Les métadonnées: définition et présentation générale

Anthony Betremieux
Traitement Numérique Multilingue
Institut National des Langues et Civilisations Orientales (INaLCO)
Paris, 75007, France
anthony.betremieux@laposte.net

Résumé

Ce document présente un aperçu de la notion de métadonnées, en partant de sa définition, de l'usage qui est fait de ces métadonnées sur les services web, pour mettre en exergue un des points cruciaux des métadonnées dans la préservation des documents numériques, la notion d'interopérabilité.

1. Introduction

la généralisation de l'internet et la multiplication des supports numériques pour la rédaction de documents recquiert la mise en place d'outils qui en permettent une gestion et une utilisation ultérieures à des fins spécifiques de traitement ou d'archivage. Le document électronique pose d'autant plus de difficultés qu'il représente un objet fragile, soumis à de nombreuses contraintes (origines diverses, formats et contenus, contenu mouvant, accessibilité plus ou moins large dans le temps, etc.).

Devant toutes ces contraintes, la recherche d'un document pertinent, par exemple sur le web via un moteur de recherche, peut vite s'avérer des plus laborieux. L'utilisation de données sur des données (métadonnées) permet de résoudre un tel problème, en accompagnant le document de toutes sortes d'informations permettant de l'identifier plus facilement, comme l'auteur, la date de publication, l'éditeur, etc. Cette méthode permet l'ajout d'un contenu structurel et cognitif, sémantique, au document, permettant aux moteurs de recherche de ne pas se limiter au strict contenu de la page d'un document.

2. Présentation

2.1. Définition

le terme de “métadonnées” est construit sur la base de “données” auquel on lui préfixe “meta”, signifiant en grec “ce qui dépasse, englobe un objet, une science.” (*le Petit Robert*). Littéralement une métadonnée est donc une donnée à propos d'une autre donnée, ou, pour reprendre la définition qu'en donne Patrick Peccatte, *“un ensemble structuré d'informations décrivant une ressource quelconque.”*

La métadonnée représente de l'information structurée décrivant, expliquant et localisant la ressource et en en facilitant la recherche, l'usage et la gestion.

2.2. Typologie des métadonnées

De manière générale, on peut distinguer les métadonnées externes des métadonnées internes, selon que l'on a à faire à des métadonnées créées *a priori* en accompagnement de la ressource électronique (interne) ou que l'on a à faire à des métadonnées qui sont retrouvées et combinées *a posteriori* par des systèmes de recherche.

Selon une typologie de Marie-Elise Fréon, il existe 4 sortes de métadonnées qui s'inscrivent à l'intérieur de ce schéma métadonnées internes/externes:

Externes:	Internes:
- de types de bases de données externes aux ressources, utilisées séparément pour la recherche. - externes à la ressource mais fournies en même temps.	- encapsulées, fournies dans la ressource (ex/ le Dublin Core). - englobantes, pouvant inclure la ressource elle-même (entièrement ou partiellement).

3. Intérêt des métadonnées

Devant la multiplication de la publication des documents numériques, l'apparition des métadonnées présente de nombreux intérêts, parmi lesquels:

- faciliter la gestion et l'archivage de l'information (informations sur le cycle de vie des documents, gestion des collections de ressources, gestion des archives électroniques).
- Gestion et protection des droits (les droits de propriété intellectuelle, les droits d'accès à des pages web).
- Authentification d'un texte (encoder une signature électronique pour valider un texte sur Internet).
- Faciliter l'interopérabilité (partage et échange des informations).
- Faciliter la recherche d'information (décrire le contenu et les relations entre les fichiers d'un site, classer le contenu suivant un degré de difficulté ou un public cible, mieux référencer un site ou une page web).

3.1. référencement web et notion de méta-tag (html)

Obtenir un bon référencement sur Internet est une chose essentielle, améliorant la pertinence et l'exhaustivité des recherches, le tri et le filtrage des données. Le référencement suppose une indexation de toutes les pages du web.

Les termes crawler, spiders et robots, se rapportent tous à un logiciel d'indexation qui est conçu pour passer en revue des sites Web et pour télécharger l'information contenue dans ces sites.

L'information qu'il télécharge est le code source HTML.

Dans le cas des moteurs de recherche, ce code source est stocké dans une grande base de données et plus tard analysé et classé dans l'index des moteurs de recherche.

Le robot ne va pas accorder la même importance au texte brut contenu dans la page html, et aux balises META, qui sont des métadonnées placées dans l'en-tête de la page html, à l'intérieur de l'élément head. Cette implémentation dans html, bien que représentant une recommandation ancienne de Dublin Core (un des jeux de métadonnées) est néanmoins toujours valable.

On distingue deux types de méta tags :

Les métas NAME, permettant de décrire la page HTML :

```
<meta name="Nom du tag"
CONTENT="Attribut">
```

Les métas HTTP-EQUIV, permettant d'envoyer des informations supplémentaires au navigateur via le protocole HTTP :

```
<META HTTP-EQUIV="Nom du tag"
CONTENT="Attribut">
```

Il est possible de renseigner plusieurs métas les uns après les autres dans l'en-tête de la page.

Toutefois, cette utilisation des métadonnées souffrent de plusieurs inconvénients, comme une standardisation insuffisante des usages et valeurs des éléments, et une sous utilisation, tant par les webmasters que par les outils de recherche (à savoir les principaux moteurs de recherche).

Après avoir été utilisées de manière intensive, il est à noter que ces balises META sont de moins en moins prises en compte par les moteurs de recherche, devant faire face à une utilisation abusive de ces balises, devenant de véritables réservoirs à spam. A tel point que Google ne les utilise plus (ou si peu) dans son travail d'indexation.

3.2. Métadonnées et format xml

Les métadonnées sont toujours implémentées dans un langage structuré (X)HTML, XML, RDF, et échangées par des protocoles (HTPP par exemple).

Aujourd'hui, les recommandations encouragent une implémentation des métadonnées en XML. XML (*eXtensible Markup Language*) est un développement de la norme internationale SGML avec les apports du langage hypertexte de HTML. Mais, à la différence de HTML, XML ne mélange jamais la mise en forme de la structure d'un document. A ce titre, XML est très souvent utilisé pour l'encodage de ressources, et aussi donc pour l'implémentation des métadonnées destinées à l'échange.

XML présente plusieurs avantages:

- il est indépendant des logiciels et utilise UNICODE, permettant une grande interopérabilité et une utilisation dans de multiples environnements.
- Il permet une structuration logique du contenu du document.
- Existence de modèles standards et partageables de documents, DTD et,

dernièrement, le schéma XML.

- Le processus de transformation XSLT permet d'associer une feuille de style XSL, permettant une utilisation multiple des ressources XML.
- XML apparaît comme une simplification de l'implémentation sous SGML, dont il reprend seulement 10% des éléments.

La structuration XML des métadonnées permet donc une plus grande pérennité du contenu, une meilleure accessibilité, une interopérabilté facilitant une conversion pour d'autres utilisations comme la partage de données (alimentation de portails et catalogues, d'archives, analyses statistiques).

4. la question de l'interopérabilité

L'interopérabilité est une capacité juridique du citoyen d'utiliser l'informatique sans se soucier d'aspects techniques. Cette notion d'interopérabilité touche tous les domaines techniques de l'informatique, c'est *"la capacité d'échanger des données entre systèmes multiples disposant de différentes caractéristiques en terme de matériels, logiciels, structures de données et interfaces, et avec le minimum de perte d'informations et de fonctionnalités*[*1] *"*.

L'interopérabilité des ressources se comprend dans trois contextes de réalisation technique:

- une description des ressources avec des sémantiques communes issus de différents jeux de métadonnées standardisés
- un contexte générique d'implémentation de ces métadonnées dans des langages structurés comme XML ou RDF
- des protocoles informatiques d'échange de ces données normalisées comme par exemple HTTP

	Standards récents
Jeux de métadonnées	Dublin Core MARC-XML, MODS EAD LOM, ...
Cadre générique d'implémentation	XML RDF espace de nom URL
Protocoles	HTTP OAI-PMH SRU/SRW

1*National Information Standards Organisation (NISO), Understanding Metadata, 2004, http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

Dans les cas où les données structurées et non-structurées coexistent, on considère que les parties communes sont constituées par des "métadonnées", qui à l'origine furent des mots-clés (les meta tags) introduits dans le langage de balisage tes que SGML, HTML, etc.

Et on a vu que c'est le langage XML qui aujourd'hui est vu comme le langage permettant d'accéder à l'ensemble des ressources informatiques sur le web, par ces métadonnées, dans le cadre RDF défini par le W3C en 1999.

Mais il y a un frein à cette interopérabilité, c'est l'existence de tous ces standards qui, s'ils sont utiles dans les différents contextes de création et de gestion de documents numériques, n'en constituent pas moins un obstacle théorique et pratique en termes d'interopérabilité et de partages de données.

4.1. Dublin Core

L'augmentation sensible de l'information numérique s'est accompagnée d'un développement pléthorique de schémas de métadonnées qui, s'ils s'avèrent utiles à un groupe spécifique d'utilisateurs ou à un secteur de contenus particuliers, rend cependant impossible l'utilisation de ces métadonnées dans un environnement différent.

Certains jeux de métadonnées se présentent comme la solution à l'objectif d'interopérabilité, comme c'est le cas de Dublin Core qui, à la suite de son adoption par le gouvernement canadien, a été adopté par plusieurs gouvernements dans le monde utilisant des référentiels de métadonnées basées sur le Dublin Core.

Dublin Core se définit comme un jeu de base de 15 éléments, normalisé ultérieurement par l'ISO sous le numéro 15836-2003, et comprenant des éléments descriptifs du contenu, des éléments administratifs, des éléments instanciels, une cinquantaine de qualificatifs et des schémas d'encodage, soit syntaxiques, soit sémantiques.

Plusieurs syntaxes permettent une implémentation des éléments de Dublin Core:

- HTML et XHTML (intérêt limité car non utilisé par les moteurs de recherche actuels).
- XML
- RDF (utilisé en raison d'une interopérabilité plus général de ce format)

Le fait que le Dublin Core ait fait l'objet d'une standardisation de la part de l'ISO démontre bien le large consensus que Dublin Core a créé autour de lui. On trouvera ci dessous les 15 éléments de

descriptions formels de Dublin Core tel que définis par ISO 15836:

Contenu	**Propriété intellectuelle**	**Matérialisation**
Titre	Créateur	Date
Sujet	Editeur	Type
Description	Contributeur	Format
Source	Droits	Identifiant
Relation		
Couverture		
Langage		

Toujours davantage orienté vers la notion d'interopérabilité, Dublin Core a même pris soin de préciser le sens de ce qu'est un registre de métadonnées: c'est un *"système de gestion des métadonnées, c'est-à-dire un système formel qui fournit l'information d'autorité sur la sémantique et la structure de chaque élément. Pour chaque élément, le registre en donne la définition, les qualificatifs qui lui sont associés, ainsi que les correspondances avec des équivalents dans d'autres langues ou d'autres schémas."* (février 2004)

4.2. l'exemple de la directive INSPIRE

La directive Inspire exprime clairement la nécessité de procéder, à un échelon européen, à une harmonisation des données (géographiques en l'occurence), et notamment des métadonnées.

Parmi les obligations de la directive, il y a la fourniture des données selon des règles de mise en oeuvre communes, la constitution de catalogues de données (métadonnées), l'application de règles d'interopérabilité, et (entre autres) l'accès gratuit aux métadonnées. On mesure par le biais de ce projet européen toute l'importance accordée aux métadonnées, ainsi que toute l'importance accordée à une harmonisation de ces métadonnées, pour l'échange et le traitement des informations

Les enjeux de la directive INSPIRE sont de la plus haute importance pour permettre l'accès aux informations environnementales et le co-traitement d'informations de sources publiques multiples.

Le changement apporté à long terme par INSPIRE portera dans trois directions :

- amélioration de l'information sur les données en fournissant des méta-données de façon systématique et en respectant les dispositions résultant des règles de mise en œuvre,
- faciliter les échanges de données entre acteurs : l'information géographique numérique ne prend sa véritable dimension que lorsqu'elle est échangée, partagée et enrichie par ses divers utilisateurs. Cette étape de mutualisation est ainsi une véritable source d'économies,
- moderniser les méthodes de travail en utilisant des données numériques de qualité dans leurs activités quotidiennes.

Références bibliographiques

Bibliothèque et Archives Canada

[http://www.collectionscanada.gc.ca/cdis/012033-741-f.html]

Centre national de documentation pédagogique

[http://www.cndp.fr/standards/metadonnees/general.htm]

Educnet

[http://www.educnet.education.fr/dossier/metadata/metadonnees-normes-et-standards/metadonnees]

Appropriation par la Recherche des Technologies de l' IST (*Métadonnées et interopérabilité*)

[http://artist.inist.fr/rubrique.php3?id_rubrique=41]

Karen Morgenroth, *Les métadonnées démystifiées*, forum consacré aux métadonnées au Canada Institute for Scientific and Technical Information (CISTI)

[http://www.limsi.fr/Individu/pap/inalco/meta_donnees_demysthifiees_014005-05209-a-f.pdf]

National Information Standards Organisation (NISO), Understanding Metadata, 2004

[http://www.niso.org/standards/resources/UnderstandingMetadata.pdf]

Grand Dictionnaire Terminologique

[http://www.oqlf.gouv.qc.ca/ressources/gdt.html]