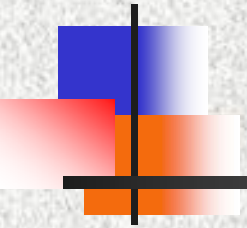
ATICA - Deuxième journée de la réutilisation des données
Mercredi 9 octobre 2002



Les métadonnées

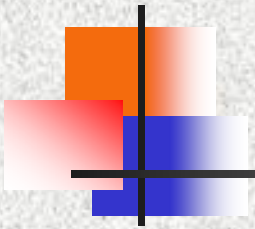
un élément clé de la gestion de contenu

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

Les métadonnées - Plan

- 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
- *Démonstration* - collecte et transformation de métadonnées

plus technique… [XML]

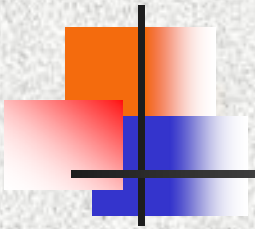


Objectifs de la présentation

- 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

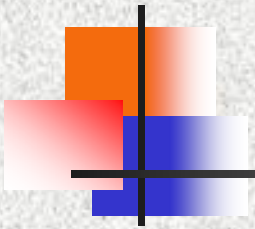
Une première définition

- Une *métadonnée* [*metadata*] est littéralement une *donnée sur une donnée*
  - Dans le domaine des *métadonnées*, 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

- 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 indispensables…
  - …pour décrire les *droits*, *relations*, *formats*, *dates*, etc. associés à une *ressource*
  - …pour la gestion et la recherche d'images, de séquences audio et video, 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 *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

Plan - 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

Démonstration - collecte et transformation de métadonnées

Métadonnées informatiques [1/6]

- Où sont les métadonnées informatiques ?
  - dans les bases de données
  - dans les données elles-mêmes [les métadonnées sont incluses 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*

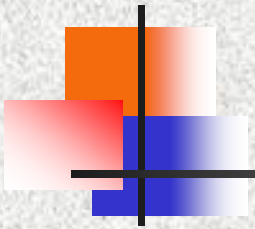
Métadonnées informatiques [2/6]

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

```
<title>Le Monde.fr : Conciliabules contre Jean-Marie Messier</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">
```

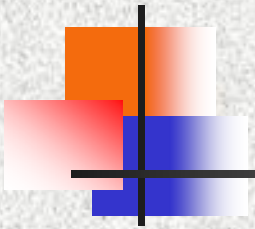
Métadonnées informatiques [3/6]

- 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]
  - Initiative 1dok.org
    Ministère allemand de l'Économie, de la Technologie et des Transports & Fondation Technologie du Schleswig-Holstein financés par la Commission Européenne
    - Implémentation d'un modèle de métadonnées extensible et orienté objet dans OpenOffice.org



Métadonnées informatiques [4/6]

- 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 [5/6]

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

Métadonnées informatiques [6/6]

- 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 *métadonnée*s ???

- 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*

Plan - 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

Démonstration - collecte et transformation de métadonnées

Dublin Core Metadata Initiative [1/4]

- Prolifération de besoins "métiers" variés

musées, bibliothèques, archives, …

- Diversité des *nomenclatures* et des *structures* des métadonnées informatiques

Une image peut posséder 5 "Descriptions" différentes

  - Recherche d'un *standard*
- 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 [2/4]

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

DC parle de **creator**, pas d'**author**

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

Dublin Core Metadata Initiative [3/4]

- Une version plus évoluée du *Dublin Core* autorise l'usage de qualificateurs
- Exemple:
  l'élément Description peut être raffiné à l'aide des qualificateurs tableOfContents et abstract

Dublin Core Metadata Initiative [4/4]

- 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"

Plan - 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**

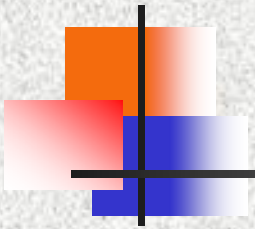
PRISM - Publishing Requirements for Industry Standard Metadata

Relations avec d'autres spécifications: NewsML, NITF, etc.

XMP - Extensible Metadata Platform

Vers le Web sémantique

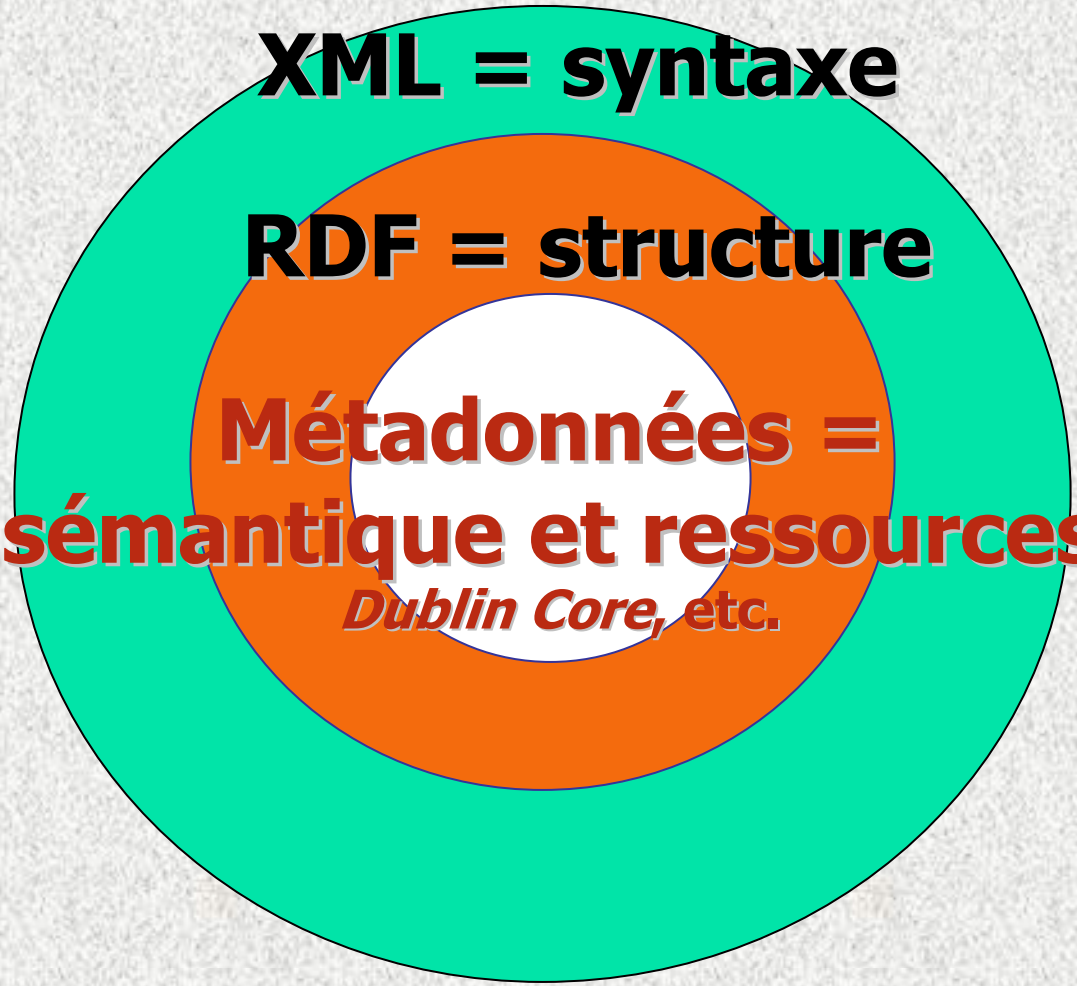
Démonstration - collecte et transformation de métadonnées



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
TéléÉducation NB

RDF – Resource Description Framework [3/9]

- **RDF** est basé sur des triplets

  sujet - **prédicat** - **objet**

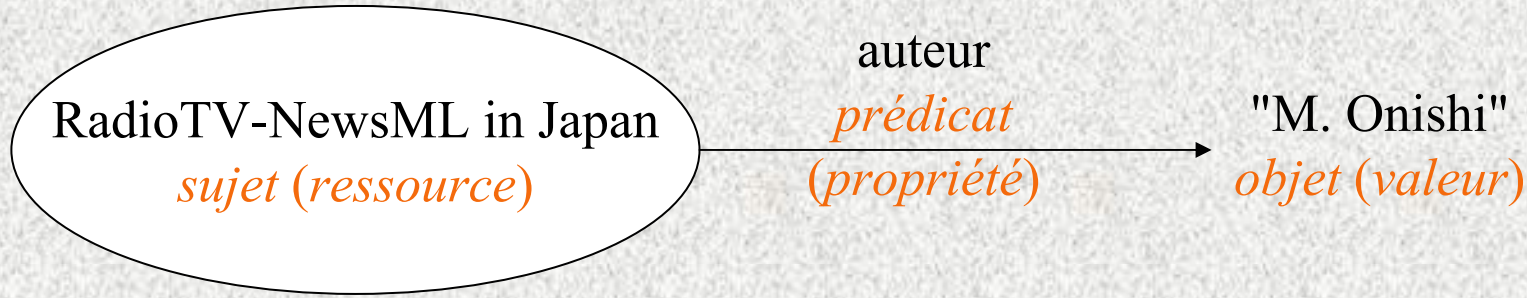
  ou ressource - propriété - valeur

- Exemple

  Le document RadioTV-NewsML in Japan a pour auteur M. Onishi

  sujet | prédicat | objet

- Modélisé à l'aide de graphes orientés étiquetés

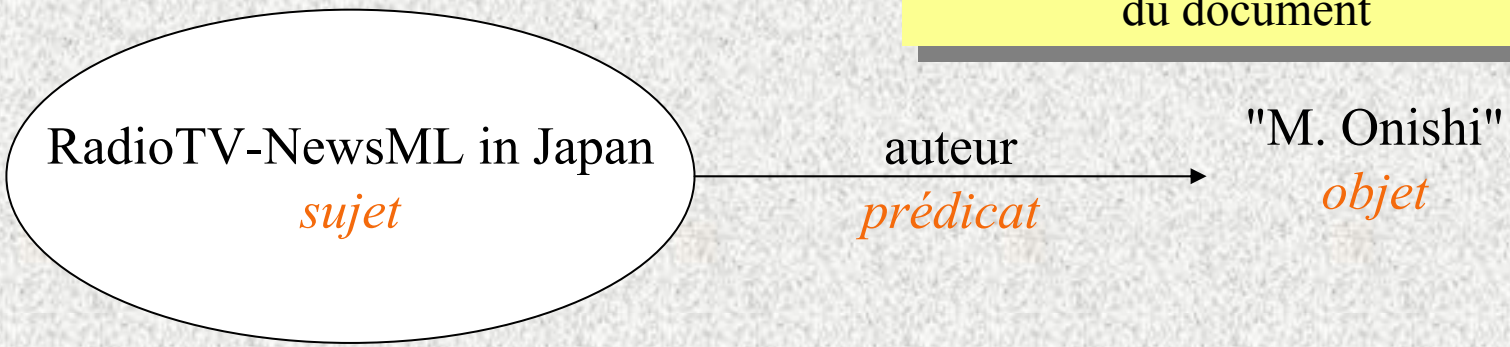


RDF – Resource Description Framework [4/9]

- 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**

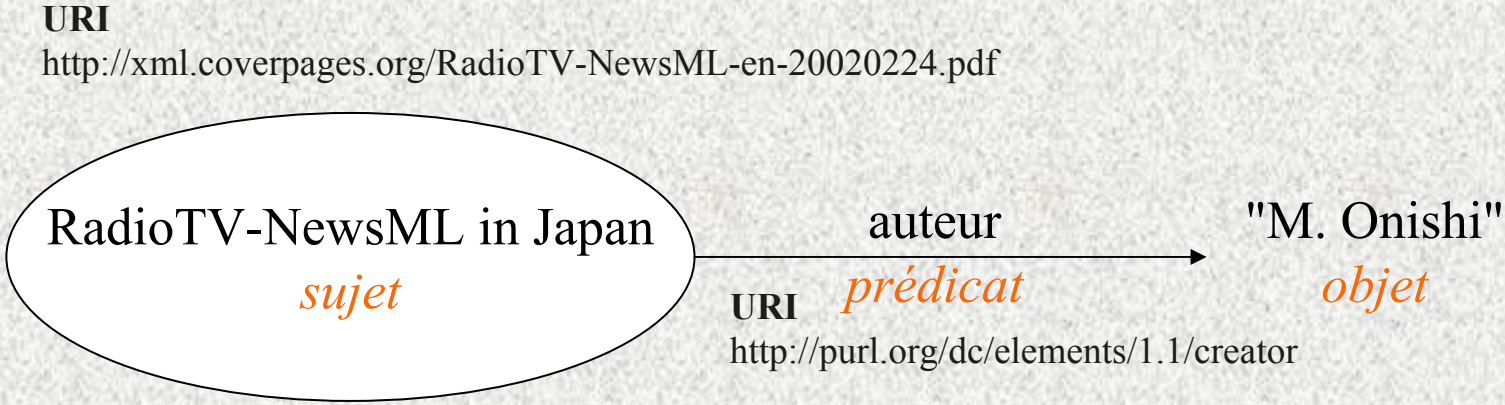
URI
http://xml.coverpages.org/RadioTV-NewsML-en-20020224.pdf

L'URI de la ressource est l'URL du document



RDF – Resource Description Framework [5/9]

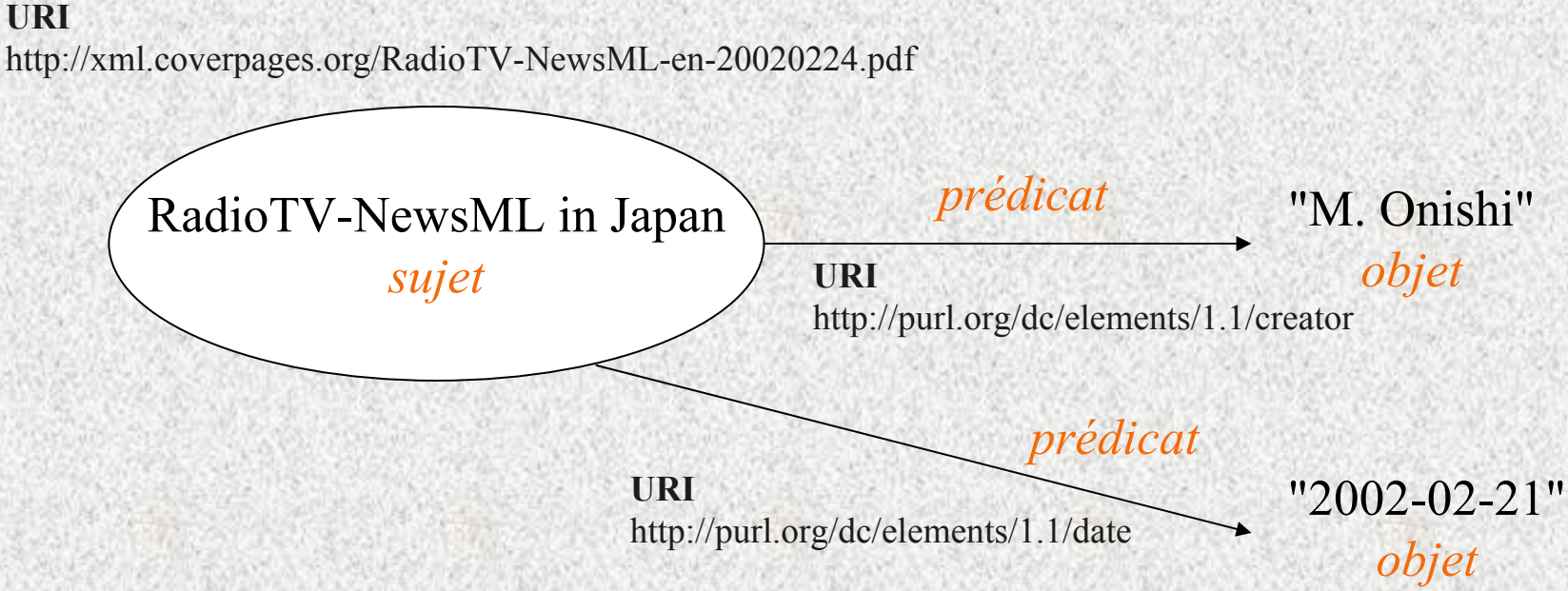
- Les *prédicats* (*propriétés*) sont également représentés par des **URI**



L'URI du prédicat est l'élément *creator* du schéma *Dublin Core* version 1.1

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*

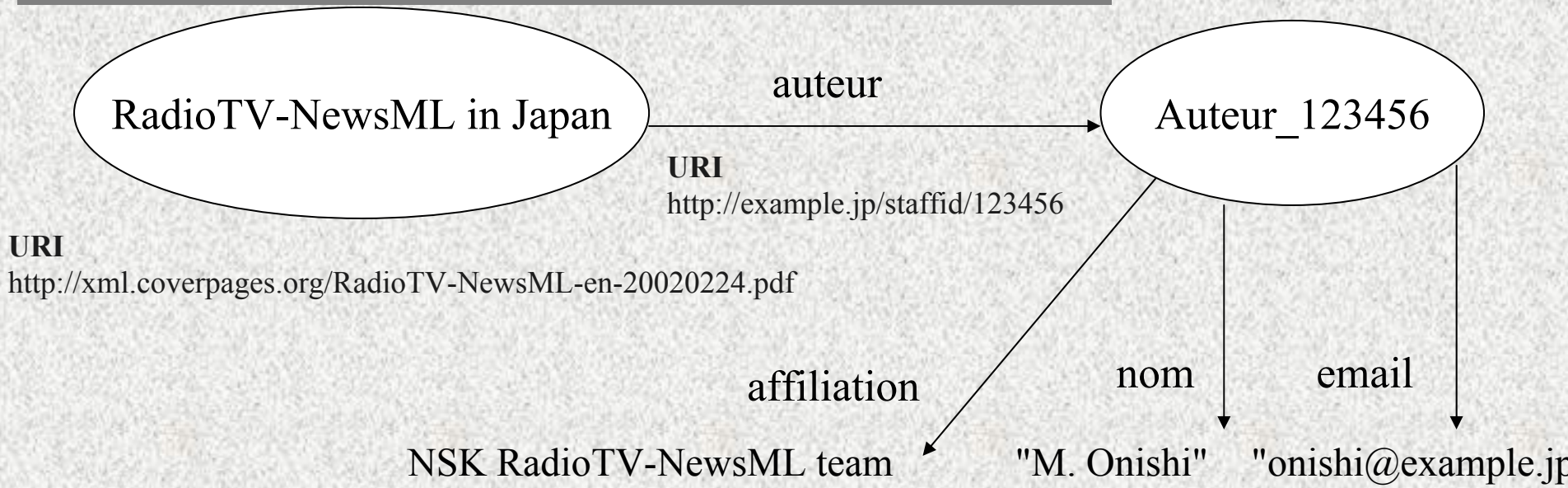
propriété *creator*

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

Plan - 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

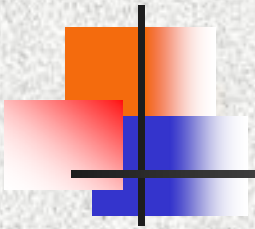
Vers le Web sémantique

Démonstration - collecte et transformation de 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 prism_2.6_1.rdf
- **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

Plan - 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

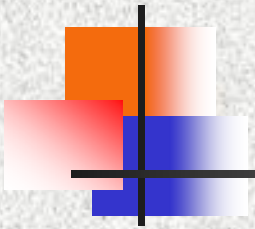
PRISM - Publishing Requirements for Industry Standard Metadata

- **Relations avec d'autres spécifications**: NewsML, NITF, etc.

XMP - Extensible Metadata Platform

Vers le Web sémantique

Démonstration - collecte et transformation de métadonnées

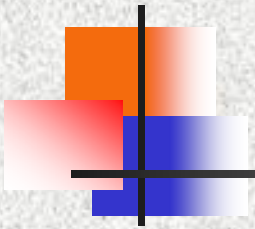


Relations avec d'autres spécifications [1/3]

- **NewsML** – spécification de l'*IPTC* pour la transmissions des articles et l'automatisation des fils d'agences
  - Bien qu'il existe certains chevauchements entre **PRISM** et **NewsML**, les deux spécifications sont largement complémentaires
  - Le vocabulaire **PRISM** a été défini de telle façon qu'il puisse être utilisé dans la partie de **NewsML** traitant des métadonnées

Relations avec d'autres spécifications[2/3]

- **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
- **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 [3/3]

- **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)

Plan - 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

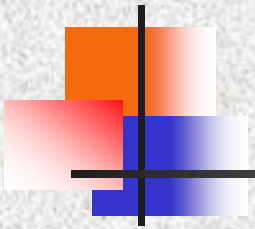
PRISM - Publishing Requirements for Industry Standard Metadata

Relations avec d'autres spécifications: NewsML, NITF, etc.

- **XMP - Extensible Metadata Platform**

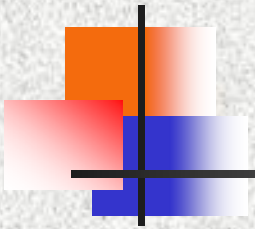
Vers le Web sémantique

Démonstration - collecte et transformation de métadonnées



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

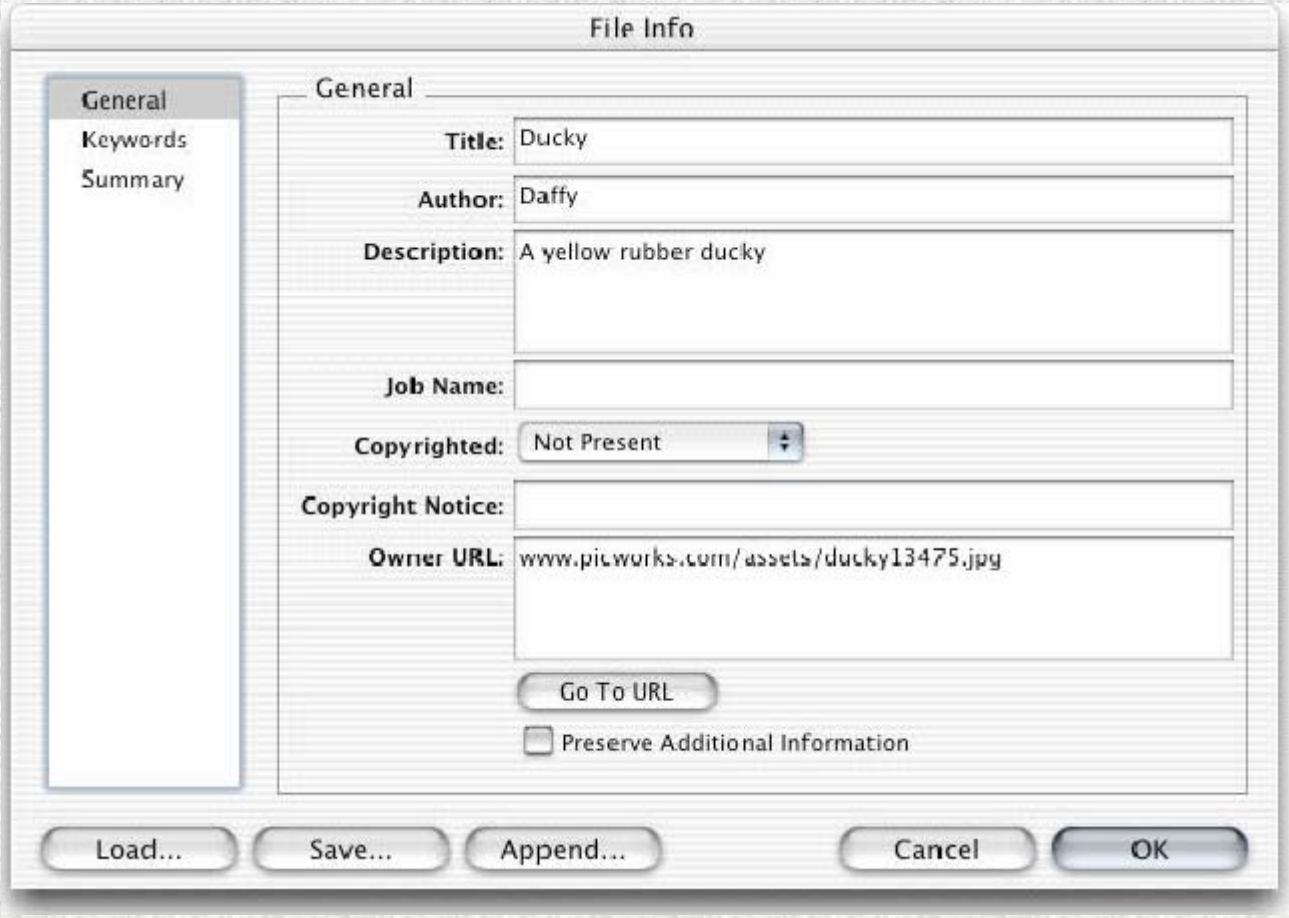


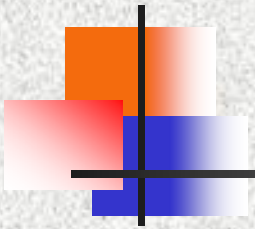
XMP – Extensible Metadata Platform [2/6]

- 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, GoLive 6, Illustrator 10, Photoshop 7
  - InDesign 2, InCopy 2, FrameMaker 7

XMP – Extensible Metadata Platform [3/6]

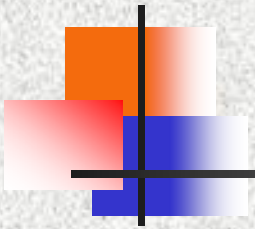
- Exemple d'interface utilisateur (ne fait pas partie de la spécification **XMP**)





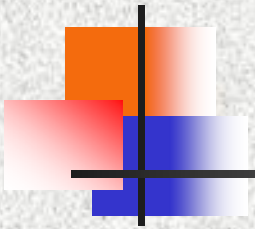
XMP – Extensible Metadata Platform [4/6]

- **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.

Plan - 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

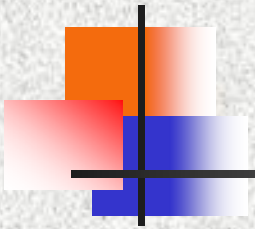
PRISM - Publishing Requirements for Industry Standard Metadata

Relations avec d'autres spécifications: NewsML, NITF, etc.

XMP - Extensible Metadata Platform

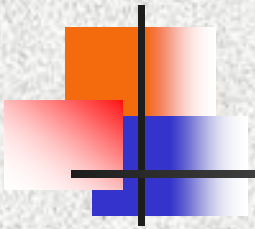
- **Vers le Web sémantique**

Démonstration - collecte et transformation de métadonnées



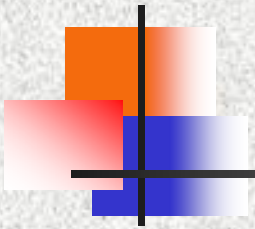
Vers le Web sémantique [1/3]

- 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)



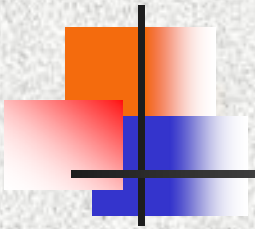
Vers le Web sémantique [2/3]

- **RDF** est une technique générale de description de ressources
  (similaire aux *mots-clés* d'une fiche de catalogage)
- Les **Topic Maps** utilisent des réseaux sémantiques
  (similaires aux *index*, *glossaires*, *thesaurus* d'un livre)



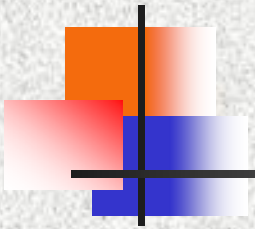
Vers le Web sémantique [3/3]

- **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
- Il existe quelques applications basées sur Topic Maps: Mondeca



Références

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



Démonstration

- Collecte de métadonnées informatiques
- Export HTML et XML
- Transformation en RDF
- Création d'un graphe RDF en SVG
- Lancer *Catalogue*