Séance 5a - Métadonnées, standards et vocabulaires

1. Pourquoi les métadonnées sont au cœur de la médiation patrimoniale

Toute démarche de médiation et de valorisation numérique du patrimoine repose sur un paradoxe fondamental : plus un objet est riche, complexe, chargé d'histoire, plus il est difficile à décrire de manière claire, partageable et intelligible. Une photographie, un objet archéologique, un moulage, un modèle 3D ou un mémoire de master ne parlent jamais d'eux-mêmes. Ils nécessitent un discours d'accompagnement, une contextualisation, une structuration de l'information.

Dans le monde numérique, ce discours ne passe pas uniquement par le texte visible à l'écran. Il repose, de manière souvent invisible pour l'utilisateur final, sur les **métadonnées**. Décrire une ressource, ce n'est pas seulement la nommer ou la commenter : c'est décider quelles informations sont pertinentes, comment elles sont organisées, et dans quel langage elles sont exprimées. Les métadonnées sont ainsi au cœur de la médiation numérique, car elles conditionnent à la fois la compréhension par le public, la circulation des données entre systèmes, et leur réutilisation scientifique.

Cette séance vise donc à montrer que les métadonnées ne sont ni un supplément technique, ni une formalité administrative. Elles constituent un outil intellectuel, un lieu où se cristallisent des choix scientifiques, disciplinaires et méthodologiques. Comprendre les métadonnées, c'est comprendre comment le patrimoine devient lisible, comparable et partageable dans l'espace numérique.

2. Qu'est-ce qu'une métadonnée ? Une définition contextualisée

Le terme « métadonnée » signifie littéralement « donnée sur la donnée ». Cette définition minimale est correcte, mais insuffisante. Dans un contexte patrimonial, une métadonnée est une **information structurée** qui permet de décrire, d'identifier, de contextualiser et d'interpréter une ressource.

Lorsqu'un objet est publié sur un site patrimonial, plusieurs niveaux d'information coexistent (cf. texte séances 2-3). Il y a ce que l'utilisateur voit directement - une image, un texte, une vidéo -

et ce qui permet à cette ressource d'exister dans un système : son titre, sa date, son auteur, son sujet, son statut juridique, ses relations avec d'autres objets. Ces informations ne sont pas accessoires ; elles constituent la trame documentaire de la ressource.

On distingue classiquement, sans que ces catégories soient étanches, des métadonnées descriptives, structurelles et administratives. Les métadonnées descriptives servent à identifier et caractériser la ressource ; les métadonnées structurelles permettent de comprendre ses relations internes ou externes ; les métadonnées administratives concernent la gestion, les droits, la conservation. Cette typologie est utile, mais l'essentiel est ailleurs : toute métadonnée est le résultat d'un choix.

Il est essentiel de comprendre que décrire n'est jamais un acte neutre. Choisir de renseigner un champ plutôt qu'un autre, décider qu'une information est centrale ou secondaire, adopter un terme plutôt qu'un synonyme, tout cela relève d'une interprétation scientifique.

Prenons l'exemple d'un objet archéologique. Le simple fait d'indiquer une « date », une « provenance » ou une « fonction » suppose :

- une méthode de datation,
- une définition du lieu (site, région, entité administrative),
- une interprétation fonctionnelle, parfois discutée.

Les métadonnées ne sont donc pas un simple habillage de l'objet : elles constituent une hypothèse formalisée, rendue explicite et partageable. Dans un environnement numérique, cette formalisation est indispensable, car elle permet la comparaison, la recherche transversale et l'analyse à grande échelle.

3. Les standards de métadonnées : parler un langage commun

Dès lors que les données sont destinées à circuler, à être partagées entre institutions ou réutilisées par d'autres chercheurs, une question devient centrale : comment s'assurer que les métadonnées produites localement seront compréhensibles ailleurs ? C'est à cette question que répondent les **standards de métadonnées**.

Un standard ne dicte pas ce qu'il faut penser ; il propose un **cadre commun d'expression**. Il définit un ensemble d'éléments, leurs intitulés, et leur signification générale. L'objectif n'est pas d'épuiser la complexité du réel, mais de rendre possible un minimum de compréhension partagée. Dans le contexte patrimonial, le standard fondamental est le **Dublin Core**. Conçu à l'origine pour décrire des ressources numériques de manière générique, le Dublin Core propose un ensemble limité d'éléments (titre, créateur, sujet, description, date, type, format, droits, etc.) capables de s'appliquer à des objets très différents.

3.1 Le Dublin Core : définition et genèse

Dublin Core Metadata Initiative (DCMI) est une organisation internationale créée en 1995 visant à proposer un schéma de métadonnées générique, interopérable et minimal, destiné à décrire des ressources numériques sur le Web. Le **Dublin Core Metadata Element Set (DCMES)** comprend **15 éléments** fondamentaux : Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. Ces éléments sont :

- **optionnels**,
- **répétables**,
- **non contraints syntaxiquement** à l'origine.

Dire qu'un élément est **optionnel** signifie qu'aucun champ du Dublin Core n'est obligatoire par principe. Une ressource peut être décrite avec très peu d'éléments, voire un seul, sans être considérée comme invalide. Ce choix découle d'une volonté initiale d'ouverture maximale : le Dublin Core devait pouvoir s'appliquer à des contextes très variés, depuis une simple page web jusqu'à des collections patrimoniales complexes. Cette souplesse a favorisé une adoption massive du standard, mais elle a aussi une conséquence directe : deux notices décrivant des objets comparables peuvent être très inégalement renseignées, ce qui complique les comparaisons et les traitements automatiques.

Le caractère **répétable** des éléments signifie qu'un même champ peut apparaître plusieurs fois pour une même ressource. Par exemple, un objet peut avoir plusieurs créateurs, plusieurs sujets, plusieurs dates ou plusieurs descriptions. Cette possibilité est essentielle pour rendre compte de la complexité réelle des objets patrimoniaux, qui sont rarement univoques. Toutefois, en l'absence de règles supplémentaires, cette répétabilité peut introduire une certaine ambiguïté : toutes les occurrences d'un même élément sont placées sur un même plan, sans hiérarchie ni distinction explicite entre leurs rôles respectifs. Un « créateur » peut ainsi désigner indifféremment un auteur intellectuel, un artisan, un atelier ou un photographe, sans que le standard lui-même permette de lever l'ambiguïté.

Enfin, le fait que les éléments du Dublin Core soient **non contraints syntaxiquement à l'origine** signifie que le standard ne fixe pas, dans sa version initiale, de règles strictes sur la forme des valeurs saisies. Une date peut être exprimée en chiffres, en toutes lettres, de manière approximative ou précise ; un lieu peut être indiqué par un nom, une abréviation ou une description libre ; un nom de personne peut suivre des conventions très différentes. Cette liberté reflète une approche pragmatique, adaptée aux usages du Web des années 1990, mais elle limite fortement la possibilité de traitement automatisé des données. Sans règles syntaxiques partagées, les machines ne peuvent pas interpréter finement les valeurs, ni les comparer de manière fiable.

3.2. Pourquoi le Dublin Core s'est imposé dans le patrimoine

Dans les années 1990–2000, les institutions patrimoniales (bibliothèques, archives, musées) font face à trois défis :

- **Hétérogénéité des données**
- **Besoin de diffusion sur le Web**
- **Manque de compétences techniques généralisées**

Le Dublin Core répond à ces contraintes car il est :

- **transversal** (indépendant des disciplines),
- **peu normatif** (contrairement à MARC, EAD, CIDOC CRM),
- **facilement exposable** via OAI-PMH et HTML meta.

Il devient ainsi un **socle minimal commun**, souvent utilisé :

- pour la **diffusion**,
- pour le **moissonnage**,
- comme **pivot d'échange**, mais **pas comme modèle scientifique profond**.

Dans les projets patrimoniaux, le Dublin Core sert principalement à :

- décrire des **objets numériques** (images, notices, PDF),
- exposer des **métadonnées minimales publiques**,
- assurer l'**interopérabilité entre plateformes**.

3.3. Le Dublin Core : un socle descriptif volontairement générique mais avec des limites

Le succès du Dublin Core tient précisément à sa simplicité. Il permet de décrire aussi bien un livre numérisé qu'une photographie, un objet de musée ou un mémoire universitaire. Cette polyvalence en fait un excellent point d'entrée pour des projets de médiation numérique, en particulier dans des contextes pédagogiques ou institutionnels.

Cependant, cette force est aussi sa principale limite. Le Dublin Core décrit, mais il ne modélise pas. Il permet de dire qu'un objet a un « créateur », une « date » ou un « sujet », sans préciser la nature exacte de ces relations. Il ne distingue pas, par exemple, entre un auteur intellectuel et un fabricant, entre une date de création et une date d'utilisation, ou entre un lieu de découverte et un lieu de conservation.

Dans un site Omeka, ces limites apparaissent rapidement dès que l'on cherche à décrire finement des objets patrimoniaux complexes. On est alors tenté d'ajouter des champs, de spécialiser les usages, voire de détourner certains éléments du Dublin Core. Cette pratique est

compréhensible, mais elle signale que l'on atteint les frontières du standard. Mais il montre rapidement ses limites. C'est précisément pour dépasser ces limites qu'émergent :

- **CIDOC CRM** (musées),
- **EAD / EAC-CPF** (archives),
- **LIDO** (exposition muséale),
- **EDM** (Europeana).

3.4. Le Dublin Core dans Omeka S : rôle et fonctionnement

Omeka Classic repose nativement sur le Dublin Core, mais avec une logique très spécifique. Dans Omeka S, le Dublin Core est un vocabulaire parmi d'autres, au même titre que CIDOC CRM, etc. Omeka S n'impose pas le Dublin Core comme modèle scientifique, mais comme socle minimal d'interopérabilité. Trois sont les raisons principales de cela :

- **Interopérabilité**
  - OAI-PMH
  - moissonnage par Isidore, Europeana, etc.
- **Lisibilité humaine**
  - titres, descriptions, droits compréhensibles
- **Compatibilité historique**
  - héritage d'Omeka Classic
  - attentes institutionnelles

Par ailleurs, les exports CSV / JSON "plats" reposent principalement sur le Dublin Core alors que la structuration scientifique réelle passe par des vocabulaires sémantiques (voir séance 6). C'est pourquoi les utilisateurs sont souvent déçus des exports ; les notices paraissent pauvres hors contexte.

Le Dublin Core n'est ni faux, ni mauvais, ni obsolète. Il est insuffisant pour penser le patrimoine, mais indispensable pour le publier. Il constitue à la fois une porte d'entrée, un langage commun, et un format de médiation. Il ne représente toutefois pas un modèle d'analyse ou encore une ontologie du patrimoine : dans cette logique, le Dublin Core relève principalement des métadonnées descriptives et administratives, et très marginalement des métadonnées structurelles, ce qui explique à la fois sa simplicité et ses limites expressives.

3.5. Le Dublin Core en contexte institutionnel : l'exemple de la BnF

L'exemple de la Bibliothèque nationale de France permet de comprendre comment le Dublin Core, conçu comme un standard volontairement simple et générique, devient réellement opérant dans un cadre institutionnel patrimonial. Sur son site, la BnF présente le Dublin Core non comme un modèle scientifique autonome, mais comme un format de diffusion et d'échange, utilisé notamment pour l'exposition de notices via des protocoles de moissonnage tels

qu'OAI-PMH. Chaque élément du Dublin Core y fait l'objet de recommandations d'usage précises (formes contrôlées pour les auteurs, recours à des vocabulaires normalisés pour les sujets, formats de date normalisés), montrant que la simplicité du standard n'exclut pas une forte discipline documentaire locale. Le Dublin Core apparaît ainsi comme un niveau minimal commun, destiné à rendre les données compréhensibles et réutilisables hors du système de catalogage interne, sans prétendre remplacer des formats plus riches ni des modèles conceptuels complexes. L'exemple de la BnF illustre donc clairement le rôle du Dublin Core dans les institutions patrimoniales : non pas un outil d'analyse ou de modélisation fine du patrimoine, mais un instrument de publication, d'interopérabilité et de médiation à grande échelle.

4. Vocabulaires contrôlés : stabiliser le sens des termes

Une autre difficulté majeure de la description patrimoniale tient au vocabulaire. Deux personnes peuvent employer des termes différents pour désigner une même réalité, ou le même terme pour des réalités légèrement différentes. Dans un système numérique, cette variabilité est problématique, car elle empêche les regroupements automatiques et fausse les recherches.

Les **vocabulaires contrôlés** répondent à ce problème en proposant des listes de termes normalisés, souvent organisés hiérarchiquement. Utiliser un vocabulaire contrôlé, ce n'est pas appauvrir le discours ; c'est au contraire stabiliser le sens des mots employés, afin qu'ils puissent être compris et réutilisés par d'autres.

Dans Omeka, l'usage de vocabulaires contrôlés permet de relier les métadonnées locales à des référentiels plus larges, qu'ils soient disciplinaires, géographiques ou institutionnels. Cette pratique prépare déjà, sans toujours le dire explicitement, une entrée dans la logique du Web sémantique.

5. Omeka et les métadonnées : un apprentissage progressif

Avec Omeka, les métadonnées ne sont pas seulement des champs de formulaire ; elles deviennent des propriétés explicitement définies, associées à des vocabulaires et à des URI. Mais avant d'en arriver là, l'apprentissage passe nécessairement par une phase plus descriptive, fondée sur des standards comme le Dublin Core.

Dans le cadre du master et des projets de valorisation du patrimoine universitaire, cette progression est essentielle. Elle permet de comprendre :

- pourquoi il faut décrire avant de relier,

- pourquoi un standard simple est souvent préférable au départ,
- et pourquoi les limites rencontrées ne sont pas des erreurs, mais des signaux d'évolution.

Conclusion : des métadonnées aux modèles conceptuels : un changement d'échelle

Nous avons donc vu que les métadonnées ne sont pas un simple outil de catalogage, mais le socle de toute médiation numérique du patrimoine. Elles conditionnent la manière dont les ressources sont comprises, partagées et mises en relation. En choisissant un standard comme le Dublin Core, on accepte volontairement une description générique, en échange d'une large interopérabilité et d'une grande facilité de mise en œuvre.

Mais à mesure que les projets gagnent en ambition scientifique - description fine des objets, prise en compte des temporalités, des acteurs, des événements, des relations complexes - les limites de cette approche deviennent visibles. Ce moment de friction est décisif : **il marque le passage d'une logique de simple description à une logique de modélisation (c'est ici que se concentre toute l'importance de votre travail sur le modèle de ressource présent dans Omeka S).**

Autrement dit, après avoir appris à décrire des ressources, se pose inévitablement la question suivante : comment exprimer formellement les relations entre ces ressources, et comment donner un sens explicite à ces relations dans un environnement numérique ? C'est précisément cette question qui ouvre la séance suivante, consacrée au **Web sémantique et aux ontologies**, et notamment au rôle du **CIDOC CRM** dans la modélisation des données patrimoniales.

Dans cette progression de la médiation à la description, de la description à la modélisation, et de la modélisation à l'interopérabilité, les métadonnées ne sont pas une fin en soi ; elles sont le **point de passage obligé** vers une compréhension plus profonde des données patrimoniales à l'ère du numérique.

Séance 5b - Formats de données et médiation numérique

Dans un projet de valorisation numérique du patrimoine, la question des formats de données surgit souvent de manière très concrète, presque pragmatique : il faut importer des données, les exporter, les transmettre, les réutiliser. Vous allez alors être confrontés à des formats comme CSV, JSON ou XML, qu'il arrive de manipuler sans toujours comprendre ce qu'ils impliquent du point de vue documentaire et scientifique.

Or, un format de données n'est jamais neutre. Il conditionne la manière dont l'information est organisée, ce qui est conservé ou perdu, et ce qu'il est possible de faire avec les données par la suite. Comprendre les formats, ce n'est donc pas apprendre une technique informatique supplémentaire, mais saisir les contraintes et les potentialités de la médiation numérique.

1. Le CSV : simplicité maximale, expressivité minimale

Le format CSV (Comma-Separated Values) est sans doute le plus familier. Il se présente sous la forme d'un tableau : des colonnes, des lignes, une valeur par cellule. Cette simplicité explique son succès. Le CSV est facile à produire, à lire, à ouvrir dans un tableur, et à échanger entre logiciels.

Mais cette simplicité a un coût. Le CSV repose sur une logique **strictement tabulaire**, qui suppose que chaque information puisse être ramenée à une cellule. Dès que l'on cherche à exprimer une relation complexe - par exemple plusieurs auteurs pour un même objet, plusieurs dates, des hiérarchies ou des liens entre ressources - le format montre ses limites. On est alors contraint de répéter l'information, de la concaténer dans une même cellule, ou de multiplier les tableaux, au risque de perdre la cohérence d'ensemble.

Dans un contexte patrimonial, le CSV est donc utile pour des opérations ponctuelles (imports, exports intermédiaires, nettoyage de données), mais il est structurellement fragile dès que l'on s'intéresse à la complexité des objets décrits.

2. Le JSON : une structure hiérarchique plus proche du réel

Le format JSON (JavaScript Object Notation) constitue une étape importante dans la structuration des données. Contrairement au CSV, il n'est pas tabulaire, mais **hiérarchique**. Il

permet d'organiser l'information sous forme de paires clé–valeur, imbriquées les unes dans les autres.

Cette organisation correspond beaucoup mieux à la réalité des objets patrimoniaux. Un objet peut avoir plusieurs auteurs, plusieurs dates, plusieurs représentations ; une notice peut contenir des sous-ensembles d'informations. Le JSON permet d'exprimer cette richesse sans recourir à des artifices.

C'est pour cette raison que de nombreux outils, dont Omeka, proposent des exports en JSON. Le format JSON permet d'organiser l'information de manière plus souple qu'un simple tableau : il peut regrouper plusieurs informations liées à une même ressource et refléter une certaine complexité.

Cependant, il est important de comprendre qu'un fichier JSON n'est pas, en soi, un modèle scientifique. Il s'agit avant tout d'un **conteneur de données**. Il indique comment les informations sont rangées, mais pas comment elles doivent être interprétées. Deux fichiers JSON peuvent ainsi être lisibles par un ordinateur tout en décrivant des objets très différents, avec des choix de champs et de structures qui ne sont pas comparables entre eux.

Autrement dit, le JSON aide à transporter et à organiser des données, mais il ne garantit ni que ces données soient décrites selon les mêmes principes, ni qu'elles puissent être directement comprises ou réutilisées par d'autres projets. Pour cela, il faut aller au-delà du format et s'appuyer sur des **standards de description** et des **vocabulaires partagés**, ce qui sera précisément l'objet des séances suivantes.

3. Le XML : structurer selon un modèle explicite

Le XML (eXtensible Markup Language) va encore plus loin dans la formalisation. Il permet non seulement de structurer les données, mais aussi de le faire **selon un modèle explicite**, souvent défini par un schéma (DTD, XSD). Dans le monde patrimonial, le XML est largement utilisé pour exprimer des standards documentaires complexes, comme l'EAD pour les archives ou le LIDO pour les musées.

Ce caractère normatif fait la force du XML, mais aussi sa difficulté. Produire ou exploiter du XML suppose d'accepter un cadre strict, parfois lourd, et une courbe d'apprentissage plus exigeante. En contrepartie, le XML permet une grande précision descriptive et une forte interopérabilité, à condition que le standard soit partagé.

Conclusions : exporter des données n'est pas publier sémantiquement

Exporter des données, même dans un format structuré comme le JSON ou le XML, ne signifie pas encore **publier sémantiquement**. Un export permet :

- de déplacer des données,
- de les sauvegarder,
- de les réélaborer ailleurs.

La publication sémantique, en revanche, suppose :

- des entités clairement identifiées,
- des relations explicites entre ces entités,
- des vocabulaires partagés,
- des identifiants stables.

Autrement dit, un fichier CSV ou JSON peut contenir des données patrimoniales, mais il ne constitue pas en soi un **discours scientifique structuré**. Il manque encore le cadre conceptuel qui donne sens aux relations.

Cette réflexion sur les formats permet de comprendre pourquoi les questions abordées précédemment - identifiants, métadonnées, standards - ne suffisent pas encore à exprimer toute la complexité du patrimoine. Elle prépare directement la séance suivante, où l'on ne se demandera plus seulement *comment structurer des données*, mais *comment modéliser formellement les relations entre les entités patrimoniales*.

C'est précisément à ce moment que s'ouvre la **séance 6**, consacrée au **Web sémantique** et aux **ontologies** (les vocabulaires d'Omeka S), qui proposent une autre manière de penser la publication des données : non plus comme des fichiers à exporter, mais comme un réseau de connaissances interopérables.