Panorama des travaux en cours dans le domaine des métadonnées

François Role

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Panorama des travaux en cours dans le domaine des métadonnées

François Role

N° 3628

février 1999

THÈME 3

ISSN 0249-6399

Panorama des travaux en cours dans le domaine des métadonnées

François Role *

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet ATOLL

Rapport de recherche n ° 3628 — février 1999 — 40 pages

Résumé : L'augmentation spectaculaire des volumes d'information disponibles sur Internet entraîne un intérêt grandissant pour les "métadonnées", c'est-à-dire des données décrivant d'autres données pour rendre ces dernières plus accessibles et plus faciles à utiliser.

Ce rapport rappelle tout d'abord quelques éléments importants concernant cette notion. Il décrit ensuite une sélection de projets et de services d'information en réseau mettant en oeuvre des métadonnées. Des pistes pour de futures recherches sont suggérées en conclusion.

Mots-clé : Métadonnées, bibliothèques numériques, Dublin Core, PICS, RDF

(Abstract: pto)

* François Role@inria.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11
Télécopie : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Survey of Current Research on Metadata

Abstract: With the explosion of electronic information available on the Web and in Digital Libraries there is an increasing interest in metadata, i. e. data that describe other data in order to make it more accessible or easier to use.

This report first gives some background on this notion in the form of a review of relevant literature. It then describes a selection of metadata-based networked information services or projects.

In conclusion future directions for metadata research are suggested.

Key-words: Metadata, Digital Libraries, Dublin Core, PICS, RDF

Table des matières

1 Introduction

L'expansion rapide des réseaux et le volume croissant des données hétérogènes accessibles entraînent un regain d'intérêt pour la notion de "métadonnées" c'est-à-dire de données associées à d'autres données pour en faciliter l'exploitation ou l'interprétation dans le cadre d'une relation utilisable par un programme informatique.

Cette notion recouvre certaines réalités connues depuis longtemps en informatique documentaire (les index, les annotations, les étiquettes PICS, les schémas de bases de données, etc.), mais c'est seulement depuis quelques années, avec le développement du multimédia et des réseaux qu'elle émerge comme un objet d'étude à part entière. Cette évolution est par exemple très perceptible à travers les nombreux travaux qui se développent au sein de la communauté Internet autour de la question de la description des ressources du Web (PICS, RDF, etc.).

L'intérêt accru pour les métadonnées est également à mettre en relation avec le développement des recherches sur le thème des bibliothèques numériques.

Depuis quelques années, ce thème commence en effet à se structurer comme un domaine de recherche autonome, ainsi qu'en témoignent les grands projets du programme fédéral américain NII (*National Information Infrastructure*),[1] les conférences organisées par l 'ACM [47] [48], l'IEE [28] , l'ERCIM, l'apparition de publications spécialisées [30] [27] [1] [17] [2].

Un des objectifs principaux des recherches dans le domaine des bibliothèques numériques est de concevoir des systèmes et des architectures permettant à l'utilisateur d'accéder de façon cohérente et uniforme à des bibliothèques de ressources hétérogènes et géographiquement distribuées. Il est clair que de tels systèmes doivent s'appuyer sur des ensembles de métadonnées complexes décrivant ces ressources et en permettant la manipulation

Notons qu'on rejoint là les préoccupations d'une partie de la communauté des bases de données, notamment pour traiter les problèmes d'interopérabilité entre bases de données hétérogènes.

Au terme de ce qui vient d'être dit, il apparaît clairement que le thème des métadonnées fait l'objet de travaux menés en parallèle dans différentes communautés (bibliothèques numériques[2], bases de données multimédia, Web, etc.) pour traiter des problèmes souvent similaires. Mais ces communautés sont issues de cultures différentes, et l'examen de leurs publications montre qu'elles communiquent assez peu entre elles d'où l'idée d'effectuer une synthèse permettant de confronter les différentes approches, d'identifier les recoupements et de mettre en lumière des questions insuffisamment abordées.

Compte tenu de l'ampleur du sujet, nous ne pouvons prétendre donner ici une présentation exhaustive des projets ou des propositions concernant les métadonnées. Nous avons donc sélectionné une palette d'exemples que nous jugeons représentatifs dans la mesure où ils

1. Ces projets en cours depuis trois ans sur plusieurs grands campus nord-américains ont pour objectif la conception et le développement de systèmes informatiques capables de gérer de très vastes volumes d'informations hétérogènes.

2. Et au sein même de la communauté "Bibliothèques numériques" [10] fait remarquer qu'on peut distinguer des équipes influencées par la culture du catalogage bibliographique traditionnel et des équipes qui regroupent plutôt des acteurs d'Internet qui ne sont pas des professionnels du catalogage.

permettent d'introduire les principaux termes et concepts utilisés dans la littérature explicitement consacrée aux métadonnées.

Cette sélection d'exemples a été aussi l'occasion de proposer une catégorisation en terme des types de documents auxquels on applique les métadonnées et des traitements que ces métadonnées sont censées supporter. Ce niveau de présentation intermédiaire n'apparaît pas dans la littérature consacrée aux métadonnées. Le plus souvent, il est question des métadonnées soit d'une façon très générale (*data about data*) soit à propos d'applications et de types de données très précis.

Nous présentons des projets où il est fait "explicitement" mention des métadonnées. Dans la mesure du possible, nous donnons des exemples d'implémentation. Nous avons essayé de donner une description aussi précise que possible, mais il faut bien garder en tête un certain nombre de contraintes. En fonction de leur thème principal, les projets que nous avons choisis de présenter illustrent plus ou moins le cycle de vie des métadonnées (création, gestion, mise à jour, suppression). Par exemple, un projet visant à intégrer des métadonnées hétérogènes se concentre sur les mécanismes d'intégration et considère les métadonnées qui font l'objet de l'intégration comme existantes, sans se poser forcément la question de leur création.

Par ailleurs, certains des travaux présentés ici sont des propositions très récentes qui ont donné lieu à des niveaux d'implémentation plus ou moins avancés. Malgré leur caractère plus théorique, ces travaux illustrent des idées ou soulèvent des questions qui nous paraissent importantes.

Comme nous l'avons déjà souligné, les travaux concernant les métadonnées mettent en oeuvre des techniques et des aspects très variés (bases de données, réseau, édition électronique, représentation des connaissances, etc.) Ce caractère encyclopédique est sans doute la raison pour laquelle on ne trouve semble-t-il pas ou peu de synthèse dans la littérature. D'un autre côté, c'est son caractère global qui fait à priori l'intérêt de cette notion.

2 La notion de métadonnées dans la littérature

Le terme anglo-saxon *metadata* est parfois traduit par le mot "métadonnées", notamment dans la littérature francophone consacrée aux bases de données.

Dans la littérature anglo-saxonne abondante sur le sujet, les métadonnées sont souvent définies de façon expéditive comme des "data about data". Par exemple, les organisateurs d'une conférence IEE tenue récemment sur ce thème définissent ainsi le sujet de la conférence :

> The IEEE Metadata Conference goal is to encourage discussion of metadata issues. Metadata is loosely defined as "data about data", or "additional information that is necessary for data to be useful".

Certains auteurs essaient de préciser cette notion en donnant une liste de ses usages possibles mais ces listes couvrent un champ si vaste que la définition reste imprécise :

> Metadata can be used in a variety of application areas; for example: in resource discovery to provide better search engine capabilities, in cataloging for describing

> the content a vailable at a particular web site or page, by intelligent software agents to facilitate knowledge sharing and exchange, in digital signatures, in content rating, and in many others (for example, metadata can be used for specialized tasks such as organizing a group of web pages for purposes of printing them as a single unit, or for producing a visualization of the link relationships between them). [35]

> In its most abstract form, metadata is "information about information." Information on the Web, known as Web resources, have many pieces of associated descriptive information which is often not explicitly represented in the resource itself. Examples of metadata include the creator of a resource, its subject, length, publisher, creation date, etc. Such descriptive metadata can be used to make information easier to locate by improving Web searches [Weibel, 1995], rate information to protect children from indecent content (e.g. the Platform for Internet Content Selection (PICS) [Miller et al., 1996]), capture copyright information, contain a digital signature, or store cataloging data. Many other uses are also possible. [60]

Dans des cas plus rares, on rencontre à l'inverse des articles où des auteurs issus du monde des bibliothèques restreignent la portée des métadonnées et ont tendance à assimiler ces dernières à de simples données catalographiques [59].

> Metadata is data about data. Its most familiar form is the descriptive catalogue record used to describe printed books defined by AACR2 (the Anglo American Cataloging Rules) and expressed in MARC format.

Les métadonnées descriptives, nous revenons plus loin sur ce point, ne sont cependant qu'un type particulier de métadonnées, comme cela est clairement exprimé dans les documents diffusés par la *Digital Library Federation*:[3]

> As DLF participants developed the plan for Making of America, Part 2; they realized that such seamless links require the creation and management of a complex set of "metadata." They distinguished descriptive metadata, such as the information encoded in bibliographic records and detailed finding aids, which serve to identify the contents of the special collections, from the structural and administrative information needed to organize and manage the collections in digital form. Structural metadata include information about page sequencing or other divisions that enables a reader to navigate a work effectively in a digital environment. Administrative metadata include information about the manner of creation, the provenance and ownership of a work that enables a digital library effectively to manage the rights in the intellectual property of a work.

3. La *Digital Library Federation* est un regroupement de bibliothèques américaines ayant pour objectif le développement d'un réseau national de bibliothèques électroniques.

> These distinctions are incorporated in the NEH proposal, but Berkeley and its colleagues in Making of America, Part 2, are already at work in an early phase of the project, which CLIR and DLF have funded to stimulate the development of practices for encoding structural and administrative metadata. [21]

Pour être complet, il faut signaler la classification proposée dans [31] et qui est souvent reprise dans la littérature consacrée aux bases de données multimédia. Cette classification distingue trois types de métadonnées :

- les métadonnées dépendantes du contenu (*content-dependent metadata*)
- les métadonnées descriptives du contenu (*content-descriptive metadata*)
- les métadonnées indépendantes du contenu (*content-independent metadata*)

Selon cette classification, les métadonnées dépendantes du contenu sont des métadonnées qui peuvent être dérivées directement du contenu. Dans le domaine des documents textuels, il peut par exemple s'agir d'index en texte intégral comme dans WAIS ou de vecteurs. Les métadonnées qui sont basées sur le contenu des données mais qui ne peuvent être dérivées directement sont considérées comme des métadonnées indépendantes du contenu.

Par exemple, la date de dernière modification d'un fichier est une métadonnée indépendante du contenu de ce fichier.

Même si cette classification correspond jusqu'à un certain point à une intuition et recoupe par certains côtés des distinctions existantes en informatique documentaire, la frontière à partir de laquelle on sort du domaine des métadonnées dérivées "directement" du contenu pour entrer dans celui des métadonnées obtenues "moins directement" semble trop floue pour rendre cette distinction utile. Tout semble en effet dépendre du niveau de raffinement des techniques d'extraction mises en oeuvre.

Au total, même si elle est utilisée dans un grand nombre d'articles, il semble ne pas exister de définition vraiment formelle et précise de cette notion, ce qui est sans doute dû à sa grande généralité.

L'idée générale qui se dégage est celle de données relatives à d'autres données et destinées à supporter des traitements impliquant ces autres données.

A la section suivante, nous proposons ci-dessous quelques critères permettant de caractériser les métadonnées en termes des types de données et des traitements dans lesquels elles sont mises en oeuvre.[4]

4. Etymologie : le mot meta en grec a un sens difficile à saisir mais il désigne généralement une transition, le passage à un autre état ou à un autre niveau, sens qui reste perceptible dans des mots français comme "métamorphose" ou "métalangage".

3 Démarche suivie et limite de cette étude

3.1 Catégorisation par types de données et types de traitements concernés

Il semble utile de catégoriser les métadonnées en termes des données auxquelles elles sont associées et des traitements auxquelles elles servent de support, les deux critères n'étant d'ailleurs pas indépendants, certains traitements ne pouvant s'appliquer qu'à certains types de données.

Même si cela peut paraître évident, l'examen des travaux en cours montre qu'il est utile d'insister sur le fait que des métadonnées peuvent être associées à des données de taille et de type arbitraires (pensons par exemple à une annotation associée à une zone de quelques pixels dans une image médicale).

Ce rappel est utile, car certaine communautés travaillant activement sur les métadonnées, par exemple celles qui s'attachent à définir des métadonnées catalographiques (au sens des catalogues de bibliothèques), ont parfois tendance à considérer que les métadonnées sont forcément attachées à un "document électronique" comme une fiche de catalogue est attachée à un livre imprimé classique.

Dans la suite, nous appelons "données" une collection quelconque de bits, ces derniers n'ayant même pas à être stockés de façon contiguë sur un support de masse ou en mémoire. Le seul critère est que cette collection puisse être désignée et accessible via un mécanisme informatique.

La catégorisation par type de données associées aux métadonnées est insuffisante. Pour une même donnée (ce terme étant pris au sens où nous l'avons défini ci-dessus) les métadonnées peuvent servir de support à un vaste éventail de traitements.

Dans certaines communautés intéressées par le thème des métadonnées, comme celle des "entrepôts de données" (*Data Warehouse*), des auteurs essaient de dresser une liste des opérations auxquelles les métadonnées peuvent servir de support dans leur domaine. Par exemple dans [22] les métadonnées sont définies comme des informations sur les données devant notamment permettre de savoir :

- ce que signifie une donnée ;
- d'où elle provient ;
- comment on la calcule ;
- quel est son domaine de valeurs ;
- où elle est stockée et dans quel format.

Ce type de liste peut fonctionner pour un domaine limité, mais il serait vain de vouloir donner une liste exhaustive de ces traitements.

En utilisant les deux critères présentés ci-dessus, nous avons essayé d'identifier les principaux secteurs où la notion de métadonnées est utilisée, et nous présentons dans les sections suivantes quelques projets que nous avons jugés représentatifs de ces secteurs.

3.2 Limites de cette étude

Compte tenu de l'ampleur du sujet, nous ne pouvons prétendre donner ici une présentation exhaustive des projets mettant de près ou de loin en oeuvre des métadonnées.

Nous avons cependant essayé d'illustrer les grandes tendances, tant en ce qui concerne les communautés ayant recours à cette notion de métadonnées que les utilisations qui en sont faites dans des projets.

Dans le tableau ci-dessous, nous donnons en colonne de gauche un récapitulatif très schématique des principales communautés mettant explicitement en oeuvre la notion de métadonnées. En colonne de droite, nous indiquons les principales utilisations qui sont faites des métadonnées par ces communautés.

Les cases de la colonne de gauche ne sont pas des domaines disjoints (pour ne citer qu'un exemple, des chercheurs en bases de données interviennent souvent dans des projets d'édition électronique, dans des applications à base d'objets distribués, etc.) Il n'en reste pas moins que les communautés scientifiques que nous avons isolées ont chacune leurs spécificités et utilisent une terminologie et des concepts qui leur sont propres.

La communauté des bibliothèques numériques, dont nous avons souligné l'émergence en introduction, peut d'ailleurs être vue comme une tentative récente pour fédérer ces différents secteurs de la recherche en informatique. Elle ne figure pas explicitement mais elle est potentiellement intéressée par toutes les applications énumérées en colonne de droite.

Il est par ailleurs clair qu'un nombre considérable d'autres communautés utilisent plus ou moins implicitement des métadonnées pour traiter bien d'autres problèmes. Mais les domaines d'application présentés ci-dessous sont ceux où la notion de métadonnées est mise en avant comme un élément central, et c'est la validité de ce choix que nous essayons d'évaluer dans ce document.

Communautés utilisant des métadonnées

Communauté	Utilisation
Acteurs du Web	– utilisation de métadonnées par des robots de recherche pour créer des index permettant d'effectuer des recherches documentaires précises (Dublin Core) – utilisation de métadonnées par des programmes de filtrage de l'information pour bloquer l'accès à certains sites (PICS) – utilisation de métadonnées pour authentifier un document (DSIG) – utilisation de métadonnées pour gérer les droits

Communautés utilisant des métadonnées

Communauté	Utilisation
Bases de données	– utilisation de métadonnées pour représenter des correspondances entre attributs et contribuer à l'interopérabilité entre bases hétérogènes – utilisation de métadonnées pour générer dynamiquement des requêtes SQL à destination de bases dont le schéma est à priori inconnu – utilisation de métadonnées pour identifier les sources intéressantes à consulter (resource discovery) – associer à une base de données multimédia des métadonnées permettant d'effectuer une corrélation entre des données de types différents – associer à des bases de données accessibles en mode client/serveur des métadonnées permettant à des clients ne connaissant pas a priori les caractéristiques de ces bases de les interroger (Z39.50/Explain)
Edition électronique	– utilisation de métadonnées pour structurer de vastes corpus de textes (par exemple les entêtes de la *Text Encoding Initiative*)
Informatique "décisionnelle" et informatique de gestion	– utilisation de métadonnées de gestion pour gérer des entrepôts de données (*Data Warehouse*)

Communautés utilisant des métadonnées

Communauté	Utilisation
Objets distribués	– utilisation de métadonnées fournissant des informations sur les méthodes de façon à permettre une invocation dynamique de ces dernières (CORBA)

On notera l'absence des communautés travaillant sur les problèmes de représentation des connaissances et les agents intelligents. Les travaux dans ces domaines abordent pourtant souvent des questions similaires et partagent un certain nombre de concepts avec la communauté "métadonnées", comme par exemple la notion d'ontologie. Cette faible présence est un point important sur lequel nous revenons à la fin de ce rapport.

De même, la notion de métadonnées, si elle est latente dans les travaux consacrés aux systèmes d'annotation (à part bien sûr certains travaux liés au Web et à PICS et sur lequel nous reviendrons) est rarement invoquée explicitement.

Dans la suite de ce document, nous présentons des projets liés aux ressources du Web (sections 4, 5, 6), et aux bases de données documentaires (section 8).

Ces travaux illustrent de façon inégale les différentes étapes du cycle de vie des métadonnées (création, gestion et mise à jour, suppression) mais au total, ils nous paraissent constituer un échantillon représentatif des travaux actuels dans le domaine des métadonnées.

4 Le Dublin Core et le Warwick Framework

La question de la description des ressources sur le Web est d'actualité (précisons ici que par "ressource du Web" nous désignons ici toute ressource accessible via le protocole HTTP).

Au cours des dernières années on a assisté au développement de nombreuses initiatives dans ce domaine, en provenance soit du consortium W3 (PICS, PICS-NG, XML-DATA, MCF, etc.) soit de communautés particulières (GILS, Dublin Core, Warwick Framework).

Nous présentons ci-dessous deux initiatives issues principalement de la communauté bibliographique et connues sous le nom de Dublin Core et de Warwick Framework.

4.1 Le Dublin Core

Des volumes d'information de plus en plus importants sont accessibles au travers des réseaux. Le développement rapide du World Wide Web est la meilleure illustration de ce phénomène.

Pour retrouver un document parmi les millions qui sont accessibles sur cette infrastructure, l'utilisateur a recours a des moteurs de recherche optimisés (Lycos, Altavista, etc.)

capables d'effectuer des recherches très rapides au sein de volumes très importants. Ces moteurs constituent leurs index en pratiquant une indexation en texte intégral des documents Web, procédure qui peut générer un bruit non négligeable lors des interrogations.

Pour améliorer les performances de ces moteurs de recherche plusieurs propositions ont été développées au sein de la communauté Internet, visant à associer des données descriptives aux documents du Web.

Le "Dublin Core" est une initiative de ce type, qui émane de sociétés ou d'institutions spécialisées dans le traitement de l'information bibliographique.

4.1.1 Motivations et concepts

L'historique et les motivations du Dublin Core sont résumés dans [57]. Il s'agit de fournir aux robots de recherche (Lycos, Alta Vista, etc.) des informations leur permettant de construire des index précis par auteur, titre, sujet, etc., ces index devant permettre des recherches par champ.

Dans la tradition des catalogues de bibliothèque informatisés, on ne cherche pas sur le document lui-même, mais sur un substitut constitué d'un ensemble de couples attributs valeurs (ce que les anglo-saxons appellent parfois un "fielded surrogate"). Cette approche correspond exactement au modèle de l'enregistrement de bases de données documentaires tel qu'il est défini par la norme Z39.50 (voir la section 8 de ce rapport).

Un ensemble de métadonnées ayant cette finalité et connu sous le nom de "Dublin Core metadata set" a été élaboré au cours d'une série de conférences internationales dont la première a été organisée par OCLC[5] en 1995 à Dublin, dans l'OHIO d'où le nom de "Dublin Core" [12] [56].

Cet ensemble de métadonnées est destiné à faciliter l'accès aux documents disponibles sur Internet. Les promoteurs du Dublin Core sont partis d'un double constat :

- Le nombre de ressources Internet disponibles augmente tous les jours, ce qui rend de plus en plus difficile la recherche d'informations précises.

- Les techniques de description actuellement mises en oeuvre sur Internet ne permettent pas de répondre efficacement à cette augmentation du nombre de documents potentiellement accessibles.

Le point de vue du Dublin Core est en effet que les index créés automatiquement, directement à partir du contenu des documents, par les services comme Lycos, WebCrawler, etc. ne sont pas assez riches pour permettre des recherches vraiment systématiques.

A l'inverse, les entreprises de catalogage manuel des ressources Internet entreprises dans certains grands organismes documentaires permettent d'associer des notices descriptives très riches mais le coût de création de ces notices est prohibitif compte tenu du nombre de documents à traiter sur Internet.

5. OCLC (*Online Computer Library Center*) est une société américaine spécialisée notamment dans le traitement informatique des notices bibliographiques.

La solution que propose le Dublin Core pour résoudre cette contradiction est d'associer aux documents du Web un ensemble d'éléments descriptifs minimum (correspondant pour l'essentiel aux éléments de données que l'on retrouve dans des notices bibliographiques simplifiées de catalogues de bibliothèques), suffisant pour enrichir les index du Web et en même temps plus simple à créer que les notices bibliographiques complètes. Cette norme de description comprend les quinze eléments figurant dans la table 2 (une traduction française de ces éléments a été donnée récemment par [54]). Dans les spécifications du Dublin Core les ressources du Web auxquelles sont associées des métadonnées sont désignées sous le nom de DLO (*Documents Like Objects*). Il s'agit, ainsi que leur nom le suggère, de documents électroniques proches du modèle de l'imprimé comme des pages HTML, des documents Postcript, etc. Nous décrivons plus loin les problèmes que pose cette définition restrictive des ressources du Web.

4.1.2 Exemple d'implémentation

En termes opérationnels, l'idée initiale des promoteurs du Dublin Core était d'inciter les auteurs de documents destinés à être placés sur Internet à créer eux mêmes les descriptions de leurs documents en utilisant les métadonnées du Dublin Core (par exemple en utilisant un éditeur HTML qui permet d'insérer facilement cette description).

Dans cette approche, les éléments du Dublin Core sont représentés à l'aide de l'élément `<META>` prévu dans l'en-tête (`<HEAD>`) des documents HTML.

Voici, à titre d'exemple l'en-tête de la page d'accueil du site d'UKOLN, un des premiers sites à avoir ajouté des éléments Dublin Core à ses pages[6] :

```
<HEAD>
<TITLE>UKOLN: UK Office for Library and Information Networking</TITLE>
<META NAME="DC.title" CONTENT="UKOLN: UK Office for Library and Information
Networking">
<META NAME="DC.subject" CONTENT="national centre, network information
support, library community, awareness, research, information services,
public library networking, bibliographic management,
distributed library systems, metadata, resource discovery, conferences,
lectures, workshops">
<META NAME="DC.description" CONTENT="UKOLN is a national centre for
support in network information management in the library and information
communities. It provides awareness, research and information services">
<META NAME="DC.creator" CONTENT="UKOLN Information Services Group">
<META NAME="DC.creator.email" CONTENT="isg@ukoln.ac.uk">
<META NAME="keywords" CONTENT="national centre, network information
support, library community, awareness, research, information services,
public library networking, bibliographic management,
```

6. Le centre de recherche d'UKOLN coordonne les projets anglais de bibliothèque électronique relevant du programme national de recherche eLib

	Nom	Signification
1	TITLE	Nom donné à la ressource par l'entité CREATOR ou PUBLISHER
2	AUTHOR	Personne responsable du contenu intellectuel du DLO
3	SUBJECT	Description du sujet ou du thème dont traite la ressource
4	DESCRIPTION	Description en texte libre du contenu de la ressource
5	PUBLISHER	Personne ou institution en charge de la diffusion du DLO
6	CONTRIBUTORS	Personne ayant apporté une contribution intellectuelle à la ressource (en plus de l'entité désignée par AUTHOR)
7	DATE	Date de publication
8	TYPE	Le genre (au sens littéraire) auquel se rattache le DLO
9	FORMAT	Format du DLO
10	IDENTIFIER	Chaîne ou nombre utilisé pour identifier le DLO
11	SOURCE	Documents (sous forme imprimée ou électronique) dont ce DLO est dérivé
12	LANGUAGE	Langue dans laquelle est exprimé le contenu intellectuel de la ressource
13	RELATION	Relations avec les autres ressources
14	COVERAGE	Emplacement physique et les caractéristiques de durée de l'objet
15	RIGHTS	Gestion des droits

TAB. 2 – *Les éléments du Dublin Core*

```
distributed library systems, metadata, resource discovery, conferences,
workshops">
<META NAME="description" CONTENT="UKOLN is a national centre for
support in network information management in the library and information
communities. It provides awareness, research and information services">
</HEAD>
```

Dans cet exemple, les noms d'éléments sont préfixés par la chaîne "DC" pour indiquer que les éléments correspondent au Dublin Core.

La page ci-dessus présente aussi des exemples d'utilisation de "qualificateurs d'éléments", caractéristique que nous décrivons à la section 4.1.4.

Nous décrivons ci-dessous un exemple d'implémentation plus sophistiqué qui est actuellement en test sur le site Web du centre d'UKOLN [53].

La procédure est la suivante. Après avoir produit sa page, l'auteur crée un fichier de métadonnées correspondant à cette page et le sauvegarde au format SOIF (Summary Object Interchange Format, le format utilisé par le système Harvest [25]) dans le même répertoire que la page.

Ce fichier a pour nom le nom du fichier HTML auquel il s'applique, auquel on ajoute le suffixe .soif. Par exemple le fichier welcome.html a un fichier de métadonnées associé nommé welcome.html.soif.

L'association entre la page et les métadonnées s'effectue dynamiquement via un script SSI (SSI pour "Server Side Include" est un mécanisme simple pour créer tout ou partie d'une page Web dynamiquement). Chaque page HTML contient en effet un appel de script comme dans l'exemple ci-dessous :

```
<head>
<title>Un exemple de script SSI</title>
<!--#exec cmd="getmeta" -->
</head>
```

Lorsqu'un robot se connecte au site, le serveur, au lieu de se contenter de lire le fichier sur le disque et de le renvoyer, analyse le fichier à la recherche d'un script SSI.

S'il en trouve un, il l'invoque en lui passant le nom du fichier HTML qu'il est en train de lire. Le script ajoute le suffixe `.soif` à ce nom et cherche un fichier de métadonnées correspondant au nom ainsi créé.

S'il en trouve un, il le convertit en éléments `<META>` qu'il renvoie au serveur. Ce dernier insère alors les éléments `<META>` dans la page qu'il renvoie au robot. Le robot peut alors exploiter les éléments Dublin Core insérés dans la page pour créer des index par auteur, titre, sujet, etc.

4.1.3 Evolutions récentes

Comme on l'a dit, le Dublin Core a été élaboré (et continue à se développer) lors d'une série de conférences étalées sur trois ans. Au cours de cette période, les promoteurs du Dublin

Core ont essayé de faire évoluer leurs propositions initiales pour améliorer la précision du système et élargir son champ d'application.

Pour des raisons de place, nous ne présentons ici que le dispositif des "qualificateurs" du Dublin Core. Pour d'autres propositions d'extension concernant par exemple la définition de hiérarchies de catégories permettant de caractériser les ressources ou la description des relations pouvant exister entre les ressources nous renvoyons le lecteur à [15] [16] [13] [9] [14].

4.1.4 Améliorer la précision des éléments

La sémantique de certains des éléments est relativement floue. Les auteurs du Dublin Core ont donc introduit récemment la notion de "qualificateurs d'éléments" (*element qualifiers*). Ces qualificateurs doivent aider les logiciels à interpréter les valeurs assignées aux éléments.

Deux qualificateurs ont été très rapidement ajoutés au Dublin Core: *scheme* et *type*. Le qualificateur *scheme* sert à préciser quelle syntaxe a été utilisée pour noter la valeur d'un élément. Par exemple, associée à l'élément Subject l'attribut *scheme* permet d'indiquer si cette valeur correspond à un mot clé libre ou à une valeur correspondant aux grands schémas d'indexation en usage dans les bibliothèques (Classification Décimale Universelle, Classification Dewey, vedettes sujet de la Bibliothèque du Congrès, etc.).

Le qualificateur *type* sert à préciser la sémantique d'un élément. Par exemple, appliqué à l'élément Date, l'attribut type sert à préciser si la date doit être interprétée comme une date de création ou de modification (ce mécanisme est proche de l'élément `<context>` utilisé dans le projet CIMI [5]).

Nous donnons ci-dessous un exemple de qualificateur conforme à la DTD HTML 3.2:

```
<META NAME="DC.subject" CONTENT="(SCHEME=LCSH) Library information
networks -- Great Britain">
```

En HTML 4.0, avec l'attribut SCHEME, une écriture encore plus explicite serait:

```
<META NAME="DC.subject" SCHEME="LCSH" CONTENT="Library information
networks -- Great Britain">
```

Dans ces exemples, on précise la valeur de l'élément DC.subject en indiquant qu'il correspond à un terme de la *Library of Congress Subject Heading* (LCSH).

Plusieurs autres qualificateurs du même genre ont été élaborés dans le cadre du projet anglais ROADS [50]. On trouve une liste complète des qualificateurs du Dublin Core à l'adresse [19] [14] .

4.2 Le Warwick Framework

La proposition initiale du Dublin Core était une simple liste d'éléments de données au sens du catalogage, et ne comportait pour ainsi dire pas d'informations quant à l'implémentation. Le seul modèle vaguement suggéré était celui d'un auteur insérant lui même dans son document des éléments Dublin Core à l'aide d'un éditeur HTML configuré pour ce faire.

L'objectif pricipal de la proposition dite "Warwick Framework" est donc de décrire les grandes lignes d'une architecture informatique de réference pour gérer des métadonnées.

Par ailleurs, il est apparu rapidement à certains des auteurs du Dublin Core qu'une liste de quinze éléments inspirés du catalogage descriptif n'épuisait pas la liste des métadonnées dont on pouvait avoir besoin pour traiter les ressources accessibles sur le Web

L'identification des limites évoquées ci-dessus ont donné lieu à un nouvel ensemble de recommandations, compatibles avec le Dublin Core, et connues sous le nom de *Warwick Framework Architecture*, du nom de la ville de Warwick (Grande-Bretagne) où ces recommandations furent formulées pour la première fois lors d'un congrès en 1996.

A la différence du Dublin Core qui définit en extension un ensemble de métadonnées descriptives, le Warwick Framework propose une architecture permettant de gérer des ensembles de métadonnées de nature variée (" ...an architecture, the Warwick Framework, for aggregating multiple sets of metadata.. ") [34] et donne aussi quelques indications sur les implémentations possibles, point qui n'est pas du tout abordé dans le Dublin Core. L'"architecture" du Warwick Framework repose sur deux notions, le package et le container.

Le package correspond à un ensemble de métadonnées d'un type donné (ainsi dans cette approche les quinze éléments du Dublin Core ne constitue qu'un package descriptif parmi d'autres, destiné à faciliter la recherche d'informations sur Internet).[7]

Le container, comme son nom l'indique, est un objet du réseau servant à regrouper un ensemble de packages que l'on souhaite associer à un objet.

Un exemple possible est celui d'un container regroupant deux packages descriptifs (un ensemble d'éléments Dublin Core et un enregistrement USMARC contenant des informations descriptives équivalentes) et un package détaillant les conditions commerciales (par exemple le tarif de consultation) auxquelles le document associé au container est accessible.

Le container peut être un objet du réseau à part entière (c'est-à-dire un objet ayant son propre URI – ce qui permet de désigner des container indirectement via leur URI) ou il peut être intégré dans un objet du réseau (typiquement le document qu'il sert à décrire – dans le cas le plus simple un document HTML). [34] suggère trois façons d'implémenter le Warwick Framework :

- Insertion d'un container de métadonnées à l'intérieur d'un document HTML à l'aide de l'élément META et de ses attributs. Cependant, la pauvreté syntaxique des types d'attributs SGML (dont HTML est tributaire) ne permet pas de représenter des structures de métadonnées complexes. Ainsi, dans le Warwick Framework un package peut théoriquement être un container, mais cette récursivité est impossible à représenter avec les valeurs d'attributs permises par HTML/SGML.

7. En fait, WF distingue trois types de packages (Metadata set : un package de ce type contient des métadonnées (un enregistrement MARC, un enregistrement Dublin Core etc.) ; Container : un package qui est lui même un container ce qui autorise la récursion ; Indirect : un package qui est une référence indirecte (via un URI) à un autre objet du réseau) mais un package WF est bien fondamentalement un ensemble de métadonnées d'un type donné.

- Mise en correspondance de packages WF avec les parties d'un message MIME "multipart". A l'intérieur de chaque partie, le package peut être codé suivant un type de contenu quelconque (par exemple text/sgml, application/usmarc, etc.)
- Représentation des containers et des packages sous forme d'un ensemble de classes utilisées dans le cadre d'une architecture d'objets distribués de type CORBA. Une classe abstraite nommée MetaDataPackage est la super-classe de toutes les classes du Warwick Framework. Les sous-classes de MetaDataPackage sont :
  - *MetaDataSet*: une classe abstraite dont les sous-classes peuvent être des ensembles de métadonnées de tout type (un package Dublin Core, un package donnant des informations sur l'historique du document, etc.)
  - *MetaDataContainer*: une classe qui représente la notion de container. Le fait que MetaDataContainer descende de la classe abstraite MetaDataPackage permet d'implémenter la récursivité (un package peut être un container qui peut lui même contenir un ou plusieurs packages).
  - *MetaDataIndirect*: une classe qui représente un package référencé indirectement. Cette classe a une méthode qui permet d'obtenir l'URI de l'objet qui contient réellement le package et qui peut être n'importe où sur le réseau. L'idée qui justifie l'existence de cette classe est que les ensembles de métadonnées ne sont pas forcément physiquement liés aux documents mais peuvent exister en tant qu'objets à part entière du réseau, ayant leur propre URI, différent de celui du document auquel on peut vouloir les associer.

Par rapport au Dublin Core, le Warwick Framework a surtout l'avantage de ne pas réduire les métadonnées aux métadonnées descriptives. Il prend en compte l'existence de métadonnées descriptives mais également d'ordre administratif, commercial, etc.

Malgré cette évolution, les deux initiatives que nous venons de décrire présentent cependant un certain nombre de limites.

4.3 Limites du Dublin Core et du Warwick Framework

La première limitation du Dublin Core tient à la nature des objets auxquels les métadonnées sont censées s'appliquer. Les éléments du Dublin Core sont des métadonnées applicables à ce que les recommandations Dublin Core appelle les DLOs (Document-like objects), c'est-à-dire des documents électroniques "atomiques" surtout textuels et assez proches des documents du monde imprimé traditionnel (pages HTML, documents PDF ou postscript, dictionnaires en ligne, etc.).[8]

Il est clair que la palette des ressources accessibles sur Internet ne peut se réduire à cette notion simpliste de DLO. Même si l'on s'en tient aux documents HTML les plus simples,

8. Il faut signaler que dans les documents les plus récents concernant le Dublin Core le mot "resource" tend à se substituer à DLO, mais sans que soient tirées les conséquences pratiques de cette évolution terminologique.

on constate qu'ils ont une structure (qui correspond d'ailleurs à une DTD SGML précise, la DTD HTML) même si cette dernière est, pour l'instant, peu ou pas exploitée par les outils d'indexation.

Par ailleurs les documents Internet (notamment ceux du Web) sont également multimedia et hypertextuels.

L'adoption possible de XML, au moins dans certains secteurs du Web, est susceptible de renforcer encore cette tendance.

Cette triple dimension (structurelle/composite, multimedia et hypertextuelle) des documents couramment rencontrés sur le Web n'est pas réellement prise en compte par le Dublin Core, ce qui constitue une sérieuse limitation.

La seconde limitation des métadonnées du Dublin Core est leur nature presque exclusivement descriptive (les éléments du Dublin Core correspondent de près aux informations que l'on peut trouver dans des catalogues de livres imprimés - des essais ont d'ailleurs été faits à la Bibliothèque du Congrès pour convertir automatiquement les éléments Dublin Core en notices au format USMARC, le format utilisé par les bases de données de la bibliothèque du congrès) [40].

Dans l'approche Dublin Core, on ne décrit une ressource qu'en vue de la retrouver (de même qu'un catalogue n'a pour but que d'avoir accès à l'information) alors que, comme on l'a rappelé dans l'introduction, on peut avoir besoin de métadonnées pour d'autres usages (évaluation, gestion des accès, facturation, archivage et d'une façon générale administration des données).

Cette restriction montre que, même dans ses versions les plus récentes, le Dublin Core reste influencé par son origine bibliographique.

Enfin, en ce qui concerne les aspects techniques, les recommandations du Dublin Core ne fournissent aucun détail pratique sur l'implémentation possible d'un système mettant en oeuvre les métadonnées du Dublin Core. Les implémentations existantes comme celle qui est décrite à la section 4.1.2 reposent toutes sur des solutions ad hoc.

Le Warwick Framework, dont la plupart des auteurs avaient participé à l'élaboration du Dublin Core, apporte certains éléments de réponse aux faiblesses que nous venons d'évoquer.

Il semble que le principal apport du Warwick Framework soit l'accent mis sur le fait que les métadonnées n'ont pas qu'une vocation descriptive. Cette évolution est par exemple perceptible dans la citation suivante de Clifford Lynch, l'un des auteurs de la proposition Warwick Framework :

> I think too that we need to be careful about equating Dublin Core descriptive metadata and metadata in general. In fact, metadata is much broader than just the Dublin Core – it includes parental control kinds of ratings that the current version of PICS is intended to support. It includes rights management, terms and conditions; it includes evaluative information. I think that the Dublin Core is an important first step in supporting descriptive information. The Warwick Framework gives us a broader setting into which we can also slot other kinds of metadata, and I think it's very important that we get some projects moving which give us some real experience with other classes of (non-descriptive) metadata.

> We need to understand how these classes of metadata work in the information discovery and retrieval process. [39]

Cependant, la notion de DLO introduite par le Dublin Core n'est pas remise en cause sur le fond par le Warwick Framework ce qui constitue une limite au plan théorique.

Enfin, en ce qui concerne l'implémentation, nous avons certes indiqué que le Warwick Framework propose des cadres d'implémentation un peu plus concrets que le Dublin Core qui, lui, ne traite pour ainsi dire pas les aspects de représentation informatique. Mais le schéma proposé reste très général. On a pour l'essentiel une taxinomie de classes très générale, ceci pour pouvoir coder le plus large éventail possible de métadonnées. En particulier, le Warwick Framework ne propose pas de scénarios opérationnels précis concernant notamment les points suivants :

- Comment obtenir les métadonnées (comment les extraire, comment intégrer des métadonnées hétérogènes, etc.)
- Comment associer les métadonnées aux ressources du réseau.
- Comment mettre à jour les métadonnées, et d'une façon générale comment les gérer au quotidien.

C'est d'ailleurs la recherche d'une solution pratique qui a poussé récemment les promoteurs du Warwick Framework et du Dublin Core à envisager de s'appuyer sur PICS, une autre initiative importante en matière de métadonnées appliquées au Web. On a ici un intéressant exemple de convergence entre les travaux menés au sein d'une communauté plutôt spécialisée dans les aspects bibliographiques(Dublin Core) et les travaux plus généralistes menés au sein du consortium W3. [9]

Les limites évoquées ci-dessus ne doivent pas faire oublier que le Dublin Core et le Warwick Framework ont joué et continuent à jouer un rôle moteur dans le développement des travaux sur les métadonnées.

5 PICS (Platform for Internet Content Selection (PICS 1.1))

5.1 Motivations

Initialement, le système PICS a été conçu pour filtrer l'information destinée aux mineurs. En comparant des critères fixés par les parents et de métadonnées d'évaluation (les "étiquettes" PICS) associées à un document donné, le logiciel de filtrage intégré à un navigateur Web autorise ou non l'accès à ce document.

Plus récemment, les auteurs de PICS se sont rendus compte que, moyennant des modifications, PICS pouvait être un mécanisme général permettant d'associer des métadonnées de

9. Pour d'autres propositions d'extensions du Warwick Framework nous renvoyons le lecteur à [11] et [33].

n'importe quel type à des documents Web. Parallèlement, des projets particuliers, comme le Dublin Core, ont pensé pouvoir baser leur implémentation sur la technologie PICS.

Cette évolution d'un système très spécialisé à un système d'association de métadonnées très généraliste s'est traduite par la constitution en décembre 1996, au sein du W3C d'un groupe de travail intitulé PICS-NG auquel a succédé un projet à vocation plus générale nommé RDF (*Resource Description Format*) dont nous parlons plus loin.

5.2 Principaux termes et concepts

Nous présentons ici rapidement la version actuelle de PICS (PICS 1.1) décrite dans [32] et [45]. Dans la terminologie PICS, l'information associée à un ou des documents HTML est appelée "étiquette" (content label). Une étiquette est structurée en trois parties :

- l'URL du service d'évaluation ayant créé cette étiquette ;
- les options de l'étiquette ;
- l'information d'évaluation proprement dite, sous forme d'un ensemble de couples attribut/valeur.

La syntaxe complète des étiquettes PICS est donnée dans la spécification. Nous donnons ci-dessous un exemple d'étiquette tiré de cette spécification :

```
PICS-1.1 "http://www.gcf.org/v2.5"
       by "John Doe"
       labels on "1994.11.05T08:15-0500"
              until "1995.12.31T23:59-0000"
              for "http://w3.org/PICS/Overview.html"
              ratings (suds 0.5 density 0 color/hue 1)
              for "http://w3.org/PICS/Underview.html"
              by "Jane Doe"
              ratings (subject 2 density 1 color/hue 1))
```

Schématiquement cet exemple représente une évaluation portant sur les documents `Overview.html` et `Underview.html` (les deux lignes commençant par le mot réservé `for`). Les éléments d'évaluation proprement dits apparaissent à la ligne débutant par le mot `ratings`.

L'entité qui fournit les étiquettes peut être un individu, une institution publique, une société commerciale, etc. Cette entité est désignée sous le nom de "service d'évaluation" (*label service*). Ce service doit rendre librement accessible un site Web sur lequel des logiciels de filtrage peuvent trouver une description de ce service et notamment du système d'évaluation (*rating system*) qu'il utilise.

Cette information doit être fournie aux logiciels de filtrage sous forme d'un document de type application/pics-service. Ce document a pour but de permettre aux logiciels de filtrage de configurer leur interface en fonction du service d'évaluation sur lequel ils se basent.

Le système d'évaluation (*rating system*) spécifie les attributs (la spécification PICS utilise aussi le terme *dimension* ou *category*) utilisés pour l'évaluation et les valeurs possibles que ces attributs peuvent prendre. Il est important de noter que les valeurs sont limitées aux entiers et aux décimaux.

PICS définit trois modes d'association entre l'étiquette et le document auquel cette étiquette correspond.

Une liste d'étiquettes peut être insérée au sein du document HTML via la balise `<META>` comme dans l'exemple ci-dessous :

```
<head>

<META http-equiv="PICS-Label" content='
(PICS-1.1 "http://www.gcf.org/v2.5"
   labels on "1994.11.05T08:15-0500"
          until "1995.12.31T23:59-0000"
          for "http://w3.org/PICS/Overview.html"
   ratings (suds 0.5 density 0 color/hue 1))
'>
</head>
...corps du document...
```

Une liste d'étiquettes peut être renvoyée par un serveur dans l'en-tête HTTP, suite à une requête du client ayant demandé que les étiquettes associées par un service au document faisant l'objet de la requête soient renvoyées en même temps que ce document.

Dans l'exemple ci-dessous, le client demande le document `foo.html` à un serveur supportant PICS. Il demande également au serveur de lui renvoyer l'étiquette associée à ce document par le service `http://www.gcf.org/v2.5`:

```
GET /foo.html HTTP/1.0
Protocol-Request: {PICS-1.1 {params full
                    {services "http://www.gcf.org/v2.5"}}}
```

Le serveur répond en incluant l'étiquette demandée dans l'en-tête HTTP :

```
HTTP/1.0 200 OK
Date: Thu, 30 Jun 1995 17:51:47 GMT
Last-modified: Thursday, 29-Jun-95 17:51:47 GMT
Protocol: {PICS-1.1 {headers PICS-Label}}
PICS-Label:
 (PICS-1.1 "http://www.gcf.org/v2.5" labels
  on "1994.11.05T08:15-0500"
  exp "1995.12.31T23:59-0000"
```

```
  for "http://www.greatdocs.com/foo.html"
  by "George Sanderson, Jr."
  ratings (suds 0.5 density 0 color/hue 1))
Content-type: text/html
...document foo.html...
```

Une liste d'étiquettes correspondant à un document peut enfin être demandée à un bureau et renvoyée sous la forme d'un document HTML de type application/pics-label.

5.3 Exemple d'implémentation

Nous présentons ici un scénario d'utilisation où les étiquettes sont stockées indépendamment du contenu. Les trois modes d'association que nous avons décrits plus haut rendent d'autres scénarios possibles.

Les employés d'un service d'évaluation produisent une étiquette correspondant au document `exemple.html` situé à la racine du serveur `www.essai.com`. Le service d'évaluation propose sur le serveur `www.service.com` une description de ses critères d'évaluation. Cette description est sous la forme d'un document de type `application/pics-service`.

Sur un autre site, un administrateur de PC (qui dans l'approche PICS initiale peut être par exemple un parent) configure un navigateur Web en récupérant sur www.service.com le document de type `application/pics-service` décrivant le service d'évaluation. L'administrateur du PC définit en particulier des seuils de valeur pour les différentes catégories du service d'évaluation.

Un utilisateur veut récupérer le fichier `exemple.html` situé sur `www.essai.com` via le navigateur qui vient d'être configuré.

Au lieu d'envoyer un GET pour récupérer le document demandé, le navigateur émet une requête vers un bureau de distribution `www.bureau.com` qui assure la diffusion des étiquettes produites par le service d'évaluation, et demande l'étiquette correspondant au document `http://www.essai.com/exemple.html`.

Si le bureau renvoie une étiquette applicable à ce document (la spécification PICS indique un certain nombre de traitements pour les cas d'erreur ; nous supposerons ici que le service d'évaluation effectue un recensement quasi exhaustif dans un domaine donné), le navigateur compare cette étiquette aux seuils définis par l'administrateur et refuse l'accès au document `http://www.essai.com/exemple.html` si ces seuils sont dépassés.

5.4 Limites

Comme on a pu en juger d'après les exemples ci-dessus, et à la différence du Dublin Core et du Warwick Framework qui posent des problèmes d'implémentation, PICS repose sur des techniques simples et éprouvées (HTML, HTTP) et des implémentations existent pour la

plupart des navigateurs du Web. Cependant, comme dans le cas du Dublin Core, l'unité d'annotation principale reste le fichier HTML pris comme un tout. [10]

Compte tenu du domaine d'application initial de PICS, (le filtrage des informations destinées aux mineurs) ceci peut poser des problèmes pratiques. Par exemple, dans la version actuelle de PICS, une étiquette associée à un document ne s'applique pas aux images référencées dans le document.

Une autre limitation vient du fait que les valeurs des attributs PICS sont limitées à des valeurs numériques (entiers ou décimaux), pour permettre aux logiciels de filtrage d'effectuer simplement des comparaisons. Cette pauvreté du système de valeurs constitue un obstacle à l'utilisation de PICS comme système de métadonnées généraliste et des évolutions récentes tendent à atténuer ce problème.

5.5 Evolutions récentes de PICS

L'évolution progressive de PICS vers un système généraliste de gestion des métadonnées a été résumée récemment par Paul Resnick, le président du groupe de travail PICS du consortium W3 [49] :

> PICS labels can describe any aspect of a document or a Web site. The first labels identified items that might run afoul of local indecency laws. For example, the Recreational Software Advisory Council (RSAC) adapted its computer-game rating system for the Internet. Each RSACi (the "i" stands for "Internet") label has four numbers, indicating levels of violence, nudity, sex and potentially offensive language. Another organization, SafeSurf, has developed a vocabulary with nine separate scales. Labels can reflect other concerns beyond indecency, however. A privacy vocabulary, for example, could describe Web sites' information practices, such as what personal information they collect and whether they resell it. Similarly, an intellectual-property vocabulary could describe the conditions under which an item could be viewed or reproduced [see "Trusted Systems," by Mark Stefik]. And various Web-indexing organizations could develop labels that indicate the subject categories or the reliability of information from a site.
>
> Labels could even help protect computers from exposure to viruses. It has become increasingly popular to download small fragments of computer code, bug fixes and even entire applications from Internet sites. People generally trust that the software they download will not introduce a virus; they could add a margin of safety by checking for labels that vouch for the software's safety. The vocabulary for such labels might indicate which virus checks have been run on the software or the level of confidence in the code's safety.

10. En réalité, au lieu d'attribuer une étiquette à un document isolé, PICS permet d'appliquer une étiquette "générique" à un ensemble de documents pouvant aller d'un site complet, à un répertoire. Compte tenu de sa généralité, cette méthode pose cependant des problèmes pratiques non négligeables

> In the physical world, labels can be attached to the things they describe, or they can be distributed separately. For example, the new cars in an automobile showroom display stickers describing features and prices, but potential customers can also consult independent listings such as consumer-interest magazines. Similarly, PICS labels can be attached or detached. An information provider that wishes to offer descriptions of its own materials can directly embed labels in Web documents or send them along with items retrieved from the Web. Independent third parties can describe materials as well. For instance, the Simon Wiesenthal Center, which tracks the activities of neo-Nazi groups, could publish PICS labels that identify Web pages containing neo-Nazi propaganda. These labels would be stored on a separate server; not everyone who visits the neo-Nazi pages would see the Wiesenthal Center labels, but those who were interested could instruct their software to check automatically for the labels.
>
> Software can be configured not merely to make its users aware of labels but to act on them directly. Several Web software packages, including CyberPatrol and Microsoft's Internet Explorer, already use the PICS standard to control users' access to sites. Such software can make its decisions based on any PICS-compatible vocabulary. A user who plugs in the RSACi vocabulary can set the maximum acceptable levels of language, nudity, sex and violence. A user who plugs in a software-safety vocabulary can decide precisely which virus checks are required.
>
> In addition to blocking unwanted materials, label processing can assist in finding desirable materials. If a user expresses a preference for works of high literary quality, a search engine might be able to suggest links to items labeled that way. Or if the user prefers that personal data not be collected or sold, a Web server can offer a version of its service that does not depend on collecting personal information.

5.5.1 PICS et le Dublin Core

Comme nous l'avons indiqué à la fin de la section 4.3, la communauté Dublin Core/Warwick Framework a proposé parallèlement aux travaux du consortium W3 d'utiliser PICS comme mécanisme de transport pour les éléments du Dublin Core. Dans cet esprit le *Distributed System Technology Centre* (DSTC) de l'Université de Queensland a proposé d'utiliser le mécanisme des extensions prévus dans PICS 1.1 pour transporter les éléments du Dublin Core.

Pour pouvoir supporter des métadonnées textuelles (alors que les "ratings" de PICS sont basés uniquement sur des valeurs numériques) tout en s'éloignant le moins possible de PICS 1.1, le DSTC propose d'ajouter un type de données *string* et de définir deux extensions rat-inherit et sub-label dans la description des services [18].

Un client qui voit "rat-inherit" dans la description d'un service doit importer cela comme le fait que l'URL entre guillemets pointe vers un autre système d'évaluation. On peut ainsi, par héritage, baser un système d'évaluation sur un autre sytème d'évaluation.

Nous donnons ci-dessous un exemples de codage du Dublin Core mettant en oeuvre le type de données string et les mécanismes d'extension évoqués ci-dessous.

Soient les métadonnées concernant la ressource située à l'adresse `www.marliyn.net/book/` :

```
  Author:        Paul Flora, Acme Organisation, Phone +61 555 555 555
                 Jacky Crystal, jacky@crystal.com
  Title:         Photographs of Marilyn Monroe and Others
  Notes:         With an commentary by Spike Milligan from United Artists
                 Other actresses include Mae West and Brigitte Bardo
  Subjects:      Actors and actresses - Portraits
                 Women Comedians - United States
```

La représentation sous forme d'étiquettes PICS basée sur le Dublin Core est :

```
(PICS-2.0 (service "http://metadata.net/DC/V1.0/")
   (label
      (for "http://www.marliyn.net/book/"
       by "Renato Iannella <renato@dstc.edu.au>"
       on "1997.01.01T08:15-1000"
       until "2000.12.31T23:59-0000"
       lang "en-uk")

          (label  (name "title")
           type/name "Photographs of Marilyn Monroe and Others"
           language/en )
          (label  (name "creator")
           type/name "Paul Flora"
           type/affiliation "Acme Organisation"
           type/phone "+61 555 555 555"
           language/en )
          (label  (name "creator")
           type/name "Jacky Crystal"
           type/email "jacky@crystal.com"
           language/en )
          (label  (name "contributors")
           type/name "Spike Milligan"
           type/affiliation "United Artists"
           type/role "commentator"
           language/en )
          (label  (name "subject")
```

```
scheme/lcsh "Actors and actresses - Portraits"
scheme/lcsh "Women Comedians - United States"
type/keywords "Marilyn Monroe, Mae West, Brigitte Bardot"
language/en )))
```

Les "sous-étiquettes" sont séparées par des barres obliques.

6 De "PICS-NG" à RDF

Parallèlement aux travaux d'extension de PICS que nous venons de décrire, le consortium Web recevait de nombreuses propositions poursuivant des buts similaires (PICS-SE [3], XML-DATA [37], "Web collections" [26], MCF [24] [23] etc.)

Face à cet afflux de propositions , le consortium W3 a réagi en deux temps. Il a tout d'abord essayé de développer des extensions de PICS, sous le nom de PICS-NG, puis dans un deuxième temps, il a lancé un projet distinct de PICS et baptisé RDF (*Resource Description Format*).

Si RDF est aujourd'hui le chantier majeur du consortium W3 en matière de métadonnées, il reprend néanmoins de nombreuses idées proposées initialement dans le cadre de PICS-NG et une présentation rapide de cette initiative s'avère donc utile pour la clarté de l'exposé.

6.1 PICS-NG : une transition entre PICS et RDF

Pour transformer PICS, qui était initialement un système dédié à l'évaluation et au filtrage de contenus accessibles sur Internet, en un système généraliste de représentation et de gestion des métadonnées, le consortium W3 a lancé en 1997 une initiative connue sous le nom de PICS-NG [35].

PICS-NG est basé sur une modélisation orientée objet des métadonnées. Les principaux objets du modèle sont les étiquettes (*labels*).

Les étiquettes sont des collections d'attributs auxquels correspondent des valeurs. Les attributs peuvent prendre pour valeur des :

- types primitifs (string, number, boolean) ;
- étiquette ;
- une liste.

Un couple attribut/valeur est appelé une affirmation (*statement*)

En utilisant des étiquettes, on peut émettre des affirmations à propos de ressources accessibles par un URL ou à propos d'autres étiquettes. L'ensemble des attributs d'une étiquette donnée, ainsi que les types de valeurs que ces attributs peuvent prendre sont définis par un schéma auquel l'étiquette fait référence via l'URL de ce schéma.

Une application qui connaît le schéma particulier utilisé par une étiquette est par définition capable d'interpréter la sémantique de chacune des affirmations contenues dans cette étiquette.

Par contre, une application qui reçoit une étiquette dont elle ne comprend pas le schéma doit être au moins capable de l'analyser syntaxiquement en couples attribut/valeur et de la transmettre intacte à une autre application.

Si une étiquette correspond à plusieurs schémas, l'application peut choisir (héritage multiple) le premier schéma (dans l'ordre gauche-droite) dont elle a connaissance pour interpréter l'étiquette à l'aide de ce schéma.

A partir du schéma référencé dans l'étiquette par un URL, on peut accéder à une description lisible par machine de ce schéma. Ceci est destiné à permettre théoriquement à une machine d'assimiler la sémantique d'un schéma à la volée (cependant, la spécification PICS-NG ne précise pas de quelle façon cet apprentissage peut avoir lieu concrètement).

Un type est un identifiant utilisé par un schéma pour nommer un composant d'un système de typage. Le système de typage par défaut de PICS-NG contient les types suivants (librement inspirés de ceux que l'on trouve dans les dialectes LISP) :

String, Symbol, Integer, Float, Range, Number, Boolean, List, Label, URL, ISODate, Any (qui représente l'ensemble de tous les autres types).

Les attributs peuvent avoir des valeurs multiples (type, list).

L'objet auquel s'applique une affirmation s'appelle le "referent". PICS-NG distingue trois types de référents:

- Referent Value : les affirmations s'appliquent à l'objet nommé par le référent.
- Indirect Referent Value : les applications s'appliquent au référent du référent. Si le référent est une étiquette décrivant un ensemble, les affirmations valent pour chacun des éléments de l'ensemble.
- Immediate value : les affirmations s'appliquent à l'objet référent lui-même. Si le référent est une étiquette, les affirmations concernent l'étiquette elle même.

Une étiquette fait référence à un ou plusieurs schémas pour donner une sémantique aux affirmations qui la composent. Chaque implémenteur est libre de définir le ou les schémas correspondant aux attributs dont il a besoin.

Pour amorcer le système, PICS-NG définit cependant un ensemble d'attributs de base susceptibles d'être utilisés par n'importe quelle étiquette, ces attributs ne pouvant être modifiés ou redéfinis dans d'autres schémas. Plus précisément, la spécification définit les six attributs suivants : `schema`, `for`, `for-indirect`, `for-immediat`, `id`, `dsig` (l'utilisation de certains de ces attributs est illustrée dans les exemples qui suivent).

Enfin deux syntaxes sont proposées par la spécification, soit des s-expressions soit une syntaxe concrète basée sur XML.

Nous donnons ci-dessous un exemple d'étiquette applicable à un document situé sous le répertoire `Lassila/` et interprétable à l'aide du schéma accessible à l'adresse :

```
"http://www.w3.org/authors-and-stuff"
```

Les attributs précédés par une étoile correspondent aux attributs censés être connus de tous les programmes.

```
(pics-2.0
      (label *schema "http://www.w3.org/authors-and-stuff"
             *for "http://www.w3.org/People/Lassila/"
             author "Ora Lassila"))
```

Une étiquette peut être appliquée à une autre étiquette comme dans l'exemple ci-dessous :

```
(pics-2.0
      (label *schema "http://www.gcf.org/v.2.5"
             *for-immediate (label *schema
    "http://www.w3.org/authors-and-stuff"
                                   *for
    "http://www.w3.org/soap.html"
                                   author "Ora Lassila")
             suds 1
             density 0
             color/hue 0))
```

Comme nous allons le voir, la spécification PICS-NG peut être vue comme un précurseur de la proposition RDF.

6.2 RDF

RDF (Resource Description Format) a son origine dans les projets d'extension de PICS que nous avons décrits à la section précédente. La spécification emprunte également des idées aux propositions XML Web Collections, XML/MCF, XML-Data et Dublin Core/Warwick Framework dont nous avons également parlé précédemment.

RDF a été conçu pour supporter les activités suivantes :

- représentation lisible en machine de la "sémantique des métadonnées" ;
- interopérabilité entre métadonnées ;
- recherche d'informations hétérogènes ;
- évaluation (au sens de PICS) des ressources accessibles sur le Web.

La spécification RDF se distingue des propositions précédentes dans le sens où elle propose un modèle de données clairement indépendant de la syntaxe (alors que les propositions précédentes concernant les métadonnées du Web mettaient souvent excessivement l'accent sur les problèmes de codage).

RDF comporte donc :

- un cadre théorique abstrait pour définir et utiliser les métadonnées ;

– une ou des syntaxes concrètes (pour l'instant une représentation en XML a été proposée, mais d'autres syntaxes sont théoriquement possibles).

La spécification RDF comprend deux parties, l'une dédiée à la description du modèle de données RDF [36], l'autre dédiée à l'expression des "schémas RDF" [4], notion que nous présentons ci-dessous.

6.2.1 Le modèle de données RDF

Formellement, le coeur du modèle de données RDF est constitué de trois ensembles :

– un ensemble de noeuds appelé Nodes ;

– un sous-ensemble de Nodes appelé PropertyTypes ;

– un ensemble de triplets, appelé Triples et dont les éléments sont appelés `properties`.

A partir de ce modèle de base, il est possible de décrire les attributs des ressources du Web ou des relations entre ces ressources.

Par exemple, le triplet :

```
(Auteur , [http://www.bib.fr/Les-Contemplations.html] , "Victor Hugo")
```

permet d'exprimer que le document accessible à l'URL :

```
http://www.bib.fr/Les-Contemplations.html
```

a pour auteur Victor Hugo.

Le premier élément de ce triplet correspond à un noeud du sous-ensemble PropertyTypes (le sous-ensemble des noeuds RDF qui sert à désigner des propriétés), le second élément est un noeud quelconque (il s'agit ici d'un noeud correspondant à un document du Web, mais ceci n'est pas une obligation, car comme nous l'indiquons plus loin, l'une des forces de RDF est de permettre d'associer une ou des propriétés à des noeuds quelconques, y compris des noeuds représentant des propriétés), le troisième élément représente la valeur de la propriété (qui est ici une chaîne mais qui pourrait aussi correspondre à un noeud).

Pour se ramener à une représentation courante, le triplet ci-dessus peut être vu comme un arc orienté reliant deux noeuds d'un graphe.

```
[http://www.bib.fr/Les-Contemplations.html] ---- auteur ----> "Victor Hugo"
```

La source de l'arc est la ressource dont on décrit une propriété. La cible est la valeur prise (valeur qui dans l'état actuel de la spécification RDF peut être atomique, ou être un noeud de l'ensemble Nodes). L'arc est étiqueté par un noeud appartenant au sous-ensemble de noeuds représentant les propriétés.

La spécification RDF indique également comment réifier une propriété RDF (c'est-à-dire un triplet). Cette opération de réification permet d'exprimer une propriété sous forme d'un noeud et donc potentiellement d'attacher des propriétés à d'autres propriétés. On peut ainsi introduire un niveau méta supplémentaire et par exemple exprimer son avis ou son degré de certitude vis-à-vis d'une propriété.

La spécification prédéfinit également trois noeuds de base `RDF:Seq`, `RDF:Bag` et `RDF:Alt` qui comme leurs noms le suggèrent peuvent être utilisés pour créer des collections de noeuds.

6.2.2 Les schémas RDF

Parallèlement au modèle de données décrit ci-dessus, le consortium W3 a commencé à travailler sur la notion de "schéma RDF".

Un schéma RDF est une collection d'informations relative aux classes de noeuds RDF. L'objectif est d'une part de pouvoir exprimer que certaines classes sont des sous-classes ou des super-classes d'autres classes, et d'autre part de pouvoir spécifier certaines contraintes (caractère optionnel ou obligatoire des propriétés, contraintes sur le nombre d'occurrences, etc.) s'appliquant aux instances de ces classes.

Un schéma peut être exprimé en utilisant le modèle de données décrit ci-dessus. Par exemple la version actuelle de la spécification sur les schémas stipule l'existence d'une propriété (au sens où nous avons défini ce terme quand nous avons présenté le modèle de données) nommée `RDFS:subClassOf`. Comme son nom l'indique cette propriété RDF sert à spécifier une relation sous-classe/super-classe entre deux classes de noeuds RDF.

Les travaux sur les schémas RDF n'en sont qu'à leurs débuts, et de nombreuses évolutions sont à prévoir.

6.2.3 Syntaxe concrète

RDF n'impose pas une syntaxe concrète particulière pour manipuler et/ou échanger les métadonnées représentées à l'aide du modèle. La spécification comporte néanmoins de nombreux exemples de codage en XML.

6.2.4 Bilan

RDF est l'aboutissement des nombreuses initiatives en matière de métadonnées présentées dans les sections précédentes.

RDF a clos au sein de la communauté Web le débat sur la distinction entre données et métadonnées. Comme on l'a vu, pour RDF cette distinction n'a pas de sens : via la réification une propriété peut se voir attacher des propriétés au même titre qu'une ressource Web au sens classique du terme.

RDF a aussi le mérite de s'inspirer de certains travaux en représentation des connaissances, et d'élever le niveau théorique des débats sur les métadonnées du Web.

7 Bilan et perspectives concernant les méta-données du Web

De nombreuses propositions ont été faites depuis deux ans concernant les métadonnées du Web. Ces différentes propositions n'abordent pas la question des métadonnées sous le même angle mais un certain nombre de points de consensus tendent à se dégager de tous ces travaux :

- Au plan théorique, l'unité élémentaire à laquelle on associe des métadonnées est la "ressource" qui peut être soit un document au sens classique du terme soit un ensemble de données lisibles par une machine (human readable versus machine readable).
- Les métadonnées sont des données à part entière auxquelles peuvent être attachées des métadonnées (on notera que les promoteurs du Warwick Framework, une des initiatives que nous avons décrites dans les sections précédentes, ont longtemps buté sur cette question).
- Au plan pratique, les métadonnées concernant un document peuvent être fournies sous trois formes:
  - Les métadonnées peuvent être contenues dans ce document ;
  - Les métadonnées peuvent être stockées dans un autre document dont l'URI est fourni ;
  - Les métadonnées peuvent être transférées en même temps que le document (par exemple dans un champ de l'en-tête HTTP). PICS est une bonne illustration de ce principe.

8 La fonction Explain de la norme Z39.50 : un exemple de métadonnées relatives à des bases documentaires

Nous entendons ici par "bases documentaires" les bases de données telles qu'elles sont modélisées dans la norme Z39.50 version 3 [46].

Une base, au sens de cette norme, est un objet très simple (une base est simplement définie comme une collection d'enregistrements), clairement défini, et la norme, à travers une de ses fonctionnalités, a prévu explicitement les métadonnées qui pouvait être attachées à ce type d'objet.

L'intérêt de ce type de base de données est aussi qu'elles sont souvent utilisées par des équipes impliquées dans les grands projets de bibliothèques électroniques américains que nous avons mentionnés dans l'introduction, et nous avons dit plus haut qu'une partie de l'intérêt récent porté aux métadonnées vient justement de l'émergence de ces projets.

8.1 Quelques rappels sur Z39.50

Z39.50 est une norme spécifiant une interaction en mode client/serveur entre un logiciel dit "source" (schématiquement un client) et un logiciel "cible" (schématiquement un serveur donnant accès à une ou plusieurs bases de données documemtaires).

Dans la norme Z39.50, une base de données est définie comme une collection d'enregistrements (*database records*). A la base de données sont associés des ensembles d'attributs (*characteristic of a search term*) qui peuvent être spécifiés dans une requête adressée par le client au serveur. Une fois qu'il a identifié un enregistrement, suite à une requête adressée au serveur, le client peut demander le transfert de cet enregistrement ou de certains de ses éléments, selon une certaine syntaxe.

8.2 La fonction Explain

Explain est une fonctionnalité de Z39.50 qui permet à un client d'obtenir des informations sur différents aspects de l'implémentation de la cible (bases de données interrogeables ; jeux d'attributs et jeux de diagnostics utilisés par la cible ; schéma, syntaxe d'enregistrement, définitions de spécification d'éléments supportés pour le transfert sur le poste client). Les cibles qui supportent la fonctionnalité Explain donnent accès (via les services Z39-50 "de Recherche" et "de Présentation") à une base de données nommée IR-Explain-1 (connue sous le nom de "base de données Explain") qui contient toutes ces méta-informations.

A titre d'exemple, la consultation de la base IR-Explain-1 peut aider un utilisateur humain à formuler sa requête ou un logiciel client à se configurer dynamiquement.

Explain distingue de nombreuses catégories d'information, chacune étant susceptible de fournir aux clients Z39.50 qui l'interrogent des informations relatives à un aspect particulier de la cible Z39.50 et des bases de données supportées par cette cible.

Pour des raisons de place, nous ne pouvons toutes les décrire en détail ici, certaines étant d'ailleurs liées à des spécificités de la norme et ne présentant pas un intérêt général. Nous donnons cependant ci-dessous un aperçu des principales catégories de métadonnées contenues dans la base Explain :

- informations relatives à la cible dans son ensemble (catégorie 'TargetInfo') ;
- informations relatives à une base de données particulière (catégorie 'DatabaseInfo') ;
- informations relatives à un schéma spécifique (catégorie 'SchemaInfo');
- informations concernant les syntaxes d'enregistrement supportées par la cible (catégorie 'RecordSyntaxInfo') ;
- informations concernant les ensembles d'attributs spécifiques (catégorie 'AttributeSet-Info') ;
- informations concernant les listes de termes utilisés dans les index d'une base de donnée (catégorie 'TermList-Info') ;

– informations relatives aux combinaisons d'attributs qui peuvent être utilisées quand on interroge une base de données (catégorie 'AttributeDetails').

Ce type d'information existe peu ou prou dans des environnements de bases de données généralistes (on peut par exemple avoir courammment des informations sur les tables en SQL) mais ce qui fait l'intérêt de l'approche Explain est son aspect hautement structuré et son caractère exhaustif. Chaque niveau d'information est pris en compte et clairement identifié (le serveur dans son ensemble, les bases accessibles sur ce serveur, chacune des composantes de ces bases, etc.) et c'est cette démarche analytique et classificatoire très poussée qui nous semble intéressante dans la fonctionnalité Explain.

9 Conclusion

Compte tenu du grand nombre d'initiatives ou de projets mettant en jeu des métadonnées, nous n'avons pu ici que sélectionner quelques projets jugés représentatifs.

Pour donner une idée de l'ampleur de la question, indiquons à titre d'exemple que, même si nous nous étions limités au domaine de la recherche d'informations sur Internet, nous aurions pu encore citer les notices IAFA (dont on trouve une présentation détaillée dans [10]) , Harvest, les profils applicatifs de Z39.50, les travaux sur la structuration des répertoires de services, etc.

En essayant d'avoir une vision globale et en décrivant des projets issus d'univers de recherche divers, nous avons poursuivi plusieurs objectifs. Nous avons tout d'abord voulu montrer comment, malgré les frontières disciplinaires, ces différents projets se sont parfois mutuellement influencés.

A cet égard, les circonstances compliquées de la naissance de RDF, son apparition sous l'influence conjuguée de PICS et des projets de métadonnées à vocation bibliographique comme le Dublin Core, présentent un intérêt tout autant épistémologique que technique.

Nous avons aussi essayé de mettre en évidence la généricité des acquis de la recherche et des questions à résoudre. Comme exemple d'acquis sur lequel un consensus s'est à peu près établi quelles que soient les communautés, nous renvoyons le lecteur à la section 7.

De la même façon on note qu'un certain nombre de questions fondamentales sont évoquées aujourd'hui par toutes les communautés, et ceci sans qu'il y ait forcément communication directe au travers de coopérations scientifiques mais simplement parce que la réalité qu'ils ont à prendre en compte présente des caractéristiques communes.

Au premier plan figure la question de l'hétérogénéité:

– Comment utiliser des métadonnées pour gérer des masses d'information hétérogènes?

– Comment faire "interopérer" des ensembles de métadonnées hétérogènes mais que l'on souhaite mettre en oeuvre de concert pour atteindre un objectif donné (faciliter la recherche d'informations, faciliter la gestion d'informations, certifier l'origine d'informations, etc.)?

Ces recherches font l'objet de travaux intenses notamment dans la communauté des bases de données.

Parmi les grandes questions à traiter, citons également les problèmes liés à l'imprécision qui entoure la définition des ressources électroniques auxquelles les métadonnées sont censées s'appliquer. Dans un univers de plus en plus hypertextuel, la notion de "ressource" est difficile à appréhender. Elle n'a pas ce caractère monolithique des documents du monde imprimé, où chaque entité bibliographique peut être facilement appréhendée dans sa totalité et donner par exemple lieu à une description sous forme bibliographique.

Un examen même superficiel montre clairement que la plupart des ressources électroniques (même les pages HTML simples que l'on trouve sur le Web actuel) ont une structure pouvant être complexe.[11]

Cette complexité suscite deux types de question :

- Comment et à quel niveau de granularité associer des métadonnées aux éléments d'une structure?
- Quelle relation établir entre les métadonnées associées aux composants d'un objet structuré et cet objet lui même?

Les extensions du Dublin Core que nous avons citées à la section 4.1.3 commencent certes à aborder ce type de question, mais beaucoup reste à faire dans ce domaine.

Par ailleurs, les ressources électroniques, notamment dans les projets de bibliothèques numériques, sont rarement des objets isolés. Ils existent au sein de fonds, de collections qu'il est nécessaire de décrire à l'aide de métadonnées complexes. Par conséquent, de façon similaire à ce qu'on a vu pour les documents et leurs composants se pose aussi la question du rapport entre les métadonnées du niveau collection et les métadonnées du niveau document.

Pour rester dans le domaine des bibliothèques numériques, citons enfin la question du rapport entre, d'une part, les métadonnées créées par l'institution qui gère et rend accessibles les ressources et, d'autre part, les métadonnées personnelles créées par un utilisateur dans le cadre de son activité intellectuelle. La capacité à intégrer ces deux sources de métadonnées est essentielle pour élaborer des environnements de travail personnels du type PLAO ([55] [41]).

S'agissant des questions que nous venons d'évoquer, des avancées réelles ne sont possibles que si sont pris en compte les acquis de la recherche dans le domaine de la représentation des connaissances. Les choses semblent évoluer peu à peu dans ce sens notamment avec RDF dont certains auteurs sont issus du domaine de la représentation des connaissances ou, de manière plus ciblée, avec la prise en compte de la notion d'ontologie pour traiter des métadonnées bibliographiques [58].

Nous terminons ce rapport sur quelques considérations générales sur la notion de métadonnées et son utilité. Faute de place, nous n'avons pu examiner ici que quelques exemples de travaux concernant les métadonnées. De cet examen partiel, nous tirons cependant la conclusion que ce qui semble faire l'intérêt de cette notion est qu'elle permet de considérer

11. Ce fait va apparaître de façon encore plus claire avec la diffusion de documents XML.

de façon globale tous les dispositifs de gestion et d'accès à l'information, indépendamment des implémentations et des mécanismes d'accès sous-jacents.

Cette démarche d'abstraction semble être indispensable pour surmonter les problèmes posés par l'accès en réseau à de vastes collections d'informations hétérogènes.

Il n'est donc pas étonnant que la notion de métadonnées soit particulièrement mise en avant dans les secteurs où ces problèmes se posent de façon aiguë (bases de données multimédia, certains grands projets de bibliothèques électroniques américains, *data warehouse*, etc.)

Références

[1] Communications of the ACM : Special issue on digital libraries vol. 38 no 3, May 1995.

[2] Ariadne, an electronic journal published by the United Kingdom Office for Library Information and Networking (UKOLN).
http://www.ukoln.ac.uk/ariadne/.

[3] Ingo Braun and Andreas König. PICS-SE : A proposed standard for the annotation of Internet documents using a string extension to PICS, 1997.
http://www.kulturbox.de/aid/pics-se/Feb97/.

[4] Dan Brickley and R. V. Guha. *Resource Description Framework (RDF) Schemas*. W3C, 1998. http://www.w3.org/TR/WD-rdf-schema/.

[5] Consortium for the Computer Interchange of Museum Information.
http://www.cimi.org/.

[6] CNI white paper on networked information discovery and retrieval. http://www.cni.org/projects/nidr/www/toc.html, 1997. Incomplete Draft by C. Lynch and Avra Michelson.

[7] CNI/OCLC workshop on metadata for networked images - executive summary. http://www.oclc.org:5046/research/dublin_core/summary.html, 1997.

[8] Federal Geographic Data Committee. Content standards for digital geospatial metadata.
http://geochange.er.usgs.gov/pub/tools/metadata/standard/metadata.html.

[9] Simon Cox, Rebecca Guenther, and Diann Rusch-Feja. Dublin Core metadata workshops : Resource types, 1998.
http://www.agcrc.csiro.au/projects/3018CO/metadata/dc_tf/type_3.html.

[10] Roland Dachelet. Description des ressources électroniques : quelques éléments. In J. C. Le Moal et Bernard Hidoine, editor, *Créer et maintenir un site Web : cours INRIA IST98*, pages 175–200, Pau, 28 septembre - 2 octobre 1998. Editions de l'ADBS.

[11] Ron Daniel and Carl Lagoze. Extending the Warwick Framework: from metadata containers to active digital objects. *D-Lib Magazine*, November 1997. http://www.dlib.org/.

[12] Dublin Core metadata. http://purl.org/metadata/dublin_core/, November 1997.

[13] Dublin Core: relations working group. http://purl.org/metadata/dublin_core/wrelationdraft.html.

[14] Provisional report of the Dublin Core subelement working group. http://purl.org/metadata/dublin_core/wsubelementdraft.html.

[15] Dublin Core resource types. minimalist draft: July 17, 1997. http://sunsite.Berkeley.EDU/Metadata/minimalist.html.

[16] Dublin Core resource types. structuralist draft: July 24, 1997. http://sunsite.Berkeley.EDU/Metadata/structuralist.html.

[17] D-lib magazine. http://www.dlib.org/dlib/.

[18] DSTC. PICS extension proposal to support text-based metadata, 1997. http://www.dstc.edu.au/RDU/PICS/proposal03.html.

[19] Dublin Core qualifiers. http://www.roads.lut.ac.uk/Metadata/DC-Qualifiers.html.

[20] EC metadata workshop - 1-2 december 1997, 1997. http://hosted.ukoln.ac.uk/ec/metadata-1997/.

[21] Digital Library Federation. Home page. http://www.clir.org/diglib/dlfhomepage.htm, 1998.

[22] Jean-Michel Franco. *Le Data Warehouse*. Eyrolles, 1997.

[23] R. V. Guha. Meta content framework, 1997. http://mcf.research.apple.com/hs/mcf.html.

[24] R. V. Guha and Tim Bray. Metacontent framework using xml, 1997. http://developer.netscape.com/mfc.html.

[25] The Harvest information discovery and access system. http://harvest.transarc.com.

[26] Alex Hopmann. Web collections using XML, 1997. http://www.w3.org/pub/WWW/TR/NOTE-XMLsubmit.html.

[27] IEEE Computer : Theme issue on the US digital libraries initiative, May 1996.

[28] IEEE metadata.
http://www.llnl.gov/liv_comp/metadata/index.html, 1998.

[29] IFLA - Digital libraries : Metadata resources.
http://www.ifla.org/II/metadata.htm.

[30] International Journal on Digital Libraries.
Springer Verlag.

[31] V. Kashyap, K. Shah, and A. Sheth. Metadata for building the multimedia patch quilt. In V. S. Subrahmanian and Sushil Jajodia, editors, *Multimedia Database Systems : Issues and Research Directions*, pages 297–319. Springer, 1996.

[32] Tim Krauskopf, Jim Miller, Paul Resnick, and Win Treese. *PICS Label Distribution Label Syntax and Communication Protocols. Version 1.1.* W3C, 1997.
http://www.w3.org/PICS/labels.html.

[33] Carl Lagoze. From static to dynamic surrogates : Resource discovery in the digital age. *D-Lib Magazine*, June 1997.
http://www.dlib.org/.

[34] Carl Lagoze, Clifford A. Lynch, and Ron Daniel Jr. The Warwick Framework: A container architecture for aggregating sets of metadata. Technical Report 96-1593, Cornell University,
http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593, 1996.

[35] Ora Lassila. PICS-NG metadata model and label syntax, 1997.
http://207.201.154.232/murray/specs/WD-pics-ng-metadata-970514.html.

[36] Ora Lassila and Ralph R. Swick. *Resource Description Framework (RDF) Model and Syntax.* W3C, 1999.
http://www.w3.org/TR/PR-rdf-syntax/.

[37] Andrew Layman, Jean Paoli, Steve de Rose, and Henry S. Thompson. Xml-data, 1997.
http://www.sil.org/SGML/xml-data9706223.html.

[38] Tim Berner Lee. Axioms of Web architecture, 1997.
http://www.w3.org/pub/WWW/DesignIssues/Metadata.html.

[39] Clifford Lynch and Lorcan Dempsey. Clifford Lynch in interview. *Ariadne*, July 1997.
http://www.ariadne.ac.uk/issue10/clifford/.

[40] Metadata, Dublin Core and USMARC: a review of current efforts, library of congress : Marbi discussion paper 99.
gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc, January 1997.

[41] Catherine C. Marshall. Making metadata: a study of metadata creation for a mixed physical-digital collection. In Rob Akscyn Ian Witten and Frank M. Shipman, editors, *Proceedings of the Third ACM Conference on Digital Libraries*, pages 162–171, Pittsburgh, PA, June 23-26 1998.

[42] Metadata coalition.
http://www.he.net/ metadata/papers/intro97.html.

[43] Alain Michard. *XML : langages et applications.* Eyrolles, 1998.

[44] Eric Miller. An introduction to the resource description framework. *D-Lib Magazine*, May 1998.
http://www.dlib.org/.

[45] Jim Miller, Paul Resnick, and David Singer. *Rating Services and Rating Systems (and Their Machine Readable Descriptions). Version 1.1.* W3C, 1997.
http://www.w3.org/PICS/services.html.

[46] National Information Standards Organization (NISO). *Information Retrieval : Application Service Definition and Protocol Specification.* NISO Press, Bethesda, 1995.

[47] *Proceedings of the 1st ACM International Conference on Digital Libraries*, 1996.

[48] *Proceedings of the 2nd ACM International Conference on Digital Libraries*, 1997.

[49] Paul Resnick. Filtering information on the Internet. *Scientific American*, pages 106–108, March 1997.

[50] The ROADS project: Resource organisation and discovery in subject-based services.
http://ukoln.bath.ac.uk/roads/, December 1996.

[51] Terence R. Smith. The meta-information environment of digital libraries. *D-Lib Magazine*, July/August 1996.
http://www.dlib.org/.

[52] C. M. Sperberg-McQueen and L. Burnard. *Guidelines for Electronic Text Encoding and Interchange (TEI P3).* ACH-ACL-ALLC Text Encoding Initiative, Oxford and Chicago, 1994.

[53] UKOLN center home page.
http://ukoln.bath.ac.uk/.

[54] Anne Marie Vercoustre. Eléments de métadonnées du Dublin Core.
http://www-rocq.inria.fr/ vercoust/METADATA/DC-french.html, 1998.

[55] J. Virbel. Reading and managing texts on the Bibliotheque de France station. In P. Delany and G. Landau, editors, *Text based computing in the Humanities*, pages 31–52. 1993.

[56] Stuart L. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin Core metadata for resource discovery - rfc 2413, September 1998.
ftp://ftp.isi.edu/in-notes/rfc2413.txt.

[57] Stuart L. Weibel and Carl Lagoze. An element set to support resource discovery. *International Journal on Digital Libraries*, 1(2):176–186, September 1997.

[58] Peter C. Weinstein. Ontology-based metadata : Transforming the MARC legacy. In Rob Akscyn Ian Witten and Frank M. Shipman, editors, *Proceedings of the third ACM International Conference on Digital Libraries*, pages 254–263, Pittsburgh (PA), June 23-26 1998.

[59] Sue Welsh. OMNI : alternative approaches to Internet metadata. In *Online Information 96 Proceedings*, pages 379–385, 1996.

[60] Jim Whitehead. A proposal for Web metadata operations, 1997. http://www.ics.uci.edu/ ejw/authoring/proposals/metadata.html.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399