# L'usage des métadonnées dans la description et la recherche des ressources sur le WEB

CHAWK Mohamad
Université Lille III, Laboratoire CERSATES
chawk@univ-lille3.fr
BEN ABADALLAH NABIL
*Université de Bourgogne, LIMSIC (Laboratoire Image, Médiations, Sensible, en Information-Communication)*
*n.ben_abdallah1@tiscali.fr*

**Mot-clés :**
Description de ressources / Métadonnée / Ontologie / OWL/ Document numérique / Marqueur sémantique / Recherche d'information

**Keywords :**
Description of resources / Metadata / Ontology/ OWL / Numerical document / Semantic marker / Retrieval information

**Résumé :**
Le rôle prépondérant de l'Internet en tant que canal de diffusion de l'information est aujourd'hui indéniable. La question est de savoir comment peut-on exploiter au mieux les diverses ressources disponibles sur le Web sachant que les machines actuelles sont incapables d'atteindre la richesse sémantique de ces ressources. Allons-nous nous contenter de transférer, vers ces machines connectées aux réseaux, des pratiques manuelles définies et façonnées par un contexte restreint d'utilisation ou devons-nous développer des nouvelles approches de recherche et de description des ressources pour approprier davantage ce nouveau «media» ? Nous essayons, dans le cadre de ce travail, de définir les conditions nécessaires à une meilleure standardisation de la production des métadonnées indispensables pour une exploitation optimale des ressources Web. Dans telles conditions, l'usage classique des metadonnées comme moyen de représentation du contenu va-il se transformer ? Les avantages et les limites des schémas actuels sont décrits et discutés pour déterminer les apports et dégager les tendances en matière d'enrichissement sémantique des ressources Web.

**Abstract :**
The dominating role of the Internet as a channel of diffusion of information is undeniable today. The question is to know how can one as well as possible exploit the various resources available on the Web knowing that the current machines are unable to reach the semantic richness of these resources. Will we be satisfied to transfer, towards these machines connected to the networks, from the manual practices definite and worked by a restricted context of use or must us develop new approaches of research and description of the resources to adapt this new "media more"? We are trying in the work to define the necessary condition to improve the standardization of the production of metadata which are essential elements for the optimal conception and exploitation of resources Web. Will the classic usage of these metadata as a tool for the presentation of content be transformed in these conditions? The advantages and the limit of the actual schema will be described and discussed to determine the contribution for the improvement of semantic resources Web.

# 1 Introduction

Le Web actuel renferme une masse importante d'informations inaccessibles et mal exploitées. Cette situation est due essentiellement aux facteurs suivants : la description des ressources n'est pas appropriée, l'hétérogénéité des formats et des informations, l'imprécision dans la recherche d'information – les interfaces ne sont pas adaptées aux multiples situations de recherche- et la variété de services sur le Web. Dans ce contexte, le Web sémantique se présente comme une alternative : nous passons d'un Web axé essentiellement sur des protocoles de mise en ligne et d'échange de ressources multiples et hétérogènes à un Web axé sur le contenu. Les applications de la «nouvelle génération» doivent permettre d'atteindre la richesse sémantique des ressources Web en facilitant leur exploitation. Selon Tim Berners-Lees, le Web sémantique permettra de rendre le contenu sémantique des ressources Web interprétables non seulement par l'homme mais aussi par la machine. Les multiples «facettes» de la gestion du contenu des ressources imposent au domaine de recherche du Web sémantique une approche pluridisciplinaire croisant diverses disciplines telles que les sciences cognitives, l'ingénierie documentaire, l'ingénierie de connaissances, les systèmes muti-agents, le traitement automatique des langues, etc.

Le Web sémantique est actuellement un vaste projet de recherche dont l'acteur principal est le W3C (Word Wide Web Consortium). Le projet vise à représenter le contenu sémantique des documents sur le Web, à augmenter la précisons et la pertinence de la recherche d'information, à intégrer des ressources d'information hétérogènes, à doter les machines par des moyens leur permettant de faire des raisonnements complexes sur ces données et enfin à assister l'utilisateur dans sa recherche d'information et à le décharger de certaines tâches. Les applications visées par les travaux du Web sémantique sont diverses : Recherche d'information, veille technologique, gestion de l'information scientifique et technique, e-learning, e-commerce, catalogues : gestion, construction et consultation, ontologies et serveurs de connaissances, Interopérabilité et coopération entre applications, mémoire d'entreprise, etc.

Les travaux sur le Web sémantique s'articulent essentiellement sur deux axes : l'élaboration des langages spécifiques et le développement des outils appropriés. Ces langages et outils, sans un certain consensus entre les différents acteurs impliqués dans leur développement, ne peuvent être performants et évolutifs. Ce consensus doit se traduire par un ensemble de standards et de normes régissant les différents «composants» du Web sémantique. Dans ce qui suit, nous présenterons quelques langages du Web sémantique.

# 2 Langages du Web sémantique

Les langages du Web sémantique doivent permettre : d'exprimer les données et les métadonnées des différentes ressources du Web, d'exprimer les ontologies –nous y reviendrons plus loin- et de décrire les différents services. Nous avons trois types de langage : les langages d'assertions (cartes topics et RDF), les langages de représentation d'ontologies (OWL, ..) et enfin les langage de description et de composition de services (UDDI, XDD, etc.). Nous nous intéressons, dans notre travail, aux deux premiers types qui sont en relation avec l'utilisation des métadonnées.

## 2.1 RDF (Resource Description Framework)

Le langage RDF se base sur la syntaxe du métalangage XML. RDF est un langage formel qui permet de définir des relations entre des « ressources » et d'annoter ou d'associer des métadonnées à des documents écrits avec des langages non structurés. Par exemple, la syntaxe RDF permet de définir la structure des métadonnées qui peuvent être associées à des documents HTML.

Le modèle de données élémentaire du langage RDF est basé sur trois types d'objets : ressources, propriétés et déclaration. La déclaration est la valeur assignée à une propriété de la source, elle est

représentée sous forme de triplets (sujet, prédicat et objet). La valeur d'une propriété peut être littérale ou une autre ressource. Graphiquement, une déclaration RDF est représentée sous la forme de graphes étiquetés. Les nœuds sous forme d'ovales représentent les ressources et les arcs représentent les propriétés déclarées. Les nœuds représentant des chaînes littérales seront sous forme de rectangles.
RDF dispose de trois types d'objets conteneurs : Bag <rdf : Bag>, Sequence <rdf : seq>, Alternative <rdf :Alt>. Le conteneur Bag est utilisé pour représenter une liste non ordonnée de ressources ou de littéraux ; Sequence est utilisé pour déclarer une liste ordonnée de ressources ou de littéraux ; Alternative est utilisé pour représenter des alternatives à la valeur unique d'une propriété donnée. En RDF la signification est exprimée par la référence à un *schéma* définissant les termes qui seront utilisés dans les différentes déclarations. Dans l'exemple ci-dessous, l'élément <rdf :rdf> est utilisé pour préciser que le schéma Dublin Core sera utilisé dans les déclarations. <rdf : rdf xmlns : rdf = http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/http://www.w3.org/1999/02/22-rdf-syntax-ns-1999022# **xmlns : dc** = "http://purl.org/metadata/dublin_core#">.

### RDFs

RDF a fait l'objet d'extension à RDFs (RDF Schema) qui fournit un mécanisme permettant de définir des classes. Les ressources qui sont étiquetées par ces classes seront des instances ou/et des propriétés. Cette extension apportée à RDF ne fournit pas des mécanismes solides pour la spécification des classes et pour faire des raisonnements ou des calculs sur les données, c'est la limite dont soufre le langage RDF.

## 2.2 OWL (Web Ontology Langage)

Dans le guide d'utilisation du langage OWL, publié par le W3C en février 2004, nous pouvons lire *: «Le langage d'ontologie Web OWL sert à formaliser un domaine en définissant des classes et les propriétés de celles-ci, à définir des individus et affirmer des propriétés les concernant, et à raisonner sur ces classes et ces individus dans la mesure où la sémantique formelle du langage OWL le permet.»*. Il s'agit donc d'un langage formel d'écriture des ontologies Web.

### Qu'est ce qu'une ontologie :

Le terme ontologie est issu du domaine de la philosophie. Il signifie « explication systématique de l'existence » et désigne l'étude de ce qui existe. Ce terme a été introduit en Intelligence Artificielle (IA) avec l'émergence de l'ingénierie de connaissances qui s'intéresse aux problématiques de représentation et de manipulation des connaissances au sein des systèmes informatiques. On est donc loin du sens philosophique de l'ontologie (l'explication systématique de l'existence). Selon T. R. Gruber[1] : "*In the context of knowledge sharing, I use the term ontology to mean a specification of a conceptualization. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy.*" L'ontologie selon Gruber devient une spécification explicite de conceptualisation. Dans ce qui suit, nous admettons qu'une ontologie est construite à partir des concepts du domaine et des relations entre ces concepts. Il consiste à identifier et à modéliser les concepts et les relations conceptuelles.
Eléments du langages OWL

Le langage OWL s'appuie sur le langage DAML+OIL, produit de la combinaison de l'américain DAML (Darpa Agent Markup Language) et OIL (Ontology Inference Layer) provenant de projets

[1] T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993. http://ksl-web.stanford.edu/KSL_Abstracts/KSL-92-71.html.

européens. Le langage OWL est construit sur la base de RDF-Schéma et dispose d'une syntaxe XML. Il permet de définir des ontologies Web structurées, autorise une intégration plus riche et, garantit l'interopérabilité des données entre différentes applications. Une *ontologie OWL* renferme des descriptions de *classes*, de *propriétés* et de leurs instances. Pour une ontologie donnée, la sémantique formelle du langage OWL, issue des logiques de description, permet de déduire les faits qui ne sont pas littéralement présents dans l'ontologie.
Dans OWL, il est très important de distinguer entre une *classe* et une instance. Cette distinction conditionne la réussite ou non de la conception d'une ontologie donnée. Le guide d'utilisation du langage OWL publié par le W3C précise : *«Une classe est simplement un nom et une collection de propriétés qui décrivent un ensemble d'individus. Les individus sont les membres de cet ensemble. Ainsi, les classes devraient correspondre aux ensembles de choses qui apparaissent naturellement dans un domaine d'étude et les individus devraient correspondre aux entités réelles qu'on peut regrouper dans ces classes»*. Nous reviendrons sur cette distinction dans notre exemple d'ontologie.

Le langage OWL regroupe trois sous-langages complémentaires (OWL LITE , OWL DL, OWL FULL). Les sous-langages devraient répondre aux attentes de communautés de développeurs et d'utilisateurs spécifiques. Nous présentons brièvement les constructeurs utilisés par le langage OWL[2].

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:sameIndividualAs et owl:differentIndividualFrom permettent d'affirmer que deux individus sont égaux ou différents,
Des mot-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 triplet <*s p o*> a pour conséquence <*o p' s*>); d'autres caractéristiques sont par exemple 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 (resp. 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 (resp. 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 Schema,
permet d'utiliser une classe en position d'individu dans les constructeurs,
*unionOf, complementOf* et *intersectionOf :* permettent des combinaisons arbitraires de classes et de restrictions.

[2] Pour plus d'information sur les constructeurs, nous renvoyons le lecteur au guide d'utilisation du langage OWL publié par le W3C en février 2004.

Afin d'illustrer l'usage du langage OWL, nous avons formaté un **extrait simple** d'une ontologie sur les technologies d'information et de communication (TIC) selon les spécifications OWL définie dans le guide d'utilisation publié par le W3C en février 2004. L'ontologie est en cours d'élaboration, elle pourrait être exploitée, par exemple, pour faciliter la description et la recherche des ressources dans un portail dédié à la veille des usages des TIC en entreprise.

```
<rdf : RDF
        xmlns : rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns :rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns : owl ="http://www.w3.org/2004/OWL/#"
        xmls = "http://dmoz.org/World/Fran%c3%a7ais/Informatique#"

<owl : Ontology rdf:about="">
        <owl:VersionInfo>
                Mon exemple août 2004
        </owl : VersionInfo>
</owl : Ontology>
<owl: Class rdf : ID="Logiciel">
        <rdfs : comment>Un logiciel est un programme informatique destiné à …</rdfs : comment>
        <owl : DataProperty  rdf : ID= "Nom">
                <rdfs : domain rdf : resource= "Logiciel">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty>
        <owl : DataProperty  rdf : ID=  "Editeur ">
                <rdfs : domain rdf : resource= "#Logiciel">
                <rdfs : range rdf : resource="&xsd; string"/>
        </owl : DatatypeProperty>
        <owl : DataProperty  rdf : ID=  "Version ">
                <rdfs : domain rdf : resource= "#Logiciel">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty>
</owl : Class>
<owl : Class rdf : ID="SystemeExploitation">
        <rdfs:comment>Un System d'exploitation est un ensemble de programmes … </rdfs:comment>
        </owl : DatatypeProperty>
        <owl : DataProperty  rdf : ID= "Nom">
                <rdfs : domain rdf : resource= "#SystemeExploitation">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty
        <owl : DataProperty  rdf : ID= "PlateformMateriel">
                <rdfs : domain rdf : resource= "#SystemeExploitation">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty>
        <owl : DataProperty  rdf : ID= "NumeroVersion">
                <rdfs : domain rdf : resource= "#SystemeExploitation">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty
</owl: Class>
<owl: Class  rdf : ID="Programmation">
        <rdfs : comment>Programmation est une activité informatique … </rdfs:comment>
        <owl : DataProperty  rdf : ID= "TypeProgram">
                <rdfs : domain rdf : resource= "#Programmation">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty
</owl : Class>
<owl : Class rdf : ID="Bureautique">
        <rdfs : comment>Bureautique  est une application informatique …</rdfs : comment>
        <rdfs : subClassOf rdf:resource="#Logiciel"/>
        <owl : DataProperty  rdf : ID= "ChampApplication">
                <rdfs : domain rdf : resource= "#Bureautique">
                <rdfs : range rdf:resource="&xsd; string"/>
```

```
        </owl : DatatypeProperty>
</owl : Class>
<owl : Class rdf : ID="Navigateur">
        <rdfs : comment>Navigateur est un Logiciel permettant l'affichage des document html … </rdfs:comment>
        <rdfs : subClassOf rdf:resource="#Logiciel"/>
        <owl : DataProperty  rdf : ID= "TypePlatefomr">
                <rdfs : domain rdf : resource= "#Navigateur">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty>
</owl : Class>
<owl:Class  rdf : ID="SystemeProprietaire">
        <rdfs : comment>SystemeProprietaire est une sous classe SystemeExploitaion</rdfs:comment>
        <rdfs : subClassOf  rdf :resource="#SystemeEploitation"/>
        <<owl : DataProperty  rdf : ID= "NomProprietaire">
                <rdfs : domain rdf : resource= "#SystemeProprietaire">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty>
</owl:Class>
<owl:Class  rdf: ID="SystemeLibre">
        <rdfs : comment>SystemeLibre  est une sous classe de SystemeExploitaion</rdfs:comment>
        <rdfs : subClassOf rdf:resource="#SystemeEploitation"/>
        <owl : DataProperty  rdf : ID= "NomProducteur">
                <rdfs : domain rdf : resource= "#SystemeLibre">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty
</owl:Class>
<owl:Class  rdf: ID="Langage">
        <rdfs : comment>Langage est une sous classe de Programmation </rdfs:comment>
        <rdfs : subClassOf rdf:resource="#Programmation"/>
        <owl : DataProperty  rdf : ID= "Nom">
                <rdfs : domain rdf : resource= "#Langage ">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty
        <owl : DataProperty  rdf : ID= "NumeroVersion">
                <rdfs : domain rdf : resource= "#Langage">
                <rdfs : range rdf:resource="&xsd; integer"/>
        </owl : DatatypeProperty>
</owl:Class>
<owl : Class  rdf: ID="Methode">
        <rdfs : comment>Methode  est une sous classe de Programmation</rdfs : comment>
        <rdfs : subClassOf rdf:resource="#Programmation"/>
        <owl : DataProperty  rdf : ID= "NomMethode">
                <rdfs : domain rdf : resource= "#Methode">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty
        <owl : DataProperty  rdf : ID= "TypeMethode">
                <rdfs : domain rdf : resource= "#Methode">
                <rdfs : range rdf:resource="&xsd; string"/>
        </owl : DatatypeProperty>
</owl:Class>
```

# 3 Web sémantique et veille technologique

La gestion des différents types d'information est aujourd'hui un enjeu majeur pour l'entreprise. Les ressources électroniques internes (documentations techniques, procès verbaux, procédures de productions, contrats, etc.) et externes (essentiellement les ressources disponibles via le Web) sont en augmentation permanente et posent des problèmes aux entreprises. Cette situation rend urgent l'élaboration des outils «intelligents» pour traiter ces ressources et faciliter, par conséquent leur exploitation par l'entreprise. Les besoins de cette dernière en matière de gestion d'information sont énormes : retrouver les informations pertinentes par rapport à un besoin donné, découvrir des services utiles, partager et communiquer des connaissances spécifiques, etc. Tous ces besoins et les

développements qui en résultent pourraient être un des domaines d'application du Web sémantique. Ainsi, la recherche, l'extraction, la collecte, le traitement, le stockage et la diffusion de l'information sont autant de tâches qui seront facilitées par les méthodes et les outils développés dans le cadre du Web sémantique.

En exploitant des ressources électroniques enrichies par des métadonnées, organisées sous formes d'ontologies permettant des inférences sur les données disponibles, **la veille** stratégique, économique, concurrentielle, technologique, … produira certainement des produits (ex. revue de presse, bulletin d'information, annuaire, données statistiques, etc.) plus précis et plus exhaustifs. En outre, la technologie du Web sémantique permettra à l'entreprise : d'exploiter des ontologies de son domaine d'activités, d'annoter ses ressources électroniques pour une meilleure optimisation de sa «capitale connaissances», de vérifier la validité des sources d'information, de définir des profils d'utilisateur pour des applications données, etc. En outre, l'ontologie a une place centrale dans un projet de mémoire d'entreprise.

Plusieurs projets ont été développés autour du Web sémantique d'entreprise, nous citons ici, à titre d'exemple, Le projet CoMMA (Corporate Memory Management throught Agents ) réalisé au sein l'INRIA. Le projet s'intéresse à la gestion de connaissances, la construction d'ontologie, l'apprentissage, le tout est intégré dans une plate-forme multi-agents. Le système est doté d'une ontologie O'CoMMA permettant le recueil, la structuration, la validation et la formalisation de l'information en langage RDFs. La recherche d'information se fait à partir des agents guidés par des ontologies.

Dans ce qui suit, nous nous intéresserons essentiellement aux métadonnées comme moyen d'enrichissement sémantique des ressources sur le web.

# 4 Schémas de métadonnés

Les langages formels (RDF, OWL, etc.) de représentation et de structuration de métadonnées permettent une exploitation optimale des schémas prédéfinis de métadonnées (DC, LOM, etc.). Ces schémas sont caractérisés par des éléments dont la sémantique est établie au moment de leur conception. Dans ce qui suit, nous présenterons un schéma de métadonnées utilisé dans la description des ressources électroniques sur le Web.

Les concepteurs des schémas de description affirment que l'association des métadonnées descriptives standardisées aux différents objets en réseau offre un potentiel d'amélioration substantiel des possibilités de localisation de ressources : en permettant des recherches basées sur des champs (créateur, titre, etc.), en permettant l'indexation d'objets non-textuels et en permettant l'accès à un contenu de substitution, ce qui est différent de l'accès au contenu de la ressource elle-même. Il est clair que l'objectif principal de l'utilisation des métadonnées reste l'amélioration de la qualité de recherche sur le réseau [Lago 97]. Ainsi, la volonté d'accroître la précision des moteurs de recherche sur le Web est un argument fréquemment avancé pour justifier l'élaboration de ces schémas. La détermination des métadonnées cohérentes est un des facteurs clés qui permettra d'atteindre un meilleur taux de rappel. En revanche, des métadonnées incohérentes cachent souvent (silence) les ressources désirées, ce qui donne des résultats de recherche, imprévisibles ou incomplets.

Une recherche sur des champs (données structurées) est plus facile à réaliser et à plus de chance d'aboutir qu'une recherche sur le contenu (données «brutes» ou/et semi-structurées). Nous retrouvons un des vieux paradigmes de la recherche d'informations sur les catalogues en ligne (bibliothèques ou/et bases de données) où une recherche et d'autant plus précise qu'elle combine plusieurs critères, de préférence à valeurs «certaines» (ex. Titre, auteur, éditeur, etc.).

## Dublin core (DC)

Le Dublin core résulte d'un ensemble de métadonnées communes à diverses communautés. Il s'agit d'une Initiative définie en 1995 par le NSCA (National Center for Supercomputing Applications) et l'OLC (Online Computer Library Center). Le schéma de métadonnées Dublin Core est composé d'un ensemble de 15 éléments[3] censés décrire une large variété de ressources en réseau. Chaque élément est optionnel et peut être répété, les éléments pourraient être regroupés sous trois rubriques génériques : le contenu de la ressource décrite, la propriété intellectuelle de la ressources et, les différentes versions de la ressources. Le champ sémantique des éléments a été établi par un consensus international de professionnels provenant de diverses disciplines telles que la bibliothéconomie, l'informatique, etc[4]. Le Dublin core ne décrit pas la manière selon laquelle les métadonnées doivent être représentées. Plusieurs représentations sont utilisées actuellement et d'autres sont certainement envisageables[5]. En outre, le Dublin Core n'empêche pas l'utilisation d'autres éléments nécessaires à des mises en œuvre des applications locales.

Le schéma DC a été conçu pour qu'il soit utilisé dans l'univers ouvert du Web caractérisé essentiellement par des utilisateurs qui ne sont pas forcément familiarisés avec l'usage des métadonnées et des ressources qui ressemblent, généralement peu, aux documents textuels traditionnels [Weib 98]. Ces conditions d'usage des métadonnées DC ont déterminé d'une certaine manière leurs caractéristiques. Les concepteurs ont ainsi insisté sur les aspects suivants : simplicité, extensibilité, interopérabilité sémantique, consensus international, flexibilité et procédure d'évolution continue. Par exemple, l'extensibilité est assurée par des éléments répétitifs qui peuvent être qualifiés afin de fournir des définitions plus étoffées. Chacun des aspects évoqués est justifié par les usages possibles des métadonnées.
Afin de raffiner davantage les significations des différents éléments, le Dublin Core Metadata Initiative (DCMI) a publié depuis juillet 2000 plusieurs recommandation visant à standardiser l'utilisation des qualificatifs assignés aux différents éléments des métadonnées DC. On cite deux types de qualificatifs : les qualificatifs d'élément et les qualificatifs de valeur.
Les premiers permettent de préciser le sens d'un élément pour qu'il soit plus précis. Ainsi, un élément qualifié a le même sens que l'élément non qualifié mais avec une portée plus restreinte. Par exemple, l'existence de deux ou plusieurs dates clés dans le cycle de vie de la ressource peut induire l'utilisateur en erreur. Afin d'éviter toute ambiguïté, des qualificatifs peuvent être assignés à l'élément Date. Si la ressource doit être décrite par deux dates importantes – la date de première parution et la date de publication sur le Web-, on utilise deux termes de raffinement associés à l'élément date : Created (pour première parution) et Issued (pour publication sur le Web). On a ainsi les deux expressions suivantes : DC.Date.Created et DC.Date.Issued. Une liste des qualificatifs d'éléments ou de raffinement d'éléments (Element rafinements) est définie et mise jour régulièrement par le DCMI[6]. Par exemple, accessRights est un qualificatif de l'élément Rigths proposé en février 2003 pour définir les autorisations d'accès à une ressource donnée. Dans un document publier le 19 novembre 2003 par le DCMI, on recense 32 éléments de raffinement.
Les seconds permettent de préciser la valeur attribuée à un élément de métadonnées particulier, ils peuvent par exemple préciser le mécanisme d'encodage normalisé auquel la valeur se conforme ou un vocabulaire sélectionné (ex. un système de classification ou un ensemble de vedettes matières) duquel la valeur est tirée. Par exemple, la valeur de la date suivante 2004-01-10, qui est encodée suivant la norme ISO 8601, sera lue comme suit : le 10 janvier 2004, et non le 01 octobre 2004. On recense dans

---

[3] Le Dublin Core Metadata Initiative (DCMI) propose un autre élément (audience) pour préciser le groupe de personnes à qui le document est destiné. L'audience est déterminée par le créateur, l'éditeur, ou un tiers.
[4] http://www.bibl.ulaval.ca/DublinCore/usageguide-20000716fr.htm#1.2
[5] Le DCMI a publié depuis 2001 plusieurs recommandations pour l'écritures du DC. Nous citons à titre d'exemples :

- l'écriture du DC en HTML/XML (2003/11/30)
- l'écriture du DC en RDF/XML (2002/07/31)
- l'écriture des qualificateurs DC en RDF/XML (2002/05/15)

[6] http://dublincore.org/documents/dcmi-terms/#H3

le document de DCMI cité au paragraphe précédent 17 qualificatifs de valeur (ex. MESH, DDC, DCMIType, IMT, Point, etc.) concernant huit éléments (ex. Subject, Type, Format, Source, etc.). Le qualificatif DCMIType de l'élément Type regroupe l'ensemble des types de document définit par le DC. La liste actuelle des types de document renferme 12 valeurs (ex. Collection, Service, Software et Physicalobject, etc.).
L'aventure du DC a commencé en 1995 avec la définition d'un schéma de métadonnées relativement simple composé de 15 éléments censés pouvoir décrire une large variété de ressources en réseau. Le Web de 2004 n'est plus celui de 1995 ou 1996, les ressources et les applications se sont multipliées donnant toujours naissance à d'autres besoins et, par conséquent, d'autres problèmes à résoudre. Les responsables d'«entretient» et de mise à jour du DC ne peuvent pas rester à l'écart, ils sont obligés de suivre l'évolution et d'apporter des solutions aux problèmes posés. Les nouvelles recommandations et les nouvelles propositions se multiplient, on passe d'un schéma de métadonnées simple à un schéma plus complet mais difficile à manier. On est donc forcément loin de la simplicité affichée au départ même si, on insiste sur le fait que les précisions apportées n'empêchent pas l'usage du schéma basic du DC. Les arguments avancés en faveur de l'usage possible des éléments simples des métadonnées DC ne peuvent pas continuer à convaincre si l'enjeu reste l'amélioration de l'accessibilité aux ressources en réseau. Les moteurs ou/et les agents intelligents qui sont chargés d'indexer et de retrouver les ressources dont les utilisateurs ont besoin, seront plus performants si le champ sémantique de chaque élément des métadonnées utilisées est mieux cerné.
Le DCMI a élargi ses travaux aux métadonnées des ressources pédagogiques. En août 1999 un groupe de travail (DCMI Education Working Group[7]) a été créé pour réfléchir à la manière selon laquelle le schéma DC peut être adapté à des ressources pédagogiques. L'accord passé avec le IEEE/LTSC en décembre 2001 a permis d'enrichir l'ensemble initial des éléments du DC par des éléments spécifiques aux ressources pédagogiques. Le DC Education intègre quatre autres éléments : Audience, InteractivityType, InteractivityLevel et TypicallearningTime. Des correspondances ont été établies entre les éléments du DC et ceux du Learning Object Metadata (LOM). Nous y reviendrons plus loin.

# 5 Spécificités de la recherche sur le Web

La recherche d'information sur le Web revient souvent à localiser parmi un nombre important de ressources celles qui répondent au mieux à la question posée. Cette question est généralement formulée en utilisant des critères de recherche -mot(s) ou expression(s)- reliés entre eux par des opérateurs suivant une syntaxe prédéfinie. Le moteur de recherche sollicité traite la question en appariant les termes de la requête avec ceux de l'index pour sélectionner les ressources à renvoyer à l'utilisateur. Les ressources réponse sont généralement classées par ordre de pertinence.

## 5.1 Notion de pertinence

La «fonction» de calcul de pertinence varie d'un système à un autre mais, ce qui est certain aujourd'hui c'est que cette pertinence calculée n'est pas forcément celle de l'utilisateur. Plusieurs travaux en science de l'information - Barry C.L. [Barr 94], [Bate 96], etc.- ont largement traité la notion de pertinence et plus précisément les facteurs influençant le jugement de pertinence par l'utilisateur. Pour certains d'entre eux, un document pourrait avoir deux types de pertinence : une pertinence de contenu ou/et une pertinence d'usage. On parle d'une pertinence de contenu lorsque la signification de la question trouve son équivalent dans le document retrouvé –plusieurs modes d'appariement sont à envisager-. En d'autres termes, le contenu sémantique du document s'apparie avec celui de la question. La notion de pertinence de contenu présuppose donc que la sémantique du document et celle de la question peuvent être déterminées sans ambiguïtés. Quant à la pertinence d'usage, elle s'accorde mieux avec la notion de besoins ponctuels de l'utilisateur. Ce dernier détermine la pertinence en fonction de ce que le document pourrait lui fournir comme information répondant à un besoin contextuel ou ponctuel. Un document ayant une pertinence d'usage pourrait avoir aussi une pertinence de contenu. Cette dernière n'est pas nécessaire pour que le document

[7] http://uk.dublincore.org/groups/education/

acquière une pertinence d'usage. Par exemple, un document de la base peut ne pas répondre à la question (ne vérifie pas une pertinence de contenu) mais l'utilisateur estime, toutefois, que ce même document pourrait présenter un intérêt d'usage. A l'inverse, un document vérifiant une pertinence de contenu pourrait ne pas présenter une pertinence d'usage.

Nous pouvons admettre donc que la pertinence n'est pas la propriété d'un objet ou d'un individu, mais une relation. Si, par exemple, un document est jugé pertinent par notre système ce n'est pas en tant que tel, mais relativement à un ensemble d'informations contextuelles relatives à la tâche à réaliser. Cette relation est traitée comme une fonction dont le domaine de définition n'est autre que le champ d'application d'une recherche définit par un utilisateur ou un groupe d'utilisateurs accomplissant la même tâche.

Il est donc clair que même pour un système de recherche d'information traditionnel (SRI) caractérisé par des espaces de documents, de questions et de réponses bien définies, la détermination d'une fonction de pertinence n'est pas une tâche facile à réaliser. Pour un système de recherche sur le Web, la détermination d'une telle fonction est encore plus difficile, elle nécessite davantage de connaissances de l'environnement Web.

## 5.2 Ressources et Requêtes

L'espace de documents ou de ressources ne peut en aucun cas être délimité avec précision étant donné que les ressources sont évolutives et leur nombre ne cesse d'augmenter. Aucun moteur actuel d'indexation et de recherche sur Web ne couvre toutes les ressources en réseau. En outre, on observe une grande redondance des informations présentes sur le Web. Certains voient dans la grande taille du Web et dans la redondance des informations y présentes une opportunité pour les utilisateurs du fait qu'ils peuvent trouver des ressources même si leurs requêtes sont très précises. Dans cet univers de surabondance d'information les systèmes de question-réponse pourraient représenter un alternatif aux systèmes de recherche d'information qui fournissent généralement des documents entiers dans lesquels l'utilisateur devrait localiser la réponse.
L'espace des questions est lui aussi difficile à définir du fait que l'univers du Web est conçu de telle sorte qu'il soit accessible à tout le monde, plusieurs catégories d'utilisateurs peuvent solliciter les mêmes ressources. Chaque catégorie manifeste évidemment des attentes spécifiques et, par conséquent, prévoir le genre de questions qui seront posées n'est pas évident[8].
Rechercher des ressources précises et pertinentes sur le Web suppose donc la mise en place d'une stratégie de recherche différente de celle qui peut être utilisée pour rechercher ces mêmes réponses dans une «collection» fermée de ressources. D'ailleurs, l'utilisation du terme ressource ou lieu du document reflète la complexité du problème de description et de recherche d'information sur le Web. Une ressource Web est une «entité» identifiable, généralement par une adresse unique, elle peut être un document numérique, une image, un service, une collection de ressources, etc. Certes, nous sommes tous familier avec ces différents objets de notre vie quotidienne mais quant-il s'agit de les définir pour mieux les approprier, nous ne formulons pas forcément la même définition. Le terme ressource s'impose donc, en l'absence de définitions précises, comme un «consensus».
Prenons l'exemple du document numérique, il peut être différemment défini d'une communauté à une autre. Les traits mis en évidence ne sont pas les mêmes puisqu'ils reflètent généralement l'usage qu'on fait du document. Un groupe de travail au sein du réseau thématique pluridisciplinaire du département STIC du CNRS s'est intéressé à la notion du document dans son passage au numérique. A partir de recherches privilégiant plutôt la forme (comme un objet matériel ou immatériel), le signe (comme un porteur de sens) ou le médium (comme un vecteur de communication), le groupe propose, dans un

[8] Le TREC (Text REtrieval Conference) fournit des collections Web permettant d'évaluer les techniques de recherches sur le réseau. Les collections comprennent plusieurs Giga de données issues de sites Web, d'un ensemble de requêtes et les référentiels de réponses justes (documents pertinents / non pertinents) établies par des experts du domaine. Les documents non jugés par les experts sont considérés non pertinents.
http://es.csiro.au/TRECWeb/

texte disponible en ligne[9], trois définitions possibles du document numérique. a) «*Un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs.» b) «Un document numérique est un texte dont les éléments sont potentiellement analysable par un système de connaissance en vue de son exploitation par un lecteur compétent.» c) «Un document numérique est la trace de relations sociales reconstruite par les dispositifs informatique »*
Les définitions ne s'excluent pas l'une l'autre étant donné qu'elles résultent de trois entrées (la forme, le signe et le médium) qui ne sont pas indépendantes. Chaque entrée est vue comme une dominante et non comme une dimension exclusive. Les auteurs ont cité les disciplines et les métiers qui seraient concernés par chaque entrée tout en précisant que les chercheurs qui abordent le document suivant une entrée donnée, ne négligent pas forcément les deux autres, néanmoins leur analyse et raisonnement privilégient l'entrée choisie, les deux autres restent des compléments ou des contraintes extérieures.
Ce genre de travail apporte certainement des précisions à la notion de document numérique, des précisions qui deviennent de plus en plus indispensables pour pouvoir définir les traits caractérisant les ressources en réseau. Nous ne pouvons pas continuer à gérer des objets dont nous ignorons leurs usages «techniques» et sociales. La performance du système de description et de recherche dépendra, en large partie, des connaissances que nous avons de son environnement d'utilisation.

## 5.3 Usage des métadonnées

Un schéma de métadonnées ne peut être défini indépendamment de ses futurs usagers, même s'ils ne constituent pas un groupe socioprofessionnel homogène, il doit prendre en considération leurs attentes et leurs besoins. Ainsi, les métadonnées pourront inclure les caractéristiques de la ressource mais aussi ceux des usagers. Dans les systèmes de recherche d'information, le profil (les caractéristiques) de l'utilisateur est habituellement utilisé pour évaluer les réponses du système. Par exemple, le système détermine, en fonction du niveau d'éducation de l'utilisateur, si la réponse est appropriée ou non. Dans notre cas, il ne sera pas question d'évaluer la réponse à une question mais de traiter, par un processus de filtrage, les attentes de l'utilisateur. Ainsi, la ressource et l'utilisateur (lecteur, acteur social, etc.) doivent être traités suivant une approche globale qui mettra en évidence les interactions possibles. En outre, les schémas de description des ressources peuvent être spécialisés pour des tâches spécifiques (ex. La construction ou l'utilisation de cours en ligne) ou des domaines d'application particuliers. Le système ou le dispositif destiné à exploiter ces métadonnées doit aussi intégrer dans son mode de fonctionnement des «schémas» de traitement proches de ceux des utilisateurs.

Le champ «sémantique» couvert par un réseau de ressources n'est pas facile à mettre en évidence et les liens non déclarés et qui font parti désormais du domaine de l'inférence ne facilitent certainement pas cette navigation supposée être à la porté de tout le monde. A ce niveau, il est peut être intéressant de savoir si l'usage des métadonnées peut apporter ou non une solution aux problèmes résultant de l'«hypertextualité» non contrôlée. Ajouter des métadonnées à des ressources consiste à leur associer une structure qui pourrait être utilisée dans la définition du champ sémantique couvert par un réseau. Il est évident que cette structure ne peut être exploiter que si une sémantique est associée aux domaines de valeur des différents éléments du schéma de description. La sémantique dont nous parlons ici doit être interprétable non seulement par les utilisateurs mais aussi par la machine. Ceci nous amène à parler de la gestion des métadonnées d'un point de vue inférentiel.

Si le DC ne dit rien sur ce qu'il est possible de faire de ses différents éléments, avec une ontologie écrite en OW, nous pouvons préciser par exemple que deux instances d'un élément sont synonymes. Les ontologies utilisées dans le contexte du Web sémantique permettent une modélisation sémantique de connaissances d'un domaine et peuvent, par conséquent, être utilisées pour compléter les informations apporter par un schéma de métadonnées. Ceci facilite évidemment l'intégration de multiples sources d'information, la formulation de requête très précise grâce à un vocabulaire riche et, une description plus fine du contenu des documents recherchés.

[9] *http://archivesic.ccsd.cnrs.fr/sic_00000511.html*

# 6 Conclusion

A partir de deux langages du Web sémantique (RDF et OWL) et des schémas prédéfinis de métadonnées, nous avons essayé de cerner davantage l'apport des métadonnées en matière de recherche d'information sur le Web. Le présent travail nous a permis de découvrir les différentes facettes du document Web auquel des schémas de métadonnées pourraient être rattachés. Nous avons mentionné que la notion de document numérique nécessite davantage de précisions pour pouvoir définir les traits caractérisant les ressources en réseau. La performance du système de description et de recherche dépendra donc, en large partie, des connaissances que nous avons de l'objet à gérer. Si l'utilisation des métadonnées est une pratique généralisée et justifiée pour la description des documents «traditionnels», le transfert de cette pratique vers les ressources évolutives et distribuées du Web n'est pas évident. Même s'il y a des traits fonctionnels communs, nous ne pouvons pas comparer les schémas de métadonnées existants aux métadonnées structurées (catalogues) des bibliothèques qui ont été conçues dans un contexte d'utilisation différent.

En exploitant des ressources électroniques enrichies par des métadonnées, organisées sous formes d'ontologies permettant des inférences sur les données disponibles, **la veille** stratégique, économique, concurrentielle, technologique, … produira certainement des produits (ex. revue de presse, bulletin d'information, annuaire, données statistiques, etc.) plus précis et plus exhaustifs. En outre, la technologie du Web sémantique permettra à l'entreprise : d'exploiter des ontologies de son domaine d'activités, d'annoter ses ressources électroniques pour une meilleure optimisation de sa «capitale connaissances», de vérifier la validité des sources d'information, de définir des profils d'utilisateur pour des applications données, etc… En outre, l'ontologie a une place centrale dans un projet de mémoire d'entreprise.

Notre étude nous a permis de formuler des questions autours de la problématique générale de la recherche d'information sur le Web. Avons-nous besoin d'un schéma «générique» adapté tel que le DC ou d'un schéma spécifique pour gérer différents types de ressources sur le Web ? Il y a certainement plusieurs pistes de recherche à explorer pour pouvoir apporter les améliorations et les solutions attendues. Nous citons ici à titre d'exemple, l'utilisation des critères d'usage pour la sélection des ressources en réseau, il s'agit de décrire les ressources Web par des traits qui n'ont pas forcément de lien direct avec le contenu. Le travail sur la caractérisation de l'utilisateur doit aussi aboutir à des modèles exploitables.

# Bibliographie


BARRY C.L., User-defined relevance criteria : An exploratory study. Journal of the American Society for Information Science. 1994, 45(3) : 149-159.

BATE M.J., Document familiarity, relevance and Bradford's law. Information Processing and Management. 1996, 32(6) : 697-707.

Tim Berners-Lee : Frequently asked questions
http://www.w3.org/People/Berners-Lee/FAQ.html#What2

BERNARD B., La normalisation des TIC pour l'apprentissage : enjeux, controverses et état d'avancement. Compte rendu du Petit déjeuner du FFFOD- Procope le 23 juin 2004.

CHARLET J., LAUBLET P. et REYNAUD C., Le Web sémantique (rapport final de l'action spécifique 32 CNRS/STIC), Octobre 2003.

DCMI Metadata Terms, Novembre 2003
**http://dublincore.org/documents/dcmi-terms/#H3**
Diane Hillmann. Guide d'utilisation du Dublin Core, Janvier 2001 http://www.bibl.ulaval.ca/DublinCore/usageguide-20000716fr.htm#1.2

GIRALDO G. et REYNAUD C., Construction semi-automatique d'ontologies à partir de DTDs relatifs à un même domaine. *13èmes journées francophones d'Ingénierie des Connaissances*. Rouen, 2002.

GUARINO N., *The ontological level*, in R. CASATI B. S.&WHITE G., eds., *Philosophy and thecognitive sciences*, Hölder-Pichler-Tempsky, 1994.

LAGOZE C., From *Static to Dynamic Surrogates. Resource Discovery in the Digital Age*. In: D-Lib Magazine, June 1997. http://www.dlib.org/dlib/june97/06lagoze.html.

NILSON M. et al., Semantic *Web Metadata for e-Learning – Some Architectural Guidelines*. 11th World Wide Web Conference (WWW2002), 2002 Hawaii, USA.

VIDAL P., BROISIN J., DUVAL E. et TERNIE S., Normalisation et Standardisation des Objets d'Apprentissage : l'Expérience ARIADNE. IN :
http://e-miage.ups-tlse.fr/colloque/papiers/P-Vidal-ObjetApprentissage.pdf

WEIBEL S. et HAKELA J., DC-5 : The Helsinki Metadata Workshop: A Report on the Workshop and Subsequent Developments. Official report of the Helsinki DC Meeting. In :
D-Lib Magazine, February 1998.
http://www.dlib.org/dlib/february98/02weibel.html

Document : forme, signe et médium, les re-formulations du numérique, Juillet 2003
**http://archivesic.ccsd.cnrs.fr/sic_00000511.html**

**Web Research Collections - TREC Web Track, 2003**
**http://es.csiro.au/TRECWeb/**

W3C. OWL Web Ontology Langage Guide.
http://www.w3.org/TR/2004/REC-owl-guide-20040210/ (10 février 2004)

W3C. OWL Web Ontology Overview.
http://www.w3.org/TR/2004/REC-owl-features-20040210/.(10 février 2004)

W3C. OWL Web Ontology Language Semantics and Abstract Syntax. http://www.w3.org/TR/2004/REC-owl-semantics-20040210/ (10 février 2004)

W3C Resource Description Framework (RDF) Schema Specification
http://www.w3.org/TR/1999/PR-rdf-schema-19990303 (4 mars 1999)

W3C Resource Description Framework (RDF) Model and Syntax Specification
http://www.w3.org/TR/1999/REC-rdf-syntax-19990222 (22 février 1999)
http://www.w3.org/TR/2004/REC-rdf-primer-20040210/ (10 février 2004)