La description des documents électroniques diffusés sur le web : pour une recherche pertinente

Tarek Ouerfelli

La description des documents électroniques diffusés sur le web : pour une recherche pertinente

Tarek OUERFELLI

Institut Supérieur de Documentation
Campus Universitaire de la Manouba
2010 – Tunisie

Résumé :

Dans cet article, nous mettons l'accent sur les métadonnées et leur apport dans la description des ressources électroniques pour la recherche d'information sur le web. Ces éléments permettent de décrire le contenu d'un site et ainsi de mieux le référencer sur Internet. Les éléments méta seront développés à travers le standard.Dublin Core.

I- Introduction

Face à la grande prolifération de l'information électronique disponible sur le web, le défi consiste à créer des moyens pour faciliter l'accès aux documents afin d'y rechercher et d'en extraire rapidement l'information pertinente. Pour cela, nous avançons quelques éléments de réflexion sur la description des documents diffusés sur le web par les métadonnées. Les éléments méta permettent de produire de nouvelles connaissances sur le document, concernant à la fois son contenu et sa mise en forme, en vue de son utilisation pertinente.

Dans un premier temps, nous présentons la notion du document Web et sa structuration. Ensuite, nous dressons un état sur la description par les éléments méta et les usages de ces éléments ainsi que leur rôle dans un contexte de recherche d'information sur le web.
Les éléments méta seront abordés à partir du standard Dublin Core qui est devenu récemment une norme ISO (International Standard Organization). Nous présentons le schéma descriptif de ce standard et ses spécificités.

II- Document web et structuration

2.1 Le document Web

Document web, site web, document numérique diffusé sur le web, désignent la même notion à savoir « un objet structuré où les éléments de structure sont complémentaires et liés par une relation orientée » (Gagneux 02). Ainsi, le document web est avant tout un document numérique composé d'un ensemble de pages reliées entre elles par des liens hypertexte « hyperliens » qui permettent de naviguer d'une page à une autre. Le document web peut, outre du texte, comporter des images et des composantes multimédias (vidéo, son). Les pages du document sont structurées en format HTML (HyperText Markup Language), le format le plus généralisé de codage des documents du web.

2.2 Le langage HTML

HTML est un langage de structuration de documents web. C'est un langage de balisage « markup », il repose sur les balises pour structurer les documents. Il se borne ainsi à déclarer la présentation que doit avoir le contenu d'un document. Cette déclaration est multi plate forme et peut être lue sur tout système muni d'un navigateur. En effet, l'objectif principal du langage HTML est de produire des documents destinés à être diffusés sur le web, tout en permettant une consultation facile de ces documents indépendamment des plates formes et des logiciels. Ce qui facilite l'échange d'informations entre des systèmes informatiques de natures différentes.

Tout document HTML commence et se termine par une déclaration <html></html> et il est composée de deux parties :

- Une parties en-tête encadrée par les balises d'ouverture et de fermeture <head></head>. Cette partie contient les méta informations : les informations relatives au document dans sa globalité (exemple : l'auteur du document, la date de publication, l'objet du document, etc.).
- Une partie corps du document qui est encadrée par les balises d'ouverture et de fermeture <body></body>. Cette partie contient toutes les informations qui seront lisibles par l'internaute (le contenu du document affiché sur la page écran : le texte, les images, les formulaires, etc.). Le contenu de cette partie est structuré aussi par des balises selon la fonction demandée. Par

exemple, pour mettre un mot en gras il faudra l'encadrer par les balises . Pour signaler un titre, il s'agira de l'encadrer par les balises <Hn></Hn>, dont n est le niveau de titre de 1 à 6.

Dans ce travail, nous mettrons l'accent sur la première partie encadrée par les balises <head></head> qui contient les méta informations. Nous montrerons le type de données que peut comporter cette partie, ainsi que son utilité dans la description des documents web pour un accès facile et rapide au contenu.

III- La recherche d'information sur le web

Le web est l'application la plus connue de l'Internet. Elle permet d'accèder à l'information en provenance de différents sites à travers le monde. En ce sens, Michard (99) affirme que le web apporte une solution générale aux besoins d'accès à l'information à distance.

La recherche d'informations sur le web se fait à l'aide d'outils de recherche automatiques notamment les moteurs de recherche (altavista, google, etc.). Ce sont des logiciels puissants permettant de parcourir tout le web à la recherche de nouveaux sites pour les indexer et les intégrer dans leurs bases de données. Lorsque l'internaute formule sa requête via l'interface d'interrogation du moteur de recherche, ce dernier procède, par la suite, à la recherche dans les sites référencés dans sa base, pour fournir en sortie les documents en rapport avec la question posée.

Du point de vue documentaire, la performance de ces outils de recherche est bien inférieure à leur puissance informatique dans la mesure où les résultats d'une requête de l'utilisateur pourrait engendrer du bruit (documents non pertinents retrouvés), ou bien du silence (documents pertinents non retrouvés). Ce phénomène est dû principalement à la pratique d'indexation réalisée par ces outils, qui est considérée comme une « indexation plein texte en aveugle », au sens de Michard (99).

Pour dépasser ces limites et améliorer les résultats de recherche sur le Web, une solution peut se présenter. Elle consiste à structurer le web pour rendre explicites les relations sémantiques qui existent entre les unités d'information qu'il contient (Michard 99). C'est dans cette perspective que s'inscrit l'usage des métadonnées et le Dublin Core.

IV- La description des documents web par les métadonnées

4.1 Principe et usage des métadonnées :

Les métadonnées ou *metadata* peuvent être définies comme étant des données relatives à d'autres données *data about data* ou bien des données décrivant d'autres données. Ce terme est surtout utilisé pour désigner l'information lisible par machine concernant des fichiers de données. Il désigne une information référentielle sur des données numériques (Amerouali 99).
Les métadonnées sont des outils importants pour le développement de la description de documents électroniques sur Internet. Les usages de ces éléments peuvent être répartis en :

- Usage spécifique pour la description du document lui même.
- Usage générique pour faciliter l'affichage des documents par les navigateurs et l'indexation de ces documents par les moteurs de recherche tout en permettant une normalisation de la description des ressources électroniques dans un contexte réseaux.

A partir de l'usage des métadonnées, il convient de préciser que la description des documents web par ces éléments n'est pas un objectif final mais plutôt un moyen pour faciliter l'usage de ces documents dans une perspective de recherche d'informations. Dans ce cadre, plusieurs projets ont été engagés dans un objectif d'unification de la descriptions des documents web. Parmi ces projets, on trouve le Dublin Core.

4.2 Le Dublin Core

Origine :

C'est en mars 1995 que des chercheurs en informatique avec des spécialistes venant des bibliothèques et du domaine du codage des textes se sont réunis à Dublin (Ohio – USA) avec comme objectif de définir des propriétés pour la description des documents électroniques conservés en réseau.
Ce groupe a retenu un ensemble d'éléments susceptibles d'être intégrés aux documents électroniques afin de les identifier automatiquement. Ils constituent le DUBLIN CORE META DATA ELEMNT SET. Ces éléments sont connus sous le nom de Dublin Core.

Le Dublin Core vise depuis sa création à résoudre le problème de la description unifiée des ressources d'information électroniques et de leur localisation dans un contexte réseaux (Ben Henda 99). C'est dans cette perspective qu'il est devenu une norme ISO 15836 depuis février 2003.

Le schéma descriptif de DUBLIN CORE

Selon la norme ISO 15836, le Dublin Core propose une quinzaine d'éléments descriptifs. Ces éléments sont les suivants :

Title : nom donné à la ressource[1] par son auteur ou son organisme. Le titre c'est le nom par lequel la ressource est officiellement connue

Creator : entité (personne physique ou morale) responsable de la création du contenu intellectuel de la ressource. Il s'agit de l'auteur principal dans le cas d'un document écrit.

Subject : description du domaine sémantique par des mots clés ou des phrases ou un code de classification précisant le sujet de la ressource. L'usage de vocabulaires contrôlés pour cet élément est encouragé (Vercoustre 02).

Description : Description textuelle du contenu. Généralement cet élément contient un résumé descriptif sur le contenu de la ressource.

Publisher : entité responsable de l'édition et la publication de la ressource.

Contributor : personne (physique ou morale) qui a collaboré à la production du document , exemple : illustrateur, traducteur…

Date : date de création ou de publication de la ressource conformément au format ISO 8601 (AAAA-MM-JJ) ex : 2002-12-25, ou simplifiée 2002.

Type : catégorie à laquelle appartient la ressource : roman, poème, thèse, etc. Il est recommandé de choisir la valeur du type d'une liste contrôlée, par exemple la liste de types DCT[2] de Dublin Core (Vercoustre 02).

Format : c'est la matérialisation physique ou digitale de la ressource (texte, son, image). Cet élément peut être utilisé pour préciser le logiciel ou autre équipement nécessaire pour afficher la ressource.

Identifier : identification unique de la ressource par un URI (Uniform Resource Identifier) qui peut inclure l'URL (Uniform Resource Locator) ou l'ISBN (International Standard Book Number).

Language : langue du contenu intellectuel de la ressource sous forme d'un code. La valeur de l'élément langue doit respecter les directives en vigueur, c'est pour cela qu'il est recommandé d'utiliser les codes définis par le schéma du RFC 3066[3] (ISO 15836). Ce schéma donne un code à

[1] Ressource désigne le document web.
[2] [DCT1] List of Resource Types. Dublin Core Draft Working Group Report. http://dublincore.org/documents/type-rfc-review/ (dernière visite le 10 juin 2003)
[3] [RFC3066] Tags for the Identification of Languages, Internet RFC 3066.

chaque langue à deux ou trois caractères selon la norme ISO 639, et dans certains cas il sera suivi d'un code à deux caractères pour le pays. Par exemple « ar » pour l'arabe, « fr » pour le français et « en-GB » pour l'anglais utilisé en Grande Bretagne.

Relation : identificateur d'une seconde ressource ayant une relation avec la première

Coverage : la couverture spatiotemporelle de la ressource. Il est recommandé d'utiliser un vocabulaire contrôlé pour choisir la valeur de cet élément.

Rights : cet élément couvre les droits de propriété intellectuelle (copyright). Cet élément doit être mentionné pour préserver tous les droits des créateurs de la ressource. Si cet élément est absent de la description, aucune hypothèse ne peut être faite sur l'état des droits des différents créateurs.

Source : référence à une source à partir de laquelle le document est dérivé. Il est recommandé de référencer cette source par une chaîne de caractères.

Voilà donc la liste des éléments méta retenus par le Dublin Core. Il est à noter que chaque élément est optionnel et répétitif. De plus, ces éléments peuvent apparaître dans n'importe quel ordre. Nous pouvons constater aussi que la définition des éléments est purement sémantique : elle ne fait aucune hypothèse sur les langages formels et sur les outils logiciels qui peuvent être employés pour créer des descriptions, les associer aux ressources et les exploiter dans les moteurs de recherche comme l'affirme par ailleurs Michard (99).

Strcutration des métadonnées dans un document web

Le Dublin Core peut être parfaitement exploité notamment en utilisant les éléments META du langage HTML, dans la partie en-tête du document.

Les informations méta sont structurées comme suit :

<META NAME= " … " CONTENT= " … " >

Le groupe d'information souhaité est tout d'abord spécifié avec <META NAME= " … ", puis on saisit l'information correspondante à CONTENT= " … ".

Exemple de fichier utilisant les métadonnées selon le Dublin Core

```
<HTML>
<HEAD>
<TITLE>eXtensible Markup Language (XML) 1.0 : W3C
Recommandation 10-Feb-98</TITLE>
<META NAME= "DC.creator." CONTENT="BRAY, Tim">
```

http://www.ietf.org/rfc/rfc3066.txt (dernière visite le 10 juin 2003).

```
<META NAME= "DC.creator." CONTENT= "PAOLI, Jean”>
<META NAME= "DC.title" CONTENT= ”eXtensible Markup Language
(XML) 1.0 : W3C Recommandation 10-Feb-98">
<META NAME="DC.publisher" CONTENT="W3C”>
<META NAME= "DC.subject.keywords" CONTENT= "langage XML ;
document XML">


<META NAME="DC.language" CONTENT="en-us>
<META NAME="DC.date" CONTENT="1998-02-10">
<META NAME="DC.identifier"
CONTENT="http://www.w3.org/TR/1998/REC-xml-19980210.html">
</head>
</body>
……
</body>
</html>
```

D'après cet exemple, nous pouvons noter que seuls sept champs sont remplis, puisque les éléments peuvent être facultatifs. De plus, l'élément "DC. Creator " est répétitif, il a deux valeurs. Les éléments méta définis permettent d'indexer la ressource selon les données signalées. Par exemple, l'élément méta "DC. Subject " fournit au moteur de recherche les mots clés selon lesquels le contenu du document pourrait être indexé. Ce qui permettrait de décrire aussi fidèlement que possible le contenu du document, pour assurer une pertinence lors de la réponse à la requête de l'utilisateur, et d'éviter ainsi au maximum le bruit ou le silence dans les résultats.

V- Conclusion

Le grand défi de la recherche d'informations sur le web est de pouvoir cibler au maximum les résultats de recherche suite à une requête de l'utilisateur. Ce défi peut être franchi par une structuration de la description des ressources électroniques en utilisant les métadonnées. Ces éléments jouent un rôle de plus en plus importants dans la réussite ou l'échec d'un document web. C'est sur leur qualité que repose la pertinence de la recherche et la satisfaction de l'utilisateur.

Ainsi, on peut conclure sur l'importance de la normalisation du standard Dublin Core dans une perspective d'unification de la description des ressources, à l'ère de la prolifération de

l'information sur le web. Face à ce phénomène, d'autres projets de métadonnées ont été lancés, notamment avec l'apparition du langage XML (eXtensible Markup Language) considéré comme le langage le plus adapté aux nouvelles applications du web, par rapport à HTML, qui apparaît comme un langage limité pour l'élaboration de documents complexes (Fondin 98). Le développement d'XML depuis 1998 a conduit à la création du standard RDF (Ressource Description Framework), pour la description du contenu des documents web structurés en XML. Ce développement montre l'importance de plus en plus croissante des métadonnées dans la description et la diffusion des documents sur le web.

VI- Bibliographie

[Amerouali 99] Y. Amerouali.- *métadonnées basées sur des éléments de description de ressources et profils d'utilisateur*.- in : colloque ISKO, Lyon (France), 21 – 22 octobre 1999.- pp. 43 - 48.

[Ben Henda 99] M. Ben Henda.- *L'indexation par éléments méta dans le processus de référencent du texte arabe entre HTML4, Unicode et Dublin Core*.- in : colloque ISKO, Lyon (France), 21 – 22 octobre 1999.- pp. 105 - 111.

[Cassagne 01] F. Cassagne, R. Rampnoux.- *Initiation aux langages HTML et XML*.- Paris : Ellipses, 2001.- 96p.

[Fondin 98] H. Fondin.- *Le traitement numérique des documents*.- Paris : Hermès, 1998.- 382p.

[Gagneux 02] A. Gagneux, H. Emptoz.- *Le document Web, de la structure à la lisibilité*.- in : CIFED, Hammamet (Tunisie), 21 – 23 octobre 2002.- pp. 105 - 112.

[ISO 03] *The dublin core metadata element set*.- ISO TC46/ SC 4 N 515.-26-02-2003. http://www.niso.org/international/SC4/n515.pdf (dernière visite le 5 juin 2003)

[Lubkov 97] M. Lubkov.- SGML, HTML, XML, des normes pour les documents.- in : *Archimag*, n° 107, septembre 1997.- pp. 30-31.

[Michard 99] A. Michard.- *XML langage et applications*.- Paris : Eyrolles, 1999.- 361p.

[Vercoustre 02] A. M. Vercoustre.- *Eléments métadonnées de Dublin core, version 1.1 : Description de référence*.- mars 2002.
http://www-rocq.inria.fr/~vercoust/METADATA/DC-fr.1.1.html (dernière visite le 25 mai 2003).