Table Ronde du CampusXML

vendredi 28 février 2003

Les métadonnées

Patrick Peccatte
IPA Systems S.A.
www.ipa-france.com

Les métadonnées - Plan

Deux parties

- Partie "théorique"
  - présentation des concepts essentiels
- "Travaux pratiques"
  - présentation d'outils simples pour éditer, collecter et transformer les métadonnées

Les métadonnées – Plan détaillé

- Objectifs de la présentation
- Métadonnées – définition, utilité
- Métadonnées "métiers"
- Métadonnées informatiques - exemples
- Dublin Core Metadata Initiative
- RDF - Ressource Description Framework
- PRISM - Publishing Requirements for Industry Standard Metadata
- Relations avec d'autres spécifications: NewsML, NITF, etc.
- XMP - Extensible Metadata Platform
- Vers le Web sémantique
- **Partie II, T.P.** éditer, collecter et transformer les métadonnées

plus technique... [XML]

Objectifs de la présentation

- S'orienter dans le dédale des standards, recommandations et initiatives…
  - Dublin Core ?
  - RDF ?
  - XMP ?
- Comprendre comment **XML** [**RDF**] fournit un cadre adapté à la gestion et à l'échange des métadonnées et constitue la base de la réutilisation des données
- Donner une connaissance pratique de la gestion des métadonnées informatiques associées aux fichiers

Une première définition

- Une *métadonnée* est littéralement une *donnée sur une donnée*
  - Dans le domaine des *métadonnées* [*metadata*], on parle de *données* sur une *ressource*
- Plus précisément, c'est un *ensemble structuré* de données décrivant une *ressource* quelconque
- Une *métadonnée* peut être utilisée à des fins diverses…
  - la *description* et la *recherche* de ressources
  - la *gestion* de collections de ressources
  - la *préservation* des ressources

Utilité des métadonnées [1/2]

- Les métadonnées sont *en général* constituées de mots-clés ou de texte libre Pas toujours! Les métadonnées sont en fait des données quelconques
- Ces informations peuvent être évidentes (l'*auteur*, la *date de publication*, l'*éditeur* d'un livre), ou plus complexes et moins aisément définies
  - les avis d'un collectif de lecture d'un article, par exemple, nécessitent une structure de métadonnées évoluée capable d'annoter des portions de l'article, et cela, de façon multiple
- Les métadonnées sont particulièrement importantes pour les ressources visuelles qui, sans elles, peuvent demeurer pratiquement inexploitables et impossibles à retrouver
  - les utilisateurs dépendent en effet des informations ajoutées aux images ou vidéos pour effectuer des recherches pertinentes et précises

Utilité des métadonnées [2/2]

- La recherche de documents à l'aide de leur indexation *full-text* ne suffit pas
  - Exemple (gag): rechercher sur Internet les entreprises américaines spécialistes des *portes* et *fenêtres*…
- Les métadonnées sont également indispensables d'un point de vue technique et administratif
  - pour décrire les *droits*, *relations*, *formats*, *dates*, etc. associés à une *ressource*, l'appartenance à une collection digitale, l'acquisition de la *ressource*, etc.
- Les métadonnées sont utilisées dans les systèmes de gestion de contenu [*content management*]
  - pour éditer, gérer, rechercher, réutiliser, diffuser, publier de multiples contenus (textes, images, vidéo, etc.)

Métadonnées "métiers" [1/2]

- Les *ressources* décrites par des métadonnées ne sont pas nécessairement sous forme digitale
  - un catalogue de bibliothèque ou de musée contient aussi des métadonnées
- De nombreuses communautés s'intéressent aux métadonnées
  - bibliothécaires, documentalistes, archivistes, conservateurs de musées, …
  - …gèrent de nombreux types de *ressources*

Métadonnées "métiers" [2/2]

- ressources: monographies, publications en série, articles, archives, pièces de musée, séquences audio ou vidéo, etc.
  - on ne décrit pas toutes ces variétés de *ressources* de la même façon
- apparition de standards de métadonnées "métiers"…
  - MARC (Machine-readable cataloging)
  - ISBD (International Standard Bibliographic Description)
  - Dewey Decimal Classification system
  - EAD (Encoded Archival Description)
  - CIMI consortium (Computer Interchange of Museum Information)
  - RKMS (Recordkeeping Metadata Schema)
  - MPEG-7 (Multimedia Content Description Interface)
  - LOM (IEEE - Learning Object Metadata)
  - SCORM (Sharable Content Object Reference Model)

bibliothèques
- catalogage
- indexation
- classification

archives

musées

audio

éducation

Métadonnées - où en est-on ?

Objectifs de la présentation
Métadonnées – définition, utilité
Les métadonnées "métiers"
- **Métadonnées informatiques - exemples**

Dublin Core Metadata Initiative
RDF - Ressource Description Framework
PRISM - Publishing Requirements for Industry Standard Metadata
Relations avec d'autres spécifications: NewsML, NITF, etc.
XMP - Extensible Metadata Platform
Vers le Web sémantique
T.P. éditer, collecter et transformer les métadonnées

Métadonnées informatiques [1/7]

- Où sont les métadonnées informatiques ?
  - dans les bases de données
  - dans les données elles-mêmes
    [les métadonnées sont "embarquées" dans les données]
- Considérons la *ressource* suivante:
  http://www.liberation.fr/livres/index.php
- Cette *ressource* contient plusieurs métadonnées
  - Protocole *http*
  - Site *liberation.fr*, top level domain *fr*
  - Page Web dynamique écrite en *php*, index de la rubrique *livres*
- Les noms informatiques sont des *métadonnées*
  - Plus généralement: chemin d'accès, nom, extension, taille, attributs, date de création, date de modification, propriétaire, droits d'accès, etc. sont des *métadonnées*

Métadonnées informatiques [2/7]

- Champs <title> et <meta> des fichiers HTML
- Exemple

```
<title>Le Monde.fr : Sean Penn : "Je ne souscris pas à une conception simpliste du bien et du mal</title>
<meta NAME="ROBOTS" CONTENT="INDEX,FOLLOW,NOARCHIVE">
<meta name="DESCRIPTION" content="LE MONDE, Journal Le Monde, quotidien d'information francophone / Le Monde, the french quality newspaper of record">
<meta name="KEYWORDS" content="LE MONDE, INFORMATIONS, INFOS, QUOTIDIEN, DAILY NEWS, PRESSE, PRESS, NEWS, FRANCE, FRENCH, DOSSIERS, ECONOMIE, ECONOMY, CULTURE, INTERNATIONAL, BOURSE, CINEMA, MOVIES, LIVRES, BOOKS, MULTIMEDIA, EDUCATION, FORUMS, FORUM, SERVICES, ABONNEMENTS, BOUTIQUE, EMPLOI, EXPOSITIONS, FESTIVALS, SPORT, MAGAZINE, EUROPEEN, DIPLOMATIQUE, PARTENAIRES, PUBLICITE, LETTRES D'INFORMATIONS, NEWSLETTERS, JOURNAL EN LIGNE, LE MONDE ON LINE, VERSION PALM, VERSION MOBILES, MOBILE SERVICES, METEO, ARCHIVES, DOCUMENTATION, NOUVELLES TECHNOLOGIES, HIGH TECH, TRADUCTEUR, TRANSLATOR">
```

Métadonnées informatiques [3/7]

- Propriétés des documents MS Office (Word, Excel, etc.)
  - Titre, Auteur, Sujet, Mots-clés, Commentaires, Responsable, Société, Catégorie, etc. [25 éléments]
  - Possibilité de propriétés personnalisées
- Propriétés des documents OpenOffice.org
  - Titre, Description, Sujet, Mots-clés, Créateur initial, etc. [25 éléments]
  - Possibilité de propriétés personnalisées [4 au maximum]

la terminologie est différente

Métadonnées informatiques [4/7]

- Informations sur les documents PDF
  - Titre, Auteur, Sujet, Mots-clés, Créateur, Producteur, etc. [9 éléments]
- Champs IPTC des images JPEG/TIFF
  - Titre, Source, Crédit, Copyright, Statut éditorial, Priorité, Catégorie, Mots-clés, etc. [33 éléments]
- Champs EXIF des images JPEG
  - Fabricant de la caméra, Modèle, Orientation, Temps d'exposition, Résolution en largeur, Résolution en hauteur, etc. [30 éléments]
- Champs ID3 des fichiers MP3
  - Titre, Compositeur, Auteur du texte, Durée, Copyright, etc. [74 éléments organisés en frames]

Métadonnées informatiques [5/7]

- Métadonnées informatiques non textuelles
  - Image haute définition [*ressource*]
  - Image basse définition [*métadonnée 1*]
  - Vignette [*métadonnée 2*]

Métadonnées informatiques [6/7]

- Métadonnées spécifiques à chaque plate-forme…
  - Macintosh
    Famille (Essentiel, Important, En cours, Personnel, etc.) et *Commentaires*
  - Windows 2000 & XP
    Propriétés associées à un fichier quelconque (Titre, Sujet, Catégorie, Mots-clés, etc.)

Métadonnées informatiques [7/7]

- Estampillage électronique [*Watermarks*]
  - But - authentifier un document (garantie de non-falsification) et prouver l'appartenance d'une œuvre à son propriétaire
  - Moyen - Filigrane, tatouage, estampillage, etc. insertion d'informations numériques dans les fichiers binaires que sont les images, sons, vidéo

dans le domaine des *métadonnées*???

- Stéganographie
  - science qui consiste à cacher de l'information dans un quelconque medium de façon à ce que seul un utilisateur muni du secret adéquat puisse retrouver cette information

n'est pas dans le domaine des *métadonnées*

Métadonnées - où en est-on ?

Objectifs de la présentation
Métadonnées – définition, utilité
Les métadonnées "métiers"
Métadonnées informatiques - exemples
Dublin Core Metadata Initiative
RDF - Ressource Description Framework
PRISM - Publishing Requirements for Industry Standard Metadata
Relations avec d'autres spécifications: NewsML, NITF, etc.
XMP - Extensible Metadata Platform
Vers le Web sémantique
T.P. éditer, collecter et transformer les métadonnées

Dublin Core Metadata Initiative [1/5]

- Prolifération de besoins "métiers" variés (musées, bibliothèques, archives,...)
- Métadonnées informatiques: diversité et non-interopérabilité des *nomenclatures* et des *structures*
  - Recherche d'un *standard* (une image peut posséder 5 "Descriptions" différentes)
- NCSA (National Center for Supercomputing Applications)
- OCLC (Online Computer Library Center)
  réunis en 1995 au siège de l'OCLC à Dublin, Ohio
  - Définition d'un ensemble de métadonnées communes à diverses communautés:
    le *Dublin Core Metadata Initiative* (DCMI).

Dublin Core Metadata Initiative [2/5]

- Le *Dublin Core* est un ensemble de 15 éléments de métadonnées ayant trait:
  - au Contenu
    Title, *Description*, *Subject*, *Source*, *Coverage*, *Type*, *Relation*
  - à la Propriété intellectuelle
    ~~*Creator*~~, *Contributor*, *Publisher*, *Rights*
  - à la Version
    Date, *Format*, *Identifier*, *Language*

Creator et non pas **Author**

DC définit un vocabulaire de métadonnées *commun* à plusieurs communautés

Dublin Core Metadata Initiative [3/5]

- Une version plus évoluée du *Dublin Core* autorise l'usage de qualificateurs
- Exemples:
  - l'élément Description peut être raffiné à l'aide des qualificateurs tableOfContents et abstract
  - Date peut être raffiné à l'aide des qualificateurs Created, Valid, Available, Issued, Modified

Dublin Core Metadata Initiative [4/5]

- Les éléments du *Dublin Core* peuvent être encodés dans des balises HTML <meta>
- Exemple

CISMeF [Catalogue et Index des Sites Médicaux Francophones]

```
<meta name="DC.Language" content="fr">
<meta name="DC.Title" content="CISMeF">
<meta name="DC.Title.Subtitle" content="Catalogue et Index des Sites Médicaux Francophones ; Catalog and Index of French-speaking Health Resources">
<meta name="DC.Type" content="texte.guide ressources">
<meta name="DC.Subject.Keywords" content="(SCHEME=MeSH) France ; Internet ; médecine ; santé ; medicine ; health">
<meta name="DC.Creator" content="équipe CISMeF : Benoit Thirion ; Stéfan Darmoni ; Florence Baudic ; Magaly Douyère ; Jean-Philippe Leroy ; Josette Piot">
```

DC "simple" DC "qualifié"

Dublin Core Metadata Initiative [5/5]

- Le *Dublic Core* ne prétend pas répondre aux besoins et à la complexité de tous les métiers
- Le *Dublin Core* est un ensemble simple et très utilisé de métadonnées (en cours de normalisation ISO 15836), mais il n'est pas suffisant
- Dans la plupart des besoins professionnels, il doit être complété par d'autres schémas de métadonnées

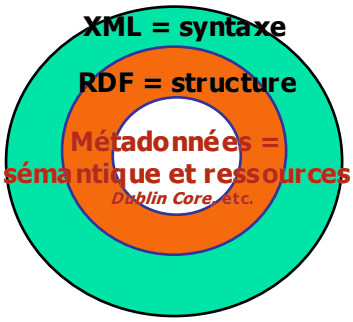
Métadonnées - où en est-on ?

Objectifs de la présentation
Métadonnées – définition, utilité
Les métadonnées "métiers"
Métadonnées informatiques - exemples
Dublin Core Metadata Initiative
RDF - Ressource Description Framework
PRISM - Publishing Requirements for Industry Standard Metadata
Relations avec d'autres spécifications: NewsML, NITF, etc.
XMP - Extensible Metadata Platform
Vers le Web sémantique
T.P. éditer, collecter et transformer les métadonnées

RDF – Resource Description Framework [1/9]

- **RDF** est un moyen d'encoder, échanger et réutiliser des métadonnées structurées
- **RDF** est un idiome **XML** développé par le W3C (Recommandation en 1999)
- **RDF** ne précise pas la sémantique des ressources décrites par les différentes communautés d'utilisateurs de métadonnées
  - **RDF** est un cadre [*framework*] de description des ressources pour n'importe quel domaine d'application
- **RDF** est un langage *extensible*

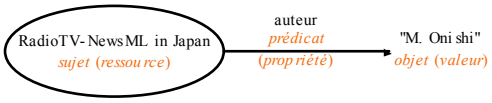
RDF – Resource Description Framework [2/9]



d'après
Julia Innes
Rory McGreal
Toni Roberts
Téléé**ducation NB**

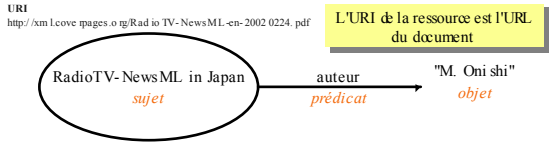
RDF – Resource Description Framework [3/9]

- **RDF** est basé sur des triplets
  sujet - **prédicat** - **objet**
  ou ressource - propriété - valeur
- Exemple
  Le document RadioTV-NewsML in Japan a pour auteur M. Onishi
  sujet | prédicat | objet
- Modélisé à l'aide de graphes orientés étiquetés



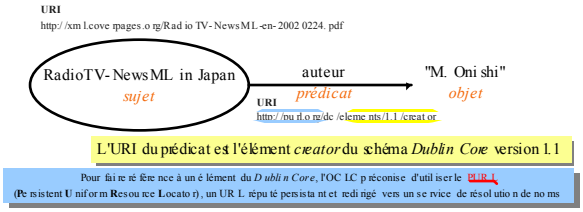
RDF – Resource Description Framework [4/9]

- Les *ressources* sont identifiées par des **URI** (*Unified Ressource Identifier*)
- Les **URI** sont un "stock de noms" utilisés pour désigner des choses ou des concepts
- Les **URL** sont des **URI**



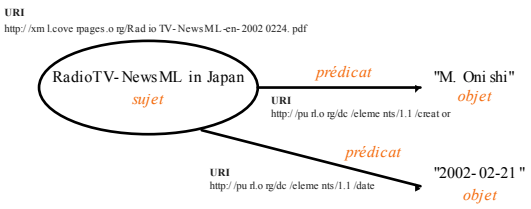
RDF – Resource Description Framework [5/9]

- Les *prédicats* (*propriétés*) sont également représentés par des **URI**



RDF – Resource Description Framework [6/9]

- Un *sujet* (*ressource*) peut posséder plusieurs *prédicats* (*propriétés*)



RDF – syntaxe XML [7/9]

conteneur *RDF*

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/">
    <rdf:Description
rdf:about="http://xml.coverpages.org/RadioTV-NewsML-en-
20020224.pdf">
        <dc:creator>M. Onishi</dc:creator>
        <dc:title>RadioTV-NewsML in Japan</dc:title>
        <dc:date>2002-02-21</dc:date>
        <dc:type>Text</dc:type>
        <dc:format>application/pdf</dc:format>
    </rdf:Description>
</rdf:RDF>
```

utilisation des espaces de noms *rdf* et *dc*

about précise la **ressource** à décrire

valeur *M. Onishi*

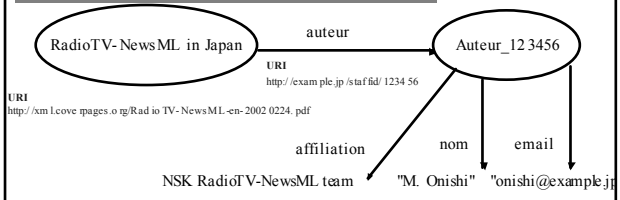
propriété *creator*

RDF – Resource Description Framework [8/9]

- Les ressources décrites peuvent être imbriquées

l'auteur n'est pas une valeur mais une autre **ressource**

la **ressource** *auteur* est identifiée par une **URI** propre à l'entreprise



RDF – Schémas [9/9]

- Un **Schéma RDF** permet de décrire un vocabulaire et une sémantique des types de *propriétés* utilisées par une communauté d'utilisateurs
- Un **Schéma RDF** précise les *propriétés* valides pour une description RDF particulière, ainsi que les caractéristiques et contraintes du vocabulaire descriptif
- Distinguer…
  - **Schéma XML** (≈ DTD) - contraintes sur la structure et la syntaxe XML
  - **Schéma RDF** - contraintes sur la sémantique des expressions d'un modèle RDF
- Exemple – le **schéma RDF** du Dublin Core

Métadonnées - où en est-on ?

- Objectifs de la présentation
- Métadonnées – définition, utilité
- Les métadonnées "métiers"
- Métadonnées informatiques - exemples
- Dublin Core Metadata Initiative
- RDF - Ressource Description Framework
- **PRISM** - Publishing Requirements for Industry Standard Metadata
- Relations avec d'autres spécifications: NewsML, NITF, etc.
- XMP - Extensible Metadata Platform
- Vers le Web sémantique
- **T.P.** éditer, collecter et transformer les métadonnées

PRISM – Publishing Requirements for Industry Standard Metadata [1/3]

- **PRISM** est un idiome **RDF** extensible permettant de décrire les métadonnées utilisées dans la presse
- **PRISM** a été initié par un groupe de travail *IDEAlliance* (International Digital Enterprise Alliance) fondé en 1999
- **PRISM** est un "vocabulaire commun" destiné à décrire les contenus, l'origine de ces contenus, les droits associés, etc.

PRISM – Publishing Requirements for Industry Standard Metadata [2/3]

- **PRISM** utilise une version simplifiée du langage **RDF**
- Les métadonnées définies à l'aide de **PRISM** doivent pouvoir être traitées par les processeurs **RDF** (l'inverse n'est pas vrai)
- **PRISM** utilise le *Dublin Core* comme fondation et recommande l'utilisation du vocabulaire *DC*

PRISM – Publishing Requirements for Industry Standard Metadata [3/3]

- **PRISM** étend le vocabulaire du *Dublin Core*
  - Exemple
    dc:coverage et *dc:subject* sont complétés par *prism:event*, *prism:industry*, *prism:location*, *prism:person*, *prism:organization*, *prism:section*
  - Exemple de codification *PRISM*
- **PRISM** recommande d'utiliser des *vocabulaires contrôlés*
  - utiliser par exemple un **Thésaurus** de noms géographiques au lieu de spécifier en toutes lettres un nom de lieu

Métadonnées - où en est-on ?

Objectifs de la présentation
Métadonnées – définition, utilité
Les métadonnées "métiers"
Métadonnées informatiques - exemples
Dublin Core Metadata Initiative
RDF - Ressource Description Framework
PRISM - Publishing Requirements for Industry Standard Metadata
Relations avec d'autres spécifications: NewsML, NITF, etc.
XMP - Extensible Metadata Platform
Vers le Web sémantique
T.P. éditer, collecter et transformer les métadonnées

Relations avec d'autres spécifications [1/6]

- **NewsML** est une spécification de l'*IPTC* pour la transmissions et l'échange des informations d'actualités (articles, images, multimédia)
- Bien qu'il existe certains chevauchements entre **PRISM** et **NewsML**, les deux spécifications sont largement complémentaires en ce qui concerne les métadonnées
- À la différence de **PRISM**, la partie de **NewsML** concernant les métadonnées ne s'appuie pas sur **RDF**

Relations avec d'autres spécifications [2/6]

- Le vocabulaire **PRISM** a été défini de telle façon qu'il puisse être utilisé dans la partie de **NewsML** traitant des métadonnées qui comprend trois catégories majeures
  - *AdministrativeMetadata*
  - *RightsMetadata*
  - *DescriptiveMetadata*
- **NewsML** permet d'étendre le jeu des métadonnées prédéfinies ainsi que l'utilisation de vocabulaires contrôlés pour spécifier certaines métadonnées
- A cette fin, **NewsML** préconise l'utilisation de l'IPTC Subject Reference System (*Catégories*) pour décrire les informations échangées

Relations avec d'autres spécifications [3/6]

- **NITF** (News Industry Text Format) – description des articles de presse
  - **NITF** possède quelques éléments permettant de décrire les métadonnées associées à un article ou à ses composants
  - comme pour **NewsML**, ces éléments ne s'appuient pas sur **RDF**
- **XMLNews** – scindé en deux parties
  - XMLNews-Story
    sous-ensemble de *NITF* pour la description des articles
  - XMLNews-Meta
    format *RDF* simplifié pour la description des métadonnées

Relations avec d'autres spécifications [4/6]

- **DIG35** spécifie un langage XML (mais non RDF) permettant de décrire un jeu complet de métadonnées pour les images digitales
- Les métadonnées **DIG35** sont regroupées en 5 blocs complétés par un bloc commun définissant les types de données utilisés
  - *Basic Image Parameter*
  - *Image Creation*
  - *Content Description*
  - *History*
  - *Intellectual Property Rights*
  - *Fundamental Metadata Types and Fields*
    bloc de définitions communes aux autres blocs
- **DIG35** définit également une technique d'encapsulation dans les fichiers JPEG et TIFF
- À l'heure actuelle, il n'existe pratiquement aucun produit supportant **DIG35**

Relations avec d'autres spécifications [5/6]

- JPEG 2000 est un nouveau format d'image (extension JP2) défini par le comité JPEG pour remplacer à terme le format JPEG
- Le format JPX est un format de fichier optionnel conçu comme une extension du format JP2; il permet de définir un conteneur pour l'image JP2 et pour les métadonnées associées
- JPX s'inspire de DIG35 et spécifie un langage XML (mais non **RDF**) permettant d'exprimer un jeu complet de métadonnées regroupées en 4 blocs complétés par un bloc commun définissant les types de données utilisés
  - *Image Creation metadata*
  - *Content Description metadata*
  - *History metadata*
  - *Intellectual Property Rights metadata*
  - *Fundamental Metadata Types and Elements*
    bloc de définitions communes aux autres blocs

Relations avec d'autres spécifications [6/6]

- **RSS** (RDF Site Summary). Existe en deux versions
  - 0.91 – largement utilisé pour la syndication de sites Web, basé sur **RDF** mais non exactement conforme
  - 1.0 – plus complexe, conforme à **RDF**, permet la description de métadonnées arbitraires
- **OCS** (Open Content Syndication) – description des sources de syndication
- Et aussi…
  - **XrML** (eXtensible Rigths Markup Language)
  - **ICE** (Information and Content Interchange)

Métadonnées - où en est-on ?

XMP – Extensible Metadata Platform [1/6]

- Créé par *Adobe* (septembre 2001)
- Utilise une version simplifiée de **RDF**
- Utilise le schéma *Dublin Core* comme fondation
- *DC* est étendu par d'autres schémas
  - Core Schema
  - Media Management Schema
  - Support Schema
  - Basic Job Tickets Schema
  - Rights Management Schema
- **XMP** est extensible - l'utilisateur peut définir ses propres schémas de métadonnées
- Exemple

XMP – Extensible Metadata Platform [2/6]

- Définit un mécanisme appelé **XMP Packet** permettant d'encapsuler les métadonnées **XMP** dans les fichiers des applications
- **XMP Packet** est supporté par les applications *Adobe* récentes
  - Acrobat 5, Illustrator 10, InDesign 2, InCopy 2, GoLive 6, LiveMotion 2, FrameMaker 7, Photoshop 7, Graphics Server, Document Server
- Ces applications utilisent 3 schémas de métadonnées
  - Dublin Core (espace de noms dc:)
  - PDF (espace de noms pdf:)
  - Photoshop (espace de noms photoshop:)

XMP – Extensible Metadata Platform [3/6]

- Exemple d'interface utilisateur (ne fait pas partie de la spécification **XMP**)

XMP – Extensible Metadata Platform [4/6]

- **XMP Packet** permet d'accéder aux métadonnées en lecture et écriture même en l'absence d'applications capables de comprendre le format de fichier
- Lorsque ce n'est pas possible d'implémenter **XMP Packet** dans un format de fichier propriétaire, les métadonnées **XMP** peuvent être stockées dans un fichier séparé

XMP – Extensible Metadata Platform [5/6]

- La technique **XMP Packet** est définie par *Adobe* pour les formats suivants: JPEG, TIFF, GIF, PNG, HTML, PDF, XML/SVG, PDF, AI, EPS
- Un fichier JPEG - par exemple - contenant un **XMP Packet** doit pouvoir être traité sans changement par les applications ne supportant pas **XMP**

XMP – Extensible Metadata Platform [6/6]

- **XMP** est moins orienté vers le Web que la plupart des applications **RDF**
- **XMP** est destiné à gérer et préserver les métadonnées tout au long de la chaîne éditoriale
- **XMP** gère les versions de documents, les changements de formats (*renditions*), les documents composites (dont les constituants doivent conserver leurs propres métadonnées)
- Supporté par Artesia, Documentum, IBM, Interwoven, Kodak, MediaBin, Profium, etc.

Métadonnées - où en est-on ?

Objectifs de la présentation
Métadonnées – définition, utilité
Les métadonnées "métiers"
Métadonnées informatiques - exemples
Dublin Core Metadata Initiative
RDF - Ressource Description Framework
PRISM - Publishing Requirements for Industry Standard Metadata
Relations avec d'autres spécifications: NewsML, NITF, etc.
XMP - Extensible Metadata Platform

- **Vers le Web sémantique**

T.P. éditer, collecter et transformer les métadonnées

Vers le Web sémantique [1/5]

- Le Web sémantique (*Semantic Web*) est la vision d'un Web structuré de telle façon que l'on puisse automatiser, intégrer et réutiliser les données au travers d'applications variées
- Deux technologies candidates
  - **RDF** (Ressource Description Framework)
  - **Topic Maps** - SGML initialement (ISO 1999) puis porté en XML (XTM)

Vers le Web sémantique [2/5]

Comparaison RDF / Topic Maps

- **RDF** est une technique générale de description de ressources
  - similaire aux *mots-dés* d'une fiche de catalogage
- Les **Topic Maps** utilisent des réseaux sémantiques
  - similaires aux *index*, *glossaires*, *thesaurus* d'un livre
  - il existe quelques applications basées sur Topic Maps: Mondeca

Vers le Web sémantique [3/5]

Comparaison HTML / RDF / Topic Maps

- **HTML** relie des données de pages Web entre elles
- **RDF** relie des ressources quelconques entre elles, qu'elles soient des données, des concepts ou des objets, basés ou non sur le Web
- **Topic Maps** structure et organise des connaissances, associe des sujets et des occurrences d'objets ou de concepts

Vers le Web sémantique [4/5]

Classifications - Hiérarchies

- XFML (eXchangeable Faceted Metadata Language) est un format XML permettant l'échange de métadonnées sous la forme de classifications à facettes (taxonomies)
- XFML est un sous ensemble de XTM (TopicMaps). Tout document XFML peut s'exprimer en XTM
- Exemples de classification à facette
  - Dublin Core exprimé en XFML
  - IPTC Subject Codes exprimé en XFML

Vers le Web sémantique [5/5]

Annotations

- Ajouter un contenu sémantique à un site Web à l'aide d'une technique d'Annotation
- Exemple de page Web annotée à l'aide de balises DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer)
  - Is Participation in the Semantic Web too Difficult.htm

Le Web sémantique est-il une utopie ?

- *Metacrap: Putting the torch to seven straw-men of the meta-utopia*
  Cory Doctorow, August 2001
- 7 raisons expliquant pourquoi le web sémantique est une utopie
  - les gens sont menteurs [les métadonnées ne sont pas fiables]
  - les gens sont paresseux [ils ne renseignent pas les métadonnées]
  - les gens sont stupides et inattentifs
  - on ne sait pas voir les choses et encore moins les décrire
  - les schémas ne sont pas neutres et l'on n'arrive jamais à s'entendre sur quoi que ce soit
  - la manie qui consiste à tout quantifier produit des résultats biaisés
  - on ne peut pas tout décrire d'une seule manière

Références

- Dublin Core Metadata Initiative (DCMI)
  www.dublincore.org
- RDF sur le site du W3C
  www.w3.org/RDF/
- XMP
  www.adobe.com/products/xmp/main.html
  Extensible Metadata Platform (XMP), sur Oasis-Open
- Métadonnées: une initiation
  (Dublin Core, IPTC, EXIF, RDF, XMP, etc.)
  article sur www.softexperience.com

"Travaux pratiques" - Plan

Présentation d'outils simples pour éditer, collecter et transformer les métadonnées

- Propriétés des documents MS Office et OpenOffice.org
- Propriétés des fichiers Windows 2000 / XP
- Les fichiers Macintosh
- NTFS, un système de fichiers "orienté métadonnées"
- Champs IPTC & EXIF des images JPEG/TIFF
- Champs ID3 des MP3
- Informations sur les documents PDF
- Collecte et transformations des métadonnées informatiques

"Travaux pratiques" – MS Office [1/2]

- Propriétés des documents MS Office (Word, Excel, etc.)
  - la terminologie Microsoft. Propriétés = Métadonnées
  - deux types de propriétés
    - propriétés générales
      Titre, Auteur, Sujet, Mots-clés, Commentaires, Responsable, Société, Catégorie, etc. [25 éléments]
    - propriétés personnalisées
  - on peut afficher et éditer ces deux types de propriétés sans qu'Office soit installé sur le poste
  - visualisation des propriétés à l'aide des *tooltips* et des colonnes de l'Explorateur en mode détail

"Travaux pratiques" – MS Office [2/2]

- Propriétés des documents MS Office (Word, Excel, etc.)
  - Possibilité de demander le renseignement des propriétés du document lors de l'enregistrement
    - *menu* Outils / Options, *onglet* Enregistrement, *cocher* Demander les propriétés du document
  - Aucun contrôle sur les propriétés renseignées
  - Il est nécessaire de développer des macros pour contrôler les propriétés renseignées (champs obligatoires, lexiques d'autorité, synonymes, etc.)

"Travaux pratiques" – Ooo [1/3]

- Propriétés des documents OpenOffice.org
  - deux types de propriétés
    - générales (Description)
      Titre, Description, Sujet, Mots-clés, Créateur initial, etc. [25 éléments]
    - possibilité de propriétés personnalisées [4 au maximum]
  - on ne peut pas afficher et éditer ces deux types de propriétés sans qu'OOo soit installé sur le poste

"Travaux pratiques" – Ooo [2/3]

- Propriétés des documents OpenOffice.org
  - Possibilité de demander le renseignement des propriétés du document lors de l'enregistrement (*menu* Outils / Options, Chargement/Enregistrement, Général, *cocher* Editer les propriétés avant l'enregistrement)
  - Aucun contrôle sur les propriétés renseignées
  - Il est nécessaire de développer des macros pour contrôler les propriétés renseignées (champs obligatoires, lexiques d'autorité, etc.)

"Travaux pratiques" – Ooo [3/3]

- Propriétés des documents OpenOffice.org
  - Initiative 1dok.org
    Ministère allemand de l'Économie, de la Technologie et des Transports & Fondation Technologie du Schleswig-Holstein financés par la Commission Européenne
    - Vers un format de document ouvert
    - Implémentation d'un modèle de métadonnées extensible et orienté objet dans OpenOffice.org (non basé sur RDF)

"Travaux pratiques" – Windows 2000/XP [1/2]

- Les deux types de métadonnées en Windows 2000/XP
  - propriétés MS Office
  - propriétés associées à un fichier quelconque
    - Titre, Objet, Auteur, Catégorie, Mots-clés, Commentaires, Source, Numéro de révision
    - pas de propriétés personnalisées

"Travaux pratiques" – Windows 2000/XP [2/2]

- Les métadonnées "non Office" en Windows 2000/XP
  - sont facultatives et peuvent être associées à tout type de fichier
  - existent uniquement sur les volumes NTFS
  - sont perdues quand on…
    - copie le fichier sur une autre plate-forme
    - transmet le fichier par courrier électronique
    - copie le fichier sur un volume non NTFS
    - compresse le fichier
    - etc.

"Travaux pratiques" – Fichiers Macintosh [1/2]

Rappel

- Les fichiers Macintosh (OS 9 ou antérieur) possèdent deux composantes
  - données (*data fork*)
    - les données statiques
    - le code exécutable PowerPC
  - ressource (*resource fork*)
    - le code exécutable 68K
    - les réglages de préférences
    - des ressources spécifiques au programme (menus, polices, icônes, etc.)
    - des ressources diverses telles que les messages des boîtes de dialogue

"Travaux pratiques" – Fichiers Macintosh [2/2]

- Le système du Macintosh maintient des informations concernant le fichier dans le catalogue du disque dur. Ce sont les *informations du Finder.* Elles permettent de savoir…
  - comment lancer l'application associée au fichier
  - quelle taille mémoire allouer à un programme
  - quelle est la signature *Type/Creator* du fichier
- Les *Ressources,* les *informations du Finder* ainsi que les *Commentaires* peuvent être considérées comme des métadonnées associées à chaque fichier Macintosh

"Travaux pratiques" – NTFS [1/3]

- NTFS est le système de fichiers natif de Windows NT/2000/XP
- En NTFS, un fichier consiste en plusieurs *data streams* qui sont un peu la généralisation du concept de *fork* pour les fichiers Macintosh
- Un *stream* particulier contient les informations de sécurité (droits d'accès, etc.) et un autre *stream* appelé *standard* contient les données habituelles

"Travaux pratiques" – NTFS [2/3]

- Il peut exister aussi d'autres *streams* (*alternate streams*) liés au *stream standard* et contenant des métadonnées associées au fichier "normal"
- Les *alternate streams* sont invisibles depuis l'*Explorateur* de Windows ainsi que depuis la quasi-totalité des applications Windows
- Les *alternate streams* peuvent être considérés comme des sous-fichiers du *stream standard* et sont introduits par le caractère deux-points

"Travaux pratiques" – NTFS [3/3]

- Les *alternate streams* sont utilisés…
  - pour stocker les propriétés des fichiers "non Office"
  - pour stocker les informations spécifiques aux fichiers Macintosh (ressources, informations du Finder, Commentaires)
  - pour stocker des vignettes d'images (Windows 2000 seulement, XP utilise un fichier caché *Thumbs.db*)
  - pour encrypter des contenus
- Les *streams* de NTFS en font un système de fichiers "orienté métadonnées"

"Travaux pratiques" – IPTC & EXIF [1/2]

- Champs IPTC des images JPEG/TIFF
  - Titre, Source, Crédit, Copyright, Statut éditorial, Priorité, Catégorie, Mots-clés, etc. [33 éléments]
- Champs EXIF des images JPEG
  - Fabricant de la caméra, Modèle, Orientation, Temps d'exposition, Résolution en largeur, Résolution en hauteur, etc. [30 éléments]
- Propriétés Windows 2000 / XP pour les images
  - Windows 2000: Titre, Objet, Auteur, Mots-clés, Commentaires, Catégorie
    - stockés dans des *alternate streams*
  - Windows XP: Titre, Objet, Auteur, Mots-clés, Commentaires, [pas de Catégorie en XP]
    - stockés dans des structures EXIF propriétaires, en Unicode

"Travaux pratiques" – IPTC & EXIF [2/2]

- Gestion des champs IPTC à l'aide de…
  - Adobe PhotoShop
  - Fotoware FotoStation
  - PixVue
  - etc.

"Travaux pratiques" – Champs ID3 des MP3 [1/1]

- Champs ID3 des fichiers MP3
  - Titre, Compositeur, Auteur du texte, Durée, Copyright, etc. [74 éléments organisés en frames]
- Windows XP permet d'éditer certains de ces champs
  - Artiste, Titre de l'album, Année, Numéro de piste, Genre, Paroles, Titre, Commentaires

"Travaux pratiques" – Informations PDF [1/2]

- Informations sur les documents PDF
  - Titre, Auteur, Sujet, Mots-clés, Créateur, Producteur, etc. [9 éléments]
  - Affichage dans Acrobat Reader
- Adobe Acrobat n'est pas le seul outil permettant de produire des documents PDF
  - GhostScript version 8.00
    - ajout d'informations PDF à l'aide de code PDFMarks (pas très souple)
  - FOP (Formatting Objects Processor) version 0.20.5
    - impossible d'ajouter des informations PDF directement à l'aide du processeur FOP

"Travaux pratiques" – Informations PDF [2/2]

- Édition des informations sur les documents PDF
- Le programme PDFExplorer

"Travaux pratiques"

Collecte et transformations de métadonnées [1/1]

- Extraction des métadonnées MS Office, OpenOffice.org, Windows NTFS, Macintosh, HTML, IPTC, PDF, XMP à l'aide du programme Catalogue Files MetadataMiner
- Transformation de l'export XML selon une XSLT