FUNDP NAMUR

crids CENTRE DE RECHERCHE INFORMATION, DROIT ET SOCIÉTÉ FUNDP - UNIVERSITÉ DE NAMUR

Archives de l'État www.arch.be

L'écrit électronique : les règles de l'art
Vendredi 1er juin 2012 @ BRUXELLES – Fondation universitaire, Rue d'Egmont 11

De la matière première bien raffinée : métadonnées pour la description de l'écrit électronique

James M Turner, professeur

Université de Montréal
Faculté des arts et des sciences
École de bibliothéconomie et des sciences de l'information

Plan

- Introduction
- La description
- Découverte et accès
- Conclusions

Introduction

- Contexte technologique
- Environnement réseauté
- Documents structurés
- La matière première
- Prolifération
- Haut degré de structure

Contexte technologique

- Machine à écrire électrique (~1961)
- Machine à écrire électronique (~1989)
- Texteur dédié (années 1970, 1980)
- Micro-ordinateur, logiciels texteurs (~1980)

Environnement réseauté

- Suite à la 2e Guerre mondiale, l'écrit électronique s'installe peu à peu
- Suite au début de l'internet avec la normalisation du protocole TCP/IP (~1982), un essor
- L'arrivée du WWW (~1994) change la donne
- Depuis, rien n'est comme avant

Documents structurés

- Quoi, pourquoi, comment
- Trois composants
- Points de repère

Quoi, pourquoi, comment

- Spécificité : la structure du document est rendue explicite
- But : permettre aux machines d'effectuer l'analyse grammaticale des documents
- Moyen : balises, par exemple :

  <nom de ville>Bruxelles</nom de ville>
  <restaurant>Bruxelles</restaurant>
  ou encore orchestre, légume, chanson…

Trois composants

- Structure, déclarée en entête (DTD, puis schéma)
- Présentation, traitée par feuilles de style (*style sheets*)
- Données, entourées de balises (métadonnées)

Échantillon de texte balisé

```
            <div id="preamble">
                <h3><span>Le chemin vers l'&eacute;dification</span></h3>
                <p class="p1"><span>
sp&eacute;cifique aux navigateurs, des <acronym title="Document Object Model">DOM</
acronym>s incompatibles, et du manque de support des <acronym title="Cascading Style
Sheets">CSS</acronym> encombrent un long chemin sombre et morne.</span></p>
                <p class="p2"><span>Aujourd'hui, nous devons nous clarifier l'esprit et
nous d&eacute;barassez des pratiques pass&eacute;es. La r&eacute;v&eacute;lation de la
v&eacute;ritable nature du Web est maintenant possible, gr&acirc;ce aux efforts infatigables
des gens du <acronym title="World Wide Web Consortium">W3C</acronym>, du <acronym
title="Web Standards Project">WaSP</acronym> et des cr&eacute;ateurs des principaux
navigateurs.</span></p>
                <p class="p3"><span>Le Jardin Zen css vous invite &agrave; vous relaxer et
&agrave; m&eacute;diter sur les le&ccedil;ons importantes des ma&icirc;tres. Commencez
&agrave; voir clairement. Apprenez &agrave; utiliser ces techniques (bient&ocirc;t
consacr&eacute;es par l'usage) de mani&egrave;re neuve et revigorante. Ne faites qu'Un avec
le Web.</span></p>
            </div>
        </div>

        <div id="supportingText">
            <div id="explanation">
                <h3><span>Alors, de quoi s'agit&#8211;il?</span></h3>
                <p class="p1"><span>Il y a clairement un besoin pour les graphistes de
prendre les <acronym title="Cascading Style Sheets">CSS</acronym> au s&eacute;rieux. Le
Jardin Zen vise &agrave; exciter, inspirer, et encourager la participation. Pour commencer,
voyez quelques concepts choisis dans la liste. Cliquez sur n'importe lequel pour le charger
sur cette page. Le code HTML demeure le m&ecirc;me, et seule la feuille de style
ext&eacute;rieure change. Oui, vraiment.</span></p>
```


Un même fichier interprété par 3 feuilles de style

Points de repère

- SGML (ISO 8879:1986)
  - HTML (~1991), exprès pour le web, indiscipliné
  - XML (~1998), simplifier et faciliter la mise en oeuvre de SGML sur le web, favoriser l'intéropérabilité
  - XHTML > HTML5 (~2000-2008), permet syntaxes XML et HTML
  - XML est devenu la *lingua franca* des métadonnées

La matière première

- Vers la normalisation
- Les sciences de l'information
- Ensembles de métadonnées

Vers la normalisation

- Évolution des langages de marquage
- Encodage de texte (ASCII > Unicode)
- Grands projets (Gutenberg, Text Encoding Initiative, W3C)

Les sciences de l'information

- Traditionnellement, bibliothèques et archives
- Développement de codes de catalogage, de classification vers le début du 20e siècle
- Aujourd'hui, sauf exception, tous les travaux focalisés sur le web
- Ainsi, les métadonnées deviennent la matière première et l'ordinateur l'outil principal de travail

Ensembles de métadonnées

- Une communauté qui a développé beaucoup d'ensembles de métadonnées
- Exemples :

  MAchine Readable Cataloging (MARC)
  Metadata Authority Description Schema (MADS)
  Metadata Object Description Schema (MODS)
  Metadata Encoding and Transmission Standard (METS)
  Encoded Archival Description (EAD)

Prolifération

- Plus que jamais, il y a prolifération et convergence de technologies, chaque instance exigeant la gestion de ses contenus :

  téléphones, téléviseurs, iPod, iPad, jeux...

- Ainsi, une course constante pour organiser tout cela

Haut degré de structure

- Cette réalité se solde par de l'information de plus en plus structurée
- Dans ce nouveau monde et cette mer de données, le besoin de décrire l'écrit aux fins de découverte, préservation et accès
- Tout cela passe par des métadonnées

La description

- Catalogage
- Le Dublin Core
- Étiquetage, texte libre

Catalogage

- Très bref historique
- Catalogage et indexation
- Description de l'objet
- Livres, manuscrits, images, vidéo, son

Très bref historique

- Registres en format de livres
- Fiches en carton
- Le format MARC (années 1960) et l'automatisation
- Les OPACs
- Le web, la norme Z39.50

Catalogage et indexation

- Strictement parlant, description de l'objet
- Plus large : description et indexation (« catalogage par matières »)
- Strictement parlant, l'indexation est l'analyse intellectuelle du contenu
- Explication : l'utilisation de fiches de bibliothèque

Autre opération : la classification

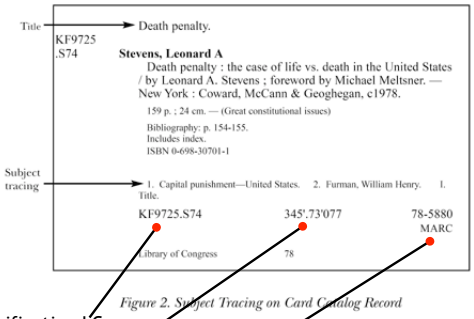

Figure 2. Subject Tracing on Card Catalog Record

Classification LC
Classification Dewey
Notice disponible en format MARC

Description de l'objet

- Petite anecdote OCLC/BNC
- Catalogage, indexation, classification : les trois exprimés sous forme de métadonnées maintenant
- Les notions demeurent distinctes :

  catalogage = description

  indexation = analyse, piste qui pointe

  classification = éléments ordonnés, ontologie

Divers objets

- Les normes de catalogage couvrent tout ce qu'on peut cataloguer
- Les normes bibliothéconomiques et archivistiques ont des chapitres pour images fixes, images animées, enregistrements sonores, fichiers informatiques
- En muséologie, normes pour la description de toutes sortes d'objets

Ensembles de métadonnées

- Tout cela s'organise aujourd'hui sous forme d'ensembles de métadonnées
- Divers intervenants en élaborent pour répondre aux besoins de leur communauté, par ex.

  Categories for the description of works of art (CDWA)
  VRA Core (Visual Resources Association)
  Encoded Archival Description (EAD)
  MPEG7 (images animées)
  Media Art Notation System (MANS)
  Darwin Core (sciences naturelles)

Le Dublin Core

- Un ensemble de métadonnées « passe-partout »
- 15 éléments, dont créateur, titre, langue, format
- Peut être utilisé en toute situation, pour tout objet
- Peu de formation nécessaire pour utiliser
- S'intègre aux grandes normes, dont le RDF
- Un noyau seulement, facilite découverte et accès

Étiquetage (*tagging*)

- Évolution
- Participation
- Rigueur
- Résultat
- Propagation automatique

Évolution

- L'étiquetage représente une évolution du traitement documentaire en environnement web
- Les gens ajoutent des mots-clés, leurs amis aussi
- Sujet de recherche fertile en sciences de l'information

Participation

- Participation d'internautes au catalogage, indexation, classification
- Projets pilotes dans musées, autres sites, maintenant assez courants
- Diverses formules, par ex. ajouts à l'indexation professionnelle, jeux d'étiquetage
- Amélioration de l'accès

Rigueur

- Peu rigoureux mais marche quand même assez bien
- Marche très bien, même, pour certains types d'informations :

  objets
  images de tous les jours
  images documentaires

Propagation automatique

- En combinant des approches, on peut propager automatiquement des étiquettes, par ex.

  St Denis de Paris, Portail de la Vierge

- Reconnaissance automatique d'une image semblable, ajout des étiquettes
- Via traducteurs automatiques, propager l'étiquette en plusieurs langues

Résultat

- Changement fondamental dans le traitement documentaire, exigé par la masse disponible
- L'ajout de métadonnées autrement omises
- Favoriser la découverte, l'accès
- Information accessible à une population beaucoup plus importante qu'avant

Découverte et accès

- Outils du web sémantique
- Atomiser, recombiner
- Où s'arrêter ?

Outils du web sémantique

- Découverte et accès
- Interopérabilité
- Harmonisation

Découverte et accès

- La croissance du web complexifie la découverte et l'accès
- Ainsi, les travaux focalisés sur le développement du web sémantique (rigueur, structure)
- Outils particuliers à chaque domaine, discipline, communauté
- Une difficulté importante : tout bouge, constamment

Interopérabilité

- On cherche à favoriser l'échange de fichiers entre usagers, systèmes
- Automatiquement, dans la mesure du possible
- Une certaine stabilité, postcompatibilité nécessaires
- Pour y arriver, harmonisation

Harmonisation

- On a tout intérêt à harmoniser les pratiques, méthodes
- Toutefois, cela n'est pas toujours faisable
- On le fait dans la mesure du possible, puis par la suite, on se fie à de multiples couches de métadonnées, filtres, traducteurs, passerelles
- Quelques outils : noyaux (*cores*), espaces de nommage (*namespaces*), grands contenants

Noyaux

- Ensembles de métadonnées critiques pour différentes communautés
- Quelques exemples :

  Dublin Core (général, toute documentation)
  IPTC Core (International Press Telecommunications Council, agences de presse, de photo)
  Darwin Core (informatique en biodiversité)
  VRA Core (Visual Resources Association, histoire de l'art)

Noyaux vs ensembles

- Un noyau comprend les éléments *essentiels* seulement
- Un ensemble de métadonnées est plus élaboré, peut comprendre des milliers d'éléments
- Les critiques des noyaux oublient souvent cette distinction
- La confusion provient en partie du fait qu'un noyau peut constituer l'ensemble au complet dans certaines situations

Espaces de nommage

- But : désambiguïsation d'identificateurs homonymes
- Un contenant qui sert de registre des espaces de nommage
- Espaces de noms XML (*XML namespaces*) permettent d'identifier outils précis lors de la description de ressources

Exemples de l'utilité

- Indiquer que le terme d'indexation « verre » provient de tel dictionnaire de matériaux de construction
- Indiquer que le « Montréal » en question est celui au Québec (il y en a en France, aux États-Unis, ailleurs)
- Désambiguïser auteurs avec un même nom

Grands contenants

- Outils comme le Resource Description Framework (RDF) et Material eXchange Format (MXF) permettent de ramasser bien des couches de métadonnées
- Comme un oignon ou une poupée russe, métadonnées gigognes

Atomiser, recombiner

- Avec les documents structurés et toutes ces balises, les documents peuvent être atomisés
- La musique et la vidéo aussi, pour les mêmes raisons
- Ainsi, on peut cibler des éléments intéressants, les extraire d'un document, les recombiner pour faire d'autres documents
- Édition, redocumentarisation, *remix*, *mashup*, etc.

Où s'arrêter ?

- La redocumentarisation : « retraiter un document ou un ensemble de documents numérisés de façon à les enrichir de métadonnées nouvelles et à réarranger et relier leurs contenus » (Salaün)
- Pour le traitement documentaire, comment faire maintenant, voilà la question !

Conclusion

- Le changement constant nous oblige à veiller constamment, être flexibles
- Suite à 15 ans d'expérimentation et recherches, il y a quand même une certaine stabilité maintenant
- Toujours vers le web sémantique, mais peut-on y arriver un jour ?

Références

Jacquet, Christophe. 2010. Métadonnées et Dublin Core. *OpenWeb Group, pour les standards du web*. <http://openweb.eu.org/articles/dublin_core>.

Jardin Zen CSS : la beauté de la conception CSS. 2012. <http://www.csszengarden.com/tr/francais/>

Salaün, Jean-Michel. 2008.03. Web, texte, conversation et redocumentarisation. *Actes des 9èmes journées internationales d'analyse statistique des données textuelles, Lyon, 12-14 mars 2008.* Presses universitaires de Lyon. <https://papyrus.bib.umontreal.ca/jspui/handle/1866/2226>

Merci

james.turner@umontreal.ca
http://mapageweb.umontreal.ca/turner/