Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère
Culture

PROGRAMME NATIONAL DE NUMÉRISATION ET DE VALORISATION DES CONTENUS CULTURELS

RECOMMANDATIONS TECHNIQUES POUR LES MÉTADONNÉES ET STANDARDS

VERSION N°1 – 2017

Ministère de la Culture
Secrétariat général

Dans le cadre des groupes de réflexion pour le Programme National de Numérisation et de Valorisation des contenus culturels (PNV), ces recommandations ont été élaborées avec la précieuse collaboration de :

- **Anila Angjeli**, Bibliothèque nationale de France, Département des métadonnées,
- **Laurine Arnould**, ministère de la Culture, Direction Générale des Médias et Industries Culturelles/Département des bibliothèques,
- **Rodolphe Bailly**, Philharmonie de Paris, Département Éducation et Ressource,
- **Olivier Baude**, TGIR Huma-Num,
- **Katell Briatte**, ministère de la Culture, Direction Générale des Patrimoines/Département des systèmes d'information patrimoniaux,
- **Sarah Brunet**, ministère de la Culture, Direction Générale des Médias et Industries Culturelles/Bureau du financement des industries culturelles,
- **Bertrand Caron**, Bibliothèque nationale de France, Département des métadonnées,
- **Jean Davoigneau**, ministère de la Culture, Direction Générale des Patrimoines/Mission inventaire,
- **Thibault Grouas**, ministère de la Culture, Délégation générale à la langue française et aux langues de France,
- **Marie-Véronique Leroi**, ministère de la Culture, Secrétariat Général/Département de l'innovation numérique,
- **Carine Prune**t, ministère de la Culture, Direction Générale des Patrimoines/Service des Musées de France,
- **Romain Wenz**, ministère de la Culture, Direction Générale des Patrimoines/Service Interministériel des Archives de France.

Le traitement documentaire

Introduction

On entend par la notion de «traitement documentaire» l'ensemble des choix et actions qui vont déterminer la description et l'accès au fonds numérisé. Une opération de numérisation suppose notamment une sélection du fonds à numériser, l'usage qui en sera fait, le choix des standards qui seront appliqués au fonds numérisé et à ses métadonnées. Ces métadonnées sont indispensables pour la compréhension de l'objet ou du fonds numérisé.

Les métadonnées permettent de décrire et/ou de représenter le contenu d'un document ou d'un fonds.

Description

Les métadonnées

Tout type de ressource numérisée peut être représenté par un ensemble structuré de métadonnées. Celles-ci permettent de décrire le contenu pour le comprendre, d'administrer le document numérique et de le rendre accessible tout en garantissant sa pérennité, son authenticité et sa traçabilité (voir la fiche «Métadonnées administratives et de structure»).

Le regard métier

Ce jeu de métadonnées ne sera pas formalisé de la même manière selon les standards employés. Un même objet peut ne pas être décrit de la manière selon la perspective « métier » portée sur lui. Le regard « métier » structure la donnée. A titre d'exemple le traitement documentaire appliqué à une collection de cartes postales ne sera pas le même selon que celui-ci est opéré par un musée ou un service d'archives. Les archivistes s'attacheront à retrouver les toponymes là où les musées relèveront plutôt des détails ayant trait à l'histoire de l'art.

Ouverture et interopérabilité

Le traitement documentaire doit s'inscrire dans des logiques d'ouverture et d'interopérabilité. La qualité des données et métadonnées conditionne les réutilisations possibles, il en est de même pour le degré d'ouverture des ressources et de leurs métadonnées.
Le choix des métadonnées qui seront produites dans le cadre d'un projet de numérisation peut répondre à des usages clairement identifiés en amont du projet. Le fait de s'appuyer sur des standards favorise l'interopérabilité et peut permettre des usages autres que ceux attendus.

Approches participatives

Une approche participative peut venir compléter le traitement documentaire. Cette approche participative peut prendre diverses formes : collecte, enrichissement de métadonnées, annotations, transcriptions collaboratives… Il est préférable d'envisager cette approche collaborative en amont ou en parallèle du projet.
Le porteur de projet doit être conscient des risques induits pour la qualité et la fiabilité des données et la nécessité de gérer et animer les communautés d'utilisateurs selon le type d'approche choisi.

Enjeux

- Favoriser la compréhension d'un objet numérisé
- Favoriser l'interopérabilité en faisant des liens interinstitutionnels et transsectoriels et en s'inscrivant dans des portails locaux, nationaux et européens
- Envisager l'usage attendu du contenu numérisé et favoriser les usages futurs en privilégiant la rigueur et la complétude des métadonnées
- Améliorer l'accès à l'objet numérisé pour les utilisateurs et faciliter les traitements automatiques

Recommandations

Critères	Niveaux d'exigence
Définir en amont du projet de numérisation la description et l'usage de la collection	Obligatoire
Définir les éléments de la collection à numériser devant porter un identifiant (voir recommandation sur l'identification)	Obligatoire
Sélectionner le ou les standards pertinents à appliquer (voir recommandation sur les standards)	Obligatoire
Fournir une licence claire ou une déclaration précisant les droits d'accès et de réutilisation des ressources et métadonnées (voir livrables du GT2 : les métadonnées doivent être placées sous une licence ouverte)	Obligatoire
Sélectionner le ou les référentiels pertinents pouvant être utilisés (voir recommandation sur les référentiels)	Recommandé
Évaluer d'éventuelles approches participatives	Recommandé
S'intégrer dans des portails et agrégateurs	Recommandé

En savoir plus

- *« Understanding Metadata : what is metadata and what is it for ? »*, Jenn Riley, NISO http://www.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf

- *« Manuel de la numérisation »*, sous la direction de Thierry Claerr et Isabelle Westeel, Paris, Éditions du Cercle de la librairie, 2011.

Les identifiants

Définition

Un identifiant est une chaîne de caractères alphanumérique qui a pour fonction d'identifier un document, une ressource ou une entité quelle que soit sa nature.

Un identifiant est dit « pérenne » dès lors que l'organisation qui l'attribue s'engage à en assurer la gouvernance et la bonne gestion.

Description

Dans tout projet de numérisation se pose la question du nommage des fichiers qui seront produits à l'issue de cette numérisation. Il est préférable de se conformer au plan de nommage en vigueur au sein de son institution ou le cas échéant d'en définir un qui sera propre à l'opération de numérisation ou généralisable.

Ces problématiques de nommage trouvent un écho tout particulier dans le contexte du web sémantique dont l'un des principes fondamentaux réside dans l'utilisation d'identifiants pérennes.

Une grande partie des producteurs de données a recours à l'utilisation d'identifiants locaux pour identifier leurs ressources. Par exemple, les identifiants peuvent être des cotes issues du cadre de classement d'un service d'archives ou d'une bibliothèque, mais aussi des identifiants de fichiers numériques, des identifiants d'enregistrements dans des bases de données, ou encore des URL (Uniform Ressource Locator - localisateur uniforme de ressource) de pages web.

Dans le web sémantique, toute ressource doit bénéficier d'un identifiant qui respecte une syntaxe particulière et qui peut être utilisé pour accéder à la ressource. Ces identifiants sont les URI (Uniform Ressource Identifier - Identifiant uniforme de ressource). Ils partagent la même syntaxe que les URL mais avec, de surcroît, une exigence forte de pérennité. Sans URI, aucune ressource ne peut être produite ou utilisée dans le web des données liées.

Enjeux

La définition par une institution culturelle d'identifiants pérennes pour ses ressources numériques constitue une première étape vers le Web des données liées. L'intérêt des identifiants pérennes réside dans la possibilité de désigner de façon unique et pérenne une ressource, l'ensemble des métadonnées qui lui sont associées ou encore les concepts, périodes ou lieux qui lui sont associées.

Les ressources numériques ayant recours à des identifiants pérennes peuvent être découvertes et retrouvées plus facilement.

L'utilisation d'identifiants pérennes permet de garantir la réutilisation des ressources numériques et leur visibilité en offrant la possibilité de les lier à d'autres ressources et ainsi démultiplier les parcours de lecture ou chemins d'accès.

Les identifiants pérennes sont un impératif pour l'interopérabilité des métadonnées et ressources.

Recommandations

Critères	Niveaux d'exigence
Se conformer au plan de nommage de l'institution ou du domaine ou le cas échéant déterminer un plan de nommage spécifiant les éléments de la numérisation à doter d'un identifiant	Obligatoire
Garantir l'unicité des identifiants	Obligatoire
Veiller à la non-ré attribution des identifiants de la ressource concernée	Obligatoire
En cas de suppression ou de dépublication, en renseigner les circonstances et le cas échéant pointer vers une ressource de substitution	Recommandé
Publier sa politique de gestion des identifiants	Recommandé
Donner des identifiants non signifiants pour éviter toute ambiguïté liée aux langues naturelles	Recommandé
Dans le cas de l'utilisation d'identifiants pérennes, prévoir des URI HTTP (c'est-à-dire que les URI donnant directement accès à l'usager, via un navigateur, à une vue de la ressource).	Recommandé

En savoir plus

- *« Identifiants pérennes pour les ressources numériques, vade-mecum pour les producteurs de données »* : http://www.culturecommunication.gouv.fr/content/download/146600/1577205/version/1/file/Vademecum_Identifiants.pdf

- *« Data on the web, best practices »* (en anglais) : https://www.w3.org/TR/dwbp/#DataIdentifiers

Les standards de métadonnées

Définition

Dans le domaine des métadonnées, un standard offre un cadre sur lequel s'appuyer pour produire et traiter les métadonnées associées aux ressources documentaires, numérisées ou non.

Description

Un standard de métadonnées propose un ensemble d'éléments (par exemple titre, auteur, date de production) à utiliser pour décrire une ressource. Il précise les contraintes qui peuvent s'appliquer à ces éléments (par exemple, élément obligatoire ou à valeur unique). Il spécifie également la syntaxe à utiliser pour chaque élément. (par exemple pour une date : aaaa-mm-jj selon la norme ISO-8601).

On trouve des standards génériques qui permettent de décrire, à minima, toute ressource numérisée tels que Dublin Core. La version étendue de Dublin Core (DCTerms) permet de décrire plus finement les ressources en ajoutant des éléments adaptés au type de ressource à décrire (photographie, périodique, document audio, etc.).

En grande majorité, les standards ne sont pas issus d'une entreprise privée qui voudrait imposer sa vision et ses produits, ils sont au contraire le fruit d'un rigoureux travail communautaire et établi par des experts du domaine concerné à l'échelle nationale, européenne ou internationale. Certains standards accèdent au statut de norme dès lors qu'ils sont pris en charge par des organismes de normalisation certifiés tels que l'ISO (Organisation Internationale de Normalisation) ou l'AFNOR (Association Française de Normalisation).

Enjeux

L'utilisation de standards est fortement conseillée à toutes les étapes de traitement des métadonnées, qu'elles soient descriptives, juridiques ou techniques. Il existe des standards pour faciliter la production et le stockage de métadonnées ou encore leur échange, leur mise en ligne et leur liage avec des référentiels ou données externes.

L'utilisation de ces standards permet de gagner du temps dans les étapes du projet de production des métadonnées. En bénéficiant du cadre établi par le travail des auteurs du standard, on trouve réponse à toutes les questions que l'on peut se poser lorsque l'on veut décrire des ressources numérisées.

Les standards très utilisés par les acteurs de la communauté se retrouvent en général au coeur des outils logiciels de production et gestion des métadonnées, qu'ils soient open source ou produits d'éditeurs de logiciel. Le choix et l'utilisation d'un standard facilite donc le choix de l'outil logiciel de gestion de la collection numérisée. Ces outils proposent des fonctionnalités d'export des métadonnées qui permettront, par exemple, de les ré-utiliser dans un moteur de recherche commun à plusieurs sources de données.
En outre, l'utilisation d'un standard permettra de rendre le résultat du projet compatible, compréhensible et réutilisable (interopérable) par d'autres projets ou systèmes d'information.

Recommandations

Critères	Niveaux d'exigence
Utiliser un standard pour décrire une ressource numérisée	Obligatoire
Pour la description des ressources, utiliser à minima les éléments fournis par Dublin Core	Obligatoire
Pour la structuration des données, utiliser des standards favorisant le liage des données (RDF, SKOS, ...)	Recommandé

En savoir plus

Quelques exemples de standards de métadonnées fréquemment utilisés lors de projets de numérisation :

- Pour la description générique de ressources : Dublin Core elements, DCTerms, SCHEMA.ORG, etc.
- Pour la description de ressources propre à une communauté professionnelle : UNIMARC, RDA, MODS, EAD, LIDO, etc ;
- Exposition des métadonnées sur le web de données : RDF, SKOS, etc ;
- Structure du document numérisé : METS, SEDA, etc ;
- Modèles conceptuels : LRM (anciennement FRBR), CIDOC-CRM, FRBRoo, HADOC, etc ;
- Les référentiels de numérisation de la BnF : http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.numerisation_referentiels_bnf.html

Les référentiels

Définition

Dans le domaine documentaire le terme « référentiel » est employé conventionnellement pour désigner de manière générique des groupements d'informations, dites « données de référence » ayant un haut niveau de « partageabilité », qui sont fréquemment utilisées ou consultées au sein du même système d'information ou par plusieurs systèmes et qui sont stables dans le temps. De ce fait, ils constituent des éléments structurants des divers domaines de connaissance. Ils servent en premier lieu aux producteurs de données pour produire une information normalisée et cohérente au sein d'un domaine de connaissance voire entre plusieurs domaines. Ils permettent d'offrir aux utilisateurs des points d'entrée fiables dans les corpus. Ils constituent enfin les piliers de l'interconnexion sémantique des jeux de données sur la toile (Linked Open Data).

Description

Dans le cadre d'un projet de numérisation, le terme « référentiel » s'applique aux dispositifs de type vocabulaires contrôlés et « systèmes d'organisation des connaissances » (Knowledge Organisation Systems - KOS). Le spectre recouvre des outils comme les thésaurus, les systèmes de classification, les nomenclatures, les taxonomies, les fichiers d'autorité, les listes de valeurs normalisées. Selon leur périmètre, ils peuvent porter sur des concepts ou des notions, des personnes et groupes humains, des lieux, des périodes, des œuvres, des objets, etc. Il peut s'agir de bases de connaissance complexes, crées par des efforts mutualisées et de longue haleine par des spécialistes dans leurs domaines respectifs d'expertise, telles que le Référentiel ONOMA (en cours de construction) pour les acteurs[1], ISNI[2] pour l'identification de tout contributeur à un contenu créatif, le répertoire RAMEAU d'autorités matière[3], mais aussi de vocabulaires métier, comme le vocabulaire RDA pour les types de médias[4], le thésaurus MIMO[5] pour les instruments de musique, la classification ICONCLASS[6] pour l'art et l'iconographie, etc.

Enjeux

Une politique solide et volontaire en matière de référentiels accroît de manière considérable la valeur des métadonnées des collections culturelles numérisées.

- Soigner les référentiels et systématiser autant que faire se peut leur utilisation dans le traitement documentaire est un gage de cohérence et de qualité des métadonnées associées aux ressources culturelles : chaque « chose » est décrite et identifiée de manière fiable et durable à travers toute la collection numérique et à travers diverses collections.

- Le principe de mutualisation des efforts et des expertises est à la base de la construction et de l'utilisation des référentiels. De ce fait, les référentiels contribuent à une économie de production des métadonnées, car ils permettent d'éviter les saisies multiples, les incohérences, etc. Par conséquent, ils contribuent également à une économie collaborative de production des savoirs.

- Les référentiels accroissent la visibilité, l'accessibilité numérique et la valeur d'usage des collections numériques. Ils permettent de construire des fonctionnalités avancées de recherche et d'exploitation, tant dans les systèmes de recherche traditionnels pour lesquels ils ont été initialement créés que dans la « toile sémantique » qui s'appuie majoritairement sur eux.

- Enfin, par le jeu des alignements qu'ils permettent, les référentiels sont la clé de l'interconnexion sémantique des collections culturelles et par là de l'interopérabilité entre jeux de données.

Recommandations

Critères	Niveaux d'exigence
Utiliser des éléments issus de schémas de métadonnées standards (par exemple dc-terms:subject, skos:prefLabel, foaf:name, issus respectivement des schémas Dublin Core, SKOS, FOAF) pour encoder les données issues des référentiels.	Obligatoire
Tout concept, notion, personne, lieu etc. issu d'un référentiel doit être identifié par un URI. Voir la recommandation sur les Identifiants pérennes[7].	Obligatoire
Utiliser les référentiels qui font autorité dans le domaine.	Recommandé
Privilégier l'enrichissement des référentiels existants à la création de nouveaux référentiels.	Recommandé
Publier sur la toile sémantique sous licence ouverte tout référentiel spécifique, pour en favoriser le partage, voire la réutilisation.	Recommandé

En savoir plus

En plus des référentiels déjà mentionnés dans cette fiche, d'autres référentiels sont disponibles en ligne. On citera, en particulier, les vocabulaires disponibles sur la plate-forme de données ouvertes du ministère de la Culture :

https://data.culturecommunication.gouv.fr/explore/dataset/les-vocabulaires-du-ministere-de-la-culture-et-de-la-communication/

1 - Personnes ou groupes mentionnés dans les ressources documentaires relatives aux biens culturels.
2 - http://www.isni.org
3 - http://rameau.bnf.fr
4 - www.loc.gov/standards/valuelist/rdacarrier
5 - http://www.mimo-international.com/MIMO/instrument-families.aspx
6 - www.iconclass.nl
7 - http://www.culturecommunication.gouv.fr/Thematiques/Innovation-numerique/Donnees-publiques/Identifiants-perennes-pour-les-ressources-numeriques

Métadonnées administratives et de structure

Définition

Les métadonnées administratives et de structure sont des informations, de préférence structurées et lisibles par machine, décrivant d'autres données ; en l'occurrence des objets patrimoniaux numérisés.

Description

Si les métadonnées administratives sont généralement structurées, standardisées et, de ce fait, lisibles par machine, elles ne sont pas nécessairement fournies individuellement et dans un formalisme spécifique. Ainsi, certains choix implicites faits au moment de la production (notamment les consignes de prise de vue, les traitements réalisés sur les objets numérisés, etc.) doivent être documentés mais, l'information étant homogène et commune à l'ensemble des produits d'une prestation, on pourra se contenter de la préserver dans la documentation du marché.

Dans le cas de métadonnées fournies unitairement (pour chaque objet numérique), il est recommandé qu'elles soient transmises dans un format de fichier texte structuré (formats XML, formats tabulés de type CSV ou TSV, voire texte brut structuré par un formalisme constant) et encodées suivant le codage de caractères UTF-8. Une fois produites, leur qualité et la cohérence avec les données qu'elles décrivent doivent être contrôlées, manuellement ou automatiquement, au moins par échantillonnage.

Métadonnées de structure

Les métadonnées de structure doivent détailler la structure de l'objet numérisé en vue de maintenir sa logique interne et de permettre son exploitation et son affichage. Le plus souvent, ces relations se limiteront à une liste ordonnée de fichiers composant l'objet numérisé, à afficher séquentiellement. Néanmoins, des objets numérisés à la structure plus complexe peuvent nécessiter l'identification via des métadonnées

- des composantes de l'objet numérisé, permettant ainsi d'associer à chacune d'entre elles des métadonnées descriptives ;

- de la nature des relations de tout à partie – les composantes doivent-elles être exploitées séquentiellement, simultanément, alternativement ? ;

- de la nature des relations horizontales entre les composantes : relations de dérivation – par exemple, fichiers OCR produits à partir de fichiers images – ou de dépendance entre une composante et son environnement technique.

Métadonnées techniques

Internes (embarquées dans le fichier lui-même, par exemple formats EXIF, XMP, IPTC, JFIF, TIFF) ou externes (transmises dans un fichier autonome), les métadonnées techniques peuvent être génériques, c'est-à-dire communes à tout type de fichier, ou spécifiques à un type de contenu voire à un format de fichier. Parmi les métadonnées techniques génériques, les métadonnées d'intégrité (taille, empreinte numérique) sont fondamentales afin de contrôler que les fichiers produits n'ont pas été altérés lors d'un traitement ou d'un transfert.

Parmi les métadonnées spécifiques à des types de contenu, on citera à titre d'exemple :

- pour le texte : l'encodage, la conformité à un schéma dans le cas de texte structuré, etc. ;
- pour l'image : la résolution, le profil colorimétrique, la profondeur d'encodage, etc. ;
- pour le son : le débit, le codec, la fréquence d'échantillonnage, etc. ;
- pour la vidéo : le nombre d'images par seconde, le profil colorimétrique, la durée, etc.

Métadonnées de droits

Les métadonnées de droits conditionnent les opérations que l'institution commanditaire et les utilisateurs finaux sont autorisés à réaliser sur les objets numérisés. Si l'absence d'une documentation précise de ces droits peut conduire à restreindre l'usage des objets numérisés, elle peut également compromettre la capacité de l'institution à assurer leur préservation. Les métadonnées de droits viseront donc à décrire des règles définies par

- les actions autorisées (ou à l'inverse interdites) sur l'objet numérisé ;
- les contraintes et restrictions portant sur ces permissions ;
- les agents auxquels les permissions ou interdictions s'appliquent ;
- les agents ayant contribué à la définition des règles (donateurs, ayants droit ou simples contacts) ;
- la base juridique sur laquelle s'appuient ces règles, qu'elle soit fondée sur une loi, une licence, une convention ou une politique spécifique à l'institution.

Métadonnées de provenance

es métadonnées de provenance documentent le contexte de production, l'historique de transmission, de préservation et éventuellement d'enrichissement de l'objet numérisé.

Elles jouent un rôle fondamental dans
- la gestion de la qualité en permettant d'identifier des processus défaillants ;
- la fiabilité et la traçabilité de l'objet numérisé en traçant les opérations l'ayant affecté et les transferts de responsabilité ;
- l'amélioration de la collection numérisée, en identifiant les traitements de correction ou d'enrichissement déjà réalisés sur l'objet numérisé.

Dans ce but, les métadonnées de provenance enregistreront
- des événements successifs intervenus sur l'objet ou l'une de ses composantes (notamment la date, le résultat, les paramètres, etc.) ;
- des agents impliqués dans les événements en question, qu'ils soient des humains, des organisations, des logiciels ou des matériels (notamment leur nom, leur version, leur marque, etc.).

Enjeux

Faire produire et préserver des métadonnées administratives de qualité répond à plusieurs besoins des gestionnaires de collections numérisées :

- assurer la qualité des livrables et leur conformité aux exigences du commanditaire ;
- construire des fonctionnalités de recherche et d'exploitation ;
- gérer les droits d'accès et d'utilisation ;
- garantir l'authenticité, la traçabilité et l'intégrité des objets numérisés ;
- permettre l'amélioration de la collection grâce à une connaissance précise des caractéristiques et des traitements réalisés ;
- garantir la pérennisation de l'information, en se prémunissant contre les risques liés à l'accessibilité et à l'obsolescence technique.

Recommandations

Critères	Niveaux d'exigence
Génériques	
Utiliser un ou plusieurs formats de métadonnées standard et lisibles par machine (voir liste de formats dans la section Ressources ci-dessous).	Recommandé
Fournir une documentation sur les métadonnées produites, le cas échéant sous la forme d'un profil d'application.	Obligatoire
Utiliser des vocabulaires contrôlés pour les valeurs de champs (ex. : type MIME ou identifiant PRONOM pour le format de fichier, vocabulaire RDA pour les types de médias, cf. www.loc.gov/standards/valuelist/rdacarrier, etc.).	Obligatoire si applicable
Métadonnées de structure	
Spécifier les relations entre les composantes de l'objet numérisé (relations de tout à partie, de dérivation ou de dépendance).	Recommandé
Métadonnées techniques	
Fournir des métadonnées techniques génériques (poids, format de fichier, empreinte numérique).	Obligatoire
Fournir des métadonnées techniques spécifiques au type de format (texte, audio, image, vidéo, etc.).	Recommandé
Inclure des métadonnées techniques internes	Recommandé
Documenter les mesures techniques de protection (DRM).	Obligatoire si applicable
Métadonnées de droits	
Documenter les actions permises à l'institution commanditaire pour assurer l'utilisation et la pérennisation de l'information et les agents ayant accordé ces permissions.	Recommandé
Métadonnées de provenance	
Documenter le contexte de production (date, matériels et outils logiciels utilisés)	Obligatoire
Documenter l'ensemble des traitements successifs réalisés sur les objets numérisés jusqu'à leur livraison.	Recommandé

En savoir plus

Les formats listés ci-dessous sont des guides, non des contraintes : ils fournissent des vocabulaires et syntaxes riches et variées pour exprimer différents types de métadonnées sur différents types de contenus numériques. Une étude préalable devra définir un profil d'application décrivant l'utilisation faite par le prestataire et/ou le commanditaire d'un ou de plusieurs formats selon les besoins et les moyens.

La quasi-totalité des formats cités ci-dessous sont des formats de métadonnées externes fondés sur la syntaxe XML, sauf les formats EXIF et XMP, qui sont deux formats de métadonnées internes (ces dernières sont encapsulées dans le fichier image lui-même) et le format ODRL, qui utilise la syntaxe RDF. Le format PREMIS dispose, en plus d'une syntaxe XML, d'une ontologie permettant de l'exprimer également en RDF.

Formats de métadonnées d'empaquetage

Il s'agit de formats de métadonnées génériques permettant de décrire globalement la structure de l'objet numérisé et d'encapsuler dans le même fichier XML des métadonnées spécifiques (techniques, de droits et de provenance).

- METS (Metadata Encoding and Transmission Standard) http://www.loc.gov/standards/mets/
- SEDA (Standard d'Échange de Données pour l'Archivage) https://francearchives.fr/seda/
- Le standard IIIF (International Image Interoperability Framework) peut également être considéré comme un format de métadonnées d'empaquetage et de structure, dans la mesure où il décrit les images ou les fragments d'images qui composent l'objet numérique et l'ordre dans lequel ces derniers doivent être affichés. http://iiif.io/api/presentation/
- PREMIS (PREservation Metadata Implementation Standard) http://www.loc.gov/standards/premis complété par des vocabulaires de préservation sur http://id.loc.gov/vocabulary/preservation

Formats de métadonnées techniques

Fichiers image

- MIX (Metadata for Images in XML) http://www.loc.gov/standards/mix
- EXIF (Exchangeable Image File Format) http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf
- XMP (Extensible Metadata Platform) http://www.adobe.com/products/xmp.html

Fichiers texte

- textMD http://www.loc.gov/standards/textMD

Fichiers bureautiques (PDF, ODT, Word, etc)

- documentMD https://share.fcla.edu/FDAPublic/DAITSS/documentMD.pdf
- XMP (Extensible Metadata Platform) http://www.adobe.com/products/xmp.html

Fichiers « conteneurs » (ZIP, TAR, ARC, WARC, etc)

- containerMD http://bibnum.bnf.fr/containerMD-v1_1/

Fichiers audio

- AES57 http://www.aes.org/tmpFiles/aessc/20121211/aes57-2011-i.pdf
- MPEG-7 http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm
- audioMD http://www.loc.gov/standards/amdvmd
- PBCore (Public Broadcasting Metadata Dictionary Project) http://pbcore.org

Fichiers vidéo

- videoMD http://www.loc.gov/standards/amdvmd
- MPEG-7 http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm
- PBCore http://pbcore.org

Formats de métadonnées de droits

- ODRL (Open Digital Rights Language) http://www.w3.org/community/odrl
- MPEG-21 REL (Rights Expression Language) : http://mpeg.chiariglione.org/standards/mpeg-21/rights-expression-language
- copyrightMD http://www.cdlib.org/groups/rmg
- metsRights http://www.loc.gov/standards/rights/METSRights.xsd

Formats de métadonnées de provenance

- PREMIS http://www.loc.gov/standards/premis

Les formats de fichiers

Définition

Les standards de numérisation reposent essentiellement dans les formats techniques qui seront choisis selon les types de documents à numériser. Un format décrit la manière dont les informations sont organisées dans un fichier. Les formats techniques s'appuient sur des spécifications qui peuvent être propriétaires ou libres ou ouvertes.

Description

Les contenus numériques sont produits dans des formats techniques dédiés à l'issue d'une opération de numérisation. L'usage qui sera fait de la numérisation peut déterminer le standard de numérisation qui sera choisi : on distingue les formats de diffusion des formats d'archivage.

Les formats de fichiers sont spécifiques à la ressource qui peut être une image, un document textuel, un son ou une vidéo.

Ces différents types de fichiers peuvent être contraints par des critères spécifiques.
La compression est un critère influant sur tous ces types de fichiers.

Les fichiers « image »

Le format JPEG est le plus couramment utilisé pour la diffusion là où le format TIFF est plus répandu pour l'archivage.

À ces notions de formats techniques de fichier s'ajoutent également d'autres critères. À titre d'exemple, les questions de résolution d'image qui définissent le rapport de qualité de l'image à sa taille. La résolution de l'image se mesure en une valeur exprimée en dpi (dot per inch) ou ppp (point par pouce). Plus la résolution est importante plus l'image sera de bonne qualité. Il est important d'avoir en tête également que plus la résolution sera importante plus la taille du fichier sera importante. À titre indicatif, une résolution minimale pour un affichage sur écran d'ordinateur s'élève à 72 dpi. Les documents « image » ou « texte » de qualité normale auront une résolution de 300 dpi. Les ressources de très bonne qualité auront une résolution entre 600 et 1200 dpi.

Un autre critère est la compression des images. Un fichier au format JPEG sera beaucoup moins lourd qu'un fichier PNG en raison de la compression. LE JPEG opère une perte relative de qualité lors de la compression de l'image pour que celle-ci soit facilement diffusable tandis que le PNG compresse l'image sans altération et sans perte de qualité.

Les fichiers « texte »

Le format PDF, qui est un format propriétaire mais si largement répandu qu'il est devenu standard et reconnu par l'ISO (Organisation Internationale de Normalisation – International Standardisation Organisation), est utilisé pour les documents textuels généralement pour la diffusion.

La numérisation en mode texte permet de rechercher et/ou sélectionner du texte dans un document.

L'océrisation ou l'OCR (Optical Character Recognition – Reconnaissance Optique de caractères) est un des processus de numérisation qui permet de procéder à cette conversion en mode texte. Le document sera ensuite structuré selon un standard donné, par exemple ePub ou TEI.

Les fichiers « audio »

Les formats de fichier audio permettent de stocker et diffuser numériquement de la musique ou des captations sonores (parole ou sons). Le signal sonore peut être transformé en fichier et inversement par le biais de « codec » (abréviation de COder-DECoder). Les fichiers audio sont généralement volumineux mais la compression audio permet d'éliminer tous les sons non perceptibles à l'oreille humaine pour alléger la taille du fichier.
Plus la fréquence d'échantillonnage est élevée et plus le taux de compression sera faible, ce qui permet de

préserver la qualité sonore du fichier. Une fréquence d'échantillonnage à 192 kHz minimum permet une compression quasi sans perte de qualité. Les formats de fichiers qui dépendent en partie des codecs utilisés pour une numérisation de documents sonores peuvent être du MP3 ou du WAV.

Les fichiers « vidéo »

Les formats de fichiers vidéo permettent de stocker et diffuser des documents audiovisuels. Tout comme les fichiers « audio », les codecs audio et vidéo permettent d'encoder le signal pour le diffuser. Les notions de résolution propre à l'image ou de compression influent sur la qualité du document audiovisuel et la taille du fichier.

Les normes MP4 ou MKV sont par exemple des formats de fichier et conteneurs vidéos permettant d'encapsuler des contenus multimédias.

Enjeux

Les formats de fichiers standards jouent un rôle majeur dans le cycle de vie de la ressource numérisée et ce notamment parce qu'ils permettent de :

- garantir une bonne diffusion des ressources numérisées en privilégiant des formats techniques standards et largement répandus ;
- garantir des réutilisations par des tiers en privilégiant des bonnes résolutions d'images, des compressions sans perte ;
- garantir une meilleure interopérabilité en prvilégiant des formats de fichier non propriétaires.

Recommandations

Critères	Niveaux d'exigence
Utiliser des formats de fichiers recommandés par le RGI pour les ressources numérisées	Obligatoire
Contrôler la qualité de la numérisation (choix des formats et de la résolution)	Obligatoire
Se conformer au Référentiel Général d'Accessibilité pour les Administrations pour la diffusion des ressources numérisées	Obligatoire
Veiller à produire des ressources numérisées dans une résolution suffisamment élevée pour permettre des réutilisations	Recommandé
Utiliser des formats documentés	Recommandé
Préférer un format avec compression sans perte	Recommandé

En savoir plus

• Le référentiel général de l'interopérabilité présente une liste exhaustive et à jour des standards, notamment ceux de structuration de l'information (XML, Json, etc.) http://references.modernisation.gouv.fr/sites/default/files/Referentiel_General_Interoperabilite_V2.pdf

• Écrire un cahier des charges de numérisation de collections sonores, audiovisuelles et filmiques : https://francearchives.fr/file/9c1987da1084ddc8f81eae806042c6c57a7ec2fc/static_7929.pdf

• Les référentiels de numérisation de la BnF : http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.numerisation_referentiels_bnf.html

• La numérisation en mode image : http://www.bnf.fr/fr/professionnels/numerisation_boite_outils/a.outils_numerisation_mode_image.html

• OCR : http://www.bnf.fr/documents/ref_num_ocr.pdf

• Guide de publication Europeana : http://pro.europeana.eu/files/Europeana_Professional/Publications/Publishing_framework_summary.pdf

• Quelques formats de fichiers :

- JPEG : https://jpeg.org/
- TIFF : http://www.awaresystems.be/imaging/tiff.html
- PNG : http://www.libpng.org/pub/png/
- PDF : http://www.adobe.com/devnet/pdf/pdf_reference.html
- Epub : http://idpf.org/epub/30
- TEI : http://www.tei-c.org/index.xml
- MPEG-4: https://mpeg.chiariglione.org/standards/mpeg-4

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère
Culture

—

RECOMMANDATIONS TECHNIQUES POUR LES MÉTADONNÉES ET STANDARDS

VERSION N°1 – 2017

Ministère de la Culture
Secrétariat général
182, rue Saint-Honoré, 75033 Paris cedex 01