Université de Nice-Sophia Antipolis
Polytech' CiP1

Jeudi 15 février 2007
Durée: 1h30

Création et Manipulation de documents

(Hélène Renard / Sylvain Schmitz)

Travaux Dirigés – Séance n°3

---

1 Objectifs du TD

Métadonnées. Organisation des informations.

2 Métadonnées

2.1 Introduction

La version la plus simple pour définir les métadonnées serait de dire que ce sont des données sur les données. Un peu du charabia à vrai dire !

Les métadonnées décrivent une ressource d'information, elles permettent donc de répondre aux questions (qui, quoi, quand, où, pourquoi et comment) concernant un ensemble de données ou une ressource (par exemple : une page Web, un fichier informatique, une image, un objet). On peut alors dire que le terme de métadonnées est un terme « fashion » pour décrire le même type d'information que les bibliothécaires mettent depuis toujours dans les catalogues. Car imaginez que vous essayez de trouver un livre dans une bibliothèque sans consulter le fichier, ou bien encore un univers de 300 canaux de télévision sans guide télé.

Ces métadonnées sont structurées : une notice contenant des métadonnées est constituée d'un ensemble d'attributs, ou éléments, nécesaires pour décrire la ressource en question. Mais où trouve-t-on les métadonnées ? Reprenons notre exemple avec les bibliothécaires, et plus précisemment le catalogue de la bibliothèque. Ce catalogue contient un ensemble de notices de métadonnées comprenant des éléments spécifiques pour décrire un livre ou tout autre document que l'on trouve en bibliothèque : auteur, titre, date de création ou de publication, sujet et cote, permettant de le retrouver dans les rayonnages. Nous pouvons aussi trouver des métadonnées dans les documents XHTML, entre les balises `<head>` et `</head>`. Le lien entre une notice de métadonnées et la ressource qu'elle décrit peut donc être fait de deux façons :

- les éléments peuvent être contenus dans une notice séparée du document, comme c'est le cas pour une notice dans un catalogue de bibliothèque
- les métadonnées peuvent être intégrées dans la ressource elle-même

2.2 Exemple

Avec l'apparition de l'Internet et du Web, l'intérêt mondial pour les pratiques et standards de métadonnées a explosé. Vous avez sans aucun doute tous vu apparaître au moins une fois dans votre moteur de recherche préféré des milliers de pages en réponse à votre requête. La solution pour palier à ce problème est donc de faire une recherche avancée, par exemple basée sur des champs tels que auteur, titre etc. L'utilisation des métadonnées va donc vous permettre une recherche améliorée, une meilleure gestion du contenu ainsi qu'un meilleur accès à votre travail.

Avec l'explosion des métadonnées, il faut donc établir des règles de bases uniformes, une norme. C'est ce besoin de métadonnées descriptives standardisées que le « Dublin Core » veut combler.

3 Dublin Core

La norme de métadonnées du Dublin Core est un ensemble d'éléments simples mais efficaces pour décrire une grande variété de ressources en réseau. La norme du Dublin Core comprend 15 éléments répartis en trois catégories, tels que :

Contenu	Propriétés intellectuelles	Instanciation
Sujet	Créateur	Date
Titre	Éditeur	Format
Relation	Droits	Identifiant
Description	Contributeur	Langue
Source		
Couverture		
Type		

Prenons quelques exemples de ces métadonnées afin de mieux les expliquer :

Créateur

Nom créateur

Identifiant Creator

Définition L'entité principalement responsable de la création du contenu de la ressource.

Commentaire Exemples de Créateur incluent une personne, une organisation, ou un service. Typiquement, un nom du Créateur devrait être utilisé pour désigner cette entité.

Source

Nom source

Identifiant source

Définition Une référence à une ressource à partir de laquelle la ressource actuelle a été dérivée.

Commentaire La ressource actuelle peut avoir été dérivée d'une autre ressource source, en totalité ou en partie. Il est recommandé de référencer cette source par une chaîne de caractère ou un nombre conforme à un système formel d'identification.

Droits

Nom gestion des droits

Identifiant rights

Définition Information sur les droits sur et au sujet de la ressource.

Commentaire Typiquement, un élément Droits contiendra un état du droit à gérer une ressource, ou la reférence à un service fournissant cette information. Ces droits souvent couvrent les droits de propriété intellectuelle (IPR), Copyright, et divers droits de propriété. Si l'élément Droits est absent, aucune hypothèse ne peut être faite sur l'état de ces droits.

L'utilisation des métadonnées n'est pas sujette à un ordre établi. Chaque élément est optionnel et peut-être répété ; il possède aussi un nombre limité de qualicatifs, ces derniers permettant de raffiner la signification de l'élément.

Le but du Dublin Core est donc de rendre le travail de quiconque plus accessible à une large communauté.

4 La syntaxe

Nous allons voir dans cette partie les métabalises.

Les métabalises, en XHTML, sont constituées de deux parties pour chaque élément :

- `meta name=""`
- `content=""`

Vous trouverez ci-après quelques exemples de métabalises du Dublin Core :

- `<meta name="DC.title" content="Services - Ministère de l'Éducation, Gouvernement du Yukon" />`
- `<meta name="DC.creator" content="Gouvernement du Yukon" />`

Attention : DC.creator et non pas DC.CREATOR ou dc.CREATOR ou DC.Creator

Si des caractères non-ASCII sont requis, utiliser les mêmes conventions que dans le corps du document. Par exemple :

```
<meta name="DC.title" content="Chocolat chaud &agrave; la Cyril Lignac" />
```

Un exemple plus complet d'une page XHTML utilisant les métadonnées du Dublin Core :

```
<html>
<head>
  <link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
  <meta name="DC.identifier"
   content="http://www.normannia.info/document/vard1898.html" />
  <meta name="DC.creator" content="Vard, Adolphe" />
  <meta name="DC.title" content="FLEUR-DE-SUREAU" />
  <meta name="DC.publisher" content="Verneuil : Imprimerie Gentil" />
  <meta name="DC.date" content="1898" />
  <meta name="DC.description" content="45 pages" />
  <meta name="DC.subject" content="littérature" />
</head>
<body>En mémoire et comme gage...</body>
</html>
```

Il y a deux façons de « voir » le Dublin Core : le document numérique est une instance du document papier (ce qui est le cas de la description ci-dessus) ou la version du document numérique est indépendante de celle du papier et considérée comme une édition à part entière, ce qui modifierait le contenu des champs :

```
<meta name="DC.publisher" content="CRL de la Basse-Normandie" />
<meta name="DC.date" content="2004" />
<meta name="DC.source" content="Avranches : Imprimerie Gentil, 1898" />
```

Afin d'accroître la spécificité des métadonnées, le Dublin Core peut être qualifié. Il existe deux types de qualificatifs :

- *les qualificatifs d'élément.* Ils précisent la signification sémantique de certains éléments. Par exemple, une ressource peut présenter deux dates importantes, telles que la date de première parution et la date de publication sur le Web. Afin d'éviter toute ambiguïté, on peut attribuer les qualificatifs suivants à ces deux expressions relatives à une date : `DC.date.created` (première parution) et `DC.date.issued` (publication sur le Web).
- *les qualificatifs de valeur.* Ils précisent la valeur attribuée à un élément au sein d'un registre de métadonnées particulier. Ces qualificatifs peuvent préciser le mécanisme d'encodage normalisé auquel la valeur se conforme ou un vocabulaire sélectionné duquel la valeur est tirée. Par exemple, la valeur de la date suivante `2002-09-12`, qui est encodée en vertu de

la norme `ISO 8601`, sera lue comme suit : le `12 septembre 2002`, et non le `9 décembre 2002`.

```
<meta name="DC.date.created" content="2002-03-11" />
<meta name="DC.date.modified" content="2002-12-03" />
```

Tous les éléments du Dublin Core n'ont pas été donnés en détail ici. Afin d'obtenir de plus amples informations et pour vous aider à faire les exercices, aidez-vous des liens suivants :
`http://www.dublincore.org/`
`http://www-rocq.inria.fr/~vercoust/METADATA/DC-fr.1.1.html`

5 Exercices

Exercice n°1 : À partir du fichier `~hrenard/meta.xhtml`, extrayez :

1. l'ensemble des métadonnées
2. le titre
3. les mots-clés

Aide : `grep -o` retourne simplement la partie de la ligne qui matche l'expression rationnelle. Par exemple : `grep -o "<h1>.*</h1>"` retourne uniquement les balises `<h1>` et `</h1>` et le titre contenu. On peut extraire encore plus finement avec la commande `sed`.

Exercice n°2 :

1. Sur le site `http://opac.ge.ch/`, effectuez la recherche suivante : « des souris et des hommes steinbeck »
2. Sélectionnez la première réponse, en cliquant sur « Notice »
3. Créez les métadonnées de ce document
4. Mêmes questions en cherchant « Récit d'un rêve » sur le site `http://www.normannia.info/`

Exercice n°3 : (extrait d'un TP de l'Université de Montréal)
Vous êtes responsable de l'édition électronique pour l'IFLA (International Federation of Library Associations and Institutions). À ce titre vous êtes appelé à gérer une masse importante de documents qui, stockée sur votre serveur, est accessible au monde entier via Internet. L'emploi des métadonnées Dublin Core est l'une des solutions que vous avez retenues afin de faciliter le repérage sur le Web mais également à l'intérieur même de votre serveur qui ne compte pas moins d'un millier de documents.

1. Copiez le document `~hrenard/preservation.pdf` dans votre répertoire `CMDocs/`
2. Ouvrez ce document avec votre éditeur préféré (xpdf, acroread etc.) et faites-en une lecture documentaire. (*Une lecture documentaire n'est pas exhaustive, mais ciblée. Cela veut dire que vous ne lisez pas le document au complet mais observez plutôt certains éléments révélateurs sur le sujet, le genre et la structure du texte. Ces éléments sont : le titre, les sous-titres, l'introduction, la conclusion, les éléments mis en évidence par une typographie particulière et les illustrations, etc.*)
3. Identifiez les métadonnées pouvant décrire adéquatement le document.
4. Copiez le document `~hrenard/preservation.xhtml` dans votre répertoire `CMDocs/`
5. Insérez les métadonnées Dublin Core à l'intérieur du document XHTML `preservation.xhtml`
6. Visualisez le document `preservation.xhtml` à l'aide de votre nagivateur. (Par exemple : `firefox CMDocs/preservation.xhtml`) (Vous noterez qu'il s'agit d'une page XHTML basique, nous verrons plus tard comment la mettre en page)
7. En tant qu'archiviste, quelles métadonnées devez-vous collecter afin d'assurer la préservation de l'information contenue sur des disquettes, des CD-ROM ou tout autre support numérique ?