open XML - Document sont migrés au versement vers le format PDF et conservés comme originaux dans le Paquet d'informations..

Format reconnu par la communauté

- LaTeX

3.3. Contenus sonores

On évitera les procédés de sur-échantillonnage (*upsampling*) et d'interpolation, ainsi que tout procédé de ré-encodage d'un flux compressé dans un format non compressé.

La BnF privilégie l'encodage PCM (*pulse code modulation*, modulation par impulsions et codage), qui constitue la base des flux audio contenus dans les formats préférés par la BnF (WAVE non compressé, FLAC compressé sans perte, MPEG-1/2 layer 3 (MP3) compressé avec pertes). L'encodage PDM (*pulse density modulation*), sur lequel est basé le format DSD, est actuellement à l'étude.

Actuellement (avril 2021), la BnF numérise les contenus sonores en 24 bits (quantification) et 96 kHz (fréquence d'échantillonnage), paramètres retenus pour les fichiers masters de préservation à long terme. D'autres valeurs sont également acceptées pour des contenus sonores produits par d'autres entités :

- 16 bits / 44,1 kHz (qualité dite « CD ») ;
- 16 bits / 48 kHz ;
- 24 bits / 48 kHz ;
- 24 bits / 88,2 kHz ;
- 24 bits / 192 kHz.

3.3.1. Métadonnées techniques de caractérisation produites par la BnF

Pour les contenus sonores, le format des métadonnées de caractérisation enregistré par la BnF est le format MPEG-7[32]. Les métadonnées suivantes sont encodées selon ce formalisme :

- Format du conteneur
- Format du flux audio
- Profil du format (le cas échéant)
- Durée
- Type de débit (constant / variable)
- Débit global (moyen)
- Nombre de canaux
- Profondeur d'échantillonnage
- Fréquence d'échantillonnage
- *Timecode* de départ.

[32] *MPEG-7*, sur le site web du *Moving Picture Experts Group*, disponible sur <https://mpeg.chiariglione.org/standards/mpeg-7> (consulté le 31 mars 2021).

3.3.2. Formats

3.3.2.1. **Son numérisé**

Cette section concerne les contenus issus de la numérisation de phonogrammes analogiques ou de la dématérialisation de phonogrammes numériques.

Formats préférés

- BWF / MBWF
- FLAC
- WAVE / WAVE RF64
  - Trajectoire de préservation envisagée : la BnF étudie actuellement l'opportunité de convertir ses fichiers WAVE en FLAC grâce à l'outil flac.

Format accepté

- MPEG-1/2 layer 3 (MP3)

3.3.2.2. **Son nativement numérique**

Cette section concerne les contenus produits nativement sous forme numérique.

Formats préférés

- BWF / MBWF
- FLAC
- WAVE / WAVE RF64

Format accepté

- MPEG-1/2 layer 3 (MP3)

Format à l'étude

- DSD
- MP4 AAC

Formats reconnus par la communauté

- AIFF

3.4. Images animées

Cette section concerne les contenus constitués d'images animées, avec ou sans contenu sonore synchronisé.

On évitera les procédés de sur-échantillonnage (*upsampling*) et d'interpolation, ainsi que tout procédé de ré-encodage d'un flux compressé dans un format non compressé.

La BnF préfère obtenir les images animées dans un format orienté vers la diffusion, qui met majoritairement en œuvre des méthodes de compression et de prédiction inter-trame, plutôt que dans un format orienté vers la production, qui privilégie la compression intra-trame afin de conserver la plupart des images et ainsi la capacité de travailler à l'image près.

3.4.1. Métadonnées techniques de caractérisation produites par la BnF

Pour les images animées, le format des métadonnées de caractérisation enregistré par la BnF est le format MPEG-7[33]. Les métadonnées suivantes sont encodées selon ce formalisme :

- Format du conteneur
- Durée

Flux vidéo

- Format
- Profil du format
- Débit global (moyen)

[33] *MPEG-7*, sur le site web du *Moving Picture Experts Group*, disponible sur <https://mpeg.chiariglione.org/standards/mpeg-7> (consulté le 31 mars 2021).

- Type de débit (constant / variable)
- Gap of picture (GOP)
- Largeur et hauteur des *frames*
- Ratio image
- Espace de couleur
- Type de balayage
- Sous échantillonnage de la chrominance
- Cadence
- *Timecode* de départ

Flux audio

- Format
- Débit global (moyen)
- Type de débit (constant / variable)
- Nombre de canaux
- Profondeur d'échantillonnage
- Fréquence d'échantillonnage

3.4.2. Formats

Formats préférés

1. MP4 H.264
2. MPEG-2

Format accepté

- MOV ProRes
  - Trajectoire de préservation : les fichiers MOV ProRes sont conservés tels quels mais sont complétés au versement par une version au format MP4 H.264 à l'aide de l'outil ffmpeg, si cette dernière n'est pas fournie par le Producteur.

Format reconnu par la communauté

- Matroska FFV1

3.5. Contenus Web

Il s'agit ici de fichiers issus de la collecte de contenu web par robots, et empaquetés de sorte à être rendus exploitables par un moteur d'indexation et par un outil de visualisation dans leur contexte.

3.5.1. Métadonnées techniques de caractérisation produites par la BnF

Afin de décrire les caractéristiques du conteneur WARC et d'agréger des informations sur les fichiers contenus (poids maximal et minimal, poids total, etc.) par format, la BnF a créé un format de métadonnées adapté à la description des fichiers conteneurs, containerMD. Pour le moment limité à la description des fichiers ARC et WARC, il pourra être étendu à d'autres formats.

3.5.2. Formats

Format préféré

- WARC 1.0

Formats acceptés

- ARC 1.0, 1.1 et 2.0

4. Fiches Format

Les fiches Format suivantes ont pour objectif de décrire chacun des formats listés comme « préférés » ou « acceptés » dans la liste ci-dessus. Certains d'entre eux ne disposent pas encore d'une fiche validée dans cette version du document. Les fiches sont également consultables sur Github (https://github.com/hackathonBnF/FichesFormat/wiki), dans une version mise à jour au fur et à mesure des études et de la veille des experts formats de la BnF.

4.1. Structure des fiches

Description

La section présente succinctement l'histoire et l'usage du format, ainsi que son statut : ouvert, c'est-à-dire **documenté** (CPO-DOC) et **libre d'utilisation** (CPO-LIB), ou propriétaire. On y détaille également d'éventuels intitulés alternatifs sous lesquels est connu le format.

Sociabilité

La section évalue le niveau d'adoption du format (CPO-SOC) et mentionne sa (ses) **communauté(s) d'utilisateurs** et son niveau de support logiciel et matériel (CPO-OUT).

Relations à d'autres formats

On précise dans cette section :

- si le format est une extension, une restriction ou s'inspire simplement d'un autre format ;
- dans le cas d'un codec, s'il a une relation particulière avec un format conteneur donné ;
- si le format a des ancêtres et/ou des successeurs officiels.

Versions

La section détaille les différentes versions successives ayant un impact significatif sur la durabilité des données. On évoquera ici le niveau de **stabilité** (CPO-STA) du format. Si le format définit des profils spécifiques (par exemple, "High profile" pour H.264), en particulier s'ils permettent de déduire l'usage prévu du contenu par le créateur, on les mentionnera ici.

Contenu / Conteneur

On liste ici les **contenus additionnels embarqués** (CPO-ADD) optionnellement présents dans les fichiers du format et leurs usages. La section « Métadonnées internes » fournit plus de détails sur ce type de contenu précis.

Dans le cas d'un format conteneur, on précisera les formats ou les codecs embarqués des différents flux de données. Dans le cas d'un codec, on précisera dans quels conteneurs il peut être embarqué, et s'il existe un rapport particulier avec l'un d'eux.

Organisme de maintenance et documentation de référence

Cette section donne l'identifiant du standard / de la norme décrivant le format ou à défaut son site officiel, ainsi que, le cas échéant, de la documentation additionnelle ou des études exhaustives sur le format dans une perspective de préservation à long terme.

Identifiants

On indiquera dans cette section le(s) identifiant(s) correspondant au format dans les registres de formats *Sustainability of Digital Formats* de la *Library of Congress*, Wikidata, PRONOM de *The National Archives* et le wiki *Just Solve the File Format Problem* .

Caractéristiques techniques

On décrira ici les caractéristiques techniques ayant un impact sur la durabilité des données, en particulier sur les critères de **compacité** (CPO-COM, notamment le ou les algorithme(s) de compression

disponibles), **simplicité** (CPO-SIM), **mécanismes de protection** (CPO-PRO), **robustesse** (CPO-ROB), **transparence** (CPO-TRA) et **indépendance / autonomie** (CPO-AUT).

Métadonnées internes

Lorsque le format peut embarquer des métadonnées internes, qu'elles soient descriptives, de provenance ou techniques, on signalera dans cette section le type d'information et le formalisme de ces métadonnées.

Outils connus par la BnF

On citera dans cette section les outils que la BnF considère comme particulièrement utiles pour réaliser :

- la caractérisation ;
- la validation ;
- et, le cas échéant, la production ou la migration.

On mentionnera si l'une des versions de l'outil est préférable à une autre.

Usage ou présence dans les collections de la BnF

Cette section décrit le ou les usage(s) majoritaire(s) du format à la BnF, et le cas échéant signale dans quels cas des fichiers de ce format ont intégré les collections de l'établissement.

4.2. DAISY Digital Talking Book

À ses débuts en Suède en 1994, DAISY (*Digital Accessible Information SYstem*) était un standard propriétaire. Depuis 1997 c'est une norme ouverte recourant à des formats de fichiers standardisés (XML notamment).

Description

DAISY Digital Talking Book (ou « DAISY DTB ») est un standard développé et maintenu par le Consortium Daisy pour la production de contenus numériques accessibles pour les personnes en situation de handicap. Il met l'accent sur la description structurelle des contenus, les fonctions de navigation, l'association d'une transcription sonore au texte et la description des éléments visuels.

Un livre numérique produit selon le standard DAISY est appelé « Digital Talking Book » (DTB) dans les spécifications NISO (ANSI/NISO Z39.86-2005 (R2012) Specifications for the Digital Talking Book). On rencontre aussi l'appellation « DAISY XML » ou même « DAISY ».

DTBook (aussi appelé « XML DTBook » ou « DAISY XML ») désigne le formalisme XML défini par le standard pour encoder les contenus textuels des livres numériques produit selon le standard DAISY. Il est défini par une DTD XML (*DTBook Element Set*, dtbook.dtd), également décrite dans les spécifications ANSI/NISO Z39.86-2005.

DAISY DTB est principalement utilisé selon trois déclinaisons :

- audio avec navigation : le contenu audio est fourni, ainsi qu'une structure de navigation au sein du contenu (peut être appelé « DAISY audio »),
- audio et texte synchronisé, avec navigation : le texte complet est fourni, et il est synchronisé avec sa transcription audio (« full DAISY ») ,
- texte seul, avec navigation : le texte complet est fourni, sans transcription audio (dans ce cas, il est courant de parler de format « DAISY texte »).

Sociabilité

Le standard DAISY DTB est largement employé pour la création de contenus numériques accessibles sous la forme de livres audio synchronisés avec le plein texte.

Relations à d'autres formats

DAISY et EPUB sont liés de par leur histoire :

- le Consortium DAISY a participé à la création du format EPUB au sein de l'IDPF (désormais au sein du W3C). DTBook est un format de contenu accepté par EPUB 2.0,
- le Consortium DAISY est fortement impliqué dans la création et la maintenance du format EPUB 3, nativement accessible. Il s'est maintenant engagé à utiliser EPUB 3 comme son format principal de distribution ; le standard DTBook n'a donc pas été inclus dans l'EPUB 3.

Versions

Le format est utilisé en versions 2 (2002) et 3 (2005). Il n'est pas prévu de version 4.

Contenu

Le standard DAISY définit un ensemble de composants formant un livre numérique DAISY :

- un manifeste au format OEBPS (*Open eBook Publication Structure*) 1.2, maintenu par l'IDPF ;
- des contenus audio et images ;
- des contenus textuels conformes à la DTD XML dtbook.dtd ;
- une table de navigation au format NCX (*Navigation Control file for XML*), format également maintenu par le consortium DAISY. La table est généralement dérivée de façon automatique à partir des contenus XML ;
- un mécanisme de synchronisation audio/texte basé sur le langage SMIL (Synchronized Multimedia Integration Language) 2.0.

Organisme de maintenance et documentation de référence

Le consortium DAISY est l'organisme de maintenance. DAISY DTB est un standard NISO (ANSI/NISO Z39.86-2005 (R2012) *Specifications for the Digital Talking Book*).

Identifiants

Registre	**Identifiant**
Wikidata	Q1155804 (le format XML DTBook), Q1151676 (le livre DAISY)
Bibliothèque du Congrès	fdd000053 (version Digital Talking Book. ANSI/NISO Z39.86-2002), fdd000275 (version Digital Talking Book. ANSI/NISO Z39.86-2005)

Caractéristiques techniques

Outre les points évoqués supra, mentionnons :

- le support des formats audio MPEG-4 AAC, MPEG-1/2 Layer III (MP3), Linear PCM ;
- le support des formats image JPEG, PNG, SVG ;
- le standard ne prévoit pas de méthode de création d'un paquet numérique (telle la restriction ZIP du format EPUB).

Métadonnées internes

Des métadonnées descriptives peuvent être embarquées dans l'élément `<metadata>` du manifeste, selon le formalisme Dublin Core.

Outils connus par la BnF

Les contenus XML au format DTBook sont validés conformément à la DTD dtbook 2005-3.

L'outil Pipeline, via le script `DTBookValidator.taskScript`, vérifie la conformité du fichier DTBook à la DTD et sa cohérence (fichiers images, notes non référencées, liens sans cible, etc.).

```
pipeline.bat scripts\verify\DTBookValidator.taskScript --
input=fichier_a_valider.xml
```

Pipeline permet également de procéder à des transformations de format (XML DTBook vers HTML ou EPUB).

La caractérisation de Daisy Digital Talking Book, comme celle des autres formats XML, est réalisée par le module XML de l'outil Jhove. L'option `withTextMD` permet de produire une sortie au format textMD.

Usage ou présence dans les collections de la BnF

La BnF utilise le format DTBook, sans manifeste OEBPS ni table de navigation NCX, pour diffuser sur Gallica, sous une forme accessible, les contenus textuels des livres numériques produits au format EPUB 3. Cette production s'appuie sur un référentiel de numérisation (Référentiel DAISY)

Ces formats sont téléchargeables dans la bibliothèque numérique Gallica, en complément des livres numériques EPUB 3.

Le dispositif PLATON (PLAteforme de Transfert d'Ouvrages Numériques) accueille les fichiers d'œuvres adaptées par les organismes transcripteurs dans le cadre de l'exception handicap au droit d'auteur. Ces fichiers sont destinés à l'usage d'un public empêché de lire. Dans ce cadre, la BnF conserve et communique aux organismes habilités, par le biais de PLATON, des fichiers de type XML DTBook (3 980 en 2020), des fichiers DAISY audio en voix humaine et voix de synthèse (38 en 2020), et des fichiers dits « Full DAISY » (texte et audio, voix humaine ou voix de synthèse, 1 en 2020).

4.3. EPUB

Description

EPUB est un format ouvert de distribution et d'échange pour les publications et documents numériques. Il permet d'empaqueter sous forme d'un fichier unique compressé des contenus textuels, audio et images.

Pour répondre à l'arrivée des nouveaux petits écrans mobiles (2004 : premier e-reader ; 2007 : premiers Kindle et iPhone ; 2010 : premier iPad) EPUB, dès sa normalisation en 2007, emprunte au web ses langages et sa logique fondamentale : ses contenus sont enrichis sémantiquement grâce à une structuration (X)HTML, et mis en forme grâce à CSS pour une adaptabilité maximale au dispositif d'affichage.

Une version EPUB 2.0.1 est approuvée en 2010. Suivent EPUB 3.0 en octobre 2011, EPUB 3.0.1 en juin 2014, EPUB 3.1 en janvier 2017, et EPUB 3.2 en 2019.

En 2020, et malgré l'ancienneté notable d'EPUB 3, c'est toujours EPUB 2.0.1 qui reste majoritaire dans la production française.

Sociabilité

Dès 2010, EPUB est progressivement, puis très vite unanimement, adopté par les éditeurs français et les autres acteurs de la chaîne du livre numérique (diffuseurs et distributeurs).

Une exception est à signaler : la société Amazon, premier revendeur de livres numériques sur le marché français, exploite des formats propriétaires (dont la production repose très majoritairement sur une conversion d'EPUB).

Relations à d'autres formats

ZIP : l'extension « .epub » cache un fichier ZIP conforme. Une publication EPUB peut donc être décompressée et ses contenus ainsi rendus disponibles (en l'absence de DRM, voir *infra*).

XML et HTML : à sa racine, une publication EPUB présente un fichier « container.xml » dans un répertoire « META-INF ». Essentiellement, ce court fichier XML localise, dans l'arborescence interne, un autre fichier XML, lui aussi indispensable et dont l'extension est « .opf » (pour *Open Packaging Format*). Ligne de l'élément `<rootfile>` de « container.xml » : `<rootfile full-path="OEBPS/content.opf" media-type="application/oebps-package+xml"/>`.

Les « documents de contenu » d'EPUB 3 sont des fichiers XHTML. À ce titre, ils doivent se conformer aux syntaxes HTML et XML, et peuvent intégrer des langages XML spécifiques : MathML pour la gestion des équations mathématiques, par exemple, ou SVG (images vectorielles).

CSS : relier les fichiers XHTML à une ou des feuilles de style CSS n'est pas une obligation, mais une recommandation fonctionnelle.

Versions

Avec le passage de la version 2 à la version 3, en octobre 2011, les capacités techniques du format augmentent fortement. L'éventail des exigences éditoriales auxquelles il peut répondre s'élargit.

- Le support d'HTML 5 rend possible le codage d'une accessibilité native complète, et l'intégration de contenus audio et vidéo.
- La synchronisation audio/texte (*Media Overlays*) est un outil supplémentaire pour l'accessibilité native.
- Le support de CSS 3 améliore la gestion des flux textuels et permet la prise en charge des écritures idéogrammatiques, verticales ou de droite à gauche (arabe et hébreu par exemple) : le format s'internationalise.
- Le support de JavaScript permet le codage d'une interactivité contenus/lecteur.

La gestion des magazines, des publications scolaires, professionnelles et scientifiques reste certes encore délicate, mais EPUB 3 propose des solutions : notamment une exception au principe d'adaptabilité des contenus en permettant, lorsqu'une nécessité éditoriale l'impose, de figer les pages de la publication : on parle alors de *fixed-layout*.

Au-delà de la version 3.2 : le W3C semble satisfait de la maturité du format dans sa version actuelle, et en fait la promotion, notamment dans l'objectif d'augmenter sa représentation dans la production éditoriale (au détriment d'EPUB 2.0.1, donc).

Un format parallèle, WEB PUBLICATION, est cependant à l'étude, qui ferait le lien entre publication packagée offline et disponibilité sur le web. Le format est actuellement présenté comme « une collection de ressources, organisées ensemble par le biais d'un manifeste avec un ordre de lecture par défaut, [...] identifiable de façon unique et pouvant être présentée à l'aide des technologies *Open Web Platform* ».

Contenu

EPUB 3 peut intégrer les formats suivants *(W3C core media types)* :

- images : GIF, JPEG, PNG, SVG ;
- audio : MPEG-1/2 layer III, MP4 AAC ;
- vidéo : H.264 ;
- polices : TTF, OTF, WOFF, WOFF2.

Il est à noter qu'EPUB 3 laisse la porte ouverte à l'intégration d'autres formats, à la condition du signalement explicite dans le fichier « .opf » des formats de remplacement en cas de non prise en charge par le système de lecture (mécanisme du *manifest fallback*).

Organisme de maintenance et documentation de référence

Jusqu'en février 2017, c'est l'IDPF (*International Digital Publishing Forum*) qui maintient et promeut le format. Relais est passé depuis au W3C, avec qui l'IDPF a fusionné à cette date. Le document *EPUB 3.2 W3C Community Group Final Report* est une bonne porte d'entrée vers la très riche documentation du W3C.

Certaines versions d'EPUB ont également été portées par l'IDPF à l'ISO ; ainsi les spécifications d'EPUB 3.0 ont été publiées sous l'identifiant ISO/IEC TS 30135, et celles d'EPUB 3.0.1 sous l'identifiant ISO/IEC 23736. Une norme ISO sur l'accessibilité des EPUB est également à l'étude : ISO/IEC DIS 23761. Des spécifications techniques doivent compléter le paysage normatif autour d'EPUB : l'ISO/IEC TS 22424 définit une restriction du format EPUB pour sa préservation à long terme et son empaquetage avec des métadonnées METS et PREMIS, et l'ISO/IEC CD TS 23078 spécifie une technologie standard de DRM.

On pourra aussi consulter :

- une évaluation par la British Library de tous les formats ebook ;
- une évaluation par la British Library du format EPUB ;

Identifiants

Registre	Identifiant
Bibliothèque du Congrès	fdd000310 (famille EPUB), fdd000278 (EPUB 2), fdd000308 (EPUB 3.0), fdd000311 (EPUB 3.0.1), fdd000309 (EPUB 3.2), fdd000519 (restriction d'EPUB pour la préservation)
Wikidata	Q475488
PRONOM	fmt/483
Just Solve the File Format Problem	EPUB

Caractéristiques techniques

Une publication EPUB peut-être tatouée numériquement afin d'être rendue traçable, ou verrouillée par un DRM, qui consiste en un **chiffrement complet** de son contenu opéré à la volée au moment du téléchargement depuis un site web.

Si le verrou ACS4 d'Adobe a fait l'objet d'une adoption massive par les éditeurs français dès 2010, le consensus s'est inversé ces dernières années, et ACS4 cristallise aujourd'hui les reproches de nombreux acteurs de la chaîne du livre numérique, libraires et bibliothèques notamment : parcours de premier achat complexe, problèmes techniques liés au chiffrement, centralisation des données personnelles des usagers, coût transactionnel et coût unitaire élevés, incompatibilité avec l'accessibilité des contenus...

Plus souple, moins cher, compatible avec l'accessibilité des contenus, LCP, développé par EDRLab, est, aux yeux du monde éditorial aujourd'hui, une bonne alternative.

Métadonnées internes

C'est le fichier OPF (voir *supra* : « Relations à d'autres formats ») qui centralise les métadonnées (Dublin Core et Schema.org) dans sa première section `<metadata>`. La deuxième - `<manifest>` - détaille les ressources de la publication. La troisième - `<spine>` - fixe l'ordre de lecture des chapitres XHTML.

Outils connus par la BnF

Il serait vain ici de vouloir lister les gestionnaires de bibliothèques ou les lecteurs d'EPUB.

Méritent, à notre sens, d'être signalés :

- **InDesign**, comme logiciel PAO exportant le format ;
- **Sigil**, comme éditeur open source, permettant la correction ou la création *ex nihilo* du format ;
- **Thorium**, comme lecteur pour Windows 10, MacOS et Linux, développé par EDRLab.

Par ailleurs, il existe un outil officiel de validation - **EPUBCheck** - aujourd'hui dans sa version 4.2.5. Exécutable en ligne de commande ou utilisable comme une bibliothèque Java, il vérifie la conformité d'un fichier aux spécifications du format, dont il prend en charge les versions 2 et 3. L'outil réalise également une caractérisation qui produit nativement un rapport au format XMP.

On peut lui adjoindre un autre outil de vérification, spécifiquement du codage de l'accessibilité : Ace, développé par le DAISY Consortium.

Usage ou présence dans les collections de la BnF

Plusieurs milliers de documents patrimoniaux sont disponibles aux formats EPUB 2 et 3 dans la bibliothèque numérique de la BnF, Gallica.

Le Référentiel ePub 3 présente les attentes de la BnF en matière de production de livres électroniques au format EPUB. Il détaille notamment les modalités de conversion des contenus patrimoniaux de la bibliothèque vers ce format.

Dans le cadre de la loi relative au droit d'auteur et droits voisins dans la société de l'information dite Loi DADVSI, la BnF a pour mission de collecter, conserver et donner accès à tous les livres numériques au format EPUB diffusés au public par voie électronique en France. Ainsi, les livres numériques au format EPUB qui sont téléchargeables librement en ligne font l'objet d'une collecte automatique par moissonnage grâce à un robot-crawler, dans le cadre du dépôt légal du web. Les EPUB collectés de cette manière sont consultables grâce à l'application Archives de l'internet uniquement dans les salles de recherche de la BnF et les bibliothèques de dépôt légal imprimeur en région.

Cependant, la majorité des ouvrages au format EPUB mis à disposition du public dans le circuit commercial n'étant pas librement téléchargeables, le dépôt légal de ces livres numériques nécessite donc la mise en place d'un circuit spécifique, qui implique un dépôt de leurs publications à la BnF par les déposants. Ainsi, les livres numériques au format EPUB sans mesure technique de protection (*Digital Rights management*) collectés par dépôt des éditeurs sont accessibles grâce à l'application Gallica Intra Muros uniquement dans les salles de recherche de la BnF.

Le dispositif PLATON (PLAteforme de Transfert d'Ouvrages Numériques) accueille les fichiers source des éditeurs (11% au format EPUB, soit 6 000 fichiers en 2020) et les adaptations réalisées par les organismes transcripteurs (0,5% au format EPUB, soit 38 fichiers) dans le cadre de l'exception handicap au droit d'auteur. Les fichiers source des éditeurs sont accessibles par les organismes transcripteurs agréés via PLATON. Les adaptations sont destinées aux personnes empêchées de lire.

4.4. H.264 (codec)

Description

H.264, également nommé « MPEG-4 AVC » (*Advanced Video Coding*), ou « MPEG-4 Part 10 » est un codec normalisé de compression vidéo avec perte développé par le Joint Video Team (JVT), groupe de travail issu de l'ISO/CEI Moving Picture Experts Group (MPEG) et de l'UIT-T Q.6/SG16 Video Coding Experts Group (VCEG). La première version de la norme a été approuvée en mai 2003 et a été régulièrement révisée avec l'ajout d'extensions jusqu'à mai 2019. Ce codec vise obtenir un bon ratio entre la qualité et des débits limités (CPO-COM), tout en conservant un niveau de complexité raisonnable (CPO-SIM). La variété de ses profils lui permet de répondre à des usages différents, avec des résolutions variables allant jusqu'à la 8K.

Bien que documenté (CPO-DOC), le codec H.264 est protégé par un ensemble de brevets déposés par différents organismes (CPO-LIB). Son usage **commercial** est conditionné au paiement d'une redevance auprès de la société de gestion MPEG LA.

À l'origine, en 1998, l'UIT-T lança ce projet dans le but de créer un codec permettant de réduire de plus de moitié le débit binaire des fichiers des standards de l'époque (MPEG-2, H.263 et MPEG-4 Part 2), de créer une interface simple afin d'adapter le codec aux différents protocoles de transport (vidéophonie, streaming, télévision, HD, mobile), mais aussi de pouvoir l'implémenter à faible coût dans les appareils. Le codec H.264/AVC est donc adapté à une très grande variété de réseaux et systèmes (diffusion télévision numérique, VOD, Blu-ray, streaming RTP/IP, et systèmes de téléphonie).

Sociabilité

H.264 est un des codecs vidéo les plus répandus auprès du grand public et des professionnels de la vidéo (le rapport *Video Developer Report* de 2019 indique que 91% des développeurs utilisent ce codec). Il s'agit du format majoritaire utilisé pour la production de disques Blu-Ray ; il est également très couramment utilisé par les plates-formes de diffusion vidéo en ligne (Netflix, Amazon Prime, Youtube, etc.). C'est également dans ce standard qu'est transmise la télévision numérique terrestre (TNT) en Europe.

Relations à d'autres formats

Schématiquement, le codec H.264 est précédé par H.262, qui avait déjà pour but la transmission de vidéo à bas débit, et est suivi par H.265 (MPEG-4 HEVC) visant l'échange de vidéos de haute qualité.

Ce codec ne doit pas être confondu avec « MPEG-4 part 2 » ou « MPEG-4 visual », qui désigne un codec vidéo basé sur l'algorithme DCT (*Discrete Cosine Transform*) également employé par H.262.

Versions

La norme H.264 se caractérise par une grande polyvalence, pour des besoins allant de la téléphonie mobile au cinéma numérique. Afin d'adapter H.264 à des usages différents, la norme définit des profils spécifiques.

Dans sa version originelle de 2003 la norme définissait trois profils, ci-dessous classés par niveau de qualité croissant :

- **Baseline Profile (BP)** : principalement pour les applications à bas coût qui utilisent peu de ressources, ce profil est très utilisé dans les applications mobiles et de visioconférence.
- **Main Profile (MP)** : à l'origine, prévu pour les applications grand public de diffusion et de stockage, ce profil a perdu de l'importance quand le profil High a été ajouté avec le même objectif.
- **Extended Profile (XP)** : prévu pour la diffusion en flux (*streaming*) des vidéos, ce profil a des capacités de robustesse à la perte de données et de changement de flux.

En 2004, le JVT a développé de nouvelles extensions connues sous le nom *Fidelity Range Extensions* (**FREXt**). Ces extensions prennent en charge une quantification plus élevée (10-bit et 12-bit) et des structures de chrominance plus précises (YUV 4:2:2 et YUV 4:4:4). Elles visent de ce fait des applications professionnelles (post-production, tournage en haute définition). Plusieurs autres fonctionnalités ont aussi été adoptées pour améliorer la qualité subjective en haute définition ou pour des besoins spécifiques (codage sans perte, support d'autres espaces de couleurs). Ces extensions donnent lieu à la définition de quatre nouveaux profils :

- **High Profile (HiP)** : prévu pour la diffusion et le stockage sur disque, en particulier pour la télévision haute définition (HD DVD, Blu-ray, TNT), il s'agit du profil de loin le plus couramment utilisé.
- **High 10 Profile (Hi10P)** : ce profil va au-delà des applications grand public et s'appuie sur le profil High — ajoutant jusqu'à 10 bits de précision par pixel.

- **High 4:2:2 Profile (Hi422P)** : ce profil principalement fait pour les applications professionnelles, s'appuie sur le profil High 10 — ajoutant le support pour la quantification 4:2:2 jusqu'à 10 bits par pixel.
- **High 4:4:4 Profile (Hi444P)** : ce profil s'appuie sur le profil High 4:2:2 — ajoutant le support pour la quantification 4:4:4, jusqu'à 12 bits par pixel et en plus le support pour un mode sans perte efficace.

En outre, la norme contient quatre profils supplémentaires « AVC-I » définis comme des spécialisations de trois des sept profils précédents. Ces profils spécialisés utilisent un jeu d'outils de compression opérant uniquement en mode intra-image (chaque image est présente, entière et sans interpolation). Ce type de compression est particulièrement adapté à la post-production. Ces profils sont principalement destinés à des applications broadcast HD (actualités, publicité et cinéma) : **High 10 Intra Profile ; High 4:2:2 Intra Profile ; High 4:4:4 Intra Profile ; CAVLC 4:4:4 Intra Profile**.

En outre, les annexes de la norme définissent des extensions du format pour des besoins spécifiques.

L'annexe G de la norme décrit le Codage vidéo scalable (SVC), extension de H.264/AVC, permettant d'offrir un contenu adaptable à différents débits et qualités. La norme contient trois profils adaptables supplémentaires :

- **Scalable Baseline Profile** : principalement destiné aux applications de vidéoconférence, de téléphonie mobile et de surveillance, ce profil s'appuie sur une version limitée du profil *baseline* H.264/AVC.
- **Scalable High Profile** : ciblant principalement les applications de diffusion et de streaming, ce profil s'appuie sur le profil *high* H.264/AVC.
- **Scalable Constrained High Profile** : sous-ensemble du *Scalable High Profile* destiné principalement aux applications de communication en temps réel.
- **Scalable High Intra Profile** : ciblant principalement les applications de production, il s'agit d'une spécialisation du profil *Scalable High Profile* limité au mode intra-images exclusivement.

L'annexe H de la norme décrit le codage **MVC** (*Multiview Video Coding*), extension permettant l'encodage de vidéo stéréoscopique et spécifiant progressivement des profils pour ces nouveaux usages.

Conteneur

Il existe un lien structurel entre le format conteneur MP4 et le codec H.264 : le premier est normalisé dans la partie 14 de la norme MPEG-4 (elle-même basée sur la partie 12 : « Format ISO de base pour les fichiers médias »), le second dans la partie 10.

Néanmoins, des flux vidéo H.264 peuvent se retrouver dans de multiples autres conteneurs (MKV, AVI, MPEG-PS, Quicktime, etc.).

Organisme de maintenance et documentation de référence

Développé par le *Motion Picture Experts Group* (MPEG), le codec H.264 a pour organisme de maintenance l'Union internationale des télécommunications (UIT-T), qui en diffuse une version gratuite. Le standard a été porté à l'ISO : il s'agit de la norme ISO/IEC 14496-10 « Technologies de l'information — Codage des objets audiovisuels — Partie 10 : Codage visuel avancé ».

Un résumé conséquent de chaque partie de la norme MPEG-4 est disponible sur l'ancien site officiel du Motion Picture Experts Group.

Identifiants

Registre	Identifiant
Wikidata	Q212633
Bibliothèque du Congrès	fdd000081 pour le H.264/MPEG-4 AVC fdd000082 pour le H.264/MPEG-4 AVC Baseline Profile fdd000083 pour le H.264/MPEG-4 AVC Main Profile fdd000084 pour le H.264/MPEG-4 AVC Extended Profile fdd000215 pour le H.264/MPEG-4 AVC High Profile fdd000216 pour le H.264/MPEG-4 AVC High 10 Profile fdd000217 pour le H.264/MPEG-4 AVC High 4:2:2 Profile fdd000218 pour le H.264/MPEG-4 AVC High 4:4:4 Profile fdd000162 pour le H.264/MPEG-4 AVC Non-FREXt Extensions

Just Solve the File Format Problem	H.264

Note : le registre de formats PRONOM ne dispose pour le moment que d'une entrée succincte sur MPEG-4.

Caractéristiques techniques

Compacité (CPO-COM) : le H.264/AVC cumule de nombreuses techniques d'encodage (prédiction temporelle et spatiale, adaptation des transformées discrètes, codage entropique, adaptatif ou de longueur variable, filtrage, deux modes d'entrelacement, etc.). Ces méthodes sont efficaces, et permettent, à qualité égale, de limiter le débit par rapport à MPEG-2 H.262, autorisant ainsi l'échange de vidéo haute définition. La méthode de compression très majoritairement utilisée pour encoder en H.264 est non réversible, mais il est théoriquement possible de compresser sans perte tout ou partie des zones d'image, bien que cette méthode ait pour conséquence des volumétries beaucoup plus conséquentes.

Complexité (CPO-SIM) : Les méthodes mentionnées précédemment confèrent à H.264 une plus grande compacité que les normes précédentes, mais également une complexité supérieure. La flexibilité introduite par ses sept profils, dix extensions, et vingt niveaux y contribue également.

Robustesse (CPO-ROB) : des mécanismes de résilience sont introduits dès le profil *Baseline*, et complétés dans le profil *Extended*, afin d'augmenter la tolérance aux erreurs lors de la transmission réseau. En outre, la structuration de H.264 en deux couches (*layers*) conceptuelles améliore la résilience. La couche VCL (*Video Coding Layer*) spécifie les méthodes de compression tandis que la couche NAL (*Network Abstraction Layer*) concerne l'adaptation aux protocoles de transport. Les données sont transmises en unités NAL précédées d'un marqueur spécifique, ce qui permet aux décodeurs de se resynchroniser en cas de défaillance de la transmission réseau. Pour plus de détails, on se reportera à l'article de Till Halbach et Steffen Olsen, « *Error robustness evaluation of H.264/MPEG4 AVC* », 2004, dans Proc SPIE, 5308, p. 617-627, doi:10.1117/12.522766, consultable sur https://www.researchgate.net/publication/248595612_Error_robustness_evaluation_of_H264MPEG4_AVC.

Outre le profil, les flux H.264 sont également caractérisés par un **niveau** (*level*) correspondant à la définition d'un certain nombre de paramètres (taille d'image, débit, cadence) qui permettent aux constructeurs de dimensionner les ressources calculatoires et mémoire nécessaires aux appareils pour décoder une vidéo. Le tableau ci-dessous présente ces paramètres pour quelques niveaux significatifs (pour les paramètres de tous les niveaux, on se reportera à la page Wikipédia française du format H.264).

N° de niveau	Taille maximale de l'image	Cadence maximale (à cette taille d'image)	Débit maximum (1) (Baseline, Extended et Main Profile)	Nb maximum d'images de référence
3	720×576	25	10 Mbit/s	5
3.1	1280×720	30	14 Mbit/s	5
3.2	1280×1024	42,2	20 Mbit/s	4
4	2048×1024	30	20 Mbit/s	4
4.1	2048×1024	30	50 Mbit/s	4
4.2	2048×1080	60	50 Mbit/s	4
5	3672×1536	26,7	135 Mbit/s	5
5.1	4096×2304	26,4	240 Mbit/s	5
6	8192×4320	30,2	240 Mbit/s	5

(1) Pour les autres profils, le débit est à multiplier respectivement par 1,25 (High) ou 3 (High 10) ou 4 (High 4:2:2 ou 4:4:4)

Métadonnées internes

H.264 est un flux de données vidéo et ne porte donc pas de métadonnées descriptives ; celles-ci, lorsqu'elles existent, sont codées dans le format conteneur associé.

Outils connus par la BnF

Attention, les décodeurs ne traitent pas nécessairement tous les profils listés ci-dessus.

Un outil de manipulation et de transcodage est `ffmpeg`.

L'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7.

Usage ou présence dans les collections de la BnF

Depuis quelques années la BnF numérise certains documents avec un codec H.264 High Profile, niveau 4.1 (abrégé en `HiP@4.1L`) :

- Pour un signal SD (720×576), avec un sous-échantillonnage de la chrominance à 4:2:0 et des débits de 5, 10, 20 Mb/s, ou à 4:2:2 et 50 Mb/s.
- Pour un signal HD (1920×1080), avec un sous-échantillonnage de la chrominance à 4:2:0 et un débit de 10 Mb/s, ou à 4:2:2 à 100 Mb/s.

Par ailleurs, le H.264 est utilisé pour la diffusion sur Gallica.

Dans les collections de la BnF sont présents, au titre du Dépôt légal ou de divers dons et dépôts, des documents encodés avec le codec H.264.

4.5. JPEG File Interchange Format (JPEG)

Description

Cette fiche décrit le standard d'encodage et de compression d'images matricielles JPEG ainsi que son format conteneur le plus courant, JFIF. L'algorithme JPEG n'est applicable qu'aux images en couleur ou en niveaux de gris, non à celles binarisées. Il implique une compression forcément irréversible, appliquée à chaque enregistrement. Ce format est donc particulièrement adapté à un usage de diffusion.

Sociabilité

Le format JFIF est très largement utilisé pour la diffusion des images en ligne. À la date de publication de ce document (2019), il était néanmoins légèrement moins répandu que le format PNG (voir sur le site de W3Techs, Usage of File Formats for websites).

Versions

Il existe une méthode de compression sans perte décrite dans la norme ISO/IEC 14495, mais son usage est très limité.

Contenu

Le flux image compressé selon l'algorithme JPEG ne gère qu'une profondeur d'encodage maximale de 8 bits par couche. Par conséquent, les données encodant la couleur sur 16 bits par couche seront réduites à 8 bits lors de la migration.

Les fichiers JFIF contiennent plusieurs segments, introduits par des marqueurs sur deux octets (en hexadécimal, '0xFF' puis un octet spécifique au segment). Les images JFIF peuvent contenir une vignette dans un segment de marqueur APP1.

Documents de référence

Eric Hamilton, *JPEG File Interchange Format version 1.02*, 1er septembre 1992. Disponible sur https://www.w3.org/Graphics/JPEG/jfif3.pdf (consulté le 03/12/2020). La méthode de compression JPEG est décrite par les normes ISO/IEC 10918 et ISO/IEC 14495 (compression sans perte).

Identifiants

Registre	Identifiant
Wikidata	Q26329975
Bibliothèque du Congrès	fdd000018
PRONOM	fmt/42 (version 1.00), fmt/43 (version 1.01), fmt/44 (version 1.02)

Caractéristiques techniques

Tous les visualiseurs n'étant pas à même de traiter correctement des fichiers JPEG dans un espace couleur CMYK ou CIE `L*a*b*`, il est recommandé d'utiliser l'espace couleur RGB.

Si la BnF est amenée à convertir en JFIF des photographies numériques livrées dans d'autres formats, elle n'appliquera pas de compression en plus de celle directement issue de l'algorithme.

L'algorithme de compression JPEG permet de choisir entre plusieurs modes, notamment « Baseline » et « Progressive ». Ce dernier mode optimise l'affichage pour le web en contenant un nombre paramétrable de scans de précision croissante, qui sont successivement chargés par le navigateur.

Métadonnées internes

Le format conteneur JFIF peut embarquer des flux complémentaires au flux image JPEG (informations sur la résolution et métadonnées aux formats EXIF, IPTC et XMP).

Outils connus par la BnF

Les outils utilisés par la BnF sont les suivants :

- pour la caractérisation et la validation : module JPEG de JHOVE ;
- pour la visualisation : XnView.

Usage ou présence dans les collections de la BnF

Le format JPEG a été utilisé pendant une dizaine d'années pour la numérisation de pages couleur, il a ensuite été remplacé par le TIFF non compressé, puis par le JPEG 2000.

4.6. JPEG 2000

Description

Cette entrée décrit le standard d'encodage et de compression d'images JPEG 2000 ainsi que son format conteneur.

Le format est parfois nommé « JP2 » en référence à son extension « .jp2 ».

L'encodage JPEG 2000 supporte une compression par ondelettes sélective (les zones identifiées comme moins riches en information seront davantage compressées).

Sociabilité

Bien que les appareils photographiques ne puissent produire des images JPEG 2000, et que les navigateurs ne gèrent pas nativement ce format, les institutions patrimoniales sont nombreuses à l'avoir choisi (BnF, *Library of Congress*, *Wellcome Library*, *British Library*, etc.). En outre, sa variante pour l'image animée, Motion JPEG 2000, a été plus largement adoptée et est souvent proposée par les caméras numériques comme format de sortie.

Relations à d'autres formats

Le format JPEG 2000 n'a pas de relation particulière avec le format JFIF (JPEG), cependant les deux formats sont maintenus par le même groupe de travail international *Joint Photographic Experts Group*.

Versions

La question des versions n'est pas pertinente pour ce format.

Contenu

Le format conteneur JPEG 2000 peut embarquer des flux complémentaires au flux image JPEG 2000 (transparence, informations sur le profil couleur et autres métadonnées).

Organisme de maintenance et documentation de référence

Le format JPEG 2000 est maintenu par le groupe de travail international *Joint Photographic Experts Group*. Il a fait l'objet de la norme ISO/IEC 15444. Cette dernière comprend 16 parties en 2019, dont le coût total est estimé à environ 2500 €. Néanmoins, la principale est la première, qui définit les caractéristiques principales de l'algorithme de compression. Cette première partie est disponible gratuitement sous sa forme approuvée en 2002 sur le site de l'ITU.

La British Library a réalisé une évaluation de ce format pour la préservation à long terme.

Identifiants

Registre	Identifiant
Wikidata	Q931783
Bibliothèque du Congrès	fdd000143
PRONOM	x-fmt/392, fmt/463
Just Solve the File Format Problem	JPEG 2000

Caractéristiques techniques

Deux types de compression existent : RCT (*Reversible Component Transfer*), qui est sans perte (*lossless*), et ICT (*Irreversible Component Transfer*), qui implique une perte (*lossy*), avec chacune plusieurs niveaux de compression différents. La conversion sans perte garantit une réduction conséquente du poids des données (CPO-COM) tout en restant **totalement** réversible. La conversion avec pertes est plus efficace que la compression JPEG ; elle permet de définir le taux de compression souhaité, ce qui devient un enjeu important et doit dépendre de la nature du document reproduit.

Les taux utilisés par la BnF sont décrits dans le référentiel BnF « Format de fichier image ». Ces taux ont été définis par une étude menée à la BnF fondée sur l'évaluation du « bruit » produit par les numériseurs. Il est fortement recommandé aux institutions patrimoniales souhaitant utiliser JPEG 2000 comme format de préservation de leurs numérisations de suivre strictement les exigences exprimées par la BnF dans ce référentiel, et notamment celles sur le taux de compression, ou à défaut de n'utiliser qu'une compression sans perte.

Il est nécessaire d'intégrer aux JPEG 2000 couleur le profil colorimétrique de l'image de la manière spécifiée dans le référentiel, section 2.3.

A condition d'utiliser les marqueurs de reprise (voir le référentiel BnF « Format de fichier image », section 2.7), le format JPEG 2000 est considéré comme particulièrement résilient en cas d'altération (CPO-ROB) (voir Buonora (Paolo), Liberati (Franco), « A Format for Digital Preservation of Images: A Study on JPEG 2000 File Robustness » dans *D_Lib*, 2008, vol. 14, numéro 7/8, consultable sur http://www.dlib.org/dlib/july08/buonora/07buonora.html).

On trouvera sur le registre des risques pesant sur les formats de fichiers, maintenu par l'*Open Preservation Foundation*, quelques problèmes identifiés sur le format JPEG 2000.

Métadonnées internes

Il est recommandé d'intégrer des informations descriptives et de provenance sous la forme de métadonnées internes au format XMP. On pourra s'inspirer des pratiques BnF décrites à la section 4 du référentiel BnF « Format de fichier image ».

Outils connus par la BnF

Les outils recommandés par la BnF sont les suivants.

- Pour la production, la bibliothèque logicielle Kakadu, dans sa version 7.2.1 ou postérieure. Dans une démarche de production de fichiers JPEG 2000, la BnF recommande l'utilisation de cet outil afin de convertir les fichiers bruts après qu'ils auront subi toutes les opérations de post-production. En effet, la sortie en JPEG 2000 des logiciels de numérisation est rarement assez paramétrable pour correspondre aux exigences du référentiel BnF.
- Pour la visualisation, la bibliothèque logicielle `Kakadu` (notamment l'outil kdu_show) ou un visualiseur compatible IIIF (sous réserve que l'image soit diffusée par un serveur IIIF).
- Pour la caractérisation et la validation, l'outil `Jpylyzer`, dans sa version 2.0.0 ou postérieure. Ses rapports techniques sont convertis par la BnF vers le formalisme recommandé de métadonnées techniques pour les images : Metadata for Images in XML (MIX).

Usage ou présence dans les collections de la BnF

Le format JPEG 2000 a été adopté par la BnF comme format préférentiel pour la numérisation de documents plans à partir de l'année 2015, en remplacement progressif du format TIFF.

4.7. MPEG 1/2 Audio Layer III (MP3)

Description

MPEG-1/2 Audio Layer III, plus connu sous son abréviation « MP3 », est la spécification audio des standards MPEG-1 et MPEG-2. Il s'agit d'un format de compression audio avec perte fondé sur des recherches psycho-acoustiques menées par le *Fraunhofer Institute* à partir de la fin des années 1980. La méthode d'encodage repose sur des filtres visant à éliminer les informations sonores peu perceptibles par l'oreille humaine. Elle permet une réduction importante de la taille du flux de données audio (CPO-COM) tout en conservant une qualité de restitution couramment jugée acceptable, et donne le choix du débit selon le compromis taille-qualité souhaité. Depuis 2017, tous les brevets du *Fraunhofer Institute* portant sur ce format sont arrivés à échéance ; il est donc désormais totalement libre de droits (CPO-LIB).

Le format MP3 n'est pas recommandé comme format de données audio en vue de la préservation à long terme, car sa compression avec pertes, appliquée à chaque enregistrement, est susceptible, à l'occasion de futures opérations de préservation, de dégrader un signal déjà écrêté. A défaut de disposer des données dans un format alternatif, sa robustesse (CPO-ROB), sa simplicité (CPO-SIM) et l'arrivée à échéance des brevets portant sur lui (CPO-LIB) en font un format facilement préservable.

Sociabilité

Depuis la fin des années 1990, il est devenu l'un des formats de musique numérique les plus répandus, bien que des codecs plus performants (AAC, Vorbis, Opus, etc.) soient apparus depuis.

Il est de ce fait très largement géré par les logiciels de *streaming* audio, matériels audiovisuels et autres baladeurs, et également par tous les navigateurs. Dès le début des années 2000, des réseaux d'échange sur Internet via des logiciels de partage de fichiers tels que Napster ont beaucoup contribué à l'adoption de ce format par les consommateurs.

Il est utilisé comme format de diffusion majoritaire pour les livres audio sur support optique et téléchargement.

Contenu

Le format MP3 est une simple concaténation de *frames* audio, éventuellement suivies (ID3v1) ou précédées (ID3v2). Une extension non officielle Xing a défini un en-tête, situé dans une *frame* initiale, destiné à porter des métadonnées techniques sur les *frames* audio (nombre, longueur, etc.).

Les données audio MP3 ont des fréquences d'échantillonnage de 32 kHz, 44,1 kHz ou 48 kHz avec un nombre de canaux audio de 1 à 6. Un fichier MP3 peut être mono, double mono, stéréo, joint stéréo ou multicanal (MP3 Surround). Les canaux peuvent être groupés (cas du joint stéréo et du multicanal 5.1 surround) pour tirer parti de la corrélation intercanal et augmenter la compression. Le débit binaire peut varier entre 8 et 320 kbit/s en fonction du taux et du mode de compression choisi.

Le taux de compression varie en fonction du débit binaire (*bitrate*) choisi : plus ce dernier est bas, plus le taux de compression est élevé et plus le signal est altéré. Ce format de compression avec perte utilise différentes méthodes : algorithmes de regroupements de données identiques, filtrage des hautes fréquences, application d'une courbe en fonction de l'audition humaine, suppression de signaux quasiment inaudibles et utilisation d'un modèle psycho-acoustique de compression (« effet de masque »).

On notera également que différents modes de compression sont utilisables : CBR (*Constant bitrate*) pour un débit fixe, VBR (*Variable bitrate*) pour un débit variable visant à maintenir une qualité d'écoute théoriquement constante, et ABR (*Average bitrate*) qui est un compromis entre les deux premiers types. Si le mode de compression VBR offre le meilleur rapport entre qualité et compacité, seul le mode CBR permet de diffuser un fichier en *streaming*.

Outre les données audio, et métadonnées internes au format ID3, la version 2 d'ID3 dispose d'un champ image pour intégrer un visuel (tous les formats d'image sont acceptés, mais les formats JPEG et PNG sont recommandés pour des questions d'interopérabilité).

Organisme de maintenance et documentation de référence

Le format est normalisé par le groupe *Moving Picture Experts Group* qui maintient les standards ISO/CEI 11172-311 (MPEG-1) et ISO/CEI 13818-3 (MPEG-2).

La *British Library* a réalisé une évaluation de ce format pour la préservation à long terme.

Identifiants

Registre	Identifiant
Wikidata	Q42591
Bibliothèque du Congrès	fdd000012
PRONOM	fmt/134
Just Solve the File Format Problem	MP3

Caractéristiques techniques

Un des avantages majeurs de MP3 est sa résilience (CPO-ROB) : les fichiers dans ce format restent presque toujours lisibles malgré des altérations (fichiers tronqués, mal formés, etc.). Les pratiques d'extraction et de conversion en vue de l'échange de musique ont conduit à la circulation de très nombreux fichiers plus ou moins invalides au regard de la norme, sans que cela ait nécessairement un effet perceptible lors de leur restitution.

Bien que cela soit une pratique rare, il est possible d'intégrer à chaque *frame* une empreinte numérique CRC. En cas d'altération du flux audio, on peut ainsi identifier la partie du fichier corrompue.

Les fichiers MP3 peuvent intégrer des mesures techniques de protection (DRM), notamment ceux diffusés par iTunes de 2003 à 2009.

Outils connus par la BnF

Il n'existe pas d'outil de validation officiel ni développé par la communauté de la préservation numérique. On pourra néanmoins utiliser avec profit les utilitaires suivants.

- `MP3 Validator`, également appelé « MP3val », est un outil libre de vérification et de réparation de fichiers MP3 permettant de repérer certaines irrégularités (fichier incomplètement téléchargé, en-têtes mal formés notamment).
- `mp3_check` propose également la vérification de cohérence entre l'en-tête du fichier et les données audio.

L'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7.

Métadonnées internes

Le format MP3 permet d'embarquer des métadonnées internes au format `ID3`. Deux versions de ce format existent, qui diffèrent notamment par l'endroit du fichier où elles figurent (à la fin pour la version 1, au début pour la version 2).

Plus marginalement, on trouve également des fichiers MP3 comportant des métadonnées au format `APE`.

Usage ou présence dans les collections de la BnF

La BnF diffuse sur Gallica ses données audio au format MPEG-1 Layer III à un débit constant (CBR) de 320 kbit/s, ce qui équivaut à un taux de compression de 4,8 par rapport à son équivalent au format PCM.

Dépôt légal :

- La migration de support opérée par la BnF sur les livres audio, conférences et formations reçus par dépôt légal est à l'origine des fichiers MP3 conservés par la BnF.
- Certains flux rétrospectifs de musique dématérialisée sont attendus au format MP3, en l'absence d'une version non compressée en FLAC ou en WAVE.

Autres missions nationales : le format MP3 est représenté marginalement dans les fichiers adaptés collectés par la BnF au titre de l'exception handicap.

Dons / acquisitions / production interne **:** des fonds issus de collectes, conférences et entretiens sont présents au format MP3 dans les collections de la BnF.

4.8. MPEG-2 (conteneur) H.262 (codec vidéo) MPEG-2 layer II (codec audio)

Description

MPEG-2 est la seconde génération de la norme du *Moving Picture Experts Group*, intitulée *« Codage générique des images animées et du son associé »*, qui définit les aspects de compression de l'image et du son, ainsi que le transport des données à travers des réseaux pour la télévision numérique. Cette fiche décrit à la fois les formats conteneurs MPEG-2, le format du flux vidéo H.262 (MPEG-2 part 2) et le format du flux audio MPEG-2 layer II (MPEG-2 part 3).

Le type MIME le plus couramment utilisé, et reconnu par l'IETF, est `video/mpeg`, mais on note l'utilisation de quatre sous-types également officiels liés à chaque format conteneur (`video/mpv` pour *MPEG-1 or -2 Elementary Stream*, `video/mp2t` pour *MPEG-2 Transport Stream*, `video/mp1s` pour *MPEG-1 Systems Stream* et `video/mp2p` pour *MPEG-2 Program Stream*), ainsi que d'autres types MIME (ceux-là officieux).

Ce format est normalisé à partir de 1994 sous l'identifiant ISO/IEC 13818 (CPO-DOC). Il s'inspire largement du premier standard **MPEG-1**, mais fait appel à des outils optimisés et adaptés aux besoins de production, stockage et diffusion de programmes audiovisuels de la SD (720 x 576) à la HD (1920 x 1080).

La quasi-totalité des brevets portant sur le format MPEG-2 a expiré, en 2018 pour les derniers d'entre eux en vigueur aux États-Unis (CPO-LIB).

Le format du flux vidéo est parfois également nommé « H.262 », identifiant de la norme strictement équivalente portée par l'UIT.

Sociabilité

MPEG-2 a largement été utilisé comme format de signaux de télévision numérique transmis par ondes hertziennes, câble et satellite. Il spécifie également le format des films et programmes distribués sur DVD et d'autres disques optiques.

À titre d'exemple, la *Library of Congress* a utilisé ce format pour la numérisation de ses cassettes VHS.

Relations à d'autres formats

Le codec vidéo MPEG-2/H.262 est très similaire à celui défini dans la norme MPEG-1 partie 2. Il lui ajoute néanmoins la gestion du mode entrelacé. Il est suivi par le codec H.264 qui est adapté à un plus grand nombre d'usages, dont la transmission réseau de vidéo HD. On notera que le codec H.263 ne répond pas aux mêmes besoins que les précédents car il a pour usage principal la visioconférence.

Versions

La norme MPEG-2 définit plusieurs formats **conteneurs** selon l'usage prévu du contenu vidéo. On peut citer en particulier les conteneurs suivants.

- MPEG *Program Stream* (« MPEG-PS », « PS » ou « *Program Stream* ») pour des supports stables comme le DVD. Ce conteneur est défini par la norme MPEG-1 partie 1 et repris par la norme MPEG-2 partie 1 et son équivalent UIT H.222.0.
- MPEG *Transport Stream* (« MPEG-TS », « MTS » ou « *Transport Stream* », extensions `.ts` ou `.m2ts`) pour des usages de diffusion *broadcast* (télévision par câble ou satellite, notamment) où la transmission peut être instable. Ce conteneur est défini par la norme ISO/IEC 13818-1 (*Systems*) et son équivalent UIT H.222.0.

La norme MPEG-2 dans sa partie 2 décrit le **codec vidéo**, équivalent à l'UIT H.262, et définit six profils et quatre niveaux. Tous les profils ne sont pas utilisés à égalité ; les profils suivants semblent les plus répandus (classés par niveau de qualité croissant).

- *Simple Profile*, utilisé comme format de diffusion avec un sous-échantillonnage de la chrominance en 4:2:0. Il n'utilise pas d'image interpolée (B).
- *Main Profile* ou « MP », profil de base utilisé comme format de diffusion SD et HD ou pour la production et la transmission *broadcast*, avec un sous-échantillonnage de la chrominance en 4:2:0 avec la possibilité d'utiliser des images bidirectionnelles. Il est utilisé par le DVD et le Blu-ray, et les bouquets de programmes numériques.
- *4:2:2 Profile* ou « 422 », destiné à la production, la post-production et la transmission *broadcast*. Il permet d'encoder du 4:2:2 à haut débit et un GOP court ou un codage intra-image.

Les niveaux déterminent la définition maximale de l'image :

- *Low Level* ou « LL » : n'accepte que les images à basse définition (format SIF MPEG-1) ;
- *Main Level* ou « ML » : niveau de base de la télévision standard de 720 x 576 ;
- *High 1440* ou « H14L » : niveau correspondant à une image haute définition à 1920 x 1440 points par ligne ;
- *High Level* ou « HL » : niveau correspondant à une image HD véritable, avec une pleine définition de 1920 x 1080.

Seules douze combinaisons profil/niveau sont autorisées par la norme MPEG-2

Contenu

Cette fiche décrit les formats conteneurs MPEG-2 avec un flux vidéo au format MPEG-2 part 2 (H.262) et un flux audio au format MPEG-1/2 layer II (parfois appelé « MP2 »).

Pour mémoire, on signalera que le conteneur MPEG-PS peut également contenir :

- un flux vidéo au format MPEG-1 vidéo ou, plus rarement, au format MPEG-4 partie 2 ;
- un flux audio aux formats MPEG-1/2 layer III (MP3) ou, plus rarement, AAC. La gestion du multi-canal dépend du codec utilisé.

Les conteneurs MPEG-PS et MPEG-TS peuvent embarquer des données dans tout autre format sous la forme de *private streams* à condition d'enregistrer auprès de la *Society of Motion Picture and Television Engineers (SMPTE)* un identifiant de format (`format_identifier`). Cette méthode a notamment été utilisée pour intégrer dans des conteneurs MPEG des sous-titres ou du son au format AC-3 (Dolby Digital).

Organisme de maintenance et documentation de référence

MPEG-2 est normalisé par la norme ISO/IEC 13818 : « Codage générique des images animées et du son associé », qui comporte, en avril 2021, onze parties. Celles concernées par cette fiche sont les parties 1 à 3. Les aspects Systèmes (synchronisation, transport, stockage) décrivent les formats conteneurs et sont définis dans la partie 1 : ISO/CEI 13818-1, correspondant strictement à la norme UIT H.222.0. Le format des flux audio et vidéo, en particulier la méthode de compression, est défini dans les parties 2, ISO/CEI 13818-2 (vidéo) et 3, ISO/CEI 13818-3 (son), correspondant à la norme UIT H.262.

Les parties 1 et 2 de MPEG-2 ont été développées en collaboration avec l'UIT-T, qui en diffuse une version gratuite :

- H.222.0, équivalente à MPEG-2 partie 1 ;
- H.262, équivalente à MPEG-2 partie 2.

Identifiants

Registre	Identifiant
Wikidata	Q28018471
Bibliothèque du Congrès	fdd000335 pour la famille de formats conteneurs fdd000028 pour le codec vidéo (H.262) fdd000033 pour le profil *Simple* du codec vidéo fdd000032 pour le profil *Main* du codec vidéo fdd000034 pour le profil *4:2:2* du codec vidéo
PRONOM	x-fmt/386
Just Solve the File Format Problem	MPEG-2

Caractéristiques techniques

Mesures techniques de protection (CPO-PRO) : MPEG-2 partage avec MPEG-4 les mêmes mécanismes de protection, décrits dans le document Intellectual Property Management and Protection in MPEG Standards.

Métadonnées internes

MPEG-2 spécifie des métadonnées techniques dans les en-têtes des *macroblocks*, des images, des séquences, etc. nécessaires à la lecture du flux vidéo.

MPEG-2 n'a en revanche pas spécifié de formalisme pour des métadonnées descriptives, ce qui a conduit le groupe MPEG à normaliser MPEG-7 pour combler cette lacune.

Outils connus par la BnF

Un outil de manipulation et de transcodage est `ffmpeg`.

L'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7.

Usage ou présence dans les collections de la BnF

Avant de lui préférer le format MP4 H.264, la BnF a utilisé depuis 1999 le format MPEG-2 H.262 selon les modalités suivantes.

Avec un codec H.262 Main Profile @ Main Level (abrégé en MP@ML)

- Pour un signal SD (720×576), avec un sous-échantillonnage de la chrominance à 4:2:0 et des débits de 6, 12, 15 Mb/s, avec un GOP de 12 et en CBR. Avec un codec H.262 Main Profile @ High Level (abrégé en MP@HL)
- Pour un signal SD (720×576), avec un sous-échantillonnage de la chrominance à 4:2:0 et des débits de 25 Mb/s, avec un GOP de 1 et en CBR.
- Pour un signal HD (1440×1080), avec un sous-échantillonnage de la chrominance à 4:2:0 et un débit de 25 Mb/s, avec un GOP de 12 et en CBR.

Avec un codec H.262 4:2:2 Profile @ Main Level (abrégé en 422P@ML)

- Pour un signal SD (720×576), avec un sous-échantillonnage de la chrominance à 4:2:2 et des débits de 25 Mb/s, avec un GOP de 1 et en CBR. Les signaux audio sont encodés en MPEG-1 Layer II en 384 kb/s pour les encodages Main Profile et en L-PCM pour le 4:2:2 Profile.

Dans les collections de la BnF sont présents, au titre du Dépôt légal ou de divers dons et dépôts, des documents dans un format conteneur MPEG-2 et encodés avec le codec H.262.

4.9. MP4 (conteneur)

Description

Le format MP4, ou « MPEG-4 part 14 », est un format conteneur pour des contenus multimédia.

Le format de fichier MP4 connu sous le nom de « version 1 » a été publié en 2001 sous la référence ISO/IEC 14496-1:2001, en tant que révision du MPEG-4 Partie 1 : *Systems*. En 2003, la première version du format de fichier MP4 a été révisée et remplacée par le format de fichier MPEG-4 Partie 14 : Format de fichier MP4 (ISO/IEC 14496-14:2003), communément appelé format de fichier MPEG-4 « version 2 ». Le format de fichier MP4 a été généralisé dans le format *ISO Base Media File* (ISO/IEC 14496-12:2004 ou ISO/IEC 15444-12:2004), qui définit une structure générale pour les fichiers médias.

Le conteneur MP4 principalement spécifié par la norme de fichier média ISO/IEC 14496-12 est directement basé sur le format de fichier MOV. La structure de MPEG-4 Partie 14 est très proche de celle du format de fichier MOV, mais impose en outre la prise en charge des descripteurs d'objets initiaux (IOD) et d'autres caractéristiques MPEG.

Ce format permet d'embarquer plusieurs pistes (*tracks*) qui peuvent contenir des données de plusieurs types : audio, vidéo, images texte (en particulier, pour les sous-titres). Comme la plupart des formats de conteneurs modernes, il permet le *streaming* en ligne.

L'extension de fichier officielle est **.mp4**. D'autres extensions de fichiers sont également utilisées, telles que .m4v, .m4p, .m4b, .m4e, .m4r ou .m4a, pour signifier qu'un seul type de données est embarqué.

Sociabilité

Ce format est très largement utilisé pour diffuser de la vidéo sur le web. Normalisé, il concilie facilité de production, compression efficace et diffusion aisée. Il fait partie des formats reconnus par les navigateurs dans la balise `<video>`.

Enfin, son caractère multi-plateforme lui assure un support généralisé tant sur des ordinateurs que dans des appareils tels que les téléphones portables, les lecteurs vidéo ou les consoles de jeux.

Relations à d'autres formats

Directement inspiré du format MOV d'Apple, qui a permis l'élaboration de la norme *ISO Base Media File Format* (ISO MPEG-4 Part 12), le MP4 en est une extension, normalisé sous la dénomination MPEG-4 Part 14.

En tant que format conteneur pour les contenus multimédia, il peut embarquer un grand nombre de codecs audio et vidéo mais aussi textuels (pour les sous-titres). On peut citer notamment :

- vidéo : H.265, H.264 ou MPEG-4 Part 2
- son: : AAC, MPEG-4 Part 3 ou MP3
- sous-titres : MPEG-4 Timed Text

Versions

Formellement, il existe deux versions du format.

- Le format de fichier MP4 connu sous le nom de « version 1 » a été publié en 2001 sous la référence ISO/IEC 14496-1:2001, en tant que révision du MPEG-4 Partie 1 : *Systems*..
- En 2003, la première version du format de fichier MP4 a été révisée et remplacée par le format de fichier MPEG-4 Partie 14 : Format de fichier MP4 (ISO/IEC 14496-14:2003), communément appelé format de fichier MPEG-4 « version 2 ». C'est cette version qui est la plus utilisée et répandue.

Contenu

Les données vidéo, audio, image ou texte sont embarquées dans des « boîtes » (*boxes*, analogues aux « atomes » du MOV). Chacune d'entre elles est identifiée par un type sur quatre octets (parfois désigné sous le vocable de *fourcc*) : par exemple `avc1` pour du H.264 ou `mp4a` pour la partie son avec `esds` indiquant le codec à utiliser.

Ces boîtes sont organisées sous forme d'arborescence. Le format MP4 est conçu pour séparer les métadonnées, qui sont toutes décrites dans une arborescence qui commence à la boite racine typée `moov`, et les données qui sont incluses dans une ou des boîtes typées `mdat`, potentiellement dans des fichiers séparés.

Les différents types de boîte doivent être référencés dans le registre officiel des formats de fichier basés sur *ISO Base Media File Format*, MP4RA.

Organisme de maintenance et documentation de référence

Développé par le Motion Picture Experts Group (MPEG) en tant que groupe de travail ISO/IEC JTC-1 (JTC1/SC29/WG11), le standard est porté par l'ISO.

1e date	Date actuelle	Version	Norme	Description
2001	2010	MP4 file format version 1	ISO/IEC 14496-1:2010	MPEG-4 Part 1 (Systems), First edition
2003	2020	MP4 file format version 2	ISO/IEC 14496-14:2020	MPEG-4 Part 14 (MP4 file format), Second edition

Un résumé conséquent de chaque partie de la norme MPEG-4 est disponible sur l'ancien site du Moving Picture Experts Group.

Identifiants

Registre	Identifiant
Wikidata	Q336316
Bibliothèque du Congrès	fdd000037 pour le MPEG-4 « version 1 » fdd000155 pour le MPEG-4 « version 2 » fdd000137 pour le MPEG-4 Advanced Video Coding (Non-FRExt Extensions) Part 15 fdd000156 pour le MPEG-4 « textual format » (XMT) Part 11
PRONOM	fmt/199
File Format	MP4

Caractéristiques techniques

Le format MP4 est un format binaire, ce qui optimise sa compacité (CPO-COM).

Il est aisément extensible par l'utilisation de boîtes de type nouveau ou de type `uuid` réservés. Ces boîtes pourront être ignorées par les lecteurs ne les prenant pas en charge sans empêcher la lecture du fichier.

Le concept de boîte étant très générique, il est possible de gérer un très grand nombre de types de contenu et d'usage. Par exemple, l'ajout de boîtes de type `hint` autorise un usage en flux (*streaming*) des données.

Des **mesures techniques de protection** (CPO-PRO) peuvent être incluses en chiffrant les flux de données. Ce chiffrement est indiqué en utilisant des types de boîtes différentes (par exemple, `encv` pour les données video ou `enca` pour l'audio) et en spécifiant les méthodes en utilisant le standard IPMP (*Intellectual Property Management and Protection*). La résilience à l'erreur est facilitée par la séparation entre métadonnées et données. Elle est néanmoins très dépendante du protocole de transport ainsi que des codecs sous-jacents (CPO-ROB).

Afin de faciliter l'édition, le format permet de référencer les composants média sans les embarquer dans le ficher. Les fichiers MP4 qui utilisent cette fonctionnalité ne sont donc pas **auto-suffisants** (CPO-AUT).

Métadonnées internes

Le format définit des métadonnées internes, portées par des boîtes particulières : on citera notamment `moov/mvhd`, `moov/udta/cprt`, `moov/trak/tkhd`, `moov/trak/udta/cprt`.

Par ailleurs, il est également possible d'ajouter des métadonnées en XMP, dans une boîte de type `uuid`. Dans ce cas, il est nécessaire de synchroniser ces informations, voir en particulier le Extensible Metadata Platform (XMP) Specification: Part 3, Storage in Files.

Outils connus par la BnF

L'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7.

L'outil `ffmpeg` permet le décodage et la conversion de et vers ce format.

Usage ou présence dans les collections de la BnF

Depuis quelques années la BnF numérise certains documents avec un conteneur MP4 et un codec H.264.

Par ailleurs, le MP4 est utilisé pour la diffusion sur Gallica.

Dans les collections de la BnF sont présents, au titre du Dépôt légal ou de divers dons et dépôts, des documents au format MP4 et actuellement sauvegardées dans le Système Audiovisuel de la BnF.

Des partenaires de la BnF (Centre Pompidou Virtuel ou Centre National du Cinéma) reçoivent ou produisent ce format pour les films numériques.

4.11. Native FLAC

Description

Cette fiche décrit à la fois le format de fichier « Native FLAC » (*Free Lossless Audio Codec*) et le codec audio FLAC permettant une compression sans perte d'un flux audio LPCM (*linear pulse-code modulation*). Ce format est ouvert, documenté (CPO-DOC) et libre de droits (CPO-LIB).

Il s'agit d'un format adapté à une diffusion de type « streaming » qui est donc résilient aux altérations grâce à une structure en blocs (CPO-ROB). Chacun des blocs dispose d'une empreinte numérique embarquée de type MD5.

Sociabilité

Le format FLAC est très largement géré par les logiciels et matériels audiovisuels. Son adoption s'est développée à partir de 2017, date après les nouvelles versions de la plupart des navigateurs internet le gèrent.

Contenu

Il intègre le plus souvent des données audio LPCM avec une profondeur d'échantillonnage de 4 à 32 bits et un nombre de canaux audio de 1 à 8. Les canaux peuvent être groupés comme dans le cas de la stéréo et des canaux 5.1 *surround* pour tirer parti de la corrélation intercanal pour augmenter la compression. Les profondeurs d'échantillonnage couramment utilisées sont 16 et 24 bits, à des fréquences d'échantillonnage de 44,1 kHz à 192 kHz et en stéréo.

Il peut également intégrer des flux audio d'autres codecs comme le MQA, mais la BnF recommande que le flux audio soit encodé selon le codec FLAC.

Le format FLAC permet d'embarquer un flux image fixe pour intégrer un visuel.

Organisme de maintenance et documentation de référence

Le format FLAC est maintenu par la fondation à but non lucratif Xiph.org. Ses spécifications sont accessibles sur le site web officiel du format. En outre, le projet CELLAR vise à normaliser un format audiovisuel utilisant Matroska comme format conteneur, FFV1 comme codec vidéo et FLAC comme codec audio.

La *British Library* a réalisé une évaluation de ce format pour la préservation à long terme.

Identifiants

Registre	Identifiant
Wikidata	Q27881556 pour le format de fichier Q219848 pour le codec
Bibliothèque du Congrès	fdd000198 (version 1.1.2)
PRONOM	fmt/279 (version 1.2.1)

Caractéristiques techniques

La compression mise en œuvre par le codec FLAC permet de réduire de 30 à 70% la taille d'un flux LPCM. La réduction de taille est dépendante de la source : plus le signal est constitué d'ondes régulières (sons naturels), meilleure est la compression. Il s'agit d'une compression sans perte : si l'on encode un fichier WAVE en FLAC puis qu'on le décode à nouveau vers WAVE, le flux audio décodé est strictement équivalent (a la même empreinte numérique) que le fichier WAVE originel.

Note : afin de conserver les métadonnées RIFF d'origine dans le fichier FLAC, on veillera à employer l'option `--keep-foreign-metadata` *à l'encodage et au décodage.*

Outils connus par la BnF

La fondation Xiph.org développe également l'outil en ligne de commande `flac` qui permet d'encoder et de décoder des flux FLAC à partir de WAVE, WAVE 64, RF64 et AIFF et de valider la structure du flux FLAC. La BnF recommande l'utilisation de cet outil pour l'encodage et le décodage de fichiers FLAC, dans sa version 1.2.1

(version à partir de laquelle l'outil gère l'option `--keep-foreign-metadata` qui permet de conserver les métadonnées du fichier d'origine) ou postérieure.

L'outil d'analyse et d'extraction de métadonnées techniques est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7.

Métadonnées internes

Le format FLAC permet d'embarquer des métadonnées internes aux formats Vorbis et ID3. L'outil `metaflac` permet d'éditer ces métadonnées.

Usage ou présence dans les collections de la BnF

Pour sa **sociabilité** (CPO-SOC), son caractère ouvert (CPO-DOC, CPO-LIB) et sa **compacité (CPO-COM)**, liée à sa méthode de compression sans perte, FLAC a été adopté comme le format préféré de dépôt légal du son dématérialisé.

4.12. OpenDocument Text (ODT)

Description

OpenDocument (ODF) est un format de données pour les applications bureautiques : traitement de texte, tableur, logiciel de présentation, de diagramme, de dessin et base de données bureautique. OpenDocument Text en est la variante dédiée au texte formaté. Il s'agit d'un format ouvert (CPO-DOC, CPO-LIB) : les entreprises Sun Microsystems et IBM, qui ont contribué au format, se sont engagées à ne pas restreindre par brevet l'utilisation du format.

Sociabilité

Ce format est reconnu par la plupart des logiciels libres d'édition bureautique (OpenOffice, LibreOffice, AbiWord, KWord, …). Certains logiciels propriétaires assurent une compatibilité partielle (Microsoft Office 2007 et versions suivantes, Lotus Symphony, Google Docs, …).

En France, le format OpenDocument est le seul format recommandé comme format bureautique par le référentiel général d'interopérabilité depuis sa version 2.0 validée le 2 décembre 2015.

Un très grand nombre d'institutions de conservation (*Library and Archives Canada*, *National Archives and Records Administration* aux États-Unis, etc.) le mentionnent comme format préféré pour le contenu textuel.

Relations à d'autres formats

Attention de ne pas confondre les formats OpenDocument avec le format bureautique de Microsoft : Office Open XML, tous les deux étant des normes ISO.

Le format OpenDocument s'inspire largement du format créé pour le logiciel OpenOffice.org développé à l'origine par Sun. Il se base essentiellement sur XML pour le contenu, ce qui le rend relativement lisible par l'humain, et ZIP pour le conteneur. Il est possible d'embarquer des images dans le document. Dans ce cas la norme recommande (mais n'oblige pas) d'utiliser les formats PNG pour les images matricielles et SVG pour les images vectorielles.

Organisme de maintenance, documentation de référence et versions

Il s'agit d'un format ouvert et normalisé par OASIS (Organization for the Advancement of Structured Information Standards) depuis mai 2005 et par l'ISO depuis 2006 sous l'identifiant ISO/IEC 26300.

Version	Date OASIS	Date ISO
1.0	01/05/2005	19/07/2006
1.1	02/02/2007	11/07/2013
1.2	29/09/2011	19/06/2015
1.3	en cours	N/A

La *British Library* a réalisé une évaluation de ce format pour la préservation à long terme.

Contenu

Un document au format OpenDocument peut exister sous deux formes :

- généralement, un conteneur au format ZIP embarquant plusieurs fichiers XML et éventuellement du contenu binaire tel que des images ;
- plus rarement, un simple fichier XML dit « à plat », signalé par l'extension `.fodt`. Cette forme n'est pas toujours reconnue par les logiciels d'édition (tels que Microsoft Office 2007 à 2013).

Un fichier ODT présente obligatoirement un fichier XML : `META-INF/manifest.xml` qui contient la liste de fichiers contenus dans le conteneur, avec leur type MIME et éventuellement l'information nécessaire à leur décompression ou leur déchiffrement.

Un fichier ODT comporte optionnellement (mais sa présence est fortement conseillée) un fichier `mimetype` qui contient le type MIME du document (pour ODT : `application/vnd.oasis.opendocument.text`).

Le contenu est réparti dans les fichiers :

- content.xml : le contenu à proprement parler du document ;
- meta.xml : les métadonnées associées au document ;
- settings.xml : des paramètres destinés à l'application pour éditer le document ;
- style.xml : la partie « présentation » du document.

On peut trouver également :

- Thumbnails/thumbnail.png : une imagette enregistrée en PNG ;
- éventuellement des images dans Pictures/ : ce répertoire ne fait pas partie de la norme mais c'est une pratique généralisée ; les images contenues dans ce répertoire devraient être au format PNG ou SVG bien qu'il soit possible d'utiliser tout autre format.

Identifiants

Registre	Identifiant
Wikidata	Q27203100 (v. 1.0) Q27203404 (v. 1.1) Q27203601 (v. 1.2)
Bibliothèque du Congrès	fdd000427 (v. 1.1) fdd000428 (v. 1.2)
PRONOM	fmt/136 (v. 1.0) fmt/290 (v. 1.1) fmt/291 (v. 1.2)
Just Solve the File Format Problem	OpenDocument_Text

Caractéristiques techniques

Le format ODF gère les marques de révision et la signature numérique à partir de sa version ODF 1.2.

Basé sur le conteneur ZIP dans sa version normalisée à l'ISO, le format ODT bénéficie des mécanismes de gestion d'intégrité de ZIP. Sa **robustesse** (CPO-ROB) est donc considérée comme satisfaisante. Conformément à la norme ISO ZIP (ISO/IEC 21320-1), seule la méthode de compression sans perte *Deflate* est autorisée.

Dépendances (CPO-AUT) :

- Un document ODT peut inclure les polices utilisées dans le texte afin de le rendre autonome, mais tous les outils, par exemple Apache OpenOffice, n'autorisent pas cette opération.
- Les fichiers complémentaires (de type images ou autres) peuvent être inclus dans le conteneur ZIP ou simplement référencés. Selon la version de l'outil utilisé, un copier-coller d'une image dans un document ODT peut ne résulter qu'en une référence à l'URI de l'image, et non à l'inclusion de l'image elle-même.

Mécanismes de protection (CPO-PRO) : les fichiers ODF peuvent être aisément chiffrés en sélectionnant une option d'enregistrement dans OpenOffice.org et LibreOffice. Les fichiers ainsi chiffrés ne sont plus lisibles par les outils d'analyse cités plus bas qui renverront donc une erreur plus ou moins explicite. A partir de la version 1.2 d'ODF, plusieurs méthodes de chiffrement sont autorisées ; le fichier META-INF/manifest.xml doit indiquer laquelle a été utilisée.

Métadonnées internes

Le format ODF permet d'embarquer des métadonnées internes dans le fichier meta.xml. Le formalisme adopte Dublin Core (préfixe « dc: ») et l'étend avec un jeu de métadonnées spécifique (préfixe « meta: »). Le format accepte également des métadonnées définies par l'utilisateur, qui suivent le formalisme RDF à partir de la version 1.2 d'ODF.

Outils connus par la BnF

- De multiples outils d'**édition** existent, mais le formatage peut différer de l'un à l'autre (ainsi, Microsoft signale que « Le formatage peut être perdu lorsque les utilisateurs enregistrent et ouvrent des fichiers .odt »).
- La migration de ou vers ODT est une opération délicate ; il arrive que certaines fonctionnalités soient perdues ou mal rendues (les marques de révision notamment sont souvent mal traitées).
- L'outil ODF Validator, développé par The Document Foundation, **valide** les fichiers ODF et dispose d'une version en ligne. Il en existe d'autres, mais aucun ne propose une couverture complète du format.
- Tika permet une extraction des métadonnées internes au format XMP et du contenu textuel.

Usage ou présence dans les collections de la BnF

À la date du 16 mars 2021, le format ODT est peu représenté dans les collections de la BnF. Il est néanmoins susceptible de les intégrer par le biais de dons numériques natifs.

4.13. PDF/A

Description

PDF/A est une série de déclinaisons standardisées du format de publication PDF destiné à préserver la mise en page du document. Cette série de formats est utilisée pour conserver et échanger des documents numériques car les restrictions qu'elle impose au format PDF visent à le rendre autonome (CPO-AUT). Son développeur, l'entreprise Adobe, bien que disposant de brevets portant sur le format, accorde une licence gratuite pour la visualisation et l'édition de ces fichiers par des logiciels tiers (CPO-LIB).

PDF/A, dans ses versions 1 et 2 (la version 3, en incluant la possibilité d'embarquer n'importe quel type de contenu, comporte des risques supplémentaires), réduit sensiblement les risques portant sur le contenu des fichiers qui s'y conforment. Néanmoins, dans la mesure où il restreint fortement les contenus et les fonctionnalités des documents bureautiques, une migration vers ce format présente des risques de perte d'information qui ne peuvent être limités que par une maîtrise des formats source et cible, et un contrôle rigoureux des propriétés signifiantes après migration.

Pour une vue d'ensemble du format et de ses usages, on pourra consulter le document : *PDF/A l'essentiel 2.0 : PDF pour l'archivage à long terme*. Une introduction technique en anglais est également disponible : *A Technical Introduction to PDF/A*, PDFlib, 2017.

Sociabilité

La série de formats PDF/A est très largement utilisée et recommandée par de nombreuses organisations officielles et institutions de conservation aux États-Unis et en Europe. Les implémentations par des outils très largement répandus tels Microsoft Office, LibreOffice ou InDesign contribuent à la diffusion du format auprès du grand public.

Relations à d'autres formats

PDF/A est directement basé sur le format PDF. Il hérite de ce fait des complexités et ambiguïtés du format PDF, bien que celles-ci aient été limitées dans la version 2.0 de PDF, dont dérive PDF/A-4.

Bien que non directement lié, le format PDF/UA-1 (pour *Universal Accessibility*), normalisé comme ISO 14289-1:2014, impose et précise l'usage des balises PDF (« tags ») qui encodent la structure sémantique du PDF. De ce fait, la conformité à ce standard est recommandée par Adobe aux créateurs de PDF/A de variante « a ».

Versions

Date	Version PDF/A	Version PDF de base	Norme ISO
2005	PDF/A-1	PDF 1.4	ISO 19005-1
2011	PDF/A-2	PDF 1.7 (ISO 32000-1)	ISO 19005-2
2012	PDF/A-3	PDF 1.7 (ISO 32000-1)	ISO 19005-3
À venir	PDF/A-4	PDF 2.0 (ISO 32000-2)	ISO 19005-4 en cours

Contenu

Les principales restrictions introduites par PDF/A sont les suivantes.

- Il ne peut contenir que du texte ou des images à condition qu'elles ne soient pas compressées avec l'algorithme LZW.
- Les formats audiovisuels ainsi que les interactions (Javascript, ...) sont par conséquent interdits.
- Les polices doivent être incluses dans le fichier (on relève de ce fait des risques potentiels si les polices en question sont protégées par le droit d'auteur).
- L'inclusion des profils colorimétriques des images (généralement sous forme de profil ICC ou de spécification CIE Lab) et la présence de métadonnées au format XMP sont obligatoires.
- Le chiffrement est interdit (CPO-PRO).

Les contenus et fonctionnalités suivants sont introduits dans la version PDF/A-2 (ils étaient interdits dans la version PDF/A-1) :

- format d'image JPEG 2000 (introduit en PDF 1.5),
- gestion de la transparence,
- introduction de spécifications sur la signature électronique (déjà possible pour PDF 1.4 et donc PDF/A-1, mais sans précision sur sa forme) selon le format PAdES,
- possibilité d'inclure des fichiers PDF/A-1 ou PDF/A-2, pour permettre la création d'un unique fichier PDF/A-2 issu de plusieurs d'entre eux.

Apport de la version PDF/A-3 :

- possibilité d'inclure n'importe quel fichier associé au document ou à une de ses pages, à condition de fournir un type MIME et la relation entre le contenu associé et le contenu principal basée sur une liste fermée (« source », « data », « alternative », « supplement », « unspecified »). Les logiciels de lecture de PDF/A ne sont tenus que de permettre l'extraction du contenu attaché. L'objectif initial était de pouvoir ajouter le fichier d'origine, avant migration vers PDF/A, mais, cette liberté ayant été employée à des usages divers, elle est aujourd'hui considérée comme faisant porter un risque de préservation supplémentaire. Elle est donc généralement déconseillée par les spécialistes de la préservation numérique.

Du fait du caractère restrictif du PDF/A, le choix de la version à utiliser est directement lié au contenu que l'on cherche à porter ; les versions successives de PDF/A ne remplacent en aucune façon leur prédécesseur.

Organisme de maintenance et documents de référence

Les formats PDF/A sont maintenus par Adobe et normalisés par l'ISO, voir le paragraphe *Versions*

Requête sur les variantes du format et leur norme ISO associée sur Wikidata.

La *British Library* a réalisé une évaluation de ce format pour la préservation à long terme.

Identifiants

Registre	PDF/A	PDF/A-1	PDF/A-2	PDF/A-3
Bibliothèque du Congrès	fdd000318	fdd000125	fdd000319	fdd000360
Wikidata	Q1547957	Q26543628 (b) Q26541013 (a)	Q26546575 (b) Q26547266 (u) Q26545877 (a)	Q26548590 (b) Q26549229 (u) Q26547917 (a)
PRONOM		fmt/354 (b) fmt/95 (a)	fmt/477 (b) fmt/478 (u) fmt/476 (a)	fmt/480 (b) fmt/481 (u) fmt/479 (a)

Les variantes sont les suivantes :

- « b », pour « basic », correspond aux exigences minimales pour la restitution à long terme de l'apparence visuelle de la mise en page.
- « u », pour « unicode », ajoute à la variante « b » l'exigence de disposer pour tout le texte d'un équivalent en Unicode pour permettre la recherche et l'extraction du texte.
- « a », pour « accessible », ajoute des exigences sémantiques - « tags » - pour documenter, comme dans un format de balisage, la nature des contenus (titres, tableaux, listes, etc.), leur organisation logique au sein du document, décrit comme une structure arborescente, et son ordre de lecture, notamment en vue de favoriser l'accessibilité.

Caractéristiques techniques

Le **poids** (CPO-COM) d'un fichier PDF/A peut être sensiblement plus important que celui d'un PDF en raison des contenus obligatoires (polices, profils colorimétriques) ou des fichiers embarqués.

Certaines **dépendances externes** (CPO-AUT) peuvent malgré tout exister dans un fichier PDF/A : liens hypertextes vers une ressource web, ou encore action `GotoR` (lien vers un emplacement dans un autre fichier PDF).

Métadonnées internes

La présence de métadonnées internes au format XMP, basé sur le formalisme RDF, est obligatoire.

Outre les centaines de propriétés définies par XMP, ce format est extensible à des métadonnées déterminées par l'utilisateur, mais PDF/A impose, dans ce cas, que ces propriétés soient documentées par l'inclusion d'un schéma d'extension (*XMP Extension Schema Description*).

Outils connus par la BnF

Pour la production, les outils bureautiques (en particulier, Microsoft Word à partir de 2012, et LibreOffice Writer à partir de la version 6) sont en mesure de générer un PDF/A, moyennant une configuration lors de l'enregistrement.

L'outil de lecture Adobe Acrobat Reader sait détecter la présence de la déclaration du caractère 'A' du PDF (dans les métadonnées XMP), et le notifie par la présence d'une bannière. Tous les lecteurs de PDF sont en mesure de lire des fichiers PDF/A, mais certains ne traitent pas correctement ses particularités (non exploitation des profils colorimétriques, des polices embarquées ou des fichiers attachés, notamment). Il en est de même pour les outils de migration de ou vers PDF/A qui ne gèrent pas tous correctement les fonctionnalités et contenus du PDF/A (transparence, signature électronique, etc.). Une migration de masse PDF vers PDF/A est donc une opération périlleuse et requiert la plus grande prudence : une vérification après génération est fortement conseillée pour s'assurer qu'aucun élément n'a été supprimé ou mal représenté.

- veraPDF est un outil **spécialement** conçu pour valider le PDF/A.
- Le module PDF de JHOVE permet une caractérisation du format PDF et une détection du profil PDF/A.
- Tika permet une extraction des métadonnées internes au format XMP et du contenu textuel.

4.14. ProRes (codec)

Description

ProRes est un codec propriétaire de compression vidéo avec pertes développé par Apple Inc. pour une utilisation en post-production (non pour la diffusion, donc) qui prend en charge toutes les résolutions, de la SD à la 8K. Il est le successeur de l'Apple Intermediate Codec et a été lancé en 2007 avec le logiciel de montage, Final Cut Studio 2.

Il offre un bon compromis entre qualité et compression.

Sociabilité

Ce format est très répandu pour la post-production de contenus vidéo, notamment pour le cinéma, mais aussi pour les documentaires et autres programmes destinés à la télévision, particulièrement dans sa famille de profils 422. Quoique moins répandus, ses profils 4444 sont notamment utilisés pour la réalisation de publicités.

Relations à d'autres formats

ProRes n'est que le codec de compression vidéo ; il est quasiment toujours embarqué dans un conteneur QuickTime (extension de fichier .mov). Bien qu'il soit théoriquement possible de l'accompagner d'un flux audio compressé, la pratique très majoritaire est de l'encoder en PCM non compressé.

Il a le même usage que le codec Avid DNxHD.

Versions

On distingue plusieurs profils du codec, introduits successivement par Apple. Les différences majeures entre les profils 422 et 4444 résident dans le débit (supérieur pour les profils 4444), la profondeur couleur (10 bits pour les profils 422, jusqu'à 12 bits pour les profils 4444, ce qui donne à ces derniers une plus grande précision dans la restitution des couleurs) et le sous-échantillonnage de la chrominance (4:2:2 pour les profils 422, 4:4:4 pour les profils 4444). Les profils sont les suivants, classés par niveau de qualité croissant :

- 422 Proxy ;
- 422 LT ;
- 422 ;
- 422 HQ ;
- 4444 ;
- 4444 XQ.

Contenu

Le flux vidéo d'un fichier ProRes peut présenter toutes les tailles, de la SD (720x480 ou 720x576) à la 8K.

Organisme de maintenance et documentation de référence

Le format n'est pas documenté à proprement parler, mais Apple a dévoilé une partie de sa structure interne dans un *white paper*.

En outre, le site Multimedia Wiki fournit quelques informations techniques

Identifiants

Registre	Identifiant
Wikidata	Q47246311
Bibliothèque du Congrès	Famille des profils 422 : fdd000389
PRONOM	fmt/797

Caractéristiques techniques

Ce codec fournit une compression avec pertes mais sans pertes visuelles. La compacité du format dépendra essentiellement du profil, du nombre d'images par seconde (la cadence) et de la résolution. Dans le tableau suivant, on présente le poids estimé des fichiers selon ces paramètres. Les valeurs de ces paramètres sont limitées à celles déterminées pour le dépôt d'œuvres cinématographiques. Le poids indiqué est indicatif ; il pourra différer légèrement car le débit du flux vidéo ProRes est variable et en raison des caractéristiques du flux audio.

Profil	**Cadence en images par seconde**	**Résolution 2K cinéma (2048 x 1080)**		**Résolution 4K cinéma (4096 x 2160)**	
		Débit	**Poids estimé pour 1 heure**	**Débit (Mb/s)**	**Poids estimé pour 1 heure**
ProRes 422 LT	24	93 Mb/s	42 Go	350 Mb/s	157 Go
	25	97 Mb/s	44 Go	365 Mb/s	164 Go
ProRes 422 HQ	24	201 Mb/s	91 Go	754 Mb/s	339 Go
	25	210 Mb/s	94 Go	786 Mb/s	354 Go
ProRes 4444	24	302 Mb/s	136 Go	1131 Mb/s	509 Go
	25	315 Mb/s	142 Go	1180 Mb/s	531 Go

Le format ProRes conserve suffisamment d'information pour permettre de réaliser le montage, d'où son poids relativement important (il n'est donc pas adapté à la diffusion).

Métadonnées internes

Le codec ProRes ne porte pas de métadonnées descriptives ; celles-ci, lorsqu'elles existent, sont codées dans le format conteneur associé.

Les métadonnées techniques internes ne sont pas connues du fait du caractère propriétaire du codec.

Outils connus par la BnF

Depuis 2008, le logiciel Quicktime dispose d'un décodeur gratuit permettant de lire les fichiers ProRes sur un système d'exploitation Windows et Mac.

L'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7. Au 1er février 2021, la sortie MPEG-7 de l'outil `MediaInfo` ne contient pas toutes les métadonnées de caractérisation souhaitées par la BnF.

L'outil `ffmpeg` permet le décodage et la migration de ce format vers des formats de diffusion, comme par exemple le MP4 H.264.

Usage ou présence dans les collections de la BnF

Quelques exemples existent dans le cadre de dons numériques, actuellement dans le Système Audiovisuel de la BnF.

Des partenaires de la BnF (Centre Pompidou Virtuel, Centre National du Cinéma) reçoivent ou produisent ce format pour conserver des contenus audiovisuels nativement numériques.

4.15. Quicktime File Format (MOV) (conteneur)

Description

Le format MOV ou *QuickTime File Format* est un format conteneur pour des contenus multimédia. Créé en 1989 par Apple, il accompagnait le logiciel QuickTime de création et surtout de diffusion de vidéo. Il s'agit d'un format propriétaire maintenu par Apple (CPO-LIB), qui publie néanmoins ses spécifications (CPO-DOC).

Apple a utilisé l'intitulé « Quicktime Movie » pour désigner l'usage de ce conteneur pour embarquer des images animées, d'où l'extension majoritaire `.mov`, bien que l'on trouve également des fichiers qui portent l'extension `.qt`.

Ce format permet d'embarquer plusieurs pistes (*tracks*) qui peuvent contenir des données de plusieurs types : audio, vidéo, texte (en particulier, pour les sous-titres), timecode.

Sociabilité

Ce format reste répandu dans la communauté des utilisateurs d'ordinateurs Macintosh (qui comprend en particulier les producteurs de vidéo) mais tend, en particulier pour la diffusion, à être remplacé par le MP4, dont il est à la base et qui dispose d'une norme. Il est de ce fait moins universellement géré que MP4, en particulier par les matériels : la Playstation Portable (PSP) et plusieurs lecteurs de DVD peuvent lire du MP4 mais non du MOV.

Relations à d'autres formats

Le format MOV est la base du format conteneur MP4 version 1, défini dans la partie 1 de la norme MPEG-4 (ISO/IEC 14496-1:2001). La révision de cette partie de la norme donne naissance à la version 2 de MP4, publiée en 2003 dans la partie 14 de la norme (ISO/IEC 14496-14:2003). Par ailleurs, cette dernière est généralisée en 2004 pour créer la partie 12 (ISO/IEC 14496-12:2004, définissant l'*ISO Base Media File Format*, qui sert de base à d'autres codecs vidéo (MJPEG 2000, 3GP, etc.). La section '*History*' de la page Wikipédia anglophone sur MPEG-4 part 12 donne plus de détails à ce sujet.

La structure des conteneurs MOV et MP4 est donc très similaire, mais certaines divergences font que ces deux formats ne sont pas interchangeables.

Versions

Les spécifications de deux versions du format ont été rendues publiques par Apple :

- la version classique qui date de 2001 ;
- une version plus moderne mise à jour en 2012.

Contenu

Un fichier au format MOV est constitué d'un arbre d'atomes qui décrivent les différentes parties du fichier. Certains atomes embarquent du contenu tandis que d'autres servent à héberger des métadonnées techniques ou descriptives. Chaque atome est identifié par un type sur 4 octets (par exemple `avc1` pour du H.264 ou `mjpg` pour du Motion-JPEG). Selon la nature de son contenu, on parlera de « QuickTime Video » (ou « Quicktime movies »), de « QuickTime Audio » ou même de « QuickTime Image ».

Par exemple, un fichier vidéo au format MOV a un atome principal typé `moov` qui fait référence à des atomes techniques (par exemple, un profile `prfl`) et des atomes de pistes (`trak`) qui portent les données. Les atomes de pistes peuvent contenir directement les données ou une simple référence à ces données localisées ailleurs.

Parmi le grand nombre de codecs audio et vidéo qu'il peut embarquer, l'usage majoritaire en 2021 est l'utilisation du Apple ProRes.

Le format MOV gère les *timecodes* dans un atome spécifique typé `tmcd`.

Usages obsolètes

Le format conteneur MOV a été employé dans des usages divers, aujourd'hui (mars 2021) abandonnés :

- pour un flux audio seul au format AAC, protégé par DRM, par le logiciel Apple iTunes avec l'extension .m4p,
- pour des images fixes sous le nom QTIF et les extensions `.qti`, voire `.qtif` ou `.qif` (les produits Apple en ont désormais abandonné le support),
- à partir de 1995 pour des images virtuelles à 360°, sous le nom « QTVR » (*Quicktime Virtual Reality*).

Organisme de maintenance et documentation de référence

Ce format est propriétaire, mais Apple a publié ses spécifications sur son site developper.apple.com (voir la section Versions de cette fiche).

Identifiants

Registre	Identifiant
Wikidata	Q942350
Bibliothèque du Congrès	fdd000052
PRONOM	x-fmt/384
Just Solve the File Format Problem	QuickTime

Caractéristiques techniques

Le conteneur et sa structure sont relativement **transparents** (CPO-TRA), la complexité des fichiers étant plutôt liée à la diversité des formats de ses composants audio et vidéo (CPO-SIM).

Afin de faciliter l'édition, le format permet de référencer les composants média sans les embarquer dans le fichier. Les fichiers MOV qui utilisent cette fonctionnalité ne sont donc pas **auto-suffisants** (CPO-AUT).

La *Library of Congress* rapporte dans sa description du format le fait que des fichiers Quicktime distribués par Apple iTunes comportent des **mesures techniques de protection** (CPO-PRO) imposant l'entrée de codes pour limiter leur lecture sur trois appareils différents au maximum.

Métadonnées internes

Les métadonnées sont portées par des atomes particuliers, typés `meta`.

La liste des clés utilisées par le logiciel QuickTime d'Apple est définie ici. Toutes ces clés commencent par le préfixe `com.apple.quicktime`.

L'annexe D de la spécification indique comment sont alignées les métadonnées d'autres formats sur celles du format MOV.

Outils connus par la BnF

L'environnement QuickTime d'Apple produit et lit ces fichiers. Disponible lors de sa création uniquement sur MacOs, il fonctionne aussi sous Windows depuis 1992. Néanmoins, Apple ne maintient plus la version Windows de Quicktime depuis 2016 en raison de failles de sécurité.

La lecture est possible avec d'autres outils (VLC pour la vidéo ou XnView pour l'image) mais reste toujours partielle.

Pour l'usage vidéo, l'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7. Au 1er février 2021, la sortie MPEG-7 de l'outil `MediaInfo` ne contient pas toutes les métadonnées de caractérisation souhaitées par la BnF.

De même, l'outil `ffmpeg` permet le décodage et la migration de ce format vers des formats de diffusion, comme par exemple le MP4 H.264.

Usage ou présence dans les collections de la BnF

Quelques exemples existent dans le cadre de dons numériques, actuellement dans le Système Audiovisuel de la BnF.

Des partenaires de la BnF (Centre Pompidou ou Centre National du Cinéma) reçoivent ou produisent ce format pour les films numériques.

4.16. TIFF

Description

Le format TIFF (*Tag(ged) Image File Format*) est un format conteneur ouvert et bien documenté (CPO-DOC) pour une image matricielle (ou plusieurs, dans sa version multi-pages, voir la section « Caractéristiques techniques » de cette fiche, ci-dessous) sous une forme compressée ou non.

Développé par Aldus et Microsoft, il est désormais maintenu par Adobe. Bien qu'il ne s'agisse pas d'un organisme de normalisation, la liberté d'utilisation du format TIFF n'est plus limitée ; on le considère donc comme un format ouvert (CPO-LIB). Il s'agit d'un format adapté à la conservation mais non, du fait de son poids conséquent, à l'échange ni à la diffusion.

Sociabilité

Ce format bien éprouvé est un standard *de facto* pour la conservation de masters images, à la fois par les professionnels de l'image et les institutions mémorielles pour leurs campagnes de numérisation. Son niveau de support logiciel et matériel est très haut. Puisqu'il n'est pas un format de diffusion, sa prise en charge par les navigateurs est limitée.

Relations à d'autres formats

Plusieurs variantes (extensions ou restrictions) du format TIFF ont été définies pour pallier des limitations du format (la taille maximale des fichiers est de 4 Go, et ces derniers ne peuvent embarquer nativement des coordonnées géographiques). On citera notamment les variantes ouvertes suivantes (il existe également des variantes propriétaires) :

- GeoTIFF pour les images géoréférencées. ;
- BigTIFF pour les images de plus de 4 Go ;
- TIFF/EP (*TIFF/Electronic Photography*), format brut pour les photographies numériques..

Versions

La version actuelle de TIFF est la version 6.0, publiée en 1992, qui fait depuis lors office de version de référence, bien que l'on puisse trouver des fichiers dans les versions précédentes. Le format n'a pas connu de révision majeure depuis cette date et est donc particulièrement stable (CPO-STA). Les évolutions depuis cette date ont plutôt développé des extensions du format pour des besoins particuliers, parmi lesquels on compte les variantes citées ci-dessus.

Contenu

Outre le flux image et d'éventuelles extensions intégrant des métadonnées internes, le format TIFF est extensible et peut accueillir des extensions pouvant accueillir des données de toute nature. Bien que la spécification précise que les outils sont tenus d'ignorer les extensions non gérées, il arrive que des extensions propriétaires perturbent le traitement correct de fichiers TIFF.

Organisme de maintenance et documentation de référence

La spécification du format dans sa version 6.0 du 3 juin 1992 est disponible sur le site d'Adobe.

De très nombreuses ressources existent également, que la *Library of Congress* a listées sur sa description du format.

La British Library a réalisé une évaluation de ce format pour la préservation à long terme.

Les exigences BnF sur les images TIFF livrées à la BnF dans le cadre de ses marchés de numérisation sont décrites dans la section 4 du référentiel de numérisation des documents opaques.

Identifiants

Registre	Identifiant
Wikidata	Q215106
Bibliothèque du Congrès	fdd000022

PRONOM	fmt/353
Just Solve the File Format Problem	TIFF

Caractéristiques techniques

La spécification TIFF définit un ensemble de fonctionnalités minimales que tout lecteur doit gérer (« baseline TIFF ») ainsi que des extensions permettant de répondre à des besoins spécifiques (modèles couleurs autres que RGB, méthodes de compression, etc.). Le support logiciel de certaines extensions est inégal, en particulier

- **diverses méthodes de compression** (CPO-COM) peuvent être utilisés pour le flux image contenu dans un conteneur TIFF. Les plus communes sont LZW, CCITT Group 3 et CCITT Group 4. Une liste exhaustive des méthodes de compression plus rarement employées est proposée sur la page Wikipédia en anglais. Les méthodes moins courantes peuvent occasionner des problèmes de restitution sur certains logiciels. La BnF recommande que le flux image contenu dans le fichier TIFF **ne soit pas compressé**. Un fichier TIFF non compressé est plus robuste : une altération ne rendra illisible que la zone concernée alors qu'un TIFF compressé altéré dans la même mesure sera vraisemblablement illisible (CPO-ROB).

Note : pour les images bitonales (noir et blanc), la BnF a utilisé dans les décennies passées la compression CCITT groupe 4, très efficace. La numérisation en noir et blanc n'étant plus d'actualité, on ne devrait rencontrer cette compression que dans des données anciennes.

- **la variante multi-pages** permettant d'embarquer plusieurs images (nommées « IFD » pour *Image File Directory* par le standard) dans le même fichier. Les outils logiciels prenant en charge uniquement la version *baseline* ne sont pas tenus de traiter les images suivant la première ; certains rencontrent des difficultés pour afficher les fichiers multi-pages. Ces données images successives peuvent avoir des contenus différents ou être des représentations alternatives d'un même contenu : ainsi la sortie TIFF de certains matériels de numérisation comprend-elle un second IFD qui était une version en basse définition du premier, destiné à être utilisé comme vignette.

La complexité (CPO-SIM) et la transparence (CPO-TRA) du format TIFF dépendent de la compression utilisée. Dans sa version non compressée, le format reste simple et ne nécessite donc pas un niveau de compétence et d'équipement matériel et logiciel élevé.

Métadonnées internes

Outre le flux image, le format TIFF définit des « tags TIFF » décrivant à la fois la structure et les caractéristiques du flux image mais fournissant aussi quelques éléments de description du contenu (artiste, droits de propriété intellectuelle, date, etc.). Le site *Aware Systems* est la référence sur les tags TIFF, y compris sur ceux propriétaires.

La pratique BnF en matière d'intégration de métadonnées internes dans les images masters se limite aux tags TIFF et est décrite à l'annexe 1 du référentiel de numérisation des documents opaques.

En outre, TIFF est extensible et permet d'embarquer des métadonnées internes dans différents formats (EXIF, IPTC, XMP).

Outils connus par la BnF

- Pour la consultation, les principaux logiciels de traitement image le gèrent correctement (XnView, IrfanView, Photoshop, Gimp, ImageMagick, etc.).
- Pour la manipulation et la migration, la bibliothèque logicielle open source LibTIFF est complète et maintenue.
- Pour la caractérisation, le module TIFF de l'outil JHOVE est utilisé à la BnF et prend en charge l'extraction du profil ICC à partir de sa version 1.18. Il fournit un rapport technique complet incluant des métadonnées techniques pour les images au format Metadata for Images in XML (MIX). On peut également citer d'autres outils permettant d'extraire certaines métadonnées : Apache Tika et Exiftool.
- Pour la validation, JHOVE dispose d'un module de validation TIFF, mais celui-ci ne s'avère pas capable de repérer des fichiers tronqués. Dans le cadre du projet européen PREFORMA, un profil d'application restreint aux fonctionnalités considérées comme durables a été développé : TI/A. L'outil DPF Manager permet notamment de vérifier la conformité d'un fichier TIFF aux exigences TI/A. Néanmoins, le niveau de maintenance de cet outil reste, actuellement (juin 2020) incertain.
- Pour la consultation des tags TIFF, il existe l'outil AsTIFFTagViewer. Une composante de la bibliothèque LibTIFF, TIFFINFO permet également de récupérer des informations sur un fichier TIFF ; elle fonctionne en ligne de commande.

Usage ou présence dans les collections de la BnF

Le format TIFF a été utilisé depuis le début de la numérisation de conservation à la BnF, au début des années 1990. Il s'agit du format le plus représenté dans ses collections numériques. Depuis 2014, le format est progressivement remplacé par le format JPEG 2000 dans les nouveaux marchés de numérisation.

4.17. WARC

Description

Le format WARC est un format conteneur, utilisé principalement pour stocker les fichiers collectés sur le web au fur et à mesure qu'ils sont capturés par un robot. Il combine plusieurs ressources numériques, quel que soit leur format, dans un fichier d'archivage agrégé avec des informations de contexte. Il est utilisé à la BnF depuis le 13 octobre 2014. Il a été créé par la *California Digital Library* et *Internet Archive*. Les bibliothèques nationales d'IIPC (Consortium pour la préservation d'internet) ont contribué à ce qu'il devienne un standard ISO (CPO-DOC). La première publication de la norme date de mai 2009.

Sociabilité

Ce format est utilisé par les membres d'IIPC, réunissant une cinquantaine d'institutions et organismes autour de l'archivage du web. D'autres communautés utilisent également ce format tel que Rhizome, communauté consacrée aux arts numériques réunissant artistes et musées ou LOCKSS, sous les auspices de l'Université de Stanford, un réseau *peer-to-peer* qui développe et prend en charge un système _open source_ permettant de collecter, de préserver et de fournir à leurs lecteurs un accès au matériel publié sur le web.

Relations à d'autres formats

Le format WARC dérive de l'ancien format ARC et étend son usage pour mieux répondre aux besoins de collecte, de préservation, d'accès et d'échange des organisations d'archivage. Le format ARC a été créé par *Internet Archive* en 1996 pour archiver des copies de fichiers web. WARC permet de bien différencier les différents types d'enregistrement pour plus de structuration et de lisibilité, contrairement au format ARC.

Versions

Le format WARC 1.0 a fait l'objet d'une révision en 2017 pour préciser l'utilisation de certaines informations. Outre le contenu principal actuellement enregistré, la révision prend en charge le contenu secondaire associé, comme les métadonnées attribuées, les événements de détection de doublons abrégés et les transformations ultérieures.

Format	Date	Lien
WARC 1.0	2009	ISO 28500:2009
WARC 1.1	2017	ISO 28500:2017

Contenu

Un fichier WARC est la concaténation de plusieurs enregistrements WARC. Un enregistrement WARC se compose d'un en-tête (header) suivi d'un bloc de contenu (*payload*) ; l'en-tête comporte des champs nommés obligatoires qui documentent l'URI (identifiant de l'enregistrement), la date, le type et la longueur de l'enregistrement et qui renseignent les échanges client-serveur pour chaque ressource collectée (fichier). Il existe huit types d'enregistrement WARC : « warcinfo », « response » (réponse), « resource » (ressource) , « request » (requête), « metadata » (métadonnées) , « revisit », « conversion » et « continuation ». Chacun de ces types correspond à un résultat particulier du processus de collecte. Les blocs de contenu d'un fichier WARC peuvent contenir des ressources de n'importe quel format sous forme binaire y compris les fichiers audio ou vidéo.

Organisme de maintenance et documentation de référence

Le format WARC est maintenu par IIPC et normalisé par l'ISO depuis 2009, avec une révision en 2017 (ISO 28500:2017, Information et documentation – Format de fichier WARC). La BnF diffuse une version *draft* de la norme sur son site bibnum.bnf.fr.

IIPC propose de la documentation supplémentaire pour implémenter le WARC.

Identifiants

Registre	Identifiant

Wikidata	Q7978505
Bibliothèque du Congrès	fdd000236
PRONOM	WARC 1.0 fmt/289 WARC 1.1 fmt/1281

Caractéristiques techniques

La distribution des fichiers collectés au sein des fichiers WARC ne correspond à aucune logique documentaire. Le robot enregistre les fichiers capturés selon sa logique propre (et non inhérente au format WARC) dans les différents fichiers WARC disponibles jusqu'à la taille limite atteinte (définie à 1 Go pour la BnF). Le WARC est conçu pour être produit et consommé en flux avec des enregistrements concaténés.

Le format WARC ne définit pas de compression interne. Cependant la compression au format GZIP avec la méthode *Deflate* est recommandée. Dans une logique de production et consommation en flux la norme nous indique :

> *Comme spécifié à la section 2.2 de la spécification GZIP (voir [RFC 1952]), un fichier GZIP valide consiste en n'importe quel nombre de « membres » GZIP, chacun compressé indépendamment.*
>
> *Lorsque cela est possible, cette propriété devrait être exploitée pour compresser chaque enregistrement d'un fichier WARC indépendamment. Cela résulte en un fichier GZIP valide dont les sections (produites enregistrement par enregistrement) sont par elles-mêmes des fichiers GZIP valides.*
>
> *Les index externes au contenu du fichier WARC peuvent alors être utilisés pour enregistrer la position de départ de chaque enregistrement dans le fichier GZIP, permettant des accès aléatoires aux enregistrements individuels sans requérir la décompression de tous les enregistrements précédent.*
>
> *Notons que l'application d'une telle méthode ne produit aucun changement dans le contenu non compressé d'un enregistrement WARC individuel.*
>
> Annexe D.2, page 29

Métadonnées internes

Le WARCinfo est un champ recommandé et à la BnF, le choix a été de le rendre obligatoire.

Les fichiers WARC étant des conteneurs capables de stocker n'importe quel type de fichier – qu'il s'agisse de fichiers archivés en ligne, ou non – les métadonnées produites pour décrire la collection ou le déroulement de la collecte peuvent être empaquetées également dans un fichier WARC dit de métadonnées. Ainsi un même format conteneur peut gérer les données collectées et celles documentant la collecte, et la collection de fichiers WARC peut être ainsi auto-décrite. La communauté utilisatrice de la suite logicielle NetarchiveSuite, qui réunit une dizaine de pays autour de l'archivage du web dont la BnF, a fait ce choix-là.

Outils connus par la BnF

L'outil utilisé par la BnF pour générer des WARC est le robot de collecte Heritrix. Mais il existe d'autres outils, comme Webrecorder ou Wget qui sont également des robots de collecte, ou Warcreate (plugin de Chrome pour enregistrer une page sous forme de WARC).

Les fichiers WARC sont lus à la BnF par la Wayback machine, les JWat tools et Jhove. Mais il existe également d'autres outils de consultation comme la Python Wayback et WAIL.

Une page wiki recense les outils autour du format WARC.

Usage ou présence dans les collections de la BnF

Le format WARC est utilisé pour les données issues de l'archivage du web depuis 2014.

4.18. WAVE / WAVE RF64

Description

Le format WAVE (*Waveform Audio File Format*) dit également « WAV » en raison de son extension de fichier .wav est un format conteneur mis au point par Microsoft et IBM en 1991 afin de transporter des flux audio. Le format WAVE est construit conformément à la structure *Resource Interchange File Format*, c'est pourquoi on parle aussi parfois de « RIFF/WAVE ».

En théorie propriétaire car développé et maintenu par des entreprises privées, la liberté d'utilisation du format WAVE et de ses dérivés, eux totalement libres car développés par l'EBU (*Eureopean Broadcast Union*) est complète en raison de la disponibilité des spécifications (CPO-DOC) et l'absence de brevets portant sur ces formats (CPO-LIB).

Sociabilité

Le format WAVE est très largement géré par les logiciels et matériels audiovisuels. Il est devenu un standard *de facto* et demeure aujourd'hui encore le format de prédilection pour la production musicale et audiovisuelle *broadcast* et l'archivage du son, qu'il soit nativement numérique ou numérisé (voir les Recommandations pour la production et la conservation des objets audionumériques produites par l'IASA TC-04).

L'emploi du format WAVE a été conçu pour le système d'exploitation Windows et est supporté par Linux ; son équivalent pour les systèmes d'exploitation Apple est l'AIFF (non compressé) / AIFC (compressé).

Relations à d'autres formats

BWF (*Broadcast Wave Format*) : l'EBU a standardisé une extension du format WAVE qui lui ajoute un nouveau *chunk* (bloc) de données permettant de renseigner des métadonnées utiles à la production *broadcast* comme le *timecode*, la localisation, ou encore l'ajout de remarques sur la qualité de prise de son.

WAVE RF64, également nommé MBWF : le format WAVE (et le BWF) est limité à une taille de fichier de 4 Gio, ce qui peut être problématique dans certains domaines. Cette limite a été repoussée à 16 Eio avec le format WAVE RF64 (ou MBWF). Ce format, bien que reposant sur le format WAVE, n'est cependant pas rétro-compatible et nécessite par conséquent un lecteur adapté.

Sony WAVE64 : le nom parfois utilisé de « WAVE64 » désigne une variante **propriétaire** de WAVE développée par Sony.

Contenu

Le plus souvent, le format WAVE embarque des flux audio LPCM (*linear pulse-code modulation*) sans perte, mais il peut aussi contenir des flux audio compressés (avec perte) de type MP3, AC-3, MQA, ATRAC ou ADPCM. L'utilisation d'un conteneur WAVE pour transporter un flux audio compressé est néanmoins déconseillée : outre que le conteneur n'est pas particulièrement adapté à cette utilisation, elle peut induire en erreur l'utilisateur sur la nature du flux contenu.

Les données audio LPCM ont une profondeur d'échantillonnage de 4 à 32 bits. Les caractéristiques du flux LPCM les plus courantes sont 16 bits / 44,1 kHz (qualité CD) et 24 bits / 88,2, 96 ou 192 kHz (haute qualité).

Jusqu'à Windows 2000, le format WAVE pouvait comprendre des données audio sur un ou deux canaux. A partir de Windows 2000, le format supporte le multi-canal.

Le format WAVE peut également embarquer des flux LPCM « très haute définition » avec une profondeur d'échantillonnage sur 24 bits et des fréquences d'échantillonnage de 352,8 kHz à 384 kHz. Ce type de flux audio, nommé DXD, est souvent utilisé dans les étapes intermédiaires de production de fichiers DSD.

Le format WAVE étant basé sur RIFF, qui a une structure extensible, il peut être enrichi de *chunks* comportant des données ou des métadonnées de nature très différente, éventuellement propriétaires. Ces données ne sont pas supportées par tous les lecteurs et outils de manipulation, mais ils ne compromettent généralement pas la lecture du flux audio.

Organisme de maintenance et documentation de référence

La documentation du format RIFF/WAVE est accessible via les documents suivants :

- *Multimedia Programming Interface and Data Specifications 1.0* (voir pages 56-65);

- *New Multimedia Data Types and Data Techniques* (voir pages 12-22) ;
- *Multiple Channel Audio Data and WAVE Files* ;
- Sur le WAVE RF64 : *MBWF / RF64: An extended File Format for Audio*, spécification technique EBUTech 3306, Genève : EBU/UER, 2009 ;
- Sur le WAVE RF64 : Long-form file format for the international exchange of audio programme materials with metadata, recommandation ITU-R BS.2088, International Telecommunications Union, 2019.

La *British Library* a réalisé une évaluation de ce format pour la préservation à long terme.

Identifiants

Registre	Identifiant
Wikidata	Q217570 pour le format de fichier RIFF/WAVE Q3241497 pour le codec LPCM Q3928266 pour le format de fichier WAVE RF64
Bibliothèque du Congrès	fdd000001 pour le format de fichier RIFF/WAVE fdd000011 pour le codec LPCM
PRONOM	fmt/6 pour le format RIFF/WAVE fmt/712 pour le format RF64
Just Solve the File Format Problem	WAVE

Caractéristiques techniques

Le flux audio LPCM étant non compressé, les fichiers RIFF/WAVE sont d'une taille conséquente (CPO-COM). Le format est donc particulièrement adapté dans les cas où l'espace de stockage n'est pas contraint et où, pour des besoins d'édition audio par exemple, les processus de compression / décompression sont problématiques.

Le format WAVE contenant un flux non compressé LPCM est le format audio le plus simple (CPO-SIM), et un des plus stables (CPO-STA). Il est également assez robuste, la plupart des lecteurs étant capables de restituer un flux audio partiellement altéré. Néanmoins, contrairement aux formats destinés à être diffusés en *streaming* (comme MP3 et FLAC, il ne permet pas d'embarquer des empreintes numériques internes (CPO-ROB).

Métadonnées internes

Le format WAVE peut embarquer des métadonnées internes selon le formalisme RIFF tags.

La *Federal Agencies Digitization Guidelines Initiative* produit une recommandation visant à insérer des métadonnées minimales d'identification dans les fichiers WAVE issus d'une numérisation. Le sous-ensemble des RIFF tags défini par ces recommandations est éditable par l'outil `BWF MetaEdit`.

Le format étant extensible, il arrive parfois que l'on trouve également des *chunks* comportant des métadonnées internes aux formats ID3 ou XMP.

Outils connus par la BnF

L'outil d'analyse et d'extraction de métadonnées techniques pour la BnF est `MediaInfo`, qui dispose d'une sortie native dans le format de métadonnées de caractérisation préféré par la BnF, MPEG-7.

Le module WAVE du logiciel `JHOVE` permet de valider la conformité de la structure des fichiers RIFF/WAVE au regard des spécifications du format. Il est conseillé d'utiliser des versions du logiciel postérieures à la version 1.20, version à partir de laquelle il est capable d'identifier des fichiers WAVE tronqués. Néanmoins, le module ne contrôle pas le flux audio ; si ce dernier a été corrompu, JHOVE ne le signalera pas.

Plusieurs initiatives, comme *Lossless Audio Checker* visent à élaborer des outils capables d'analyser si un flux dans un codec audio sans perte n'a effectivement pas subi d'opérations de sur-échantillonnage (*upsampling*) d'interpolation ou de conversion depuis un flux compressé. Ces outils sont pour la plupart encore à un stade de développement bêta.

Usage ou présence dans les collections de la BnF

La migration de supports opérée par la BnF ou par des prestataires sur les CD audio (pressés) et les CD-R reçus par dépôt légal est à l'origine d'un très grand nombre de fichiers WAVE conservés par la BnF. D'autres fichiers de ce type peuvent aussi avoir été reçus en dépôt légal sous forme nativement dématérialisée.

La première campagne de numérisation du plan de sauvegarde de la BnF débuté en 1997 était conservée sur CD-R (Gold). L'ensemble de ces documents numérisés au format « CD » (LPCM - 16 bits - 44,1 kHz) a été par la suite transféré au format WAVE et stocké sur cartouches LTO.

Des fonds originaux de la BnF d'enregistrements musicaux, collectes, conférences et entretiens ont été enregistrés au format WAVE. Ils peuvent être issus de numérisation ou de recopie de supports numériques, comme par exemple de cassettes R-DAT.

Depuis 2009, l'ensemble des fichiers audio de numérisation patrimoniale sont au format WAVE (LPCM - 24 bits - 96 kHz). Pour des enregistrements de longue durée, les fichiers sont au format WAVE RF64. Pour des enregistrements devant conserver des informations de synchronisation par *timecode*, le format de fichier BWF a été retenu.

4.19. XML-ALTO

Description et sociabilité

ALTO (*Analyzed Layout and Text Object*) est un schéma XML standardisé, qui permet de stocker les informations relatives à la structure physique et au texte extrait par OCR (*Optical Character Recognition* : reconnaissance optique de caractères) d'une page d'un document numérisé (livre, revue ou journal).

Très adapté à la conservation à long terme de ces données, ALTO a été adopté par de nombreuses institutions dans leur processus de conversion en mode texte de documents numérisés ; institutions au rang desquelles on compte la *Library of Congress*, l'Université de Harvard, les Bibliothèques nationales du Danemark, Finlande, France, Nouvelle-Zélande, Pays-Bas, Singapour, etc. (liste complète ici).

ALTO est au demeurant très utile pour l'exploitation et la valorisation des collections numériques. Les coordonnées matricielles qu'il contient peuvent par exemple servir de support à la génération de PDF multicouches ou au développement d'applications web de surlignage d'occurrences.

Relations à d'autres formats

Un fichier ALTO est un fichier XML.

Il est conçu pour être utilisé comme un complément du schéma XML METS (*Metadata Encoding and Transmission Schema*). METS renseigne sur la structure logique de la page - nature sémantique des blocs de texte par exemple (titre, partie d'article, légende d'illustration, etc.) -, tandis qu'ALTO localise des contenants (blocs, lignes, etc.) sur la matrice de l'image de la page.

Lorsqu'ils coexistent, METS intègre dans sa `<structMap>` des références directes à des éléments de l'ALTO, des `<TextBlock>` par exemple. Mais les instances ALTO peuvent exister en tant que documents autonomes, indépendamment de METS. Sur l'utilisation conjointe d'ALTO et METS, on consultera cette page.

Versions

Les schémas ALTO sont mis à jour avec des nombres entiers quand leurs modifications interrompent la compatibilité descendante (version 1 à 2, par exemple), et des décimales quand elles ne l'interrompent pas (4.0 à 4.1, par exemple). Depuis juillet 2020, le schéma est dans sa version 4.2.

Versions antérieures :

- 4.1 (mai 2019) ;
- 4.0 (janvier 2018) ;
- 3.1 (janvier 2016) ;
- 3.0 (août 2014) ;
- 2.1 (février 2014) ;
- 2.0 (janvier 2010) ;
- 1.4 (août 2007).

Contenu

Un fichier XML ALTO est composé de trois sections principales, sous la racine `<alto>` :

- `<Description>` (voir *infra*, Métadonnées internes)
- `<Styles>`
- `<Layout>`

La section `<Styles>` liste les polices de composition reconnues par le moteur OCR, avec certaines de leurs caractéristiques (taille, couleur, graisse...), et des informations sur la morphologie des paragraphes.

La section `<Layout>` expose le contenu lui-même, par page (élément `<Page>`). Chaque page se compose d'un `<PrintSpace>` encadré par des marges haute, droite, basse et gauche : cinq aires rectangulaires qui ne peuvent pas se chevaucher, où se répartissent les objets graphiques (`<Illustration>` et `<GraphicalElement>`) ou textuels identifiés par le moteur OCR lors de ses opérations de segmentation. Les `<TextBlock>` sont les unités essentielles de ces derniers. Ils sont divisés en lignes (`<TextLine>`) où le caractère espace (`<SP>`) sert à séparer les chaînes de caractères (`<String>`) et où les hyphens (`<HYP>`) sont identifiés comme séparateurs de chaînes devant être fusionnées.

Organisme de maintenance et documentation de référence

Depuis 2009, c'est un Comité éditorial mis en place par la *Library of Congress* qui est l'agent officiel de maintenance du schéma. Le Comité enrichit ce dernier et agit pour étendre son utilisation dans la communauté des bibliothèques numériques.

La documentation de référence est disponible sur le GitHub du Comité et les versions du schéma sont hébergées par la *Library of Congress*.

Identifiants

Registre	Identifiant
Wikidata	Q2819247

Caractéristiques

Un fichier XML ALTO contient deux types d'informations géométriques sur les aires rectangulaires structurant la page océrisée : les coordonnées matricielles de leur coin supérieur gauche (attributs `HPOS` et `VPOS`) et leurs dimensions (attributs `HEIGHT` et `WIDTH`). L'exploitabilité de ces informations nécessite que l'ALTO soit accompagné, lors de son archivage, de l'image océrisée correspondante dans ses dimensions natives. Pour assurer a minima un lien avec cette dernière, il doit fournir dans sa section `<sourceImageInformation>` - au sein de `<Description>` (voir *infra*, Métadonnées internes) - un lien pérenne de récupération de cette image.

Métadonnées internes

La section `<Description>` du fichier XML ALTO, enfant en première position de la racine `<alto>`, contient des métadonnées auto-descriptives ; entre autres : décompte d'éléments, logiciel OCR utilisé, et « taux OCR brut ». Ce taux exprime la confiance auto-évaluée du moteur à reconnaître les caractères et reconstituer les chaînes ; techniquement, il est la moyenne des valeurs des attributs `WC` (pour *Word Confidence*) de toutes les `<String>` et se renseigne dans une balise `<processingStepDescription>`.

Outils connus par la BnF

ALTO étant un format XML, il est indispensable de disposer d'un éditeur XML complet pour le manipuler, le valider et l'éditer. La validation est réalisée à l'aide du schéma XSD officiel.

Comme outils de production, on citera ABBYY FineReader Engine®, qui équipe la chaîne interne d'océrisation de la BnF, et Tesseract (à partir de la version 4), libre et *open source*.

La caractérisation d'ALTO, comme celle des autres formats XML, est réalisée par le module XML de l'outil Jhove. L'option `withTextMD` permet de produire une sortie au format textMD.

Usage ou présence dans les collections de la BnF

Le format ALTO est demandé par la BnF dans le cadre des prestations d'océrisation externes (cf. son référentiel OCR). C'est aussi le format produit par sa chaîne interne d'océrisation, active depuis avril 2019. Sa bibliothèque numérique Gallica l'exploite pour la recherche plein texte et la mise en surbrillance des mots recherchés.

La BnF a développé son propre schéma, dérivé de la version 2 d'ALTO et conforme à ce dernier. Documentation ici.

4.20. XML-METS

Description et sociabilité

Le format METS (pour *Metadata Encoding and Transmission Standard*) a été créé en 2001 par la *Digital Library Federation* afin de réunir dans un même fichier XML toutes les métadonnées nécessaires à la description d'un document numérisé. Depuis lors, son périmètre s'est étendu aux objets nativement numériques.

Pensé à l'origine dans une logique d'interopérabilité, son usage comme format d'échange entre bibliothèques numériques s'avère marginal. En revanche, il est massivement utilisé par les bibliothèques et les archives comme format conteneur de métadonnées pour décrire les paquets d'informations conservés dans un système de préservation numérique. La liste des profils d'application METS officiels fait état d'une cinquantaine d'implémentations, chiffre très en dessous de la réalité car de nombreuses institutions utilisatrices n'ont pas fait la démarche d'enregistrer leur profil. Parmi les principales implémentations METS comme fichier conteneur de métadonnées pour la préservation à long terme, on peut citer

- Archivematica, logiciel *open source* de préservation numérique,
- SPAR, le magasin numérique de la BnF,
- le projet européen e-Ark qui vise à définir des spécifications communes pour les paquets d'informations et à développer sur ce socle des outils de préservation numérique.

Enfin, il a été adopté par Europeana dans le cadre du projet Europeana Newspapers et par la BnF à sa suite pour encoder la disposition des contenus détectée grâce à des méthodes de reconnaissance optique de la mise en page (OLR).

Relations à d'autres formats

Les fichiers METS sont des fichiers XML.

Dans son usage d'encodage de l'OLR, les spécifications d'Europeana et de la BnF ont fait le choix d'articuler METS et XML-ALTO : le fichier METS décrit dans sa carte de structure (élément `<structMap>`) des contenus intellectuels et fait référence à des éléments XML de l'ALTO qui décrivent des zones d'une page et donnent leur localisation dans les fichiers images.

Versions

Depuis sa création en 2001, le format METS n'a pas subi de révision majeure. Les évolutions successives (au 13 mai 2020, la version en cours est la 1.12.1) ont été mineures et rétrocompatibles (tous les fichiers METS produits depuis la naissance du format sont donc valides au regard de la version actuelle).

Contenu

Les sections de métadonnées du format METS (métadonnées descriptives `<dmdSec>`, de provenance `<digiprovMD>`, de droits `<rightsMD>`, techniques `<techMD>` et de source `<sourceMD>`) sont extensibles : on peut y placer des métadonnées dans n'importe quel format XML, voire dans un format non XML comme du MARC ISO 2709. En particulier, il a été pensé en complémentarité directe avec d'autres standards développés par la communauté des bibliothèques numériques :

- pour les métadonnées de description du contenu intellectuel : Dublin Core, EAD, MODS,
- pour les métadonnées de préservation (techniques et de provenance) : PREMIS,
- pour les métadonnées techniques de caractérisation :
  - MIX (images),
  - audioMD et videoMD (contenus audiovisuels),
  - textMD (contenus textuels).

Concernant le lien avec les fichiers de contenu (images, texte, contenus audiovisuels, etc.) décrits par le fichier METS, deux options existent, bien que la première option soit très majoritaire :

- pointer, dans le fichier METS, vers les fichiers par le biais d'un URI,
- encapsuler du contenu en XML ou en base64 directement dans les éléments `<file>`.

Via des éléments `<mptr>` (*METS pointer*), des fichiers METS peuvent également pointer sur d'autres fichiers METS. Cette fonctionnalité permet de décrire des collections référençant les documents numériques qui les composent.

Le format METS pouvant référencer de nombreux contenus externes (CPO-AUT), il est nécessaire de veiller à la pérennité des liens et à la préservation de ces contenus. La solution la plus couramment adoptée est d'empaqueter dans un fichier conteneur le fichier METS et ses contenus référencés.

Organisme de maintenance et documentation de référence

Le format METS est maintenu par un groupe international d'experts, le *METS Editorial Board*. Son site web et tous les contenus liés à METS sont hébergés par la *Library of Congress*.

La documentation de référence du format METS est constituée par le *METS Primer* (manuel), basé sur la version 1.6. Une introduction en français aux principales caractéristiques du format est également disponible.

Les deux usages principaux de METS à la BnF sont décrits par les référentiels de numérisation : le référentiel d'enrichissement des métadonnées METS pour les fichiers de métadonnées de numérisation (« manifestes ») et le référentiel d'enrichissement du texte pour l'encodage de l'OLR.

Identifiants

Registre	Identifiant
Wikidata	Q1640667
Just Solve the File Format Problem	METS

Caractéristiques

L'unique format de sérialisation disponible pour METS est XML, bien qu'une expression en RDF soit à l'étude depuis 2011.

Comme pour tout autre format de métadonnées, chaque organisation productrice est fortement encouragée à définir et à documenter son utilisation à travers un profil d'application. Cette bonne pratique est si pertinente pour METS en raison de sa flexibilité que le comité éditorial METS a proposé un formalisme de description en XML des profils d'application du format. La version 2 de ce formalisme propose d'embarquer dans la description du profil des contrôles exécutables par machine.

Bien que théoriquement transparent pour un utilisateur humain (CPO-TRA), le format METS est considérablement complexifié par l'usage d'identifiants internes au fichier pour faire le lien entre les composantes de la structure du document numérique décrites dans la section `<structMap>`, leurs métadonnées dans les sections `<dmdSec>`, `<rightsMD>`, `<techMD>`, `<digiprovMD>` et `<sourceMD>` et leurs représentations numériques dans la section `<fileSec>`. Cette particularité rend la lecture et le contrôle manuel des fichiers METS laborieux.

Métadonnées internes

Le format METS peut embarquer des métadonnées auto-descriptives (date de création, de dernière modification, agents ayant contribué à l'élaboration du fichier) dans son élément initial `<metsHdr>` (en-tête).

Outils connus par la BnF

METS étant un format XML, il est indispensable de disposer d'un éditeur XML complet pour le manipuler, le valider et l'éditer.

Fichiers METS manifestes de Paquets d'informations

Certains utilisateurs ont développé des outils spécifiquement pour METS, mais en raison de la flexibilité du format, ces derniers restent généralement très liés au profil d'application pour lequel ils ont été conçus. On peut cependant citer pour la visualisation l'outil METS Flask développé par la société Artefactual dans son produit Archivematica. Un *fork* de cet outil a été faite pour l'adapter au profil BnF : METS Viewer.

Fichiers METS d'OLR

L'outil Structify, développé par l'université d'Innsbruck et utilisé par Europeana dans le cadre du projet Europeana Newspapers, permet de réaliser la correction des contenus identifiés par OLR, d'y rajouter des métadonnées et d'exporter le résultat au format METS. La plateforme de transcription et d'annotation Transkribus permet également d'importer et d'exporter des documents METS.

Caractérisation

La caractérisation de METS, comme celle des autres formats XML, est réalisée par le module XML de l'outil Jhove. L'option `withTextMD` permet de produire une sortie au format textMD.

Validation

La validation est réalisée à l'aide du schéma XSD officiel et de tous autres schémas de métadonnées contenues dans le fichier METS. Un profil d'application METS complète généralement ces exigences ; les contraintes qu'il édicte peuvent être encodées dans le langage schematron et embarquées dans un profil METS « actionnable » (c'est la méthode adoptée par la BnF, voir les profils METS BnF pour la numérisation).

Usage ou présence dans les collections de la BnF

Depuis la mise en production de son magasin numérique SPAR en 2010, la BnF utilise METS comme format pivot de métadonnées pour enregistrer toutes les métadonnées techniques, de structure et de provenance utiles à la préservation à long terme au sein d'un paquet d'informations.

Depuis 2015, l'utilisation de METS s'est étendue :

- pour les nouveaux marchés de numérisation, en remplacement de l'ancien format maison « refnum », comme format conteneur de métadonnées de numérisation ;
- pour l'encodage de l'OLR appliqué à ses périodiques numérisés.

4.21. XML-tdmNum

Description

Le format tdmNum est développé et utilisé par la BnF pour la transcription des tables et index présents dans les ouvrages numérisés. Cette transcription reproduit la structure logique des contenus, identifie les intitulés des entrées ainsi que leurs liens vers le corps des ouvrages. Elle permet également de repérer les entités (personne, toponyme, organisation) éventuellement nommées dans les intitulés.

Sociabilité

Ce format n'est utilisé que dans le contexte des activités de numérisation de la BnF et de ses partenaires. Pour ces derniers, le respect de ce format garantit que leurs tables et index seront intégrés dans la chaîne d'entrée de documents numériques de la BnF et valorisés dans Gallica.

Relations à d'autres formats

Le format tdmNum est une restriction du format TEI, lui-même schéma de données structurées en XML. La sémantique du jeu de balises utilisé est décrite dans le référentiel de numérisation (voir ci-après).

Versions

Seule la version 1.1 du format tdmNum est utilisée.

Organisme de maintenance et documentation de référence

Le format est maintenu par la BnF, diffusé sur le site bibnum.bnf et décrit dans un des référentiels de numérisation de la BnF, Traitement des tables.

Identifiants

Ce format local n'a pas reçu d'identifiant international.

Métadonnées internes

Le format tdmNum n'embarque pas de métadonnées internes, si ce n'est l'identifiant du document numérique dont il décrit la table des matières ou l'index.

Outils connus par la BnF

La validation du format est réalisée par le moyen d'un schéma XSD.

La caractérisation de tdmNum, comme celle des autres formats XML, est réalisée par le module XML de l'outil Jhove. L'option `withTextMD` permet de produire une sortie au format textMD.

Usage ou présence dans les collections de la BnF

Ce format est en usage dans le cadre des marchés et programmes de numérisation de la BnF. Il sert à transcrire les tables et index des ouvrages numérisés ainsi que les plages des enregistrements sonores.

La fonctionnalité de navigation au sein des documents de Gallica (exemple sur un ouvrage imprimé, exemple sur un document sonore) utilise les transcriptions produites selon ce format.

5. Pour aller plus loin

Une liste de ressources a été produite par la *Library of Congress* : Related Resources for Digital Format Sustainability, disponible sur https://www.loc.gov/preservation/digital/formats/intro/resources.shtml.

5.1. Sources d'information sur les formats de données dans une perspective de préservation

- *Library of Congress, Format Descriptions* : https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml
- Wiki, *Solve the file format problem* : http://fileformats.archiveteam.org/wiki/Main_Page
- *Digital Preservation Coalition, Endangered Species* : https://www.dpconline.org/our-work/bit-list
- *Harvard University Libraries, Format Assessments* : https://wiki.harvard.edu/confluence/display/digitalpreservation/Format+Assessments
- *British Library, File Formats Assessments* : http://wiki.dpconline.org/index.php?title=File_Formats_Assessments
- *National Archives and Records Administration* (États-Unis) : https://github.com/usnationalarchives/digital-preservation
- *File-extensions.org* (liste d'extensions associées à des outils logiciels pour lire, écrire, convertir, etc. les fichiers) : https://www.file-extensions.org/.

5.2. Documents de politique sur les formats de fichier publiés par les institutions de conservation

- Bibliothèque et archives nationales du Québec : https://www.banq.qc.ca/documents/archives/archivistique_gestion/publications_proposees/Guide-formats-BAnQ_Final.pdf
- Centre de coordination pour l'archivage à long terme de documents électroniques (CECO, Suisse) : https://kost-ceco.ch/cms/kad_main_fr.html
- *Libraries and Archives Canada* : https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Pages/guidelines-file-formats-transferring-information-resources-enduring-value.aspx
- *Library of Congress* : https://www.loc.gov/preservation/resources/rfs/
- *National Digital Library* (Finlande) : http://digitalpreservation.fi/files/File-Formats-1.6.1-en.pdf
- *National Archives and Records Administration* (États-Unis) : https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html.
- *National archives of the Netherlands*, version 1.0, novembre 2016. <https://www.nationaalarchief.nl/sites/default/files/field-file/National%20Archives%20of%20the%20Netherlands%20preferred%20and%20acceptable%20formats.pdf>
- Référentiel général d'interopérabilité version 2 (bien qu'orienté vers l'interopérabilité et la diffusion) : http://references.modernisation.gouv.fr/sites/default/files/Referentiel_General_Interoperabilite_V2.pdf.
- *Stanford Libraries* : https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-formats.
- *Smithsonian Institution Archives*, Recommended Preservation Formats for Electronic Records. https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records