



PROGRAMME NATIONAL DE NUMÉRISATION ET DE VALORISATION DES CONTENUS CULTURELS

LES ENJEUX DES MÉTADONNÉES ET DES STANDARDS

Note de sensibilisation

Le lancement du Programme national de Numérisation et de Valorisation des contenus culturels (PNV) s'inscrit dans une stratégie numérique ministérielle résolument tournée vers les usages, la diffusion de la culture au plus grand nombre, le développement du numérique éducatif et l'émergence de nouveaux services en ligne. Ce programme¹, déconcentré au niveau des Directions régionales des affaires culturelles (DRAC), sera déployé progressivement au cours de la période 2018-2020. Dans cette perspective, le PNV devra prendre en compte les recommandations sur les métadonnées et les standards qui représentent un enjeu essentiel pour la numérisation des ressources culturelles.

Cette note s'adresse aux porteurs de projets de numérisation ainsi qu'aux personnes chargées d'évaluer ces projets. Il s'agit ici de les sensibiliser à l'enjeu des métadonnées et à l'importance des standards.

Le volume de contenus numériques est en constante croissance. Aux contenus numérisés ou en cours de numérisation s'ajoutent les contenus nativement numériques. Dans ce contexte il est crucial de faciliter l'accès aux documents numérisés.

L'enjeu des métadonnées est intrinsèquement lié à la problématique de la recherche et de l'accès à un fonds numérisé. Les métadonnées de qualité, qui respectent les standards du Web, garantissent un meilleur accès et un gain de visibilité certain pour les institutions qui les produisent. Elles améliorent ainsi les résultats de recherche en réduisant sensiblement bruit et silence. La démarche vers des métadonnées de qualité s'inscrit dans la mission de service public des institutions en faveur de la démocratisation culturelle et de la transmission des savoirs.

Les métadonnées accompagnent le cycle de vie à long terme du document numérique, depuis la production de la collection numérique et sa gestion jusqu'à son exploitation pour un usage interne ou externe.

Les métadonnées peuvent être produites automatiquement (métadonnées techniques de prises de vue par exemple) et/ou par des humains (par une institution ou de façon collaborative). Dans ce cas, elles reflètent des pratiques de traitement documentaire propres à chaque métier, qui évoluent avec les technologies.

Si les pratiques professionnelles changent, il en va de même pour les usages. Les modalités d'accès aux contenus se sont diversifiées, et les usagers recourent souvent à leurs canaux familiers d'accès à l'information (moteurs de recherche, réseaux sociaux, etc.). Ces canaux s'appuient sur des standards désormais tournés vers des logiques d'ouverture et de liage des données. Dans ce contexte, l'enjeu pour les institutions est de favoriser l'ouverture des contenus et des métadonnées qui leur sont associées, et de veiller à leur interopérabilité. Ces dynamiques d'ouverture et de liage des données portées par le ministère de la Culture² favorisent le dialogue entre les différentes logiques métiers à l'œuvre dans les institutions culturelles et tendent à « décloisonner » les données.

Dès lors, le rôle des standards de métadonnées et de formats de fichiers, qui permettent d'assurer interopérabilité, réutilisabilité et pérennité des contenus et des métadonnées, se voit renforcé.

Les projets de numérisation doivent s'inscrire dans cette dynamique. Des recommandations techniques, des orientations de formations et des critères d'évaluation, ainsi qu'un glossaire, accompagnent cette note de sensibilisation pour fournir des éléments de réflexion aux porteurs de projet et aux évaluateurs. Cet ensemble de documents vise à accompagner les acteurs culturels dans la démarche de production de métadonnées de qualité. Il vise aussi à les sensibiliser sur l'importance de la mise en place d'un contrôle qualité des métadonnées.

1 - Dans le cadre des groupes de réflexion pour le Programme national de Numérisation et de Valorisation des contenus culturels, cette note a été réalisée avec la précieuse collaboration de : Anila Angjeli, Bibliothèque nationale de France, Département des métadonnées - Laurine Arnould, ministère de la Culture, DGMIC/Département des bibliothèques - Rodolphe Bailly, Cité de la musique - Philharmonie de Paris, Département Éducation et Ressource - Olivier Baude, TGIR Huma-Num - Katell Briatte, ministère de la Culture, DGP/Département des systèmes d'information patrimoniaux - Sarah Brunet, ministère de la Culture, DGMIC - Bertrand Caron, Bibliothèque nationale de France, Département des métadonnées - Jean Davoigneau, ministère de la Culture, DGP/Mission inventaire - Thibault Grouas, ministère de la Culture, DGLFLF - Marie-Véronique Leroi, ministère de la Culture, SG / Département de l'innovation numérique - Carine Prunet, ministère de la Culture, DGP/Service des Musées de France - Romain Wenz, ministère de la Culture, DGP/Service Interministériel des Archives de France.
2 - <http://cblog.culture.fr/projet/2013/11/07/groupe-de-travail-metadonnees-culturelles/>
<http://cblog.culture.fr/wp-content/uploads/2013/07/Feuille-de-route-open-data-MCC.pdf>

Glossaire

Bruit : On parle de bruit (par opposition au silence) quand les résultats, proposés par un système d'information suite à une recherche dans une base de données, sont très nombreux et contiennent beaucoup de réponses non pertinentes masquant les réponses pertinentes. Voir « Silence ».

Collection numérique : Ensemble cohérent de documents ayant une valeur intrinsèque au-delà de la valeur de chaque document pris individuellement, et étant sous forme immatérielle, en général en vue de sa diffusion. Les documents peuvent être nativement numériques ou bien issus d'une numérisation.

Données ouvertes : Données qu'un organisme met à la disposition de tous sous forme de fichiers numériques afin de permettre leur réutilisation. Plus précisément, une donnée dite ouverte devrait répondre à dix critères : complète, primaire, opportune, accessible, exploitable, non discriminatoire, non propriétaire, libre de droits, permanente, et gratuite.

Fonds numérisé : Ensemble d'objets et de documents ayant fait l'objet d'une numérisation.

Formalisme : Voir « Modèle de données ».

Fichiers d'autorité : Les fichiers d'autorité servent à :

- normaliser la forme des noms de personnes, de collectivités, de titres d'oeuvres, des notices bibliographiques ;
- retrouver, par exemple, toute la production d'un auteur ou toutes les éditions d'une même œuvre quelle que soit la forme du nom ou la variante apparaissant sur les documents ;
- identifier sans ambiguïté l'entité.

Identifiant : Chaîne de caractères alphanumériques qui a pour fonction d'identifier un document, une ressource ou une entité quelle que soit sa nature.

Identifiant pérenne : Identifiant dont une organisation s'engage à assurer la gestion sur le long terme afin d'assurer à l'utilisateur un accès ou, à défaut, une information sur la ressource en question. La pérennité d'un identifiant est généralement déterminée par le mode de gouvernance du système dans lequel il a été attribué, c'est-à-dire par l'organisation choisie par le producteur de données pour attribuer et maintenir ses identifiants.

Indexation : L'indexation correspond à la représentation d'un texte ou d'un document par un index ou un ensemble de mots clés, avec l'aide ou non d'un langage documentaire, en vue d'en faciliter le repérage et la consultation.

Interopérabilité : Aptitude d'organisations disparates et diverses à interagir en vue de la réalisation d'objectifs communs mutuellement avantageux, arrêtés d'un commun accord, impliquant le partage d'informations et de connaissances entre elles grâce à l'échange de données entre leurs systèmes d'information et de communication respectifs.

Modèle de données : Un modèle de données est une description abstraite et formelle de l'organisation d'un ensemble de données. Un même modèle de données peut être mis en œuvre de différentes manières suivant les technologies utilisées (langages de balisage de texte, bases de données relationnelles, bases de données objet, etc.).

Open Data : voir « Données ouvertes ».

Référentiel : Groupement d'informations ou métadonnées ayant un haut niveau de partageabilité et constituant des éléments structurants d'un ou de plusieurs domaines de connaissance.

Référentiel d'autorité : Référentiel largement partagé et faisant référence dans le domaine concerné au niveau national ou international.

Référentiel de données d'autorité : Référentiels sur les auteurs d'œuvres.

Référentiels propres : Référentiels utilisés par une institution et faisant référence au sein de son système d'information.

Ressource : Toute entité identifiable et accessible dans un système d'information, par exemple un document électronique, une image ou une collection d'autres ressources.

Silence : On parle de silence (par opposition au bruit) lorsque des réponses pertinentes ne sont pas proposées par le système d'interrogation de la base de données, alors qu'elles existent dans cette base. Voir « Bruit ».

Standard : Dans ce contexte on entend par « standard » un ensemble structuré de recommandations sur lequel s'appuyer pour produire et traiter les métadonnées de description des ressources numérisées.

Taxonomie : Méthode de classification permettant de nommer, de décrire, de regrouper et d'identifier des éléments d'un domaine

Thésaurus : Un thésaurus est une liste organisée de termes contrôlés et normalisés (descripteurs et non descripteurs) représentant les concepts d'un domaine de la connaissance, et reliés par des relations sémantiques (équivalence, hiérarchie, association, synonymie...). Un thésaurus peut constituer un référentiel.

Toile : Système de navigation hypertextuelle entre des documents (ou d'autres types de ressources) réparti géographiquement et structurellement sur le réseau Internet

URI (Uniform Resource Identifier) : Chaîne de caractères respectant une syntaxe particulière, servant à identifier de manière pérenne une ressource sur la toile et utilisée pour y accéder.

URL (Uniform Resource Locator) : Chaîne de caractères utilisée pour localiser les ressources d'internet (document html, image, son, fichier, etc.).

W3C (World Wide Web Consortium) : Le World Wide Web Consortium, abrégé par le sigle W3C, est un organisme de standardisation à but non lucratif fondé en octobre 1994, chargé de promouvoir la compatibilité des technologies de la toile.

Web : Voir « Toile ».

Web sémantique : Extension du web régie par les standards du W3C. Ces standards encouragent l'utilisation de formats de données et de protocoles favorisant l'ouverture et le liage des données entre elles.

Web des données liées : Voir « Web sémantique ».

Web 3.0 : Voir « Web sémantique ».



LES ENJEUX DES MÉTADONNÉES ET DES STANDARDS

VERSION N°1 – 2017

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