http://www.techniques-ingenieur.fr/page/h7155v2niv10002/description-des-ressources-sur-le-web.html
Référence H7155 | Date de publication : 10 nov. 2007

Métadonnées, ontologies et documents numériques

Hélène RICHY, Sylvie DESPRÉS

- **Hélène RICHY :** Maître de conférences, université de Rennes-1.
- **Sylvie DESPRÉS :** Maître de conférences, université Paris-Nord.

1. DU CATALOGAGE AUX MÉTADONNÉES
1.1 Qu'est-ce que le concept de métadonnées ?
1.1.1 Comment trouver une information ?
1.1.2 Que contiennent les métadonnées ?
1.1.3 Qui détermine les métadonnées ?
1.1.4 Sur quoi portent les métadonnées ?
1.1.5 Où placer les métadonnées ?
1.2 Métadonnées traditionnelles
1.2.1 Notices
1.2.2 Formats normalisés
1.3 Documents structurés
1.3.1 Syntaxe
1.3.1.1 SGML
1.3.1.2 XHTML
1.3.1.3 XML
1.3.2 Modèles
1.3.2.1 Métadonnées
1.3.2.2 Textes littéraires
1.3.2.3 Archives
1.3.2.4 Éditeurs
1.3.2.5 Bibliothèques
1.3.2.5.1 MODS
1.3.2.5.2 METS
1.3.2.6 Archives ouvertes

Préambule

Le mot « Web » peut être considéré comme intégré à la langue française, nous l'utiliserons donc de préférence au terme « toile », choisi par les Québécois. Le terme « métadonnées » est utilisé pour désigner toutes les informations, tous les renseignements qui peuvent être associés à une ressource réelle ou virtuelle (œuvre littéraire ou artistique, image, film, page Web...).

1.1 Qu'est-ce que le concept de métadonnées ?

Le concept de métadonnées a évolué avec l'utilisation de l'Internet. Elles étaient initialement limitées aux informations d'archivage, permettant de retrouver un ouvrage ou un document dans une bibliothèque ou un service d'archives. Les métadonnées peuvent aussi répondre à d'autres objectifs : savoir quel usage peut être fait d'un ouvrage ou découvrir ce que contient effectivement une ressource.

Avec l'informatisation de la gestion des établissements publics ou commerciaux, les métadonnées ne sont plus seulement utilisées pour retrouver un ouvrage ou un enregistrement sonore, mais aussi pour stocker d'autres informations, invisibles au lecteur : l'état du stock, le nombre d'emprunts ou le

prix d'achat par exemple. En particulier, lorsqu'il s'agit de documents à diffusion limitée, des métadonnées peuvent être pertinentes pour la sécurité ou la fiabilité des informations.

1.1.1 Comment trouver une information ?

On peut considérer qu'il y a actuellement des millions de pages accessibles par leur adresse sur le Web. Cette adresse joue le rôle d'un numéro de référence dans une gigantesque bibliothèque. Mais, pour connaître les adresses des pages cherchées, il faut généralement interroger les sites qui proposent des moteurs de recherche, tels que Exalead, Google, Lycos ou Yahoo. Ces sites utilisent un robot qui recherche sur tout l'Internet, regardant le contenu des documents à la place du lecteur.

La plupart des moteurs de recherche ignorent les métadonnées. Ils se contentent généralement d'examiner le texte contenu dans le titre, l'adresse, l'en-tête ou le début d'un document et d'y appliquer des méthodes statistiques (fréquence des mots, pondération selon l'emplacement dans le document, fréquence des accès, proximité, etc.) : Lycos, par exemple, ne retient qu'une centaine de caractères pour décrire une page Web. À l'inverse, Yahoo qui n'utilise pas de robot, mais fait appel à des indexeurs humains pour cataloguer les informations du Web en fonction de rubriques très générales, peut être considéré comme un annuaire.

Le manque de précision des documents sélectionnés par les moteurs de recherche, tant pour leur contenu que par la quantité des adresses proposées, incite à envisager d'autres solutions. Alors que les bibliothécaires n'ont pu se décider à utiliser un format commun pendant tant d'années, il ne paraît pas envisageable d'imposer pour l'Internet un ensemble commun de métadonnées. Toutefois, l'usage d'une même technologie peut contribuer à faciliter les échanges, malgré cette absence de standard.

1.1.2 Que contiennent les métadonnées ?

À la différence des informations de catalogage traditionnelles qui sont fournies généralement par des spécialistes à l'intention de lecteurs humains, les métadonnées sur les pages Web sont destinées à être traitées par des machines. Le format MARC, la Dublin Core (§ 1.2.2), l'en-tête de la TEI (§ 1.3.2.2) ou simplement les propriétés d'un fichier MS Office, sont des métadonnées conçues pour être comprises par l'ordinateur. Toutefois, il s'agit essentiellement de métadonnées descriptives du contenu.

Plus généralement, on peut considérer que les métadonnées peuvent fournir toutes sortes d'informations relatives à une ressource ou à son usage :

- des métadonnées descriptives (du contenu, de l'origine de l'information, bibliographique...) ;
- des métadonnées administratives (juridiques, commerciales...) ;
- des métadonnées structurelles (relations entre composants d'une collection fractionnement...).

1.1.3 Qui détermine les métadonnées ?

À la différence des informations de catalogage qui sont fournies par des spécialistes, les métadonnées pour des documents sur l'Internet peuvent provenir de plusieurs sources comme, par exemple :

- de celui qui détient tous les droits sur le document (propriétaire) ;
- de celui qui détient le document (gestionnaire) ;
- de celui qui en fait commerce ;
- de celui qui l'évalue (pédagogue, critique, intermédiaire) ;
- de celui qui l'utilise (lecteur, abonné, client, acheteur).

L'utilisation d'une technologie qui offre la possibilité d'intégration de plusieurs sources revêt donc une importance primordiale dans le cas de sources multiples, personnelles ou publiques. De plus, il est essentiel dans un contexte comme l'Internet que l'origine des informations puisse être certifiée.

1.1.4 Sur quoi portent les métadonnées ?

Dans le domaine documentaire, les ressources électroniques sont rarement des objets isolés : on parle de fonds documentaires, de corpus, de collections ou d'œuvres qui partagent certaines caractéristiques. Dans une page Web, on peut également identifier des objets de plus petite taille, comme par exemple une image, une citation ou une annotation. Quelle que soit la granularité de la ressource, il faut pouvoir la décrire avec des métadonnées.

Le traitement de métadonnées définies à plusieurs niveaux peut donc devenir assez complexe. Le recours à des modèles de représentation de connaissance ou des ontologies peut en faciliter l'interprétation.

1.1.5 Où placer les métadonnées ?

À la différence des informations de catalogage qui sont stockées généralement indépendamment du document lui-même, les métadonnées décrivant des ressources de l'Internet peuvent faire l'objet d'une localisation plus variée :

- encapsulation des métadonnées dans la ressource, comme, par exemple, les métadonnées contenues dans l'en-tête d'un document codé avec la TEI ;
- association de métadonnées externes à la ressource, dans un document séparé (MARC) ;
- métadonnées indépendantes, reliées au document par une URI (HTML, XML) ;
- groupement de métadonnées dans une base de données de catalogage qui donne accès à un ensemble de ressources, sur le modèle des notices bibliographiques de l'archivage traditionnel ;
- encapsulation de la ressource dans les métadonnées la décrivant.

Ces différentes localisations peuvent être envisagées selon la technologie numérique mise en œuvre sur le Web.

1.2 Métadonnées traditionnelles

Que ce soit dans le domaine des archives, des bibliothèques ou des musées, les systèmes d'archivage traditionnels ont évolué avec le développement de l'informatique et de l'Internet. Toutefois, en ce qui concerne la description des contenus, les notices qui figurent dans les catalogues des bibliothèques ou les registres d'archives peuvent être considérés comme des précurseurs. Les normes internationales y jouent un rôle important.

1.2.1 Notices

Le catalogage moderne date du XIX^{e} siècle. Les catalogues sont généralement remplis par des documentalistes. Chaque document entré dans un catalogue fait l'objet d'une « notice » bibliographique complète constituée de deux parties, la description bibliographique et les points d'accès :

- la description bibliographique est créée en suivant les règles de catalogage, elle contient le titre et d'autres informations extraites du document, ainsi que des informations relatives à l'éditeur ou à la description physique de l'ouvrage ;
- les points d'accès sont contrôlés par des thésaurus ou des formes d'autorités pour les noms propres et les noms géographiques, notamment. Des dictionnaires externes, ou « fichiers d'autorités », contiennent les formes autorisées.

Une cote donnant la localisation physique de l'ouvrage, sert de lien entre le document et ces informations. Les notices sont groupées dans des catalogues qui sont consultés pour donner accès à l'ouvrage qui y est répertorié. Ces catalogues peuvent être utilisés par les documentalistes pour répondre à la demande de lecteurs, ou directement par certains lecteurs. Ces catalogues existent indépendamment des documents référencés.

Attributs du Dublin Core

DC	DC français	Contenu
Title	Titre	Titre de la ressource
Creator	Créateur	Nom et coordonnées du créateur
Subject	Sujet	Mots ou phrases clés
Description	Description	Texte de description ou résumé
Publisher	Éditeur	Nom et coordonnées de l'éditeur
Contributor	Collaborateur	Noms et coordonnées des contributeurs
Date	Date	Date de création, date d'édition
Type	Type	Nature du contenu de la ressource
Format	Format	Format multimédia de la ressource
Identifier	Identifiant	URL de la ressource
Source	Source	Origine qui a fourni la ressource
Language	Langue	Langue d'expression de la ressource
Relation	Relation	URL des autres ressources en relation
Coverage	Couverture	Description du temps et lieu du contenu de la ressource
Rights	Droits	Texte de copyright
[1] *[2]* *[3]*		

- [1] -ou son URI
- [2] -valeurs définies avec raffinement
- [3] -texte, image, son, logiciel...

Avec l'informatisation de la gestion des bibliothèques, les catalogues se sont transformés en collection de métadonnées. Puis avec le développement d'Internet et la politique de numérisation des documents, il est devenu possible de consulter à distance, non seulement les notices, mais aussi les documents numérisés. Différents formats ont été développés ou étendus pour représenter ces notices.

1.2.2 Formats normalisés

Les normes de catalogage ont été fixées par des standards internationaux ISBD (International Standard Bibliographic Description). Les notices bibliographiques ont ensuite été informatisées en utilisant le **format normalisé MARC** (Machine-Readable Cataloging) (ISO 2709) [22]. Son objectif initial était de faciliter la description informatique des documents en associant des étiquettes numérotées à toutes les informations de catalogage. Plus d'une vingtaine de versions du format MARC ont été créées (UKMARC, INTERMARC, USMARC...). Afin d'unifier les catalogues européens,

UNIMARC a été adopté par la Communauté européenne (figure 1). Mais cette disparité entre les versions nationales a rendu impossible tout échange direct d'informations bibliographiques.
Alors que les informations de ces notices bibliographiques étaient limitées aux seules informations nécessaires pour le catalogage, des systèmes de métadonnées plus complets, comportant des informations sur l'usage ou le contenu, ont été proposés.
Par exemple, le Dublin Core (DC) créé en 1995 à l'initiative du Dublin Core Metadata Initiative (DCMI) décrit non seulement l'identification et le contenu d'une œuvre, mais aussi des éléments sur la propriété intellectuelle. Chacun des éléments (tableau 1) est optionnel et peut être répété. Des évolutions de cette description initiale continuent à être proposées par le DCMI. Et de nombreux travaux ont permis d'adapter cette norme aux nouvelles technologies : les concepts du DC peuvent être représentés selon différentes syntaxes (§ 1.3.2.1) et intégrés aux documents numériques échangés sur le Web.

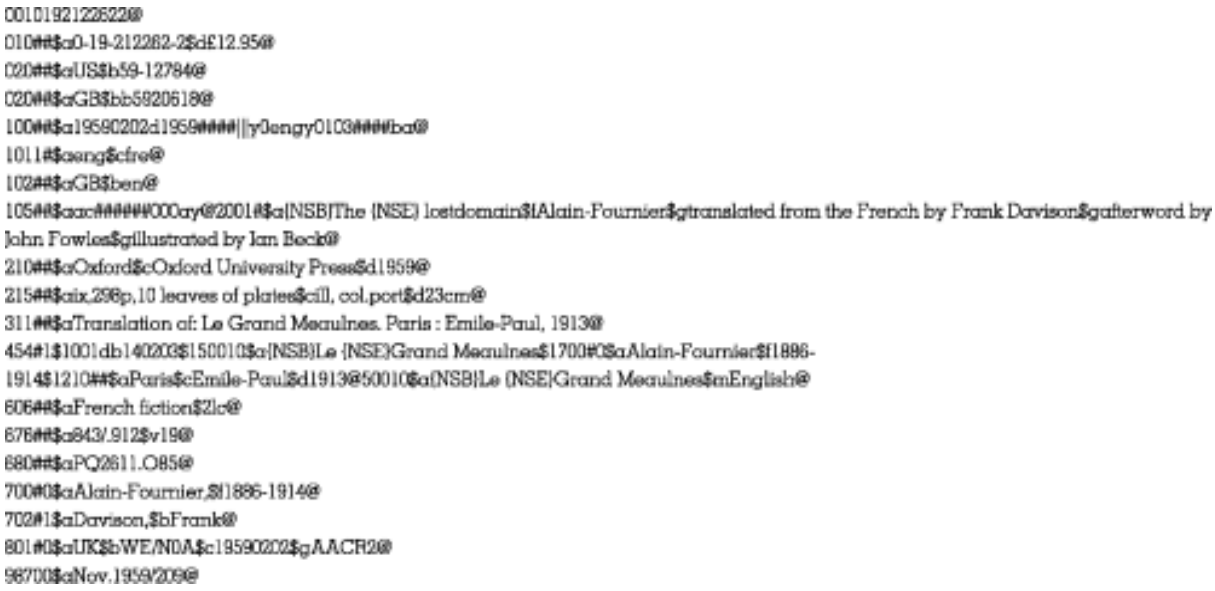
00101921226220
010##$a0-19-212262-2$d£12.95@
020##aUSb59-12784@
020##aGBbb59205180
100##$a19590202d1959####||y0engy0103####ba@
101#$aeng$cfre@
102##aGBben@
105##$aac#####000ay@2001#$a[NSB]The [NSE] lostdomain$1Alain-Fournier$gtranslated from the French by Frank Davison$gafterword by John Fowles$gillustrated by Ian Beck@
210##$aOxford$cOxford University Press$d1959@
215##$aix,296p,10 leaves of plates$cill, col.port$d23cm@
311##$aTranslation of: Le Grand Meaulnes. Paris : Emile-Paul, 1913@
454#1$1001db140203$150010$a[NSB]Le [NSE]Grand Meaulnes$1700#0$aAlain-Fournier$f1886-1914$1210##$aParis$cEmile-Paul$d1913@50010$a[NSB]Le [NSE]Grand Meaulnes$mEnglish@
606##$aFrench fiction$2lc@
676##$a843/.912$v19@
680##$aPQ2611.O85@
700#0$aAlain-Fournier,$f1886-1914@
702#1$aDavison,$bFrank@
801#0aUKbWE/NDA$c19590202$gAACR2@
9870O$aNov.1959/209@

Figure 1 - Exemple UNIMARC

1.3 Documents structurés

Dans le domaine documentaire, on assiste à la modification du support et de la diffusion des documents. Cette évolution touche aussi bien les documents administratifs, techniques que les documents commerciaux, les œuvres littéraires ou scientifiques, dans le domaine professionnel comme dans l'environnement personnel. L'informatisation des catalogues, puis la généralisation de la notion de document numérique et surtout l'accès à distance aux documents contribuent à leur dématérialisation. Mais la première évolution majeure apportée par l'informatisation des documents concerne la possibilité de structurer les documents selon leur contenu.

1.3.1 Syntaxe

Dans le monde documentaire, la structuration des ressources numériques a commencé avec SGML, bien avant l'apparition de XML. Elle s'est renforcée d'autant plus depuis que XML est apparu comme un standard pour structurer des documents.

1.3.1.1 SGML

Le standard SGML (Standard Generalized Markup Language) est publié depuis 1986 (ISO 8879) (voir l'article *SGML* [H 7 138]). Il permet à la fois de structurer un document et de le décrire à l'aide de métadonnées. La structuration des documents selon leur contenu facilite non seulement le

processus éditorial, mais aussi la gestion des connaissances. Un processus automatique peut ainsi assurer le formatage du document pour divers supports ou extraire des informations facilitant la gestion ou l'accès au document.
Le système d'encodage d'un document SGML est défini par une description, la DTD (Document Type Definition). Différentes descriptions ont été spécifiées (TEI, EAD, HTML, etc) et certaines ont particulièrement détaillé les informations descriptives du contenu.

1.3.1.2 XHTML
Le format HTML (HyperText Markup Language) est bien connu comme étant le format d'échange de pages sur le Web : une page Web contient du texte structuré par des balises HTML. Le langage HTML est défini par un modèle de document SGML. Suite à l'évolution des standards et pour favoriser l'émergence de la technologie XML, il est conseillé d'utiliser plutôt XHTML [18] pour structurer les pages Web. Les balises sont analogues à celles du HTML, mais la construction des documents XHTML respecte la syntaxe XML : balises équilibrées et bien imbriquées. Comme le HTML, le modèle XHTML prévoit que des éléments liens (*links*) et des éléments *meta* puissent être placés dans l'en-tête (*head*) des documents. Ces deux éléments peuvent être utilisés pour introduire des métadonnées, soit directement dans le document, soit liées au document. Généralement les éléments *meta* n'apparaissent pas durant l'affichage ou l'impression d'un document. Ils peuvent être utilisés pour décrire le contenu du document à l'intention d'applications d'analyse. Ils sont également utilisés par les éditeurs de page Web pour indiquer le codage des caractères. Toutefois, peu de moteurs de recherche utilisent ces informations.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
      "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
 <meta http-equiv="content-type" content="application/xhtml+xml; charset=UTF-8" />
 <title>W3C HTML Home Page</title>
 <meta name="keywords" content="HTML, HTML 4, HTML 4.01, HTML 4.0, XHTML, XHTML 1.0,
XHTML1.1, XHTML Basic, Modularization of XHTML, XML Events, XHTML-Print, XHTML 2.0, HTML Activity, HTML Working Group"/>
 <meta name="description" content="This is W3C's home page for the HTML Activity.  Here you will find pointers to our specifications
for HTML/XHTML, guidelines on how to use HTML/XHTML to the best effect, and pointers to related work at W3C."/>
 <link rel="stylesheet" type="text/css" href="markup.css" />
...
</head>
<body>
...
</body>
</html>
```

Figure 2 - Métadonnées contenues dans une ressource XHTML

L'exemple de la figure 2 reproduit un document XHTML. Les balises meta introduisent simplement une description et des mots-clés. Chaque balise meta porte un attribut name et un attribut content. Les valeurs données à l'attribut name de l'élément meta permettent de décrire l'origine du document, sa diffusion ou comment il a été produit, etc. Ces valeurs sont quelconques. Elles peuvent éventuellement faire référence à la Dublin Core.

1.3.1.3 XML
XML (eXtensible Markup Language) est un sous-ensemble de SGML, qui permet de définir toutes sortes de documents structurés (voir les articles *XML* [H 7 148], *XML : Syntaxe* [H 3 500] et *XML : gestion de contenus Web* [H 3 502]) [19]. Plus simple à utiliser que le SGML, XML s'est rapidement substitué à SGML dans le domaine de la documentation structurée. XML offre une syntaxe souple pour structurer des documents, en fournissant un ensemble de règles de création de

vocabulaires (noms de balises et d'attributs). Les schémas XML peuvent ensuite être utilisés pour autoriser la composition de vocabulaires XML.

De plus, XML peut être considéré comme la technologie de base qui permet de partager et de structurer à la fois les documents et les données sur le Web. Le codage des documents XML en Unicode conforte cette destination internationale. Associé aux autres recommandations du World Wide Web Consortium (W3C) concernant l'exploitation des documents, XML offre un environnement applicatif très riche.

La technologie XML présente de nombreux avantages, rappelés ici :

- les documents XML peuvent être décrits par des modèles de documents, appelé « schémas ». Un schéma décrit non seulement la terminologie (les noms des balises), mais aussi des contraintes d'utilisation (structure, type de contenu) ;
- le mécanisme des « espaces de noms » permet d'apporter des extensions à un schéma en déclarant un nouveau vocabulaire : de nouveaux noms d'éléments ou d'attributs ;
- le langage de transformation XSLT (eXtensible Stylesheet Language Transformations) permet de convertir efficacement des documents XML en documents XHTML, de les restructurer, de construire automatiquement des index, d'en extraire des informations, etc. (voir *XSLT. Principe et applications* [H 7 160]). La conversion peut être effectuée dynamiquement, lors de la consultation des documents, soit sur le site du serveur, soit sur le site du client. Mais peu de navigateurs offrent actuellement cette dernière possibilité ;
- la désignation d'un fragment de document est possible en utilisant XPointer (XML Pointer Language) [15]. C'est le langage de base de description d'une identification de ressource. XPointer utilise le langage XPath pour sélectionner précisément des éléments dans la structure d'un document XML ;
- le langage XPath (XML Path Language) est utilisé par XPointer et par XSLT : en s'appuyant sur la structure logique du document, sur le type des éléments, sur les valeurs des attributs, sur les caractères contenus ou sur les positions relatives, XPath permet de se déplacer dans la structure d'un document XML. Un puissant mécanisme de comparaison permet de retrouver des éléments selon leur motif, c'est-à-dire non seulement en fonction de leur contenu, mais aussi de leur structure ;
- le langage de requête XQuery [14] permet d'extraire des informations de documents XML. La désignation dans le document XML utilise la syntaxe XPath. Le résultat de la requête est structuré en XML.

En utilisant les technologies XML mentionnées ci-avant, il est donc possible d'extraire automatiquement des métadonnées d'un document XML. Il est également aisé de transformer les documents XML ou des métadonnées dans d'autres formats.

En revanche, XML n'impose aucune contrainte sémantique sur la signification des documents. C'est en ajoutant progressivement des descriptions aux données et aux documents déjà existants sur le Web, que XML [19], RDF [9] et OWL [10] permettront au Web d'être une infrastructure globale pour partager des informations et les rechercher de manière plus efficace : ce que le W3C désigne sous le vocable de Semantic Web [16].

1.3.2 Modèles

De nombreuses descriptions de documents spécifiées en XML prévoient des métadonnées figurant dans l'en-tête du document structuré ou décrivant des images (XML, IPTC, MIX). Chaque domaine d'application propose son modèle : la DocBook, pour les documents techniques, la TEI pour les œuvres littéraires, l'EAD pour les archives, etc. Mais la plupart de ces modèles intègrent les concepts du DC et éventuellement ajoutent des métadonnées complémentaires.

1.3.2.1 Métadonnées

Les concepts du Dublin Core (§ 1.2.1) ont été étendus afin de compléter certains éléments descriptifs : des qualificatifs permettent de préciser le sens d'un élément (raffinement) ; d'autres qualificatifs identifient des schémas d'encodage (définition de vocabulaire, notations, règles

d'interprétation). Des extensions pourraient encore être envisagées afin d'apporter plus d'informations concernant l'usage et la gestion des documents, notamment.

Les concepts du DC peuvent être représentés selon différentes syntaxes : en HTML, en XML, dans l'en-tête de la TEI ou en RDF/XML.

De nombreux logiciels sont disponibles (DC *tools*) pour faciliter la création des métadonnés du DC : saisie dans des formulaires pour HTML, extraction, conversion de formats (HTML, TEI, USMARC, MARC 21, UNIMARC), production d'un fichier RDF en XML, etc.

1.3.2.2 Textes littéraires

Les travaux de la TEI (Text Encoding Initiative)(voir note bas de document), commencés dans les années 1980, ont abouti à la définition d'un format standard destiné à faciliter traitement, échange et partage de textes numérisés, principalement des textes littéraires de structure complexe. Ses principes initiaux de modularité, flexibilité et extensibilité, permettent de l'adapter aussi bien au SGML qu'au XML. Un consortium TEI, créé en 1999, assure le développement et la promotion de ce standard.

Tout document conforme à la TEI comprend deux parties (voir l'article *TEI* (*Text Encoding Initiative*) [H 7 158] : un en-tête (*TeiHeader*) de nature documentaire et le contenu textuel du document. Les informations de l'en-tête sont destinées à faciliter la gestion du document, son indexation et son catalogage. L'en-tête (figure 3) est composé de quatre parties qui décrivent :

- le document électronique : son créateur, la source du document s'il y en a une, etc. (cette partie est obligatoire) ;
- les principes éditoriaux mis en œuvre pour coder le document électronique ;
- le contenu du document : sa langue, le sujet dont il traite, etc. ;
- l'historique des révisions.

Les travaux de la TEI ont inspiré des évolutions notables sur les liens en XML (XLink et XPointer) et, en s'inspirant de la structure de l'en-tête, sur la description des métadonnées.

1.3.2.3 Archives

La démarche des services d'archives américains (Society of American Archivists) est caractéristique de l'évolution du domaine des documents structurés ces dernières années. S'appuyant tout d'abord sur une technologie SGML, les développements se sont rapidement portés sur XML. Le modèle de description de documents d'archives EAD (Encoded Archival Description) est utilisé internationalement dans de nombreuses bibliothèques d'archives.

```
<publicationStmt>
  <publisher>Éditions Gallimard</publisher>
  <pubPlace>Paris</pubPlace> <date>1993</date>
  <idno type=ISBN> 2-07-011336-1</idno>
  <idno type=numero edition>64107</idno>
  <idno type=numero impression>I3-1903</idno>
  <idno type=depot legal>octobre 1993</idno>
  <availability>Copyright: Éditions Gallimard,
   Féerie pour une autre fois I, 1952;
   Féerie pour une autre fois II, 1954;
    Entretiens avec le professeur Y, 1955;
     Appendices, text, préface et apparat critique, 1993.
     </availability>
</publicationStmt>
```

Figure 3 - Extrait d'un en-tête de fichier encodé avec la TEI (mention de la publication de romans de Céline, dans la collection La Pléiade)

L'EAD reprend la notion d'en-tête (*header*) proposée par la TEI. Un document est constitué de deux parties : le premier segment, l'en-tête, contient les informations permettant d'aider à la recherche d'un document (son titre, sa date de création, etc.), le second segment contient des informations sur le corps du document lui-même (collection, groupe de documents, etc.). Du fait de son antériorité, ce modèle XML est très utilisé dans les milieux archivistiques de nombreux pays.

1.3.2.4 Éditeurs

Entre éditeurs, le format d'échange en vigueur sur l'Internet est ONIX (Online Information Exchange) [27] : la version 2, disponible depuis 2004, fait encore régulièrement l'objet de révisions mineures. Conçu en XML et en Unicode, ce format est utilisé par les principaux éditeurs diffuseurs (Amazon.com, John Wiley & Sons...) pour échanger des informations sur leurs publications. Les bibliothèques ont donc également intérêt à pouvoir communiquer dans ce format.

1.3.2.5 Bibliothèques

Afin de faciliter les échanges entre bibliothèques, le format normalisé MARC 21 a évolué en un format Marc21 Unicode, plus international. Ensuite, à l'initiative de la Bibliothèque du Congrès américain notamment, des conversions en XML ont été proposées. Quelques-uns de ces modèles sont décrits ci-après.

1.3.2.5.1 MODS

En 2003, la Bibliothèque du Congrès américain a proposé MODS (Metadata Object Description Schema) [26]. Cette description, proche du DC et d'ONIX, repose sur XML et le codage Unicode, et traduit en éléments les métadonnées du format MARC 21. Elle est utilisée, par exemple, pour décrire plusieurs millions de ressources (photos, vidéo, cartes, etc.) du projet American Memory.

1.3.2.5.2 METS

À l'initiative de la Fédération des bibliothèques numériques, le standard METS (Metadata Encoding and Transmission Standard) [24] a été diffusé pour permettre la description de collections, leur gestion et leur préservation. Ce standard préconise une description XML en six modules (en-tête, métadonnées descriptives, métadonnées administratives, fichiers, structure, comportement) qui intègrent les éléments du Dublin Core, MODS et MARC XML (figure 4). Il est maintenu par la Bibliothèque du Congrès et surtout utilisé pour des collections d'images, de vidéo ou de documents multimédias.

2. DESCRIPTION DES RESSOURCES SUR LE WEB
2.1 Identification d'une ressource
2.1.1 Localisation
2.1.2 Adresse persistante
2.1.3 Identifiant
2.1.4 Désignation d'objets numériques
2.2 RDF
2.2.1 Principe
2.2.2 Identification des ressources en RDF
2.2.3 Modèle
2.2.3.1 Schémas RDF
2.2.3.2 Espace des noms
2.3 Applications
2.3.1 Annotations
2.3.2 Contrôle et sécurité
2.3.3 Syndication de sites
2.3.4 Outils de recherche

Sans entrer dans une description détaillée du Web, il peut être utile, pour la compréhension de ce qui suit, de rappeler le fonctionnement du Web. Le Web peut être vu comme un réseau permettant des échanges d'information basés sur :

- un mécanisme d'identification des ressources (les URI) : une « ressource » désigne tout ou partie d'une page Web, qu'il s'agisse de texte, d'image ou de son ;
- des formats de représentation des ressources (XHTML, XML) et de codage (Unicode) : l'universalité du Web, ouvert à toutes les langues, toutes les cultures, impose de supporter différents codages de caractères, ce qui conduit à adopter le codage Unicode ;
- un mécanisme d'échange des ressources : plusieurs protocoles de communication sont utilisables, selon qu'il s'agit d'échanger des documents HTML (HTTP : HypertText Transfer Protocol), des fichiers (FTP : File Transfer Protocol) ou des messages (SMTP : Simple Mail Transfer Protocol).

Le Web offre des facilités d'échange de documents multimédias, intégrant le texte, l'image ou le son. Cette situation conduit à développer des systèmes de conversion numérique de l'information pour diffuser toutes sortes de documents numériques par ce canal. De nombreuses campagnes de numérisation du patrimoine sont en cours, dans le monde entier. Mais la situation n'est pas tout à fait aussi idyllique qu'il y paraît : une harmonisation entre les différents systèmes d'identification des ressources et de métadonnées reste une priorité pour répondre aux besoins de cette vaste diffusion des connaissances. Afin de pouvoir partager les informations, les ressources doivent pouvoir être clairement identifiées. Le W3C préconise donc pour le Web sémantique d'utiliser un mode de désignation qui garantisse l'unicité de la ressource et sa qualité, et qui permette d'agréger des métadonnés s'y référant.

Avant d'envisager les nouvelles applications favorisées par cette initiative, nous allons rappeler comment fonctionne la désignation des ressources et analyser quel est l'apport de RDF pour la description de ces ressources.

2.1 Identification d'une ressource

2.1.1 Localisation

La désignation actuelle des pages repose essentiellement sur leur localisation et ne permet pas d'identifier une ressource avec certitude. Comme il arrive souvent que des pages migrent, l'adressage d'une page par son URL (Universal Resource Locator) qui décrit le chemin d'accès à une

ressource, dans l'Internet pose souvent des problèmes : l'URL est constitué d'un nom de site à la suite duquel sont juxtaposés des noms de répertoires, puis le nom du fichier qui la contient. En cas de déplacement de la page, l'URL change. Pour pallier cette instabilité des URL, plusieurs propositions ont été élaborées, mais elles nécessitent la mise en œuvre de registres centraux ou de serveurs d'adresses, peu compatibles avec le fonctionnement décentralisé de l'Internet.

2.1.2 Adresse persistante

L'OCLC (Online Computer Library Center) a proposé un système d'adresse plus persistantes. Une adresse PURL (Persistent Uniform Resource Locator) ne désigne pas directement une ressource, mais fait appel à un service intermédiaire de résolution d'adresse afin d'obtenir l'URL du destinataire. Par exemple, l'adresse PURL suivante http://purl.oclc.org/OCLC/PURL/FAQ commence par l'invocation du protocole (http), contient l'adresse du système de résolution (purl.oclc.org) et se termine par le nom de la ressource recherchée.

Cette méthode améliore la stabilité des adresses, mais n'offre évidemment pas une totale garantie de succès. Cette solution a été déployée par l'OCLC à partir de 1996 pour répondre aux demandes de catalogage des projets américains pour l'éducation. C'est une étape vers la normalisation des adresses URN décrites ci-après.

2.1.3 Identifiant

L'IETF (Internet Engineering Task Force), organisme international qui développe des standards pour l'Internet, et le W3C ont proposé un système d'URI (Uniform Resource Identifier). Cet URI (RFC 3986) doit permettre d'identifier une ressource de manière unique par une simple chaîne de caractères, indépendamment de la localisation de la ressource. Plus généralement, l'URI permet de désigner des ressources abstraites quelconques (sur l'Internet ou non). L'URL peut être considéré comme le type le plus courant d'URI sur l'Internet. En complément des emplacements décrits par les URL, le système des URI est basé sur des noms (URN) et des caractéristiques (URC) :

- l'URN (Uniform Resource Name), comme son nom l'indique, est un nom unique et permanent : il désigne le nom d'une ressource qui pourra être présente sur plusieurs sites (exemplaires multiples, sites miroirs). Un serveur de résolution de noms permet de connaître la ou les adresses correspondant à ce nom ;
- les URC (Uniform Resource Characteristic) regroupent les métadonnées sur les ressources et en particulier, les conditions d'accès à ces ressources. Elles ne sont pas nécessairement gérées sur le même site que les URN.

2.1.4 Désignation d'objets numériques

Une association d'éditeurs américains (Association of American Publishers) a développé, en collaboration avec le CNRI (Corporation for National Research Initiatives), un système d'identification pour les documents numériques. Il s'agit d'un système centralisé, qui concerne non seulement des livres, des chapitres ou des images, mais aussi des enregistrements sonores ou vidéo, ou toute œuvre de création. Une fondation (International DOI Foundation), créée en 1998, est responsable du répertoire central.

Un DOI comprend un préfixe et un suffixe, séparés par une barre oblique. Un préfixe est assigné à chaque éditeur (il commence par 10.) ou à toute autre entité qui le lui demande. L'éditeur ou l'entité en question assigne un suffixe pour compléter cette identification. Chaque éditeur est libre de choisir son système de création de suffixe. Bien sûr, les éditeurs qui utilisent un standard international (ISBD ou ISBN, par exemple) sont invités à l'utiliser comme suffixe. Le DOI est principalement utilisé dans le domaine de l'édition et dans le secteur gouvernemental. À la différence d'une URL, le DOI est donc permanent, unique et peut désigner plusieurs ressources.

Dans une référence bibliographique, le DOI peut remplacer l'URL, à la fin de la référence. En ce cas, il est généralement suivi de la date de consultation entre parenthèses. Par exemple, si l'adresse

complète est sous la forme http://www.doi.org/10.xxxx/xxxx, on pourra remplacer dans la bibliographie, l'URL en fin de citation par « DOI : 10.xxxx/xxxx (consultation le 12 mars 2007) ».

Le système Handle du CNRI utilise le DOI et dispose de pointeurs (*handles*) qui désignent les ressources à la manière des URN. Élaboré pour les bibliothèques numériques, ce système est utilisé par les archives ouvertes. Toutefois, pour pouvoir être utilisé par les navigateurs standards, il nécessite l'ajout d'un module d'extension.

2.2 RDF

XML, langage de balisage de données, n'offre pas le cadre nécessaire à une description des métadonnées, à savoir la description d'une terminologie. Avec XML, chacun peut créer ses propres balises pour structurer un texte ou une page. Des programmes ou des scripts peuvent utiliser ces balises pour traiter les données contenues dans ces documents. Mais les balises ne portent aucune sémantique.

En revanche, la signification des balises peut être exprimée en RDF (Resource Description Framework) : c'est un cadre général pour la description de ressources sur le Web proposé dès 1997 par le W3C [12]. Le modèle initial est générique et n'impose ni syntaxe, ni vocabulaire. Il permet de définir différentes façons de gérer des métadonnées.

2.2.1 Principe

Le principe de RDF est de définir des ressources par des propriétés. Plus précisément, les données RDF sont constituées d'un ensemble de triplets (sujet, verbe, complément d'objet) qui permettent d'exprimer que :

- certaines choses (une personne, une page ou autre chose) ;
- ont des propriétés, par exemple « est l'auteur de » ou « est le titre de » ;
- qui portent certaines valeurs (une autre personne, une page ou autre chose).

Par exemple, les déclarations suivantes « Victor Hugo est l'auteur des Misérables » ou bien « Cet article a été créé par Open Writer » ou encore « Open Writer est un logiciel de traitement de texte » s'expriment naturellement par cette sorte de triplet : une assertion RDF. Cette description est une manière naturelle de décrire la plupart des données traitées par les machines. Plus généralement, en RDF, un ensemble de ressources peuvent être associées entre elles par des types de propriétés et représentées graphiquement, en utilisant des ovales pour les ressources, des flèches pour les propriétés et des rectangles pour les valeurs.

Il est important de remarquer que RDF ne définit pas de types de propriété : n'importe qui peut définir et utiliser ses propres types de propriétés, ce que l'on appelle son vocabulaire. Il est clair qu'un vocabulaire utilisé pour décrire des livres ou des archives sera peu adapté pour décrire des films ou des services bancaires. Toutefois, on peut supposer que, pour ce qui est des institutions officielles (bibliothèques, archives, musées), il ne s'agira que d'un simple problème de traduction (différence de langue) et non d'une différence de concept : dans ce cas, une simple table de correspondance entre deux vocabulaires assure la traduction automatique d'un fichier XML d'un vocabulaire dans un autre.

```
<rdf:RDF
 xmlns="http://www.w3.org/TR/1999/PR-rdf-schema-19990303#"
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" >
<rdf:Property rdf:ID="title">
            <label xml:lang="en">Title</label>
            <label xml:lang="fr">Titre</label>
</rdf:Property>
<rdf:Property rdf:ID="creator">
            <label xml:lang="en">Author/Creator</label>
             <label xml:lang="fr">Auteur/Créateur</label>
</rdf:Property>
<!-- etc. -->
</rdf:RDF>
```

Figure 5 - Description partielle d'un schéma RDF du Dublin Core

2.2.2 Identification des ressources en RDF

Les choses ou « ressources » désignées comme sujet ou comme objet dans une assertion, sont identifiées par un identificateur de ressource universel (URI) qui peut être utilisé comme un lien vers la ressource elle-même. De même, les propriétés ou les verbes des assertions sont identifiées par des URI. Ainsi, il est possible de définir une nouvelle propriété, un nouveau verbe simplement en définissant une URI pour ce concept quelque part sur le Web.

2.2.3 Modèle

L'utilisation d'une syntaxe XML pour RDF (notée RDF/XML) [11] présente l'avantage de permettre l'échange d'informations qui seront analysables par programme et lisibles par l'Homme.

2.2.3.1 Schémas RDF

Comme cela a été présenté dans la section sur les documents structurés, la notion de schéma XML facilite le traitement des documents, qu'il s'agisse de formatage spécialisé, de vérification de validité ou de gestion de connaissances. Notamment, la garantie de conformité d'un document à un modèle assure que des programmes applicatifs pourront traiter le document. Il en va de même avec le traitement des métadonnées : le schéma RDF définit un certain nombre de propriétés et de classes d'objets pouvant être utilisés dans les descriptions.

Par exemple, l'extrait de schéma RDF de la figure 5 définit que l'élément title pourra en français être introduit par la balise Titre, que l'élément creator sera introduit par la balise Auteur/Créateur.

Pour éviter que de nombreuses normes de métadonnées ne prolifèrent, il est important de faciliter la publication des schémas. C'est le cas du schéma du Dublin Core. Il est public et peut être utilisé sur l'Internet pour décrire de simples informations bibliographiques de n'importe quel document XML (ou XHTML). De même, pour décrire des ressources audiovisuelles, MPEG-7 (ISO/CEI 15938-6) est écrit en RDF. Le partage des schémas RDF est une contribution à l'harmonisation de l'usage des métadonnées.

2.2.3.2 Espace des noms

Recommandation du W3C depuis janvier 1999 [20], les « espaces de noms » peuvent être introduits dans un document XML pour regrouper dans un même modèle des définitions appartenant à une même sémantique. L'utilisation d'espaces de noms dans un document XML facilite la modularité des définitions et le groupement de plusieurs interprétations.

En simplifiant, on peut considérer qu'un espace de noms définit un préfixe et est identifié par une URI. Il est constitué d'un ensemble de noms qui sont des noms d'éléments ou des noms d'attributs.

Ces noms pourront être utilisés comme des balises dans le document XML, précédés du préfixe défini pour cet espace. Ce mécanisme simple permet d'ajouter un vocabulaire spécifique à un domaine, de lever des ambiguïtés entre vocabulaires ou d'adapter un vocabulaire aux usages locaux.

Par exemple, sur la figure 6, l'espace de noms de préfixe rdf fait référence à la syntaxe XML définie pour le RDF. L'espace de noms de préfixe dc fait référence aux quinze éléments définis par le Dublin Core. La balise dc:source désigne un élément source tel qu'il a été défini dans l'espace des noms de préfixe *dc*.

2.3 Applications

2.3.1 Annotations

Les métadonnées constituent les annotations des documents. Ces annotations facilitent l'indexation du document auquel elles se rapportent. Dans l'exemple de la figure 7, les métadonnées font référence au DC, mais elles ne sont pas directement dans le document XML. La même syntaxe aurait pu être utilisée pour placer ces métadonnées directement dans le document XML, comme dans l'exemple de la figure 6.

```
<?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://www.techniques-ingenieur.fr/2007/metadata/">
  <dc:creator>Richy</dc:creator>
  <dc:creator>Despres</dc:creator>
  <dc:publisher>
     <rdf:Description>
       <dc:title>Techniques Ingénieur </dc:title>
       <dc:source rdf:resource="http://www.techniques-ingenieur.fr/"/>
     </rdf:Description>
  </dc:publisher>
 </rdf:Description>
</rdf:RDF>
< !-- le contenu du document -->
```

Figure 6 - Document XML faisant référence à deux espaces de noms

```
<?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 about="http://dublincore.org/tools/">
<dc:title>Dublin Core Metadata Initiative Tools and Software</dc:title>
<dc:description>A listing of tools and software to assist in the creation and maintainence of Dublin
Core Metadata Software.</dc:description>
<dc:date>2006-12-01</dc:date>
<dc:format>text/html</dc:format>
<dc:language>en</dc:language>
<dc:publisher>Dublin Core Metadata Initiative</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

Figure 7 - Fichier XML/RDF associé à la page dublincore.org

Ci-dessus : http://dublincore.org/tools

De nombreuses applications utilisent des annotations de documents, aussi bien pour annoter des textes, des documents anciens (numérisés), des images, des films, etc. De nouveaux schémas d'encodage sont proposés pour les contenus audiovisuels (MPEG-7) (ISO/CEI 15938-6) et multimédias (MPEG-21), qui intègrent aux signaux audiovisuels, des métadonnées qui permettent d'indexer et de contrôler l'usage de ces documents.

Autre exemple, l'annotation des archives d'état civil entreprise par certains services d'archives départementales, en France : ces applications reposent sur d'abondantes métadonnées qui contiennent des indications précises favorisant l'indexation et la consultation sélective des images des registres d'état civil.

2.3.2 Contrôle et sécurité

De nombreuses applications exploitent les métadonnées décrites dans différents modèles : par exemple, PICS pour contrôler l'accès aux documents, P3P pour protéger les documents ou *digital signature* pour garantir l'origine des transactions commerciales. Mais l'usage des technologies XML se substitue progressivement aux pratiques actuelles et favorise les passerelles entre les différents systèmes existants.

En ce qui concerne la recherche documentaire, les outils sont encore en pleine évolution. Comme cela a été évoqué dans la section sur les archives ouvertes, la description bibliographique du Dublin Core qui est maintenant intégrée à la plupart des modèles, et l'utilisation du protocole OAI-PMH, favorisent l'interopérabilité des applications.

2.3.3 Syndication de sites

XML est aussi utilisé comme format de syndication de contenu Web : un flux RSS (Really Simple Syndication) permet de diffuser en temps réel des informations sur les mises à jour de certains sites. Ce système est souvent utilisé pour suivre l'actualité, à l'aide de lecteurs de flux spécialisés ou de logiciels intégrés à certains navigateurs. Le fichier RSS contient une description des métadonnées du flux, comme le montre l'exemple de la figure 8.

```
<?xml version="1.0" encoding="iso-8859-1"?>
 <rss version="2.0">
  <channel>
     <title>Mes nouveautés</title>
     <link>http://www.chezmoi.fr</link>
     <description>Un site à voir</description>
     <language>fr</language>
     <item>
       <title>Bla bla bla...</title>
       <description>        en détail ..... </description>
       <guid>http://www.chezmoi.fr/accueil</guid>
     </item>
      <item>
       <title>Encore d'autres bla bla bla...</title>
       <description>        d'autres détails ..... </description>
       <guid>http://www.chezmoi.fr/bienvenue</guid>
     </item>
  </channel>
</rss>
```

Figure 8 - Structure XML d'un fichier RSS

2.3.4 Outils de recherche

La situation actuelle est encore décevante. Comme nous l'avons déjà mentionné plus haut, la plupart des navigateurs classiques n'utilisent pas les métadonnées. Le volume exagéré des réponses et le manque de pertinence ne satisfont pas les internautes : l'abondance des documents ne suffit pas à élargir la connaissance. Quel que soit le degré de numérisation, la question importante reste celle de l'acquisition des métadonnées. Les outils de recherche ne disposent généralement pas d'informations suffisantes, ni pour lever l'ambiguïté de certains termes de recherche, ni pour identifier les auteurs des métadonnées. Sur ce dernier point, la confiance en l'auteur des métadonnées est primordiale. Cette confiance ne peut exister que si de véritables signatures de documents sont utilisées et si les sources de confiance peuvent être authentifiées.

La méthode préconisée par le W3C repose sur l'usage de RDF qui peut être intégré à de nombreuses applications en XML : catalogues de bibliothèques, agrégations de contenus ou de logiciels, etc. Certains moteurs de recherche spécialisés s'appuient sur les métadonnées définies pour des domaines d'application particuliers : ce qui montre bien l'intérêt de solutions basées sur les métadonnées. Des archives ouvertes sont déployées à grande échelle dans le monde entier. Mais, dans l'ensemble du Web, le RDF qui permet une amélioration notable de la connaissance, n'est pas encore mis en œuvre.

1.3.2.6 Archives ouvertes

Afin de contribuer à la diffusion rapide des connaissances, notamment dans le domaine scientifique, un mouvement international en faveur du libre accès aux publications a proposé dès 1991 des archives ouvertes (OAI). Un protocole d'échange entre ces différents systèmes est à la base des nombreuses archives ouvertes proposées actuellement : un protocole, l'OAI-PMH (Open Archives Initiative, Protocol for Metadata Harvesting) [28] (ISO 14721) garantit l'interopérabilité entre les services d'archive : tous les échanges se font en format XML.

Comme la plupart des systèmes de métadonnées cités précédemment reposent sur le Dublin Core, ces projets ouverts reposent sur une simple description DC en XML. Toutefois, certaines descriptions DC peuvent être étendues pour être utilisables par des moteurs de recherche plus spécialisés, de ce fait moins « ouverts » que les projets initiaux.

De nombreux projets se développent autour des grandes bibliothèques ou des organismes de gestion du patrimoine culturel. Par exemple, l'American Memory de la Bibliothèque du Congrès américain contient plus de 5 millions de ressources. Des projets comparables sont également développés en France, par la Bibliothèque nationale de France (BNF) ou au Royaume-Uni par la British Library. En Europe, la numérisation du patrimoine culturel se poursuit, dans le cadre du projet Michael (inventaire multilingue du patrimoine culturel européen), les développements ont commencé en 2004 : la plate-forme Michael doit pouvoir recueillir prochainement des informations sur des collections dispersées dans toute l'Europe. Dès maintenant, un service multilingue permet d'explorer les collections numériques des musées, services d'archives, bibliothèques et autres institutions culturelles d'Italie, de France et du Royaume-Uni.

En France, la plupart des producteurs de contenu organisent des archives ouvertes. La communauté scientifique, à l'initiative du Centre pour la communication scientifique directe (CCSD) du CNRS qui diffuse le modèle HAL (Hyper Article en Ligne) depuis 2000, a maintenant de nombreux portails d'accès et de dépôt de documents scientifiques, qui donnent une visibilité publique aux travaux des chercheurs : HAL-TEL pour le dépôt des thèses, HAL-INRIA pour l'Inria (Institut national de recherche en informatique et en automatique), HAL-INSERM pour l'Inserm (Institut national de la santé et de la recherche médicale), HAL-SHS pour les sciences de l'homme et de la société, etc. Les travaux se poursuivent pour définir des classifications thématiques, des interfaces de dépôt thématiques et garantir la visibilité institutionnelle des déposants.

```
<mets:metsHdr CREATEDATE="2001-10-23T00:00:00" LASTMODDATE="2003-04-18T07:00:00">
 <mets:agent ROLE="CREATOR">
     <mets:name>Rick Beaubien</mets:name>
 </mets:agent>
</mets:metsHdr>
<mets:dmdSec ID="DM1">
 <mets:mdRef LOCTYPE="URL" MDTYPE="OTHER" OTHERMDTYPE="HTML map"
    xlink_4:href="http://sunsite.berkeley.edu/mapsaux/histopo/sfhistopo.html"
    LABEL="SF Area Quadrangles Index Map"/>
</mets:dmdSec>
<mets:dmdSec ID="DM2">
  <mets:mdRef LOCTYPE="URL" MDTYPE="MARC"
xlink_3:href=http://sunsite2.berkeley.edu:8000/WebZ/Authorize:sessionid=0:bad=html/authofail.html:n
ext=NEXTCMD%22/WebZ/QUERY:next=html/results.html:format=B:numrecs=20:entitytoprecno=1:ent
itycurrecno=1:tempjds=TRUE:entitycounter=1:entitydbgroup=Glad:entityCurrentPage=SearchRecentA
cqdbname=Glad:entitycountAvail=0:entitycountDisplay=0:entitycountWhere=0:entityCurrentSearchSc
reen=html/search.html:entityactive=1:indexA=gl%3A:termA=167937091:next=html/Cannedresultsfram
e.html:bad=error/badsearchframe.html
 LABEL="Catalog Record"/>
</mets:dmdSec>
<mets:dmdSec ID="DM3">
 <mets:mdWrap MDTYPE="MODS">
  <mets:xmlData>
    <mods:mods>
      <mods:titleInfo>
        <title>San Francisco 15-minute Quadrangle</title>
      </mods:titleInfo>
      <mods:name type="corporate" authority="naf">
        <mods:namePart>Geological Survey (U.S.)</mods:namePart>
      </mods:name>
      <mods:typeOfResource collection="yes">cartographic</mods:typeOfResource>
      <mods:genre authority="marc">map</mods:genre>
        <mods:originInfo>
          <place>
            <placeTerm authority="marccountry" type="code">dcu</placeTerm>
          </place>
          <place>
            <placeTerm type="text">Washington, DC</placeTerm>
          </place>
          <mods:dateIssued>1897-1948</mods:dateIssued>
          <mods:dateIssued encoding="w3cdtf" point="start">1897</mods:dateIssued>
          <mods:dateIssued encoding="w3cdtf" point="end">1948</mods:dateIssued>
        </mods:originInfo>
          <relatedItem type="constituent">
            <mods:titleInfo>
             <title>Quentin 7.5-minute Quadrangle</title>
            </mods:titleInfo>
          </relatedItem>
<!--- ...  ...  etc. -- >
      </mods:mods>
    </mets:xmlData>
  </mets:mdWrap>
</mets:dmdSec>
```

Figure 4 - Extrait d'un document METS, Rick Beaubien, U.C. Berkeley Library Systems Office

Ci-dessus : http://www.loc.gov/standards/mets/sfquad.xml

3. ONTOLOGIES
3.1 Définition et rôle des ontologies
3.2 Exemple de l'ontologie du CRM
3.3 Langages pour les ontologies
3.3.1 Premiers langages pour les ontologies
3.3.2 OWL
3.4 SPARQL, un langage d'interrogation

À l'origine, l'ontologie est une branche de la philosophie dans laquelle les philosophes ont tenté de rendre compte de l'existant de façon formelle. Le terme philosophique « ontologie » signifie être du grec ancien *ôn*,*onton*, participe présent de *einai*- et discours, étude, science – de *logos* (*Encyclopédie Universalis*, 2000).

Actuellement, le terme ontologie recouvre deux usages dont le premier appartient à la philosophie classique et le second, plus récent, aux sciences cognitives. La convention veut que la notation Ontologie (avec un O majuscule) soit attribuée au domaine issu de la philosophie et ontologie aux autres.

Pris dans son sens le plus large, le terme ontologie peut être défini comme une théorie ou une conception du réel. En informatique, une ontologie est comprise comme un système de concepts fondamentaux qui sont représentés sous une forme interprétable par un ordinateur.

3.1 Définition et rôle des ontologies

Une des définitions de l'ontologie les plus citées en intelligence artificielle est : « *An ontology is a formal, explicit specification of a shared conceptualization* » [1]. *Conceptualisation* fait référence à un modèle abstrait d'un phénomène dans le monde, en ayant identifié les concepts appropriés à ce phénomène. *Explicite*signifie que le type de concepts utilisés et les contraintes liées à leur usage sont définis explicitement.*Formelle* exprime que l'ontologie doit être interprétable par une machine. *Partagée* reflète l'idée qu'une ontologie constitue un modèle de connaissances partagé par une communauté de personnes [2].

Une ontologie est composée de concepts (principes, idées, catégories d'objet, notions potentiellement abstraites) et des relations. Elle comporte généralement une organisation hiérarchique des concepts pertinents et des relations qui existent entre ces concepts, ainsi que des règles et axiomes qui les contraignent.

Dans le cadre du Web sémantique, les ontologies permettent des spécifications déclaratives des concepts et des rôles dans un domaine de discours. Elles fournissent des vues structurées et partageables des ressources. Elles donnent, entre autres, un vocabulaire pour les métadonnées. Leur formalisation permet l'automatisation de certains raisonnements.

3.2 Exemple de l'ontologie du CRM

Le modèle CIDOC CRM est un modèle sémantique de référence dont l'élaboration a débuté en 1994 par le groupe de normalisation documentaire (*documentation standards group*) du Comité international pour la documentation des musées (ICOM-CIDOC). Il est publié par l'ISO comme une norme internationale (ISO 21127).

Le modèle CRM a pour finalité de fournir un langage commun à des gisements d'information hétérogènes et permettre leur intégration, par-delà leurs éventuelles incompatibilités tant sémantiques que structurelles. Il s'agit de faciliter l'échange et la recherche d'informations dans le domaine du patrimoine culturel et de permettre aux musées de rendre compatibles leurs documentations sans rien perdre de leurs spécificités ni du niveau de précision de leurs données actuelles. Le CIDOC CRM est utilisé dans des projets de recherche tels que le projet européen SCULPTEUR (2002-2005) ou le projet SIMILE du MIT.

Il s'agit d'un modèle sémantique qui constitue une ontologie de l'information relative au patrimoine culturel, c'est-à-dire une formalisation des relations qui unissent les concepts fondamentaux jugés

robustes aux changements de contexte et de perspectives tels que événements et acteurs, participation, classification, etc. (figure 9).
L'intégration des données est la finalité principale du CIDOC CRM (figure 10). Ce processus d'intégration conduit à lever les limites des documents, à relier les informations qu'ils contiennent dans un contexte plus large, à apparier les identificateurs communs et à rédiger des propositions. Il n'y a pas de sens particulier pour accéder aux sujets, relations ou classes des entités. L'ontologie peut être utilisée dans le but d'intégrer les informations administratives. Elle fournit les relations pertinentes pour permettre l'exploration des contextes via des chemins de données concernant des sources diverses de documents.

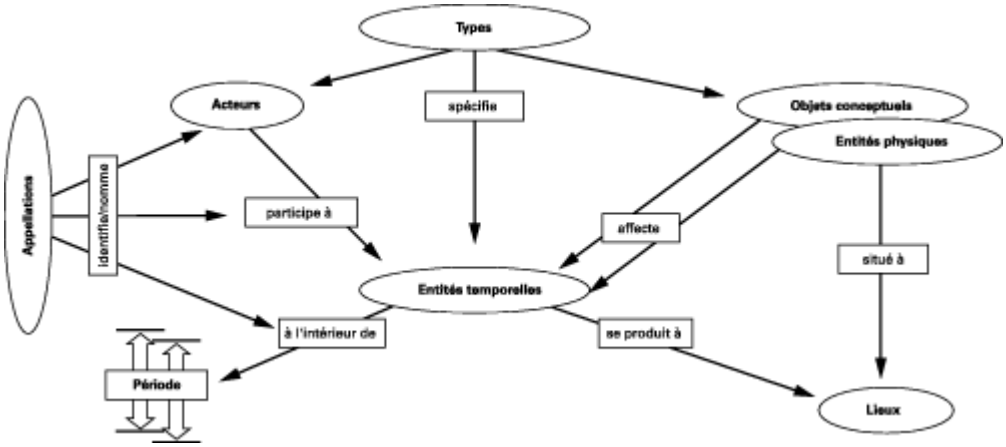

Figure 9 - Modèle conceptuel de référence de ICOM/CIDOC

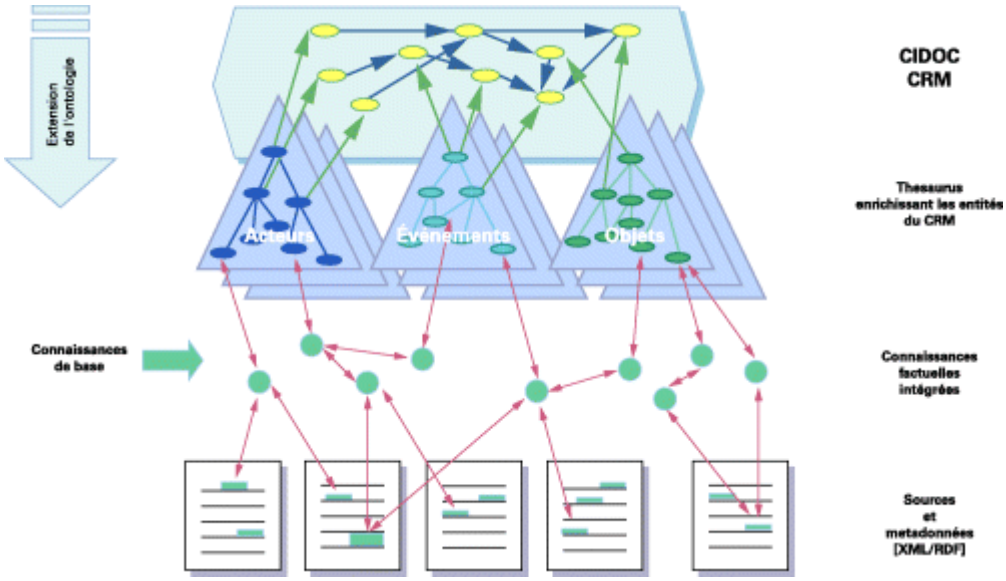

Figure 10 - Intégration d'information à partir de l'ontologie CIDOC/CRM (d'après [3])

L'ontologie est constituée d'environ 80 classes et de 130 relations. Elle est développée dans le langage de représentation des connaissances TELOS et est disponible en RDF(S) et d'autres formats. Un extrait de la hiérarchie des concepts est présenté sur la figure 11.
La description du concept document dans l'ontologie (figure 12) permet de le situer dans la hiérarchie en définissant la classe ancêtre *Information Object* et la sous-classe *Authority Document*, des notes et des exemples précisant le concept et les propriétés définissant le concept.

3.3 Langages pour les ontologies

De nombreux langages (figure 13), appelés *Web-based ontology languages* ou *Ontology Markup Language*, ont été créés pour exploiter les caractéristiques du Web. Leur syntaxe est fondée sur les *markup languages* existants tels que HTML et XML dont l'objectif n'est pas le développement d'ontologies, mais la présentation et l'échange de données.

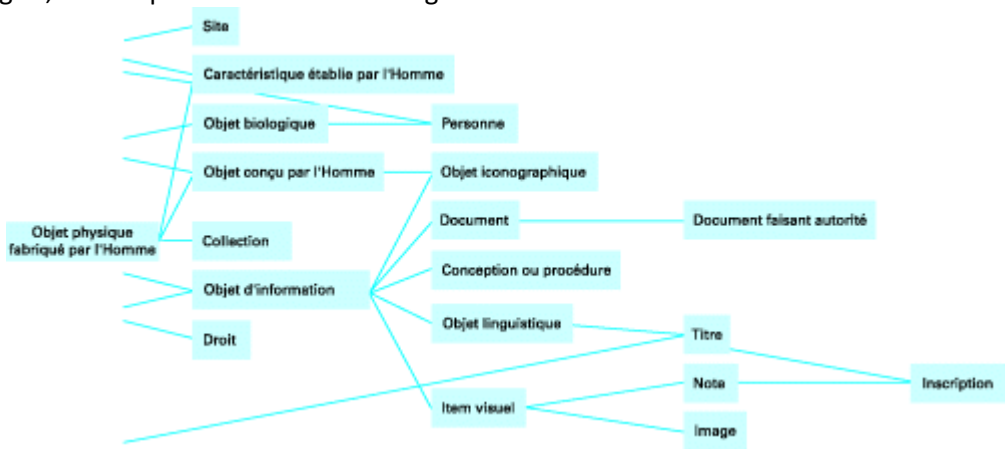

Figure 11 - Extrait de la hiérarchie des concepts du CRM

E31 Document

Subclass of:	E73 Information Object
Superclass of:	E32 Authority Document
Scope note:	This class comprises identifiable immaterial items that make propositions about reality. These propositions may be expressed in text, graphics, images, audiograms, videograms or by other similar means. Documentation databases are regarded as a special case of E31 Document. This class should not be confused with the term "document" in Information Technology, which is compatible with E73 Information Object.
Examples:	- the Encyclopaedia Britannica (E32) - the photo of the Allied Leaders at Yalta published by UPI, 1945 - the Doomsday Book
Properties:	P70 documents (is documented in): E1 CRM Entity

Figure 12 - Description du concept Document

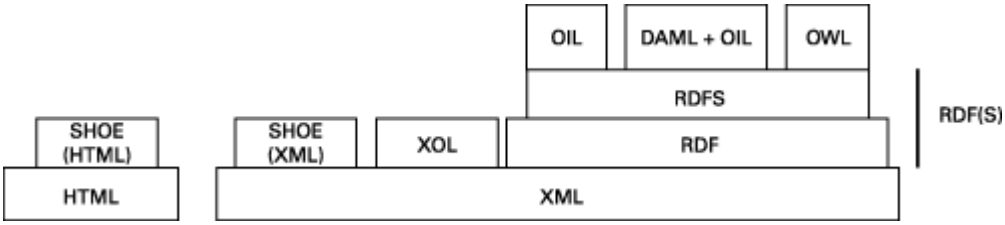

Figure 13 - Langages pour les ontologies

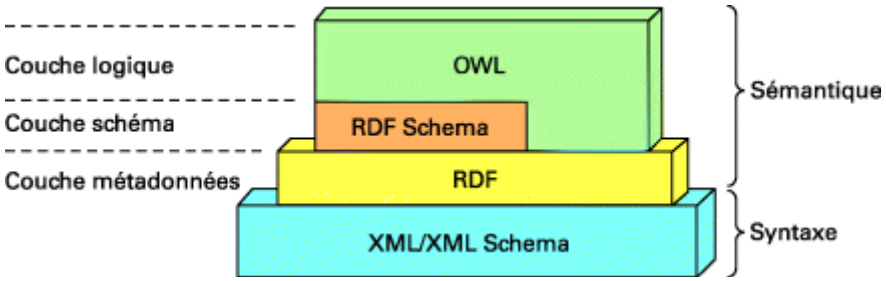

Figure 14 - OWL dans l'architecture du Web sémantique

Les langages pour les ontologies sur le Web ont été développés en respectant les contraintes suivantes :

- être facile à comprendre et à utiliser, c'est-à-dire fondé sur des idiomes familiers de la représentation des connaissances ;
- avoir une spécification formelle ;
- posséder la puissance expressive adéquate ;
- être capable de supporter un raisonnement automatisé ;
- reposer sur des standards existants sur le Web tels que XML, RDF, RDFS.

3.3.1 Premiers langages pour les ontologies

Le premier langage de ce type, SHOE (Simple HTML Ontology Extension) [5] est une extension de HTML. Il combine les *frames* et les règles. Sa syntaxe a ensuite été adaptée à XML.

L'ensemble des langages apparaissant sur la figure 13 est fondé sur XML :

- XOL [6] a été conçu comme une « XMLisation » d'un sous-ensemble de primitives du protocole OKBC ;
- RDF [12] a été développé par le W3C comme un langage fondé sur un réseau sémantique pour décrire les ressources du Web. Il a été proposé comme une recommandation du W3C en 1999 ;
- RDF Schema [13], noté RDFS, a aussi été construit par le W3C comme une extension de RDF avec des primitives fondées sur les *frames*.

Ces langages ont établi les fondements du Web. Dans ce contexte, trois langages ont été développés comme des extensions de RDF(S) : OIL, DAML+OIL et OWL.

OIL [7] a été développé au début des années 2000 dans le cadre du projet européen On-To-Knowledge. Il enrichit RDF(S) avec des primitives de représentation des connaissances fondées sur des *frames* et sa sémantique formelle est fondée sur la logique de description.

DAML+OIL [8] a été créé plus tard entre 2000 et 2001 par un comité sur *Agent Markup Language*assemblant des européens et des américains dans le contexte du projet DAML de la DARPA. Il est fondé sur les spécifications DAML-ONT et OIL. Il ajoute des primitives fondées sur la logique de description à RDF(S).

En 2001, le W3C a formé un groupe appelé Web-Ontology (WebOnt) Working Group dont le but était de concevoir un nouveau *Ontology Markup Language* pour le Web sémantique. Le résultat de leurs travaux est le langage OWL.

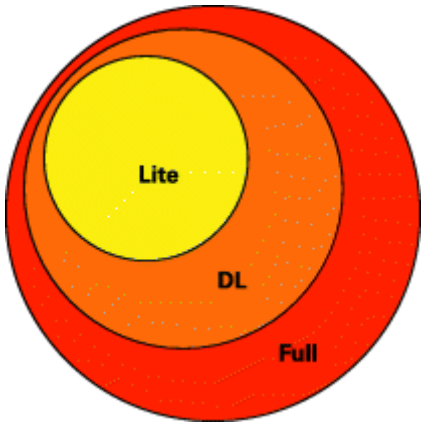

Figure 15 - Les trois sous-langages de OWL

3.3.2 OWL

Le langage OWL (Web Ontology Language) [10] est dérivé du langage DAML+OIL et construit au-dessus de RDF(S). Il enrichit le modèle des RDF(S) en définissant un vocabulaire pour la description d'ontologies complexes.

OWL et RDF(S) sont tous deux des vocabulaires RDF permettant de définir des vocabulaires. RDF(S) définit le plus petit nombre de notions et de propriétés nécessaires à la définition d'un vocabulaire simple, essentiellement :

- les notions de classe, ressource, littéral ;
- les propriétés de sous-classe, de sous-propriété, de champ de valeur, de domaine d'application.

OWL est un langage beaucoup plus riche qui, aux notions définies par RDF(S), ajoute les propriétés de classe équivalente, de propriété équivalente, d'identité de deux ressources, de différence de deux ressources, de contraire, de symétrie, de transitivité, de cardinalité, etc., permettant de définir des rapports complexes entre des ressources (figure 14).

Le langage OWL se compose de trois sous-langages qui proposent une expressivité croissante (figure 15), chacun conçu pour des communautés de développeurs et d'utilisateurs spécifiques : OWL Lite, OWL DL, OWL Full. Chacun est une extension par rapport à son prédécesseur plus simple.

Le langage OWL Lite répond à des besoins de hiérarchie de classification et de fonctionnalités de contraintes simples de cardinalité 0 ou 1. Une cardinalité 0 ou 1 correspond à des relations fonctionnelles, par exemple, une personne a une adresse. Toutefois, cette personne peut avoir un ou plusieurs prénoms, OWL Lite ne suffit donc pas pour cette situation.

Le langage OWL DL concerne les utilisateurs qui souhaitent une expressivité maximale couplée à la complétude du calcul (cela signifie que toutes les inférences seront assurées d'être prises en compte) et la décidabilité du système de raisonnement (c'est-à-dire que tous les calculs seront terminés dans un intervalle de temps fini). Ce langage inclut toutes les structures OWL avec certaines restrictions, comme la séparation des types : une classe ne peut pas être à la fois un individu ou une propriété. Il est nommé DL car il correspond à la logique de description.

Le langage OWL Full est destiné aux utilisateurs souhaitant une expressivité maximale. Il a l'avantage de la compatibilité complète avec RDF/RDFS, mais l'inconvénient d'avoir un haut niveau de capacité de description, quitte à ne pas pouvoir garantir la complétude et la décidabilité des calculs liés à l'ontologie.

Le tableau 2 résume les caractéristiques de ces trois sous-langages.

Comparatif des couches de OWL

	OWL Full	OWL DL	OWL Lite
Caractéristiques du langage	Très haute expressivité Indécidable Formalisation non standard Syntaxe RDF	Expressivité maximale Maintien de la calculabilité Formalisation standard	Hiérarchie de classes Contraintes simples
Éléments constituant les couches de OWL	Métaclasse Pas de restriction sur le vocabulaire Classes utilisées comme instances	Négation Disjonction Cardinalité complète Types énumérés	Schéma RDF Conjonction Inégalité Cardinalité 0/1 Type de données Propriété inverse, transitive, symétrique APourValeur someValuesFrom allValuesFrom

Deux classes prédéfinies :
owl : Thing et owl : Nothing

Toute classe OWL est :
– **une sous-classe de** owl : Thing
– **une super-classe de** owl : Nothing.

Les classes sont définies avec un élément owl :Class (owl :Class **est une sous-classe de** rdfs :Class)

```
<owl:Class rdf:ID="MaitreDeConference">
        <rdfs:subClassOf rdf:resource="#EnseignantChercheur"/>
</owl:Class>
```

Disjonction de classe
La classe MaitreDeConference **est disjointe de la classe des professeurs et des ingénieurs.**

```
<owl:Class rdf:about="MaitreDeConference">
        <owl:disjointWith rdf:resource="#Professeur"/>
        <owl:disjointWith rdf:resource="#Ingenieur"/>
</owl:Class>
```

On peut définir des équivalences entre classes.

```
<owl:Class rdf:ID="MaitreDeConference">
        <owl:equivalentClass rdf:resource="#Intervenant"/>
</owl:Class>
```

Figure 16 - Exemple de définition de classes en OWL

Propriété objet

```
<owl:ObjectProperty rdf:ID="estEnseignePar">
        <rdfs:domain rdf:resource="#cours"/>
        <rdfs:range rdf:resource="#enseignant"/>
        <rdfs:subPropertyOf rdf:resource="#implique"/>
</owl:ObjectProperty>
```

Propriété typée

```
<owl:DatatypeProperty rdf:ID="Age">
        <rdfs:range rdf:resource="http://www.w3.org/2001/xmlSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>
```

Figure 17 - Exemple de définition des propriétés en OWL

On peut définir des classes par combinaison ensembliste (union, intersection, complément) d'autres classes. Par exemple, on définit la classe PersonneUniversite **comme l'union de la classe** Enseignant **et de la classe** Étudiant.

```
<owl:Class rdf:about="PersonneUniversite">
        <owl:unionOf rdf:parseType="Collection"/>
          <owl:Class rdf:about="#Enseignant"/>
          <owl:Class rdf:about="#Enseignant"/>
        </owl:unionOf>
</owl: Class>
```

Figure 18 - Exemple de combinaison de classes en OWL

Les caractéristiques de OWL peuvent se résumer de la manière suivante :

- OWL est fondé sur RDF ;
- les classes, les propriétés et les individus sont disjoints ;
- un individu ne peut pas être une classe ;
- une classe ne peut pas être considérée comme un individu.

Une classe peut être définie :

- par une référence URI ;
- par l'énumération de ses instances ;
- par ses propriétés (définition intensionnelle) ;
- comme union, intersection, complément d'autres classes,

et peut être disjointe d'autres classes.

Quelques exemples présentent la définition des classes (figure 16) et des propriétés en OWL (figure 17) et les combinaisons de ces classes (figure 18).

3.4 SPARQL, un langage d'interrogation

Le Web sémantique est une évolution du Web permettant la description formelle de ses ressources afin d'automatiser des tâches et d'améliorer des traitements tels que la recherche d'information, l'intégration de sources hétérogènes, etc. Ces descriptions formelles doivent être parfois affichées par exemple pour présenter le résultat d'un appel à un moteur de recherche.

```
PREFIX rdf : http://www.w3.org/1999/02/22-rdf-syntax-ns#
PREFIX foaf : http://xmlns.com/foaf/0.1/
SELECT DISTINCT ?nom ?image ?description
WHERE{
        ?personne rdf :type foaf ;Person .
        ?personne foaf :name ?nom .
        ?image rdf :type foat :Image .
        ?personne foaf :img ?image .
        ?image dc :description ?description
```

Figure 19 - Exemple d'une requête SELECT en SPARQL

Le langage SPARQL (Simple Protocol and RDF Query Language) [17], en cours de normalisation dans le cadre de l'activité Web sémantique du W3C, définit la syntaxe et la sémantique nécessaires à l'expression de requêtes sur une base de données de type RDF et la forme possible des résultats.

SPARQL est adapté à la structure spécifique des graphes RDF, et s'appuie sur les triplets qui les constituent. En cela, il est différent du classique SQL (langage de requête qui est adapté aux bases de données relationnelles), mais s'en inspire clairement dans sa syntaxe et ses fonctionnalités. Il a aussi quelques traits de ressemblance mineurs avec le langage Prolog. Il permet d'exprimer des requêtes interrogatives ou constructives :

- une requête SELECT (figure 19), de type interrogative, permet d'extraire du graphe RDF un sous-graphe correspondant à un ensemble de ressources vérifiant les conditions définies dans une clause WHERE ;
- une requête CONSTRUCT, de type constructive, engendre un nouveau graphe qui complète le graphe interrogé.

Par exemple, sur un graphe RDF contenant des informations généalogiques, on pourra par une requête SELECT trouver les parents ou grands-parents d'une personne donnée, et par des requêtes CONSTRUCT ajouter des relations frère-sœur, cousin-cousine, oncle-neveu, qui ne seraient pas explicitement déclarées dans le graphe initial.

BIBLIOGRAPHIE

- (1) - GRUBER (T.R.) - A Translation Approach to Portable Ontology Specifications - . Knowledge Acquisition, 5(2), 199-220 (1993).
- (2) - STUDER (R.),BENJAMINS (V.R.),FENSEl (D.) - Knowledge Engineering : Principles and Methods - . IEEE Transactions on Data Knowledge Engineering, 25 (1-2), 161-197.
- (3) - DOERR (M.) - The CIDOC CRM, an Ontological Approach to Schema Heterogeneity - . Dagstuhl Seminar Proceedings 04391. Semantic InterOperability and Integration (2004). http://drops.dagstuhl.de
- (4) - GOMEZ-PEREZ (A.),FERNANDEZ-LOPEZ (M.),CORCHO (O.) - Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web - . Springer-Verlag (2004).
- (5) - LUKE (S.),HEFLIN (J.D.) - SHOE 1.01 Proposed Specification. Technical Report - . Parallel Understanding Systems Group, Department of Computer Science, University of Maryland (2000). http://www.cs.umd.edu/projects/plus/SHOE/
- (6) - KARP (P.D.),CHAUDHRI (V.),THOMERE (J.) - XOL : An XML-Based Ontology Exchange Language - . Version 0.3. Technical Report (1999). http://www.ai.sri.com/~pkarp/xol/xol.html
- (7) - HORROCKS (I.),FENSEL (D.),HARMELEN (F.),DECKER (S.),ERDMANN (M.),KLEIN (M.) - * - OIL in a Nutshell. In DIENG R., CORBY (O.) (éd.).– Lecture Notes in Artificial Intelligence LNAI 1937. Springer Verlag (2000).
- (8) - HORROCKS (I.),van HARMELEN (F.) (éd.) - Reference Description of the DAML+OIL Ontology Markup Language - . Technical Report (2001). http://www.daml.org/2001/03/reference.html
- (9) - * - HTML 4.01 Specification http://www.w3.org/TR/html401/

- (10) - OWL, Web Ontology Language - http://www.w3.org/TR/owl-features/
- (11) - RDF/XML Syntax Specification - http://www.w3.org/TR/rdf-syntax-grammar/
- (12) - RDF – Ressource Description Framework (RDF) Model and Syntax Specification - http://www.w3.org/TR/REC-rdf-syntax/
- (13) - RDF Vocabulary Description Language 1.0 : RDF Schema - http://www.w3.org/TR/PR-rdf-schema
- (14) - XQuery 1.0 and XPath 2.0 Formal Semantics - http://www.w3.org/TR/xquery-semantics
- (15) - XML Pointer language - http://www.w3.org/xptr/
- (16) - Semantic Web - http://www.w3.org/2001/sw/
- (17) - SPARQL, Query Language for RDF - http://www.w3.org/TR/rdf-sparql-query/
- (18) - XHTML 1.0, The Extensible HyperText Markup Language - http://www.w3.org/TR/xhtml1
- (19) - XML, Extensible Markup Language, XML 1.0 - http://www.w3.org/TR/xml/
- (20) - * - Namespaces in XML1.0 http://www.w3.org/TR/xml-names
- (21) - * - EAD, Encoded Archival Description, http://www.loc.gov/ead/
- (22) - MARC, Machine Readable Cataloging - http://www.loc.gov/marc
- (23) - MARCXML - http://www.loc.gov/marcxml
- (24) - METS, Metadata Encoding & Transmission Standard - http://www.loc.gov/mets
- (25) - MIX, Mediation of Information Using XML - http://www.loc.gov/mix
- (26) - MODS, Metadata Object Description Schema - http://www.loc.gov/mods/
- (27) - * - ONIX, Online Information Exchange http://www.editeur.org/onix.html
- (28) - * - OAI-PMH, Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/OAI/ openarchivesprotocol.html
- (29) - ACCARY-BARBIER (T.),CALABRETTO (S.) - XML : syntaxe - . [H 3 500], Technologies logicielles – Architectures des systèmes (2005).
- (30) - CALABRETTO (S.),ACCARY (T.) - XML : gestion de contenus Web - . [H 3 502], Technologies logicielles – Architectures des systèmes (2006).
- (31) - SONKÉ (L.) - SGML - . [H 7 138], Documents numériques – Gestion de contenu (1998).
- (32) - CHAHUNEAU (F.) - XML - . [H 7 148], Documents numériques – Gestion de contenu (2001).
- (33) - ROLE (F.) - TEI (Text Encoding Initiative) - . [H 7 158], Documents numériques – Gestion de contenu (1999).

Note : **Text Encoding Initiative (TEI)**

Depuis l'Antiquité, il est courant de marquer et d'annoter des textes en vue d'en faciliter l'étude ou la critique (pensons par exemple aux systèmes d'annotation médiévaux ou à l'appareil de symboles imaginé dès le III^{e} siècle avant J.-C. par les philologues alexandrins).

Dans l'univers numérique, le marquage électronique (défini ici comme l'insertion dans un fichier électronique de marques liées au texte mais n'en faisant pas directement partie) a longtemps servi presque exclusivement à piloter des dispositifs d'impression ou d'affichage (photocopieuses, imprimantes, écran). C'est ce marquage qu'utilisent implicitement (*) la plupart des chercheurs en sciences humaines au travers des outils PAO du commerce.

Nota :

(*) « implicitement » dans le sens où les manipulations effectuées via le clavier ou les dispositifs de pointage génèrent d'une manière ou d'une autre les informations de balisage physique sur lesquelles le logiciel de PAO s'appuie pour effectuer les opérations qu'on lui demande.

Malgré ses mérites, ce marquage est, comme nous l'avons dit, orienté vers la production ou l'affichage du texte, et n'est donc pas conçu pour faciliter une exploration intellectuelle des

documents. Peu à peu s'est donc imposée l'idée qu'il fallait recourir à un niveau de balisage moins dépendant des contraintes de production, et propice à des traitements de plus haut niveau sur les textes, parce qu'en décrivant la structure logique.

SGML (Standard Generalized Markup Language) est la norme actuellement la plus utilisée pour baliser logiquement des textes. Elle permet à tout utilisateur de définir, via l'écriture d'une DTD (Définition du Type de Document) un langage de balisage logique adapté à ses besoins.

La **Text Encoding Initiative (TEI) est une DTD SGML accompagnée par un volume de « recommandations » ; les TEI « Guidelines »** expliquant de quelle façon doit être utilisée la DTD. **Cette DTD est adaptée principalement aux besoins de la communauté des chercheurs en sciences humaines** (ou plus généralement à tout chercheur voulant explorer de vastes corpus textuels sous forme électronique). Elle permet au linguiste de baliser syntaxiquement des corpus, à l'historien de marquer dans un texte des dates, des noms de lieu ou de personnage, au chercheur en littérature d'étudier la stylistique ou la genèse d'un texte, etc.

Après quelques rappels historiques et une présentation informelle de la structure d'un texte TEI, nous décrivons les mécanismes mis en œuvre dans l'écriture de la DTD TEI (modularité, héritage, extensibilité).

Cette partie plus technique que les autres nécessite une bonne connaissance de SGML.

A la fin de cet article nous présentons quelques exemples de balisage TEI.

Les concepts et techniques liés au SGML sont exposés dans l'article « SGML » du présent traité.