Web sémantique et pratiques documentaires

Jérôme Euzenat, Raphaël Troncy

Web sémantique et pratiques documentaires

Jérôme Euzenat[1], **Raphaël Troncy**[1,2]

[1] *INRIA Rhône-Alpes*
655 avenue de l'Europe
38330 Montbonnot Saint-Martin (France)
Jerome.Euzenat@inrialpes.fr

[2] *INA,*
4 avenue de l'Europe
94366 Bry-sur-Marne (France)
Raphael.Troncy@ina.fr

RÉSUMÉ: *Le web sémantique a l'ambition de construire pour les machines l'infrastructure correspondant au web actuel et d'offrir aux humains la puissance des machines pour gérer l'information disponible dans ce web. Les technologies du web sémantique ont donc beaucoup à offrir pour assister les pratiques documentaires à venir. On présentera les technologies destinées à décrire les ressources du web et leurs ontologies dans la perspective de leur utilisation à des fins de gestion documentaires. On présentera certaines ressources déjà existantes pouvant être utilisées dans ce but ainsi qu'une application à l'indexation de données multimédia et audiovisuelles.*

MOTS CLÉ: *Web sémantique, OWL, RDF, Ontologie, Publication, Indexation, MPEG-7.*

1. LE WEB SÉMANTIQUE

Si le web actuel contient une quantité d'information formidable, il reste difficile à exploiter. Ainsi, la recherche d'un « livre sur Bertrand Russell » n'est guère aisée à l'aide des moteurs de recherche : ils commencent par supprimer le mot clé « sur » comme peu discriminant et retournent nombre de pages consacrées aux travaux du philosophe. Si l'on désire que les machines nous aident à trouver l'information, il nous faut les aider un peu plus en la leur explicitant.

Le web est constitué par un ensemble de documents, principalement textuels, formatés dans un langage particulier (HTML) permettant d'exprimer des liens entre un objet dans le document source (l'ancre) et un objet du document cible. Ce web est exploité par des dispositifs logiciels (navigateurs ou robots de recherche) qui traversent ces liens lorsqu'ils les rencontrent (ou lorsque l'utilisateur clique sur une ancre). Le travail d'exploitation de ce web est donc principalement dévolu aux utilisateurs humains qui doivent analyser le contenu des pages pour déterminer sur quel lien cliquer. Des dispositifs logiciels peuvent les aider en analysant ce contenu, mais comme on l'a vu, leur aide, bien que remarquable, reste limitée car le contenu des documents du web s'adresse aux utilisateurs humains.

De longue date, les archivistes et les documentalistes ont développé des méthodologies, des terminologies (thésaurus) et des outils pour inventorier et cataloguer les fonds documentaires dont ils ont la gestion. Retrouver des documents particuliers dans un fonds est devenu un travail de spécialiste réservé aux documentalistes. Le défi, aujourd'hui, est de faire ce travail à l'échelle

du web, c'est-à-dire un espace ouvert à tous, et non plus dans une structure plus ou moins rigide, évoluant au gré des groupes de travail internationaux.

Le principe du web sémantique, comme celui du web, est son ouverture : la possibilité de faire référence à des ressources distantes, non maîtrisées et la possibilité d'enrichir ces ressources sans être entravé par des barrières physiques, techniques ou même réglementaires. Pour hériter de ces particularités du web, le web sémantique tirera parti d'un espace d'adressage global des ressources permettant d'ajouter en permanence de l'information à ce web et d'y avoir accès sans entrave. Il offrira des langages permettant de partager et d'échanger la description de ces ressources. Une application du web sémantique est donc une application qui se nourrit des descriptions des ressources du web et qui, en retour, produit de telles descriptions.

En première approximation, le but du web sémantique est de développer un web dont le contenu s'adresse, au moins pour partie, aux machines, afin qu'elles puissent aider les utilisateurs humains [Charlet 2003]. Si l'on cherche à préciser, un tel web doit doter ses ressources (documents, service…) d'annotations dont le but n'est pas d'assurer l'affichage des documents mais l'appréhension de son contenu par divers outils logiciels. Le web sémantique doit donc être une infrastructure juxtaposant au web actuel des documents structurés par des langages pour exprimer la connaissance, pour décrire les relations entre les connaissances, pour décrire les conditions d'utilisation et pour décrire les garanties et les modes de payement, et des dispositifs permettant de trouver les ressources.

Dans un cadre documentaire, on voudra utiliser en premier lieu des informations sur les documents, qu'ils soient disponibles sur le web ou non. On s'intéressera donc ici aux techniques permettant de décrire ces documents à l'aide des outils du web sémantique et au langage permettant de le faire : RDF. Cependant, il est nécessaire de préciser le vocabulaire utilisé pour décrire les documents (un ordinateur ne peut deviner qu'une autobiographie est une sorte de biographie et une biographie un livre). D'autre part, la recherche de documents dépend d'informations sortant du domaine des documents, par exemple des informations sur les personnes ou sur les mathématiques. Pour cela, on développe un modèle conceptuel des objets considérés qui circonscrit le vocabulaire propre au domaine et contraint le sens des concepts et de leurs relations. Ce modèle, qualifié d'ontologie, dispose d'un langage adapté pour s'exprimer dans le web sémantique : OWL. Enfin, il sera nécessaire de tirer parti de ces descriptions et de ces ontologies pour retrouver les documents (à l'image de l'autobiographie de Bertrand Russell). On évoquera donc brièvement les langages de requêtes et les inférences permises par les langages considérés.

Notre but est de présenter rapidement les technologies du web sémantique actuellement disponibles et d'illustrer en quoi elles peuvent être utiles à la pratique documentaire.

Nous allons donc commencer tout naturellement par présenter le langage RDF qui permet le partage de l'information entre humains et machines (§1). Nous montrerons comment il est possible d'y exprimer des notices signalétiques et de retrouver l'information. Mais exprimer de l'information n'est pas suffisant si l'on veut améliorer l'appréhension de ces structures par la machine. Disposer d'ontologies permettra aux ordinateurs de faire automatiquement des inférences dont nous sommes coutumiers. Nous introduirons à cet effet le langage d'ontologies

pour le web OWL en prenant un exemple de construction d'ontologie des publications (§3). Nous montrerons très simplement comment celui-ci permet d'inclure d'autres ontologies dans une modélisation, profitant ainsi d'efforts de modélisation extérieurs mais aussi des sources d'information utilisant ces modélisations.

Nous présenterons ensuite des ressources déjà disponibles utilisant les technologies du web sémantique et permettant d'aller au-delà de ce qui est actuellement offert dans les centres de documentation. On évoquera divers scenarii dans lesquels elles peuvent intervenir (§4).

L'ensemble de l'exposé sera illustré par l'utilisation des technologies du web sémantique pour la gestion et la manipulation de fonds documentaire. Nous le terminerons par une application dédiée plus particulièrement à l'indexation de documents multimédia (§5).

2. PRODUIRE DES DESCRIPTIONS RDF

RDF est le langage dans lequel sont décrites les ressources du web sémantique. On présente donc ci-dessous les rudiments de RDF et son utilisation dans un cadre de signalisation documentaire.

Représenter des notices avec RDF

Comme on a pu le voir avec l'exemple de l'autobiographie de Bertrand Russell, l'annotation des ressources à l'aide de simples mots-clé, voire de catégories, n'est pas suffisant. Il faut être capable d'exprimer l'information relationnelle : qu'un objet « livre » peut avoir un « auteur » qui est une « personne » et un « sujet », ou qu'une « autobiographie » est une « biographie » dont l' « auteur » est le « sujet ». Il est donc naturel que le premier langage pour le web sémantique, RDF mette l'accent sur les relations. Une notice bibliographique pourra être exprimée en RDF à la manière de la figure 1.

RDF ("Resource Description Framework") [Lassila 1999, Champin 2000, Klyne 2004] est un langage, recommandé par le W3C, fondé sur les notions de ressources et de relations entre ressources. Un triplet $<s, p, o>$ exprime une relation p entre un sujet s et un objet o. Les relations et certaines ressources sont identifiées par des URI ("Uniform Resource Identifiers") [Berners-Lee 1998] dont l'exemple le plus connu est celui des URL, qui constituent les « adresses » des pages du web. Les ressources peuvent être identifiées ou rester anonymes, elles peuvent également être typées en utilisant la relation `rdf:type`. Notons que certaines ressources sont externes (`foaf`), et que d'autres sont spécifiques à notre application (`bib`). Les objets peuvent être des littéraux (comme une chaîne de caractères ou un nombre entier).

La figure 1 présente une partie d'un document RDF sous forme graphique. Les objets d'un triplet qui sont des littéraux sont représentés dans un rectangle (ici, `"The Autobiography of Bertrand Russell"` ou `760`). Le sommet non étiqueté représente une variable. Un document RDF constitue donc un graphe étiqueté sur ses arêtes et ses sommets (plus précisément un multi-graphe orienté étiqueté) où les éléments apparaissant comme sujet ou objet sont les sommets, et chaque triplet est représenté par un arc dont l'origine est son sujet et la destination son objet.

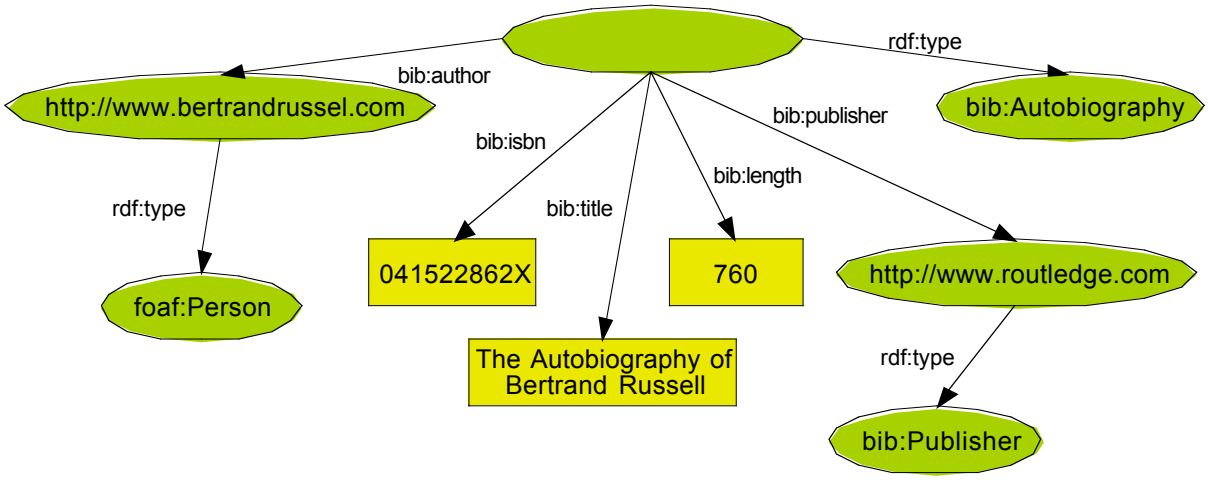

Figure 1 : Graphe RDF représentant une notice bibliographique.

Intuitivement, ce graphe peut se comprendre comme « il existe un livre qui est une autobiographie, dont l'auteur est Bertrand Russell, le numéro ISBN est 041522862X, le titre est "The Autobiography of Bertrand Russell", le nombre de pages 760, et qui est publié par Routledge ».

Ce document sera codé en machine par un document RDF/XML [Beckett 2004] ou N3. Voici l'exemple original exprimé en RDF/XML.

```
<RDF xmlns="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:bib="http://www.thebibliography.org/"
     xmlns:foaf="http://xmlns.com/foaf/0.1/">

   <Description rdf:ID="041522862X">
     <rdf:type rdf:resource="&bib;Autobiography"/>
     <bib:author>
        <foaf:Person rdf:about="http://www.bertrandrussell.com"/>
     </bib:author>
     <bib:title>The Autobiography of Bertrand Russell</bib:title>
     <bib:isbn>041522862X</bib:isbn>
     <bib:length>760</bib:length>
     <bib:size>1.67x7.90x5.14in</bib:size>
     <bib:publisher>
        <Publisher rdf:resource="#Routledge"/>
     </bib:publisher>
     <bib:date>
        <Date year="2000"/>
     </bib:date>
   </Description>
   <Description rdf:ID="Routledge">
     <address>
        <Address city="London" country="UK"/>
     </address>
   </Description>
</RDF>
```

Figure 2 : Description d'une notice bibliographique en RDF.

Un graphe à l'échelle du monde

Comme on l'a vu, les objets, ou les ressources, sont identifiés par des URI. À l'instar de ce qui se passe dans le web, ceci présente l'intérêt que l'objet référencé n'a pas besoin d'être ou d'être décrit dans le document lui-même. Les documents RDF constituent donc un immense graphe à l'échelle du monde. RDF peut-être utilisé pour annoter des documents écrits dans des langages non-structurés, ou comme une interface pour des documents écrits dans des langages ayant une structure similaire (des bases de données, par exemple).

L'utilisation des URI permet le partage des descriptions entre plusieurs documents. Ainsi, l'ouvrage dans ce qui concerne son contenu peut être décrit sur le site de l'éditeur alors que la description de l'exemplaire se trouvant dans le fonds peut-être décrit localement. De plus, si la société des amis de Bertrand Russell publie sa page FOAF ou que l'auteur a une page FOAF, mieux vaut la référencer plutôt que de la recopier ou la réinventer.

De même, si plusieurs exemplaires sont disponibles, il n'est pas nécessaire de dupliquer l'information. La figure 3 montre trois annotations de l'ouvrage cité plus haut. Le premier concerne un site de bibliothèque et indique dans quel rayonnage se trouve l'ouvrage, le second provient d'un site marchand et indique son prix et son type de reliure, le troisième provient de nouveau d'un service de documentation et indique les catégories de la classification Dewey auxquelles l'ouvrage appartient.

```
<Ouvrage rdf:ID="ExUS4533b">
   <sto:oeuvre rdf:about="http://www.routledge.com/#041522862X"/>
   <sto:size>1.67x7.90x5.14in</sto:size>
   <sto:rayonnage>USUEL-Biographie</sto:rayonnage>
</Ouvrage>

<Book rdf:ID="http://www.routledge.com/041522862X">
   <rdf:type rdf:resource="#Paperback"/>
   <price>$16.95</price>
   <rating>4</rating>
</Book>

<Autobiography rdf:ID="BA4498">
   <owl:sameAs rdf:resource="#041522862X"/>
   <rdf:type rdf:resource="#Paperback"/>
   <dc:identifier>041522862X</dc:identifier >
   <dewey>107</dewey>
   <dewey>190</dewey>
   <dewey>160</dewey>
</Autobiography>
```

Figure 3 : Descriptions alternatives de celles de la figure 2.

Il subsiste une différence entre les trois descriptions : la première fait référence à la description initiale à l'aide de l'attribut `sto:oeuvre` alors que les suivantes la complètent. En effet, les attributs `price` et `dewey` sont virtuellement intégrés dans la description initiale. En réalité, ils ne modifient en rien la description première mais, pour le site de l'application, la ressource `http://www.routledge.com/#041522862X` a les propriétés `price` et `dewey`.

De cette manière, on peut dire que les technologies mises en œuvre dans le web sémantique sont réellement ouvertes : il est toujours possible d'ajouter des propriétés à une ressource, qu'elle ait été créée par nous ou non.

En fait, dans les deux derniers cas, l'application n'a que les attributs `price` et `dewey`. Car il n'y a pas obligation pour une application RDF d'aller chercher la ressource correspondante.

Bien pire, il ne faut pas lire les URI comme des URL, il n'y a aucune obligation que l'identificateur `http://www.routledge.com/#041522862X` donne accès à un quelconque document. Contrairement aux locateurs, les URI ne sont que des identificateurs : ils permettent de spécifier de quoi l'on parle. On peut faire une analogie avec les numéros ISBN qui permettent à deux personnes de s'assurer qu'elles parlent du même ouvrage. Ainsi, si toutes les descriptions se trouvent dans un même environnement (par exemple, une application de documentation) les descriptions de `Book` et de `Autobiography` seront fusionnées avec celles de l'éditeur.

Cette seconde particularité – la non-obligation pour une URI de donner accès à la ressource – contribue à la robustesse du web sémantique. Les documents RDF peuvent être consultés hors-ligne et il n'y a plus de liens cassés, comme il peut y en avoir avec le web actuel.

Comme on peut le voir sur le dernier exemple, une même ressource peut avoir plusieurs URI, leurs descriptions étant fusionnées à l'aide de la primitive `sameAs` de OWL (voir ci-dessous).

Manipuler les graphes RDF

Cette sémantique « intuitive » ne suffisant pas à un traitement automatique, il faut munir les documents RDF d'une sémantique formelle, que peut suivre un programme informatique, ce qui fut fait récemment. La sémantique d'un document RDF est exprimée en théorie des modèles [Hayes 2004]. L'objectif est de donner des contraintes sur les mondes qui peuvent être décrits par un document RDF. Parallèlement à la validité d'un document par rapport à un schéma, la consistance d'un document par rapport à la sémantique du langage correspond à l'existence d'un modèle. La sémantique permet de déterminer ce qu'est une conséquence d'un document (c'est-à-dire un triplet satisfait dans tous les modèles du document).

Une fois les ressources annotées, il reste à vérifier que nous sommes effectivement capables de les retrouver en interrogeant la base de descriptions. Il est possible de retrouver les documents dont la requête est une conséquence de la description. Par exemple, nous pouvons formuler la requête suivante :

> *« retrouver les ouvrages de type Autobiographie dont l'auteur est une personne qui a comme nom Bertrand Russell, ainsi que leur numéro ISBN ».*

Cette requête se traduit dans le langage RDQL ("RDF Data Query Language") [Seaborne 2004] comme indiqué ci-dessous :

```
SELECT ?x, ?y, ?z
WHERE (?x, <rdf:type>, <bib:Autobiography>),
      (?x, <bib:author>, ?y),
      (?y, <rdf:type>, <foaf:Person>),
      (?y, <foaf:name>, Bertrand Russell),
      (?x, <dc:identifier>, ?z)
USING rdf FOR <http://www.w3.org/1999/02/22-rdf-syntax-ns#>,
      bib FOR <http://www.thebibliography.org/>,
      foaf FOR <http://xmlns.com/foaf/0.1/>,
      dc FOR <http://dublincore.org/2001/08/14/dces#>
```

Figure 4 : Une requête en RDQL.

En particulier,

```
BA4498, http://www.bertrandrussell.com/, 041522862X
```

est une réponse pertinente pour la requête ci-dessus. De surcroît, en tirant parti de ce qui vient d'être expliqué, si l'on recherche un ouvrage de Bertrand Russell auquel s'applique la catégorie 107 de Dewey, cette description est aussi une réponse (c'est-à-dire que la réponse peut être distribuée sur le web).

Les capacités de RDF utilisées jusque-là ne sont pas vraiment révolutionnaires pour celui qui connaît les outils de recherche documentaire. Cependant, on verra qu'elles contiennent, en germe des facilités impressionnantes. Mais d'abord, nous allons présenter la capacité de modélisation offerte par le web sémantique. C'est effectivement elle qui permettra de répondre aux requêtes telles que posées dans l'introduction.

3. MODÉLISER SON DOMAINE

Même si RDF est un format très ouvert et malléable (comme on vient de le voir, il est possible d'ajouter des propriétés à propos de n'importe quel élément), si l'on désire interroger une base documentaire ou le web, il est nécessaire d'identifier les termes sur lesquels l'identification doit porter. Ces termes identifiant les grandes classes d'objets impliqués et leurs relations constituent un modèle conceptuel du domaine [Brodie 1984] et sont décrits dans une "ontologie" [Staab 2004]. Les ontologies fournissent le vocabulaire propre à un domaine et fixent – avec un degré de formalisation variable – le sens des concepts et des relations entre ceux-ci. Ces concepts (et ces relations) sont généralement organisés par une relation de spécialisation.

Exprimer de la connaissance sur le web est l'ambition du web sémantique. Au-delà de ce simple mot d'ordre, diffuser des ontologies sur le web est le moyen de permettre à d'autres de se les approprier, de les étendre et de les réutiliser. On va donc devoir décrire les ontologies concernant les informations présentes dans le web sémantique.

Il existe une tradition de développement de langages d'expression d'ontologies en représentation de connaissance. Mais si les ontologies doivent s'échanger librement sur le web, il est nécessaire de les intégrer plus aux langages du web. Pour cela on dispose de langages compatibles avec RDF et permettant de le contraindre. Ces langages sont RDFS et OWL qui sont présentés dans la suite.

RDF Schéma : les premiers pas

À partir de RDF, le langage RDF Schéma (RDFS) a été développé [Brickley 1999; 2004]). Il a pour but d'étendre RDF en décrivant plus précisément les ressources utilisées pour étiqueter les graphes. Pour cela, il fournit un mécanisme permettant de spécifier les classes dont les instances seront des ressources, comme les propriétés. RDFS s'écrit toujours à l'aide de triplets RDF, en définissant la sémantique de nouveaux mots-clés comme:

- `<#041522862X rdf:type Autobiography>` la ressource `#041522862X` a pour type `Autobiography` (qui est donc une classe) ;

<Autobiography rdfs:subClassOf Biography> la classe Autobiography est une sous-classe de Biography, toutes les instances de Autobiography sont donc des instances de Biography ;

<bib:publisher rdf:type rdfs:Property> affirme que bib:publisher est une propriété (une ressource utilisable pour étiqueter les arcs) ;

<bib:publisher rdfs:range Publisher> affirme que toute ressource utilisée comme extrémité d'un arc étiqueté par bib:publisher sera une instance de la classe Publisher.

Ces primitives constituent la base de tout langage d'ontologie, permettant de signifier l'appartenance d'un objet à une catégorie, de déclarer la relation de généralisation entre catégories et de typer des objets reliés par une relation. RDFS est cependant plus complexe car tous les termes y sont eux-mêmes définit dans le langage : tout y est ressource ; Relation, Resource et Class sont des classes ; type et subClassOf sont des relations… Ceci en faut un langage difficile à interpréter car réflexif. Nous nous concentrerons donc uniquement sur les primitives.

Ainsi que le montre l'extrait suivant, il est possible de décrire une hiérarchie de classes de références (Reference est plus général que Book qui est plus général que Biography qui l'est plus que Autobiography). Il est aussi possible de décrire les types d'objets attendus aux extrémités des arcs : ainsi la propriété author s'applique à une Reference et a pour valeur une Person. On notera ici que si un ouvrage a plusieurs auteurs, ceux-ci seront exprimés à l'aide d'arcs multiples.

```
<rdfs:Class rdf:ID="Reference" />

<rdfs:Class rdf:ID="Book">
  <rdfs:subClassOf rdf:resource="#Reference" />
</rdfs:Class>

<rdfs:Class rdf:ID="Biography">
  <rdfs:subClassOf rdf:resource="#Book" />
</rdfs:Class>

<rdfs:Class rdf:ID=" Autobiography ">
  <rdfs:subClassOf rdf:resource="#Biography" />
</rdfs:Class>

<rdf:Property rdf:ID="bib:author">
  <rdfs:domain rdf:resource="#Reference"/>
  <rdfs:range rdf:resource="&foaf;Person"/>
</rdf:Property>

<rdf:Property rdf:ID="bib:title">
  <rdfs:domain rdf:resource="#Reference"/>
  <rdfs:range rdf:resource="&xsd;string"/>
</rdf:Property>

<rdfs:Class rdf:ID="Publisher">
  <rdfs:subClassOf rdf:resource="&foaf;Organization" />
</rdfs:Class>
```

Figure 5 : Une taxonomie en RDFS.

RDFS ne fournit que des mécanismes très primitifs pour spécifier ces classes. Indépendamment, différents langages de description d'ontologie ont été développés pour le web. Le projet européen ontoknowledge avait développé le langage OIL comme une extension d'XML Schéma (proche de RDFS) offrant des primitives inspirées des logiques de descriptions. Le programme américain DAML a pour sa part proposé le langage DAML-ONT, fondé sur RDF, et plus proche des langages objets. Ces deux langages ont été fusionnés en un langage connu sous le nom de DAML+OIL qui a servi de base à l'élaboration du langage d'ontologies pour le web OWL [Dean 2004].

OWL fournit un grand nombre de constructeurs permettant d'exprimer de façon très fine les propriétés des classes définies. À l'instar de RDF, OWL est une recommandation du W3C depuis février 2004.

On peut appréhender le langage OWL en observant la définition d'un livre et d'une biographie présentée ici. Disons qu'en plus des primitives de RDFS, OWL permet de contraindre plus précisément la description des classes (en les décrivant comme union, intersection, complémentaire d'autres descriptions ou comme l'ensemble d'un certain nombre d'individus), des domaines de relations (en spécifiant le type de toutes leurs valeurs, ou d'un certain nombre de leurs valeurs) ou des relations (en les déclarant transitives, symétriques ou en spécifiant leur inverse). Par ailleurs, il est possible de déclarer que deux classes ou ressources sont équivalentes ou, au contraire, différentes.

Ainsi, les documents précédents peuvent être régis par l'ontologie OWL contenant les classes suivantes:

```
<owl:Class rdf:ID="Book">
  <owl:intersectionOf>
    <owl:Class rdf:resource="#Reference" />
    <owl:Restriction>
      <owl:onProperty rdf:resource="#title" />
      <owl:minCardinality
rdf:datatype="&xsd;Integer">1</owl:minCardinality>
    </owl:Restriction>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#publisher" />
      <owl:allValuesFrom rdf:resource="#Publisher" />
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>

<owl:Class rdf:ID="Biography">
  <owl:intersectionOf>
    <owl:Class rdf:resource="#Book"/>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#object"/>
      <owl:allValuesFrom rdf:resource="&foaf;Person"/>
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>
```

Figure 6 : La première expression OWL s'interprète comme « la classe Book est l'intersection de la classe Reference, des objets dont la propriété title a au moins une valeur et la propriété publisher a pour valeur un Publisher » ; la seconde signifie que « Les Biography sont des Book dont l'objet est une Person»

Dans les ontologies comme dans les descriptions, l'ouverture est de mise. Ainsi, les ontologies peuvent faire référence à des concepts introduits dans d'autres ontologies. On peut le voir dans les descriptions précédentes où la classe `Person` n'est pas définie dans l'ontologie bibliographique mais dans l'ontologie FOAF[1] ("friend-of-a-friend"). Similairement, on a pu voir sur la figure 5 que la classe `Publisher` est une sous-classe de la classe FOAF `Organization` : il est donc possible, non seulement d'utiliser mais de prolonger une ontologie existante (pour ses propres besoins, bien entendu). Il est même possible d'aller plus loin et d'introduire des classes au sein d'une hiérarchie existante (par exemple, dans une ontologie du droit, de déclarer que la classe `PersonneMorale` est une superclasse des classes FOAF `Person` et `Organization`). Il est tout aussi possible de déclarer deux classes ou deux relations d'ontologies différentes comme équivalentes ou au contraire incompatibles. Enfin, il est possible de contraindre certaines caractéristiques de classes extérieures à l'ontologie. Ainsi, il est possible d'exprimer que l'image de la propriété `publisher` est un objet de la classe FOAF `Organization` qui possède au moins une `address` !

La syntaxe d'un document OWL est donnée par celle des différents constructeurs utilisés dans ce document. Elle est le plus souvent donnée sous la forme de triplets RDF.

Pour des raisons de modularité, OWL a été fractionné en trois langages distincts :

- OWL Lite ne contient qu'un sous-ensemble réduit des constructeurs disponibles, mais son utilisation assure que la comparaison de types pourra être calculée ;
- OWL DL contient l'ensemble des constructeurs, mais avec des contraintes particulières sur leur utilisation qui assurent la décidabilité de la comparaison de types. Par contre, la grande complexité de ce langage semble rendre nécessaire une approche heuristique ;
- OWL Full, sans aucune contrainte est la fusion de RDFS et de OWL DL (c'est-à-dire qu'il abolit la distinction forcée entre classes et ressources); le problème de comparaison de types (savoir si un type est plus général qu'un autre) y est vraisemblablement indécidable.

Nous donnons ci-dessous un panorama des constructeurs utilisés dans OWL, dans une syntaxe simplifiée (les mots-clés réservés de OWL sont préfixés de `owl:`) :

OWL Lite

- reprend tous les constructeurs de RDF (c'est-à-dire fournit des mécanismes permettant de définir un individu comme instance d'une classe et de mettre des individus en relation),
- utilise les mots-clés de RDFS (`rdfs:subClassOf`, `rdfs:Property`, `rdfs:subPropertyOf`, `rdfs:range`, `rdfs:domain`), avec la même sémantique,
- permet de définir une nouvelle classe (`owl:Class`) comme étant plus spécifique ou équivalente à une intersection d'autres classes,
- `owl:sameAs` et `owl:differentFrom` permettent d'affirmer que deux individus sont égaux ou différents,
- des mots-clés permettent d'exprimer les caractéristiques des propriétés : `owl:inverseOf` sert à affirmer qu'une propriété *p* est l'inverse de *p'* (dans ce cas, le

[1] http://www.foaf-project.org

triplet <*s p o*> a pour conséquence <*o p' s*>) ; d'autres caractéristiques sont la transitivité (`owl:TransitiveProperty`), la symétrie (`owl:SymmetricProperty`),

— `owl:allValuesFrom` associe une classe *C* à une propriété *P*. Ceci définit la classe des objets *x* tels que si <*x P y*> est une relation, alors la classe de *y* est *C* (quantification universelle de rôle en logique de descriptions). `owl:someValuesFrom` encode la quantification existentielle de rôle,

— `owl:minCardinality` (respectivement `owl:maxCardinality`) associe une classe *C*, une propriété *P*, et un nombre entier *n*. Ceci définit la classe des objets *x* tels qu'il existe au moins (respectivement au plus) *n* instances différentes *y* de *C* avec <*x P y*>. Pour des raisons d'efficacité algorithmique, OWL Lite ne permet d'utiliser que des entiers égaux à 0 ou 1. Cette restriction est levée dans OWL DL.

OWL DL

— reprend tous les constructeurs d'OWL Lite,

— permet tout entier positif dans les contraintes de cardinalité,

— `owl:oneOf` permet de décrire une classe en extension par la liste de ses instances,

— `owl:hasValue` affirme qu'une propriété doit avoir comme objet un certain individu,

— `owl:disjointWith` permet d'affirmer que deux classes n'ont aucune instance commune,

— `owl:unionOf` et `owl:complementOf` permettent de définir une classe comme l'union de deux classes, ou le complémentaire d'une autre classe.

OWL Full

— reprend tous les constructeurs d'OWL DL,

— reprend tout RDF Schéma,

— permet d'utiliser une classe en position d'individu dans les constructeurs.

Nous n'avons pas cité ici certains constructeurs, qui peuvent être trivialement implémentés grâce à ceux que nous avons évoqués (par exemple `owl:sameClassAs`, servant à affirmer que deux classes sont identiques, peut être écrit grâce à deux `rdfs:subClassOf`).

Apports

Cette fois encore, OWL est doté d'une sémantique en théorie des modèles permettant de spécifier tout ce qui est conséquence d'un ensemble d'assertions de OWL. Elle est directement issue des logiques de descriptions [Patel-Schneider 2004]. La sémantique associée aux mots-clés de OWL est plus précise que celle associée aux documents RDF représentant une ontologie OWL. Elle permet donc plus de déduction, comme de retourner des `Autobiography` à une requête concernant les `Biography`. Ainsi, plus on fait intervenir d'ontologies — pertinentes — dans l'évaluation d'une requête, plus les réponses sont pertinentes.

L'intérêt d'utiliser une ontologie est de pouvoir valider ses sources : par exemple, être sûr que l'éditeur d'un ouvrage est bien un `Publisher` et que son adresse est connue. Mais c'est surtout d'augmenter le nombre d'inférences possibles à partir des données. Plus précisément, une telle modélisation devrait permettre de répondre à des requêtes complexes utilisant les modèles pour compléter la connaissance disponible.

Ainsi, si l'on étudie la vie de Bertrand Russell, on voudra trouver les Biographies de tous les coauteurs et élèves de Bertrand Russell. Bien entendu une telle requête devra retourner les Biographies et les Autobiographies ; elle devra aussi trouver les personnes dont le doctorat a eu comme superviseur Bertrand Russell. Ces deux informations n'ont pas à être dans la requête, il suffit qu'elles soient présentes dans l'ontologie pour être exploitées.

L'information sur les élèves d'un auteur n'est typiquement pas celle que l'on trouve dans les notices bibliographiques actuelles mais que l'on peut trouver sur les pages web des auteurs comme sur les sites qui leur sont consacrés. Leur encodage en RDF et OWL permettra de les exploiter.

Des moteurs d'inférence, encore trop peu nombreux, ont déjà été implémentés pour des sous-ensembles significatifs de OWL DL (dans le cadre des logiques de descriptions) et sont utilisés dans divers outils (OilEd, Protégé…).

L'étape suivante conduirait à exprimer des règles dans le web sémantique afin de permettre l'expression des mécanismes de fonctionnement, des contraintes ou d'introduire un aspect opportuniste dans le web sémantique. Il y a suffisamment de volonté sur ce thème pour que l'on anticipe un tel langage mais rien n'est pour l'instant particulièrement formalisé [Horrocks 2004].

```
<r:Rule>
  <r:premisse>
    <Autobiography rdf:ID="?x">
      <author rdf:resource="?y"/>
    </Autobiography>
  </r:premisse>
  <r:conclusion>
    <Autobiography rdf:ID="?x">
      <dc:subject rdf:resource="?y"/>
    </Autobiography>
  </r:conclusion>
</r:Rule>
```

Figure 7 : On peut facilement imaginer un langage de règles permettant d'écrire que « les Autobiographies ont pour objet leur auteur ».

Autres langages

Un langage alternatif est celui des cartes topiques ("Topic maps") proposé par l'ISO [Biezunski 2000]. Il est basé sur trois types d'entités : les thèmes (ou "topics"), les associations et les portées. Les thèmes correspondent aux ressources, les associations aux relations et les portées sont des ensembles de thèmes qui permettent de circonscrire le contexte dans lequel une assertion est valide. À ceci on peut ajouter les noms permettant d'identifier les thèmes (l'approche est donc multilingue d'emblée) et les occurrences agissant comme des médiateurs entre les cartes topiques et le monde extérieur (un type de donnée particulier ou un objet identifiable par un URI). Les cartes topiques sont un modèle très versatile et peuvent donc s'adapter à différentes situations. Cependant, l'absence d'une sémantique claire du formalisme

rend difficile son appréhension. Par ailleurs, si ce langage est utilisé dans certains cercles [Garshol 2003], il n'a pas su s'imposer au sein du web sémantique.

On peut se demander pourquoi ne pas se contenter d'utiliser XML pour construire ce web sémantique [Euzenat 2003], ou plutôt XML à la place de RDF et XML Schéma [Thompson 2001] à la place de OWL. Il y a plusieurs raisons à cela. La première tient à l'ouverture : introduire la possibilité d'étendre tout document XML, ainsi que présenté ci-dessus (tout attribut pourrait avoir n'importe quelle valeur), va à l'encontre de la philosophie de XML Schéma qui tend plutôt à contraindre. Par ailleurs, ces deux langages ne possèdent pas de sémantique ce qui rend difficile la justification des inférences que l'on pourrait y faire. Enfin, un schéma XML n'est pas une ontologie car son but est de valider un document, pas d'en définir les conséquences. Il est donc difficile de l'utiliser pour faire des inférences.

4. EXEMPLES DE CADRES APPLICATIFS LIÉS À L'INGÉNIERIE DOCUMENTAIRE

Après cette rapide introduction aux langages du web sémantique utilisables dans la pratique documentaire, on s'intéresse ici à leur utilisation pouvant être mise à profit dans cette pratique. On présentera d'abord trois types de ressources utilisant les technologies du web sémantique pouvant être utiles dans les tâches de documentation et comment elles peuvent être exploitées.

Dublin core

Le Dublin Core[2] [DCMI 2003] est un vocabulaire (ou une ontologie) minimal pour l'indexation des pages web. Il a été défini sous l'égide de l'"Online Computer Library Center", et maintenant d'un forum ouvert. L'ensemble d'éléments du Dublin Core est une version très réduite des notices MARC qui ne comprend que 15 éléments :

`Title`	Un nom donné à la ressource.
`Creator`	Une entité ayant créé la ressource.
`Subject`	Un thème de la ressource.
`Description`	Une description du contenu de la ressource.
`Publisher`	Une entité responsable de la publication de la ressource.
`Contributor`	Une entité ayant contribué au contenu de la ressource.
`Date`	La date d'un événement dans le cycle de vie de la ressource (ISO 8601).
`Type`	La nature ou le genre du contenu de la ressource (DCMI).
`Format`	La manifestation physique ou numérique de la ressource (MIME).
`Identifier`	Une référence non-ambiguë pour la ressource dans le contexte courant (URI, DOI, ISBN).
`Source`	Une référence à une ressource de laquelle la ressource décrite est dérivée.
`Language`	Un langage dans lequel est exprimé le contenu de la ressource (ISO 3166).
`Relation`	Une référence à une ressource liée.
`Coverage`	L'étendue du contenu de la ressource.
`Rights`	Des informations sur les droits sur la ressource.

[2] http://dublincore.org

Utiliser ces éléments pour décrire une ressource est une garantie d'être (partiellement) compris par l'ensemble des programmes qui supporte le schéma simple du Dublin Core. Par ailleurs, on peut remarquer que le vocabulaire à utiliser dans ces champs réutilise très souvent d'autres standards, ce qui constitue une bonne pratique.

L'ensemble des éléments du Dublin Core fut une des premières « ontologies » décrites en RDF [Beckett 2002].

Creative commons

Notre deuxième exemple est celui de Creative commons[3], une initiative pour proposer aux créateurs de contenu qui veulent le distribuer de manière ouverte un moyen de produire très rapidement une licence d'utilisation. L'intérêt de cette licence est qu'elle est simultanément engendrée en trois formats à partir d'une source unique : une version à l'aide de pictogrammes pour les gens normaux, une version en RDF destinée à être traitée automatiquement et une version en langage juridique pour les avocats.

```
<rdf:RDF xmlns="http://web.resource.org/cc/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<Work rdf:about="">
   <dc:type rdf:resource="http://purl.org/dc/dcmitype/Text" />
   <license  rdf:resource="http://creativecommons.org/licenses/by-nc-
nd/2.0/" />
</Work>

<License rdf:about="http://creativecommons.org/licenses/by-nc-nd/2.0/">
   <permits rdf:resource="http://web.resource.org/cc/Reproduction" />
   <permits rdf:resource="http://web.resource.org/cc/Distribution" />
   <requires rdf:resource="http://web.resource.org/cc/Notice" />
   <requires rdf:resource="http://web.resource.org/cc/Attribution" />
   <prohibits rdf:resource="http://web.resource.org/cc/CommercialUse" />
</License>

</rdf:RDF>
```

Figure 8 : Licence Creative commons en RDF.

[3] http://creativecommons.org

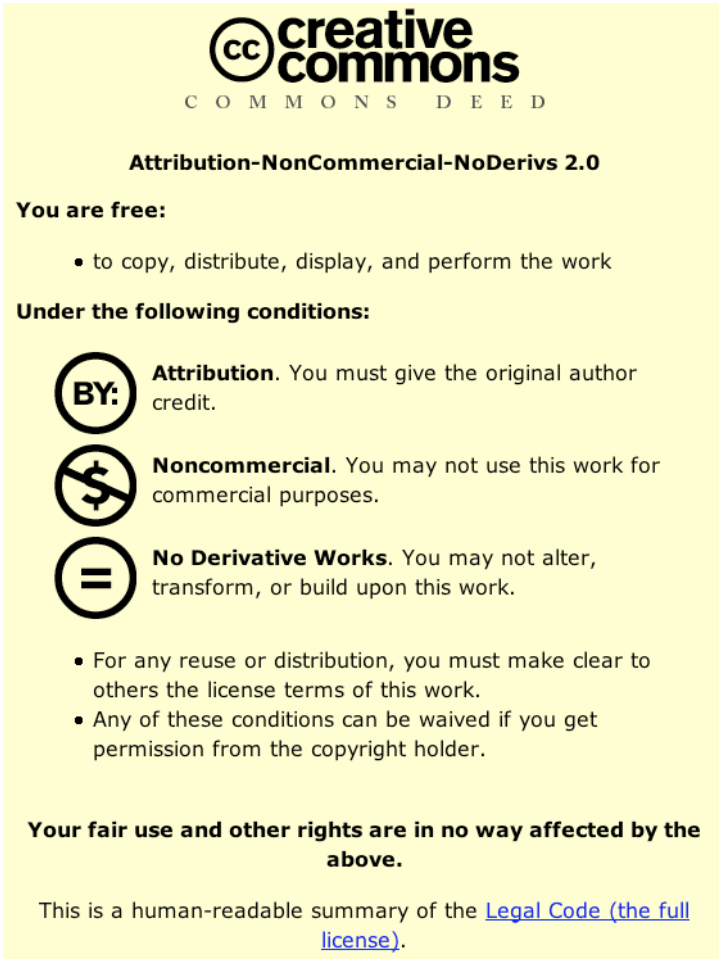

Figure 9 : Licence Creative commons destinée à la lecture.

Ainsi, lorsqu'un documentaliste accède à un document indexé à l'aide d'une licence Creative commons, il devrait savoir s'il a la possibilité de mettre à la disposition de ses usagers le document (Distribution) ou s'il doit se contenter de noter l'URL. Qui plus est, une application de recherche bien faite devrait être capable d'interpréter la licence en RDF et de ne lui proposer que les alternatives accordées.

On peut remarquer que le format utilisé en standard par Creative commons utilise les éléments du Dublin Core pour annoter les types de média.

RDF Site summary

Notre troisième exemple concerne RSS, un format pour diffuser de l'information sur le web. Son propos est de baliser une ressource de type flux d'information.

Il y a maintenant trois versions de RSS utilisées : Really Simple Syndication (RSS 0.9x et RSS 2.0) est uniquement décrite en XML alors que RDF Site Summary[4] (RSS 1.0) est décrite en XML et RDF.

Il est tout à fait imaginable que les éditeurs publient les annonces de leurs nouveaux livres ou les sommaires de leurs périodiques à l'aide d'un fil RSS. Ainsi, voici un fil imaginaire

[4] http://web.resource.org/rss/1.0

annonçant les sorties et rééditions prochaines d'un éditeur. Comme on peut le voir les informations sont très courtes, mais l'aspect relationnel permet aux lecteurs humains comme aux machines d'aller chercher plus d'information sur le site.

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns="http://purl.org/rss/1.0/"
 xmlns:dc="http://purl.org/dc/elements/1.1/">
  <channel rdf:about="http://www.routledge.com/news.rss">
    <title>Routledge news</title>
    <link>http://www.routledge.com/news/</link>
    <description>News to our costumers and friends</description>
    <items>
      <rdf:Seq>
        <rdf:li rdf:resource="http://www.routledge.com/ann041522862X"/>
        <rdf:li rdf:resource="http://example.com/2002/09/02/"/>
      </rdf:Seq>
    </items>
  </channel>
  <item rdf:about="http://www.routledge.com/ann041522862X">
     <title>Wittgenstein's Tractacus to be reprinted</title>
     <link>http://example.com/2002/09/01/</link>
     <description>The Tractatus Logico-philosophicus of Ludwig Wittgenstein
is scheduled to be reprinted next year.</description>
    <dc:description>Philosophy/Logics/Language</dc:description>
     <dc:date>2002-09-17</dc:date>
  </item>
</rdf:RDF>
```

Figure 10 : Fil RSS.

On remarquera encore ici l'utilisation des balises du Dublin Core pour permettre un filtrage plus facile de l'information disponible.

On imagine donc très bien, à partir de ces briques de base, comment une institution comme l'INRIA pourrait diffuser un fil référençant toutes les publications de ses chercheurs, indexées par les éléments du Dublin Core. On comprend toujours comment l'information sur les droits régissant les publications mises en ligne peut être insérée à l'aide des licences de Creative commons. On comprend enfin comment un centre de documentation intéressé aux sujets traités par l'INRIA peut s'abonner à ce fil, filtrer les informations et les sujets qui l'intéressent et le diffuser à ses propres membres…

En pratique…

En pratique, on peut imaginer beaucoup d'autres manières d'utiliser les technologies du web sémantique dans une activité de gestion de documentation. La raison principale est la possibilité de mettre en relation, plus qu'avant, le domaine des documents sur lequel on dispose d'une compétence bien établie avec l'ensemble des activités humaines sur lesquels d'autres ont une compétence de pointe.

Cette capacité est une caractéristique du web actuel : ce sont les internautes qui ont fait la base de référence des CD musicaux qu'aucun groupement d'éditeurs n'est capable de réaliser[5], ce

[5] http://www.gracenote.com

sont eux qui ont créé la plus grande base de films également[6], et aucun éditeur de magazine de bande dessinée n'est capable de décliner son sommaire comme sur Internet[7]. Cette capacité est possible par la collaboration de millions d'internautes sur des sujets très pointus. Ces bases de données sont structurées, et il n'est pas difficile de les exporter et les interroger en RDF ce qui permet d'utiliser ce type d'information dans d'autres contextes.

On peut imaginer un scénario différent du précédent dans lequel un centre de documentation reçoit une demande urgente pour un article publié dans les actes d'une conférence difficiles à obtenir. L'avènement du web sémantique devrait par contre permettre d'obtenir facilement le programme de la conférence en question. Se rendre compte qu'un chercheur du même institut y a présenté une communication ne devrait pas être compliqué, pas plus qu'obtenir ses coordonnées à l'aide de sa description FOAF[1] et de le contacter pour emprunter les actes demandés.

5. INDEXATION DE DONNÉES AUDIO-VISUELLES OU MULTIMÉDIA

Si le web est encore aujourd'hui principalement constitué de documents textuels, la part des documents multimédias (images, vidéos, sons) est de plus en plus importante. La fin des années 90 ayant été marquée par la numérisation progressive de la chaîne de production des contenus et l'émergence des systèmes techniques permettant de les traiter, il est désormais très facile de produire directement tout type de document numérique pour les mettre sur le web. L'audiovisuel a été le grand bénéficiaire de cette vague avec l'arrivée du numérique au cinéma et à la maison (appareil photo numérique, caméra numérique…). Mais la multiplication de ce type de ressources sur le web vient encore complexifier son exploitation générale. En particulier, l'apparition du numérique et l'association qu'il permet entre l'audiovisuel et l'information amènent de nombreuses évolutions dans le domaine de l'archivage et de la consultation des documents audiovisuels. Le système technique devenant unique, les problèmes concernent maintenant le format des descriptions et le type des informations qu'elles doivent contenir pour permettre l'instrumentation des contenus audiovisuels. Les documentalistes voient alors leurs pratiques profondément bouleversées en raison des nouvelles possibilités technologiques.

Ainsi, la question de l'accès du grand public au patrimoine audiovisuel public devient de plus en plus d'actualité à présent que les technologies rendent une telle perspective envisageable. L'INA (Institut National de l'Audiovisuel) s'est par exemple engagé à fournir d'ici à 2005 un service de vidéo à la demande sur le web pour plusieurs milliers d'heures extraites de son fonds d'archives et correspondant à des émissions diffusées à la télévision ou à la radio et libérées de droits. Dans cette optique, il est nécessaire d'explorer les solutions techniques qui pourraient rendre un tel accès possible. D'une manière générale, le système technique proposé doit pouvoir faire le lien entre le contenu audiovisuel et sa description qui devient alors la condition *sine qua non* à son accès. La mise en place de ce service nécessite une représentation de la structure et du contenu du catalogue documentant le fonds disponible, afin qu'un utilisateur puisse l'interroger ou le parcourir.

---

[6] http://www.imdb.com
[7] http://bdoubliees.com

La recherche de séquences audiovisuelles (AV) particulières ou plus généralement la manipulation du fonds (production de nouveaux documents, thématisation…) s'effectue donc grâce à la description des documents. Actuellement, plusieurs équipes de documentalistes sont chargées de décrire manuellement les émissions diffusées. Ce processus de description documentaire peut se résumer en trois étapes :

— le catalogage : il s'agit de prendre les éléments objectifs et extrinsèques au contenu d'un document (nom, auteur, producteur, durée, droits…) et de l'identifier à l'intérieur d'une programmation (titre, chaîne et heure de diffusion…) ;
— le découpage structurel : il s'agit de localiser dans le programme des entités temporelles pertinentes pour une application donnée et de leur apposer un genre audiovisuel et une thématique générale, afin de rendre compte de la structure logique du document ;
— la caractérisation des segments : il s'agit enfin de décrire le contenu proprement dit des entités repérées à l'étape précédente.

Les langages de structuration documentaire (appartenant à la famille XML) sont généralement utilisés tout le long de ce processus. En effet, souvent bien outillés techniquement, ils permettent en outre de contraindre ou d'exprimer au mieux la structure logique d'un document. L'utilisation de listes d'autorités pour caractériser les genres audiovisuels ou les thématiques générales, de mots clés issus d'un thésaurus et du texte libre pour décrire le contenu vient compléter la liste des outils documentaires mis à la disposition des documentalistes. Ces derniers peuvent ainsi, en visionnant les programmes, produire des notices documentaires qui décrivent les documents audiovisuels.

Ce cadre de description étant fixé, la recherche de séquences audiovisuelles particulières s'avère parfois difficile, notamment si elle est effectuée par des personnes non-documentalistes, et à plus forte raison non professionnelles de l'audiovisuel. En effet, la description suppose une reformulation du contenu des documents pour une exploitation. Le raisonnement est typiquement une manipulation qui permet, par exemple, de mieux satisfaire les requêtes lors de l'interrogation de la base des descriptions. Cependant, le type de langage utilisé (documentaire), qui restreint les inférences à la seule validation de structure, et l'emploi du texte libre ou de thésaurus pour décrire le contenu, qui empêche de véritablement contrôler la sémantique des descriptions, limitent sérieusement les possibilités de raisonnement. Dès lors, quel langage ou mécanisme faut-il utiliser pour pouvoir raisonner dans les descriptions documentaires ? Nous détaillons dans la suite comment combiner un langage documentaire particulier (MPEG-7) et les langages OWL et RDF pour résoudre ce problème [Troncy 2004].

Dès 1996, le comité MPEG a souligné la nécessité d'une solution puissante pour identifier et décrire les données multimédias. L'obstacle majeur mis en lumière par le comité était *le manque d'une représentation standard, compréhensible et flexible pour le multimédia*. Pour le résoudre, le comité a élaboré la norme MPEG-7 [MPEG7 2001]. Ce langage définit la notion d'*outils* de description multimédia. Dans la terminologie de la norme, les outils font référence à un ensemble de descripteurs dont les valeurs permettent de décrire des caractéristiques physiques audiovisuelles (couleur, texture, mouvement…), à un ensemble de schémas de descriptions qui permettent d'organiser les descripteurs dans des modèles pour les objets multimédias, et au

langage de définition des descriptions (DDL) qui permet d'encoder le tout. Il est à noter que les descripteurs de bas niveau (couleur dominante, mouvement de caméra, spectre sonore, mélodie…) prédominent largement dans la norme car celle-ci a, pour l'essentiel, été élaborée par la communauté de l'analyse automatique et du traitement du signal.

Le langage de définition des descriptions est une partie centrale de la norme MPEG-7 puisqu'il fournit les règles syntaxiques pour exprimer et combiner les descripteurs et les schémas de description. C'est le langage XML Schéma [Thompson 2001] qui a été retenu comme langage de définition des descriptions pour la norme. Il permet de spécifier la nature et l'organisation des éléments susceptibles d'intervenir dans une instance de document conforme à la classe qu'il est en train de définir. Synthétiquement, XML Schéma permet de déclarer les éléments (et leurs attributs) susceptibles d'apparaître dans un document XML en précisant leur ordre et leur arrangement, de différencier les types simples des types complexes (en précisant leurs usages), et de définir ces derniers, de dériver des types existants (par restriction ou par extension) en contrôlant ces dérivations, ou encore de réutiliser des définitions de type ou des déclarations d'éléments grâce au mécanisme des espaces de noms.

L'intégration des caractéristiques structurelles et sémantiques est considérée comme la contribution la plus importante du langage MPEG-7. La description structurelle est basée sur l'idée de *segment* qui est une portion spatiale, temporelle ou spatio-temporelle du contenu audiovisuel. Un segment se spécialise en différents types utilisables selon le média à décrire (audio, image, vidéo, multimédia). Ces types ajoutent les notions de *temps média*, qui permet d'obtenir un segment temporellement connecté, et de *masque* qui permet de construire des régions et des segments non connectés spatialement ou temporellement. Ils autorisent aussi certaines décompositions (dans le temps, dans l'espace, par média) selon le média auquel ils sont liés, et ils définissent alors les types résultats issus de ces découpages. La description sémantique, quant à elle, traite du monde dépeint dans le contenu audiovisuel. L'approche adoptée par MPEG-7 est un modèle centré sur l'événement interprété comme un moment où il se passe quelque chose. Les objets, les personnes et les lieux permettent de décrire cet évènement ainsi que le temps où il se produit. De plus, ces entités ont des propriétés qui les relient. Enfin, MPEG-7 a laissé la porte ouverte à la création de structures de connaissance très simple, de type thésaurus, à travers les *schémas de classification*. Ceux-ci permettent de définir des termes et de les organiser grâce à cinq relations : *plus spécifique*, *plus général*, *est lié à*, *utilise* et *est utilisé par*. Cependant, ces schémas sont vus comme des ressources externes, utilisables lors de la description pour valuer des entités, mais ils ne peuvent pas être utilisés dans un schéma pour contraindre la structure d'une classe de documents.

En première conclusion, nous remarquons tout d'abord que les descripteurs standardisés proposés par MPEG-7 sont de trop bas niveau pour prendre en compte tous les besoins de description (par exemple, ceux du type de l'INA), puisque ceux-ci sont essentiellement liés aux caractéristiques physiques des informations audiovisuelles. Ainsi, pour décrire la structure d'un document, il n'est pas possible, par exemple, de typer les segments selon leur genre audiovisuel (reportage, séquence plateau, interview…) ou selon leur thématique générale (sports, sciences, politique, économie…). De même, pour décrire le contenu, les descripteurs proposés par la

norme sont encore loin d'être suffisants pour décrire de manière fine une scène particulière. D'autre part, nous constatons qu'il est nécessaire d'exprimer la sémantique de ces descripteurs dans un langage formel et utilisable par la machine pour véritablement permettre la manipulation du contenu multimédia par les machines. Mais nous affirmons que MPEG-7 ne permet pas de jouer ce rôle puisque le langage ne possède pas de sémantique formelle et que la définition des types se restreint aux seuls mécanismes de sous-typage offerts par XML Schéma. Ce dernier permet donc d'ajouter de la structure, mais il ne peut pas exprimer sa sémantique.

La formalisation des descriptions de documents audiovisuels étant une piste pour rendre plus aisée la recherche ou plus généralement la manipulation de ces documents, nous décrivons dans la suite une architecture permettant la construction d'une base de connaissances sur laquelle il est possible d'effectuer des raisonnements tant sur la structure que sur le contenu. Plus précisément, nous montrons comment combiner les langages MPEG-7 et OWL pour produire des descriptions de documents audiovisuels. Cette architecture (figure 11) a comme base une ontologie de l'audiovisuel dont on traduit une partie dans un langage documentaire et la modélisation d'une autre ontologie de domaine pour décrire le contenu. Le découpage temporel d'une émission particulière et la description effective de son contenu génère un ensemble de faits qui viennent enrichir une base de connaissances, autorisant ainsi le raisonnement.

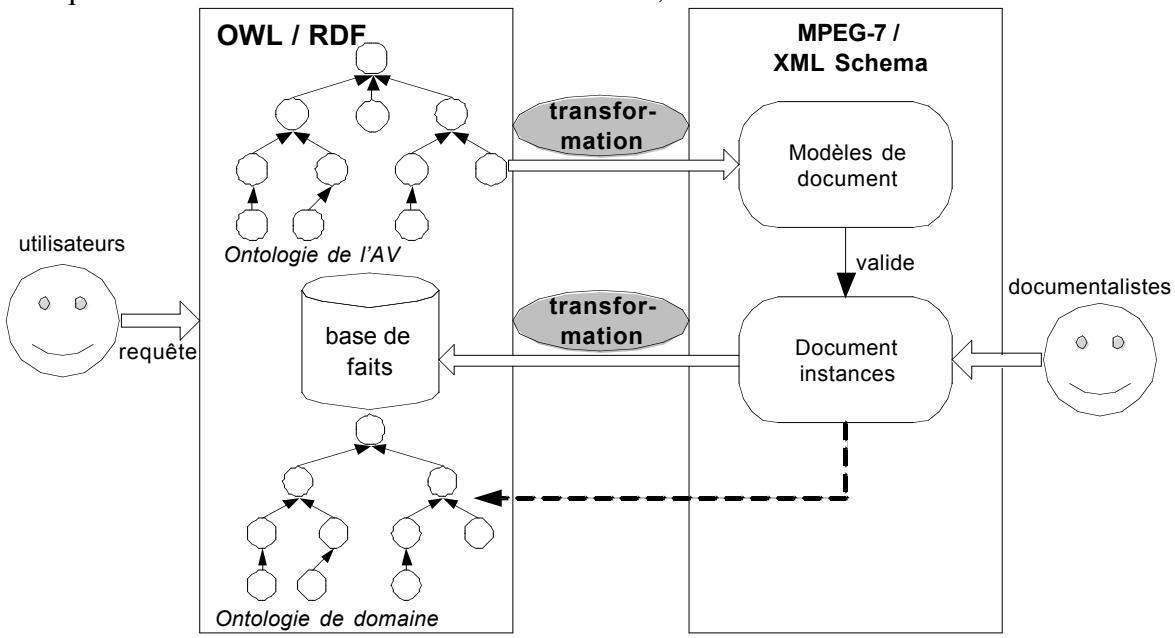

Figure 11 : Architecture pour permettre le raisonnement dans les descriptions documentaires. L'ontologie de l'audiovisuel permet de formaliser les connaissances structurelles des descriptions. Elle est traduite en types XML Schéma pour pouvoir exprimer des modèles de document et est instanciée lors de la description. Enfin, une autre ontologie de domaine, avec les faits qui y sont raccordés, permet d'exprimer la connaissance conceptuelle qui sera liée aux éléments structurels de la description.

Le document télévisuel traverse toute une série d'étapes avant d'être capté et archivé. Ainsi, les contenus audiovisuels sont d'abord produits pour être vendus à des diffuseurs qui en font des programmes. Ceux-ci s'inscrivent alors dans une grille des programmes (résultat d'une politique éditoriale) qui se transforme en un flux d'images et de sons qui parvient aux

téléspectateurs et à l'INA. L'ontologie de l'audiovisuel commence donc par distinguer l'objet audiovisuel selon la place qu'il occupe dans ce cycle de vie :

— l'*objet de production* peut être une séquence ou une émission complète à structure simple ou composite ;
— l'*objet de diffusion* permet d'inclure le programme dans une tranche horaire, et de spécifier le statut (première diffusion, multidiffusé …) et le mode de diffusion (direct, duplex, liaison téléphonique …) ;
— l'*objet d'archivage* est assimilé à la description du programme et peut s'inscrire dans une collection.

Les objets de production se spécialisent ensuite selon leur genre audiovisuel. Ainsi, le *magazine*, le *journal télévisé* ou le *best-of*, et, le *documentaire*, la *fiction* ou l'*émission plateau* spécialisent respectivement les *émissions composites* et les *émissions simples*. La figure 12 donne par exemple une définition formelle du concept émission plateau comme étant un type d'émission dont toutes les séquences sont des séquences plateau.

```
<owl:Class rdf:ID="EmissionPlateau">
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:resource="#EmissionSimple"/>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#sequence"/>
      <owl:allValuesFrom rdf:resource="#SequencePlateau"/>
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>
```

Figure 12 : L'expression OWL s'interprète comme « la classe EmissionPlateau est l'intersection de la classe EmissionSimple et des objets dont la propriété sequence prend ses valeurs dans la classe SequencePlateau ».

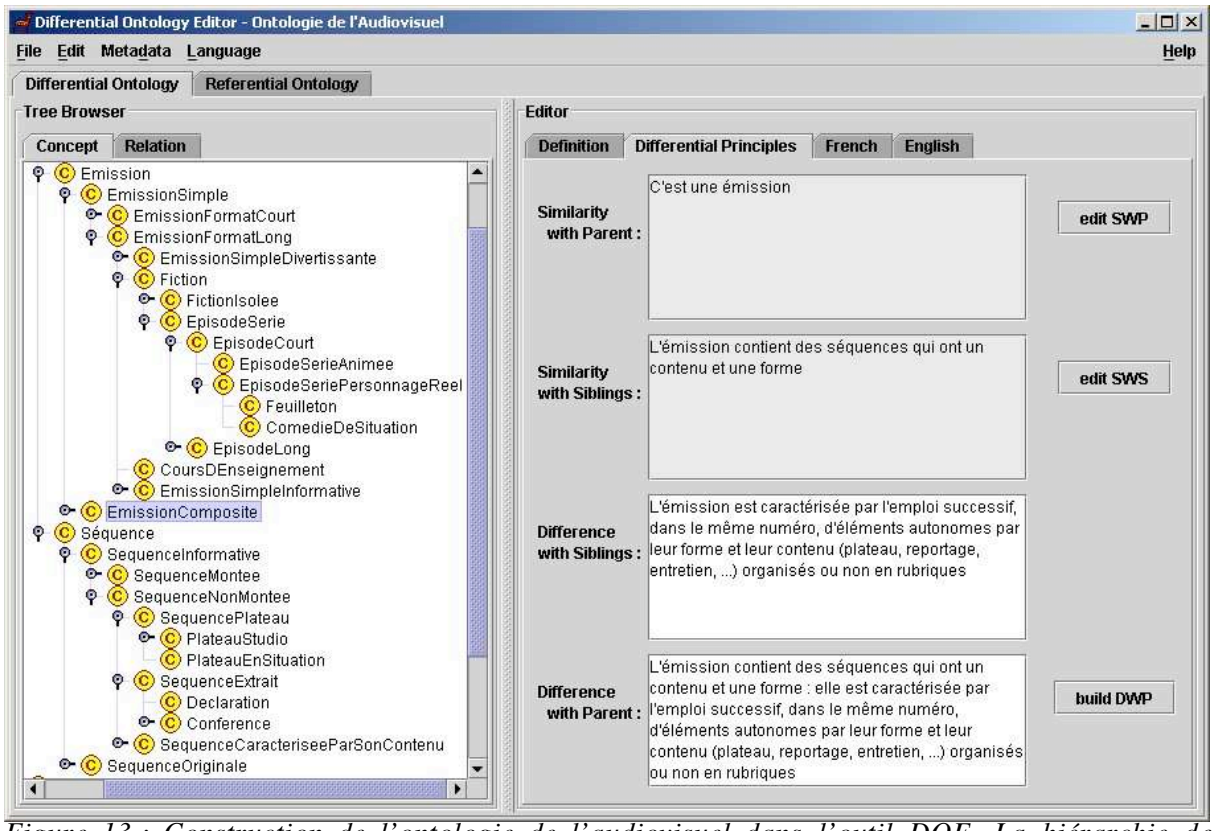

Figure 13 : Construction de l'ontologie de l'audiovisuel dans l'outil DOE. La hiérarchie de concepts y est explicitée et l'ontologie est finalement traduite dans le langage OWL.

L'ontologie de l'audiovisuel (figure 13) permet ainsi de normaliser le sens des termes couramment utilisés pour décrire la structure et la mise en forme des documents audiovisuels. Les concepts sont ensuite formalisés et leur sémantique est accessible dans un système informatique. Mais nous avons vu également l'importance de pouvoir représenter et contrôler la structure logique des documents. Le langage des schémas XML ayant été justement élaboré dans ce but, nous proposons de traduire certains concepts de l'ontologie de l'audiovisuel en types XML Schéma. La combinaison de ces types, *via* les primitives de modélisation du langage XML Schéma, permet de construire des schémas de description qui contrôleront la description de la structure logique des documents. Ainsi, il est possible d'exprimer un schéma général de description pour toute une collection d'émissions. Par exemple, la figure 14 donne la structure des émissions classifiées comme des *magazines sportifs*. Ce schéma de description indique qu'un magazine sportif commence toujours par une séquence *Plateau Début*, suivi par un certain nombre de séquences qui sont soit une séquence *Séquence Plateau*, soit un enchaînement *Plateau Lancement*, *Reportage*, et se termine par une séquence *Plateau Fin*. De plus, la hiérarchie des types conserve la modélisation ontologique du domaine de l'audiovisuel et nous indique donc que les types *Plateau Image* et *Plateau Invite* peuvent se substituer au type *Séquence Plateau*. Finalement, les *Reportage* peuvent contenir des *Interview* et des *Séquence Extrait*.

Magazine Sportif ≡ (Plateau Début,
(Séquence Plateau | (Plateau Lancement, Reportage))+,
Plateau Fin)
Séquence Plateau ≡ (Séquence Extrait | Illustration Caricature) *
Reportage ≡ (Interview | Séquence Extrait) *

Figure 14 : Structure des Magazines sportifs.

La description d'un document audiovisuel commence par la localisation d'entités d'intérêts. Il s'agit de repérer dans le temps (et l'espace) des segments dont on va caractériser la forme et décrire le contenu. Des outils sont disponibles (par exemple, SegmenTool) pour découper temporellement les émissions et produire un début de description MPEG-7. On spécialise alors les segments obtenus selon leur genre et on leur adjoint une thématique générale grâce aux types construits avec l'ontologie de l'audiovisuel. La description peut refléter la structure logique de l'émission et elle doit être validée par le schéma correspondant à la collection dont elle fait partie. Chaque séquence est alors caractérisée par un intervalle temporel sur le média et caractérisée en termes de genre audiovisuel et de thématique générale dans la description (figure 16). Comme les descripteurs utilisés ont leur correspondance dans l'ontologie de l'audiovisuel, nous pouvons engendrer des instances des concepts de cette ontologie. La figure 17 donne un exemple d'assertion RDF construite automatiquement à partir de la description MPEG-7 précédente et indiquant qu'il existe une `interview` de Sandy Casar dans un `reportage` de l'émission *Stade2*.

```
<ina:Reportage id="aa23c647c-6517-4aee-8bce-870ae52a01af">
  ...
  <ina:ReportageTemporalDecomposition>
    <ina:Interview id="adb23ab65-f8e7-4b2a-8c98-807197da600a">
      <mp7:MediaTime>
        <mp7:MediaTimePoint>T00:24:19</mp7:MediaTimePoint>
        <mp7:MediaDuration>PT00H00M07S9</mp7:MediaDuration>
      </mp7:MediaTime>
      <ina:Interviewe>Sandy Casar</ina:Interviewe>
      <ina:Thematique value="Cyclisme"/>
    </ina:Interview>
  </ina:ReportageTemporalDecomposition>
</ina:Reportage>
```

Figure 15 : Exemple de description de la structure d'une émission en MPEG-7 étendue.

```
<Interview rdf:ID="interview4">
 <hastStartTime rdf:dataType="&xsd;string">T00:24:19</hasStartTime>
 <hastDuration rdf:dataType="&xsd;string">PT00H00M07S9</hasDuration>
 <hastThematique rdf:resource="#Cyclisme"/>
 <hasParticipant rdf:resource="#Sandy_Casar"/>
 ...
</Interview>
```

Figure 16 : Exemple de triplets RDF/XML construits automatiquement à partir de la description MPEG-7 étendue.

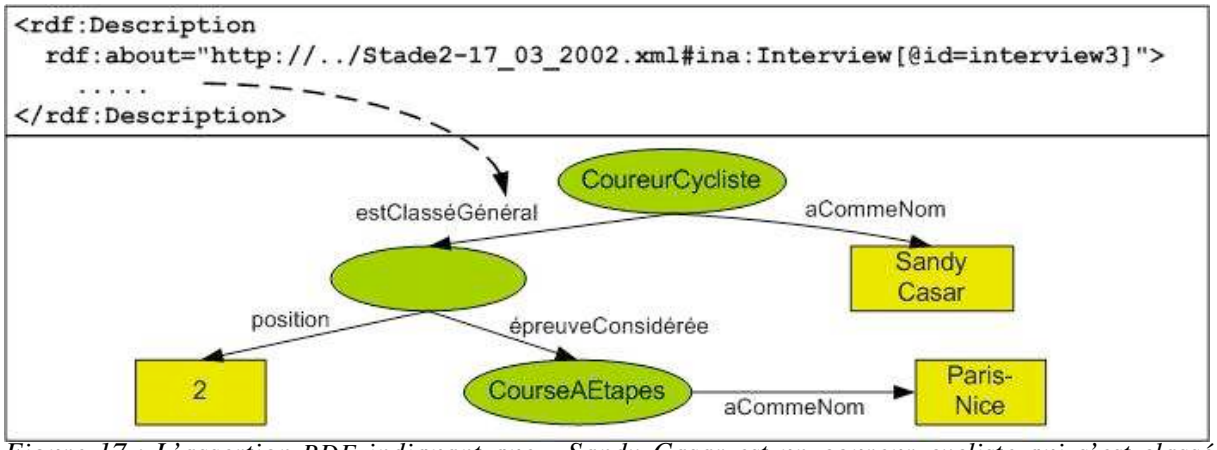

Figure 17 : L'assertion RDF indiquant que « Sandy Casar est un coureur cycliste qui s'est classé 2ème dans la course cycliste à étapes Paris-Nice » est liée à une description de la structure d'une émission TV.

Nous venons de voir comment le découpage temporel d'un programme particulier fournit l'instanciation de la structure de la description et par conséquent les connaissances assertionnelles liées au domaine de l'audiovisuel. Nous pouvons utiliser les mêmes mécanismes pour décrire le contenu proprement dit de chacune des séquences ce qui constitue la phase d'annotation proprement dite des documents. Nous pouvons modéliser une ontologie pour un sport particulier, le *cyclisme*, pour décrire les émissions traitant de ce sport (Tour de France, Magazine Sportif, Journal Télévisé…). Par exemple, la figure 17 indique que *le coureur cycliste Sandy Casar est désormais classé 2ème du classement général de la course Paris-Nice* et que cette assertion est liée – via une relation XPATH – à un segment particulier de la description du document audiovisuel. Ces faits ont également une traduction en RDF et peuvent à leur tour alimenter notre base de connaissances.

L'ensemble des assertions engendrées a donc une traduction immédiate en triplets RDF qui viennent alimenter une base de connaissances sur laquelle on peut effectuer des inférences. Ainsi, il est désormais possible de retrouver « toutes les séquences audiovisuelles décrites comme étant de genre *interview*, dont la durée est supérieure à *30s*, et dont l'interviewé est un *coureur cycliste* figurant sur le podium d'une *course cycliste à étape* ». Pour répondre à cet exemple complexe, la machine pourra inférer, grâce aux connaissances ontologiques, que « être sur un podium » revient à avoir terminé dans les trois premiers de la course au classement général. Elle pourra en outre calculer la durée des séquences audiovisuelles grâce à la donnée des index temporels de début et de fin exprimée dans la description de la structure de chaque émission. La machine pourra donc retourner la séquence décrite dans la figure 17 puisque « Paris-Nice » est bien une course cycliste à étapes et que « Sandy Casar » est bien un coureur cycliste.

6. CONCLUSION

Le web sémantique n'est pas encore un accomplissement. Cependant les technologies permettant de le construire (RDF, OWL) se mettent en place. Nous avons vu que leur utilisation dans le cadre de l'indexation documentaire et la recherche d'information était naturelle et envisageable. On a aussi présenté certaines ressources simples et pratiques (Dublin core,

Creative commons, RSS) aidant la diffusion de l'information ainsi que sa collecte automatique. Enfin, l'utilisation de ces techniques dans le cadre d'indexation de documents audiovisuels a été présentée.

Les maîtres mots de ce web sémantique sont sémantique, distribution, ouverture et partage. La sémantique permet une plus grande précision dans les requêtes et les réponses ; l'ouverture permet l'échange et le partage de ressources toujours plus complètes et toujours plus précises.

Si les fournisseurs de ressources documentaires parviennent à s'accorder pour implémenter ce partage comme d'autres communautés ont su le faire sur le web, la tâche de collecte s'en trouvera simplifiée au bénéfice de tout le monde.

7. REMERCIEMENTS

Une partie de la présentation de OWL est inspirée de la présentation de Jean-François Baget dans [Charlet 2003]. Merci à Chantal Baudin et Isabelle Rey pour leur lecture et leurs commentaires.

8. RÉFÉRENCES

[Beckett 2002] Dave BECKETT, Eric MILLER, Dan BRICKLEY (2002). Expressing simple dublin core in RDF/XML. http://dublincore.org/documents/dcmes-xml/

[Beckett 2004] Dave BECKETT, Ed. (2004). RDF/XML Syntax Specification (Revised). W3C Recommendation. http://www.w3.org/TR/rdf-syntax-grammar

[Berners-Lee 1998] Tim BERNERS-LEE, Roy FIELDING, Larry MASINTER (1998). Uniform Resource Identifiers (URI): Generic Syntax. Request for Comments 2396, IETF. http://www.ietf.org/rfc/rfc2396.txt

[Biezunski 2000] Michel BIEZUNSKI, Martin BRYAN, Steven NEWCOMB, Eds. (1999). ISO/IEC 13250:2000 Topic Maps: Information Technology — Document Description and Markup Languages. http://www.y12.doe.gov/sgml/sc34/document/0129.pdf

[Brickley 1999] Dan BRICKLEY, Ramanathan GUHA, Eds. (1999). Resource description framework schema specification. W 3 C Proposed recommandation. http://www.w3.org/TR/PR-rdf-schema

[Brickley 2004] Dan BRICKLEY, Ramanathan GUHA, Eds. (2004). RDF Vocabulary description language 1.0: RDF Schema. W3C Recommendation. http://www.w3.org/rdf-schema

[Brodie 1984] Michael BRODIE, John MYLOPOULOS, Joachim SCHMIDT Eds. (1984). On Conceptual Modelling: Perspectives from Artificial Intelligence, Databases and Programming Languages. Springer Verlag, Heidelberg (DE).

[Champin 2000] Pierre-Antoine CHAMPIN (2000). RDF tutorial. http://www710.univ-lyon1.fr/~champin/rdf-tutorial/

[Charlet 2003] Jean CHARLET, Philippe LAUBLET, Chantal REYNAUD, Eds. (2003). Web sémantique. Rapport final de l'action spécifique 32, CNRS. http://rtp-doc.enssib.fr/basedoc/rapports/ASWebSemantique2003.pdf

[DCMI 2004] Dublin Core Metadata Initiative (2003). Dublin Core Element set, Version 1.1: reference descriptioon (revised version). DCMI

http://dublincore.org/documents/dces/ (Tr. fr. http://www-rocq.inria.fr/~vercoust/METADATA/DC-fr.1.1.html)

[Dean 2004] Mike DEAN, Guus SCHREIBER Eds. (2004). OWL Web Ontology Language: Reference. W3C Recommendation. http://www.w3.org/TR/owl-ref/

[Euzenat 2003] Jérôme EUZENAT, Amedeo NAPOLI (2003). Numéro spécial XML et objets. *L'objet* 9(3).

[Garshol 2003] Lars Marius GARSHOL (2003). Living with Topic maps and RDF. Ontopia, Trondheim (NO). http://www.ontopia.net/topicmaps/materials/tmrdf.html

[Hayes 2004] Patrick HAYES, Ed. (2004). RDF Semantics. W3C Recommendation. http://www.w3.org/TR/rdf-mt/

[Horrocks 2004] Ian HORROCKS, Peter PATEL-SCHNEIDER, Harold BOLEY, Said TABET, Benjamin GROSOF, Mike DEAN, (2004). SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member submission. http://www.w3.org/Submission/2004/SUBM-SWRL-20040521/

[Klyne 2003] Graham KLYNE, Jeremy CARROLL, Eds. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation, 2003 http://www.w3.org/TR/rdf-concepts/

[Lassila 1999] Ora LASSILA, Ralph SWICK, Eds. (1999). Resource Description Framework (RDF) Model and syntax specification. W3C Recommendation. http://www.w3.org/TR/REC-rdf-syntax

[MPEG7 2001] ISO/IEC (2001). Information Technology - Multimedia Content Description Interface. Norme ISO/IEC n°15938.

[Patel-Scheider 2004] Peter PATEL-SCHNEIDER, Patrick HAYES, Ian HORROCKS, Eds. (2004). OWL Web Ontology Language: Semantics and Abstract Syntax. W3C Recommendation. http://www.w3.org/TR/owl-semantics/

[Staab 2004] Steffen STAAB, Rudi STUDER, Eds. (2004). Handbook of ontologies. Springer Verlag, Berlin (DE).

[Seaborne 2004] Andy SEABORNE (2004). RDQL - A Query Language for RDF. W3C Member submission. http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/

[Thompson 2001] Henry THOMPSON, David BEECH, Murray MALONEY, Noah MENDELSOHN, Eds. (2001). XML Schema part 1: structures, W3C Recommendation. http://www.w3.org/TR/XMLschema-1

[Troncy 2004] Raphaël TRONCY (2004). Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels. Thèse de doctorat de l'Université Joseph Fourier, Grenoble (FR).