Documentation et bibliothèques

DOCUMENTATION BIBLIOTHÈQUES

Identification des ressources sur Internet et métadonnées : diversité des standards
Identification of Ressources on the Internet and Metadata: A Diversity of Standards
Identificación de los recursos en Internet y metadatos: diversidad de normas

Catherine Lupovici

Volume 45, numéro 4, octobre–décembre 1999

Édition électronique

URI : https://id.erudit.org/iderudit/1032722ar
DOI : https://doi.org/10.7202/1032722ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Lupovici, C. (1999). Identification des ressources sur Internet et métadonnées : diversité des standards. *Documentation et bibliothèques*, *45*(4), 191–194. https://doi.org/10.7202/1032722ar

Résumé de l'article

L'identification des ressources électroniques disponibles sur Internet, ainsi que les métadonnées créées pour permettre leur découverte et leur gestion, ont profondément modifié les standards traditionnels de numérotation normalisée et de description bibliographique et documentaire. Les nouveaux standards en évolution permanente sont désormais très génériques et proches de la modélisation objet. Ils n'ignorent cependant pas l'existant et sont élaborés par une communauté beaucoup plus large que celle de l'information et de la documentation qui y apporte une contribution active.

Identification des ressources sur Internet et métadonnées : diversité des standards

Catherine Lupovici
Directrice
Département de la bibliothèque numérique
Bibliothèque nationale de France
catherine.lupovici@bnf.fr

L'identification des ressources électroniques disponibles sur Internet, ainsi que les métadonnées créées pour permettre leur découverte et leur gestion, ont profondément modifié les standards traditionnels de numérotation normalisée et de description bibliographique et documentaire. Les nouveaux standards en évolution permanente sont désormais très génériques et proches de la modélisation objet. Ils n'ignorent cependant pas l'existant et sont élaborés par une communauté beaucoup plus large que celle de l'information et de la documentation qui y apporte une contribution active.

Identification of Ressources on the Internet and Metadata : A Diversity of Standards
Internet Electronic ressources identification and metadata creation for their discovery and their management are deeply modifying traditional standards for international numbering and bibliographic description. The new emerging standards are continuously evolving. They are more generic and developed in an object-oriented approach. They are at the same time building on the existing ones and the information and documentation community is actively contributing to their elaboration within the new enlarged community involved in this standardization process.

Identificación de los recursos en Internet y metadatos : diversidad de normas
La identificación de los recursos electrónicos disponibles en Internet, así como los metadatos creados para permitir que se descubrieran y administraran, modificaron profundamente las normas tradicionales de numeración normalizada y de descripción bibliográfica y documental. Las nuevas normas, que evolucionan en forma permanente, son por el momento muy generales y se acercan a la enfoque centrado en el objeto. No obstante, no ignoran lo existente y han sido elaboradas por una comunidad mucho más grande que la de la información y la de la documentación, que aporta una contribución muy activa a ella.

Le terme de *métadonnées* désigne de manière générique les données créées pour fournir des informations sur des ressources électroniques d'information. Les métadonnées peuvent remplir différentes fonctions, telles que la gestion des ressources décrites (par exemple, le suivi du cycle de vie d'un document) ; la gestion des informations décrivant le contenu d'un document ou d'une ressource pour en faciliter la découverte ou la localisation, puis l'accès ; ou celle encore des informations relatives aux droits d'accès aux ressources. Le concept de métadonnées est une extension à l'environnement des ressources électroniques – et essentiellement, à l'heure actuelle, à celui des services Web de l'Internet – du concept d'information secondaire décrivant une ressource primaire.

Le contexte Internet

Dans le contexte classique de documents primaires décrits dans des banques de données bibliographiques et/ou documentaires, on avait affaire à deux mondes distincts ayant chacun leurs propres techniques et donc leurs propres standards. L'émergence de documents électroniques, tout d'abord sur des supports hors ligne comme le cédérom, puis sur Internet, a considérablement modifié la situation de ces deux mondes dont les techniques se sont rapprochées, avec des conséquences directes sur les standards techniques utilisés par chacun.

Tout d'abord la notion même de document et la typologie des documents ont évolué. On parle davantage aujourd'hui, dans le monde du document électronique en ligne, de ressource : celle-ci devient un concept générique désignant un ensemble de données pouvant être utilisé comme un tout et relativement à un contexte défini d'utilisation. Ainsi le terme de ressource pourra aussi bien s'appliquer à un site complet ou à une page Web que bientôt –conséquence de l'évolution que le passage du codage HTML au codage XML va imposer au Web – à une portion de page dès lors qu'elle sera autosuffisante par rapport à une utilisation particulière.

Les métadonnées, quant à elles, peuvent désormais être exprimées dans le format technique de codage de la ressource qu'elles accompagnent et être disponibles en même temps qu'elle. Cette simultanéité peut être offerte par le biais de l'application qui rapproche deux sources différentes en offrant à l'utilisateur une interface unique de recherche et de consultation pour les deux types de données ; elle peut aussi provenir de ce que les métadonnées et le document sont traités dans un même fichier informatique dès lors qu'ils sont

encodés dans le même format et gérés par les mêmes logiciels.

Enfin, on considère qu'une même ressource peut être utilisée dans tel contexte comme un ensemble de données sur des données, et comme une ressource dans tel autre contexte. Il ne s'agit plus seulement de créer *a priori* l'information secondaire qui sera ultérieurement exploitée pour trouver la ressource : c'est lors de l'utilisation et grâce à un outil approprié de traitement intelligent que les informations de base seront interprétées pour donner accès à cette ressource.

Les différents acteurs qui souhaitent exploiter Internet font par ailleurs pression pour accélérer sa normalisation technique afin de faciliter la découverte d'information grâce à des moteurs de recherche plus intelligents, et de mettre en œuvre la gestion des droits d'utilisation des ressources.

Le Resource Description Framework (RDF)

C'est pour faciliter cette normalisation technique que le W3Consortium, qui est responsable des évolutions techniques d'Internet, a proposé dès octobre 1997 un cadre général pour la description des ressources de l'Internet : le *Resource Description Framework*. Son objectif était de faire passer le Web du niveau de *machine-readable* à celui de *machine-understandable*.

Le modèle RDF ne prend en compte que les métadonnées créées dans une mention spécifique identifiée en tant que telle. Cette mention de métadonnées peut être associée à la ressource de quatre manières différentes :

- les métadonnées sont encapsulées dans la ressource : c'est le mode *embedded* ;
- les métadonnées sont externes à la ressource mais seront fournies avec elle dans le mécanisme de transfert de la réponse à une recherche d'information : c'est le mode *along-with* ;
- les métadonnées seront utilisées séparément de la ressource pour une recherche d'information, éventuellement dans des bases de données différentes : c'est le mode *service bureau* ;
- la ressource est encapsulée dans les métadonnées qui la décrivent : c'est le mode *wrapped*.

Le dispositif RDF, qui est en cours d'adoption, comprend le modèle et une syntaxe en XML (*Extended Markup Language*). Le modèle est générique et peut être exploité avec une autre syntaxe que XML. Ce dispositif comprend également des spécifications de schémas regroupant des classes de types de ressources.

RDF est défini par des groupes d'utilisateurs très divers comme les acteurs de la normalisation relative à Internet, les bibliothèques, les spécialistes de la structuration des documents (XML/SGML) et ceux de la représentation du savoir. Il bénéficie aussi d'une forte contribution technique provenant des communautés de la programmation orientée objet, des langages de modélisation et des systèmes de gestion de bases de données.

RDF est une modélisation des différentes façons de gérer des métadonnées qui intègre celles qui existent déjà dans la documentation et dans la gestion des ressources électroniques des différentes communautés participant à son élaboration. Il existe donc déjà des standards de métadonnées qui appartiennent à l'un ou à l'autre mode d'association de métadonnées et de ressource et qui sont décrits dans le modèle RDF. Voici quelques exemples de l'intégration dans le modèle RDF de standards de la documentation ou des bibliothèques.

Standards pour le mode « service bureau »

Ces standards concernent la description bibliographique ou documentaire. Ils recouvrent :

- les règles de description, c'est-à-dire la définition des éléments de description et les règles à observer pour créer ces descriptifs à partir des informations figurant dans les ressources que l'on décrit ;
- les formats de description, c'est-à-dire le codage en machine permettant de structurer les données en vue de leur exploitation dans des systèmes d'interrogation.

Certains d'entre eux se sont déjà adaptés au contexte des ressources électroniques en étendant les règles de description et les formats aux besoins de recherche et d'accès direct aux ressources électroniques. C'est ainsi que les formats MARC (*Machine Readable Cataloging*) offrent désormais la possibilité d'indiquer la localisation de la ressource électronique, ses caractéristiques techniques ainsi que le lien à partir duquel on pourra y accéder. Les formats USMARC et UNIMARC ont défini le champ 856 à cet effet.

Standards pour le mode « embedded »

Deux standards principaux ont été définis, l'un par la communauté des documents structurés, l'autre par celle de la normalisation Internet.

Le projet TEI

La *Text Encoding Initiative* (TEI) a été développée dès le début de la mise en œuvre de SGML (*Standard Generalized Markup Language*, ISO 8879) par un groupe de chercheurs en sciences humaines, littérature et linguistique s'intéressant à l'utilisation de l'informatique. Le projet TEI s'est concrétisé par une DTD (Définition de Type de Document) SGML accompagnée de recommandations pour le codage de structure et l'échange des textes. La DTD TEI s'applique à chaque document électronique encodé selon la DTD et définit pour chaque unité documentaire électronique un en-tête obligatoire comportant les métadonnées.

La sémantique de ces métadonnées peut être régie selon des règles de description standardisées ou propriétaires. Dans le cadre d'une application telle que le projet *American Memory* de la Bibliothèque du Congrès, une DTD TEI a été développée et les métadonnées sont rédigées selon les *AACR2 (Anglo-American Cataloging Rules 2)*.

La syntaxe meta HTML

La DTD HTML (*Hypertext Markup Language*) qui a été définie pour la structuration des pages Web de l'Internet comporte une codification <meta> et une syntaxe pour l'écriture des métadonnées dans les pages Web. Ces *tags* <meta> ne sont pas affichés par les navigateurs, mais ils sont utilisés par certains moteurs de recherche sur Internet de manière pondérée par rapport au reste du texte des pages Web.

Dans le cadre de cette standardisation Internet, certaines communautés d'utilisateurs, dont les bibliothèques et les producteurs de banques de données, ont travaillé à la standardisation des types de métadonnées. C'est ainsi qu'est né le *Dublin*

Core qui définit un ensemble de quinze métadonnées réparties en trois grands types :

- des métadonnées relatives au contenu intellectuel de la page Web ;
- des métadonnées relatives à la propriété intellectuelle et en particulier aux informations sur la gestion des droits d'utilisation de la page Web ;
- des métadonnées relatives à la ressource elle-même et en particulier l'identifiant de la ressource, si possible unique et persistant. Voir le tableau 1 du texte « La publication électronique des thèses » des auteurs Boulétreau, Gauvin et Ducasse, page 187.

Le succès du *Dublin Core* a été tel que les éléments de description qui y sont définis sont retenus comme sémantique de métadonnées en dehors de la syntaxe <meta> HTML ; ils ont déjà été repris dans les premières spécifications de XML, dans un schéma minimum de base.

Standards pour le mode « wrapped »

Ce mode considère les métadonnées comme l'élément englobant pour la ressource. La DTD EAD (*Encoding Archival Description*) a été développée dans le continent nord-américain pour la structuration et l'encodage en XML des instruments de recherche décrivant des collections spécialisées telles que les archives ou les collections de papiers. Ces instruments de recherche sont, par exemple, des inventaires d'archives publiques ou privées ou des catalogues de collections de documents manuscrits.

Cette DTD autorise une structuration de l'instrument de recherche en douze niveaux hiérarchiques et permet d'attacher la ressource électronique à sa description.

Elle est déjà utilisée dans des projets coopératifs de collections de ressources électroniques. L'arborescence de l'instrument de recherche peut servir de support à la navigation dans une telle collection.

La DTD EAD émane de la communauté des archives et des bibliothèques, mais elle est aussi expérimentée par des musées d'Amérique du Nord.

L'identification des ressources

Des numéros ou des codes d'identification des documents ont été créés depuis une trentaine d'années pour différents types de documents. L'ISBN (*International Standard Book Number*) pour le livre et l'ISSN (*International Standard Serial Number*) pour les publications en série, sont les plus anciens et ont acquis le statut de normes internationales ISO. L'organisation de ces systèmes de numérotation repose sur un réseau international d'agences d'attribution et d'enregistrement des numéros. Ces identifiants sont utilisés à la fois par les éditeurs pour la gestion de la commercialisation et par les bibliothèques et centres de documentation pour la gestion des commandes et l'identification bibliographique et documentaire des ouvrages et périodiques.

Ce système de codes d'identification a ensuite été étendu aux unités logiques composant un document, par exemple l'article dans un fascicule de périodique. Dans la décennie quatre-vingt-dix, cette extension a été adaptée aux besoins de la gestion, puis de la distribution de documents sous forme électronique.

Alors que le concept de ressource prend le pas sur celui de publication, une réflexion est actuellement conduite sur l'identification des ressources. Envisagée dans le contexte du Web, cette identification s'accompagne d'une dimension dynamique, car il est devenu nécessaire que le système d'identification, non content d'identifier et de localiser une ressource, y donne également accès. L'identifiant devient actif et « cliquable ».

Le système URI

L'Internet Engineering Task Force (IETF), qui développe les standards Internet, a lancé un travail de standardisation pour un système générique d'identification des ressources du réseau, dont le principe est que l'identifiant peut à la fois représenter la ressource et permettre d'y accéder : c'est le système URI (*Uniform Resource Identifier*). Il s'agit de proposer un cadre et des règles pour permettre à des organismes d'enregistrement de décrire des ressources et d'en assurer la disponibilité de manière à ce qu'elles soient finalement accessibles dans un ou plusieurs sites via leurs URL. Ce cadre d'identification s'appuie sur le dispositif suivant.

- L'URN (*Uniform Resource Name*) de la ressource est son nom (au sens Internet du terme) unique et persistant : il désigne une ressource qui pourra être présente sur plusieurs sites comme autant d'exemplaires. Ce numéro est résolu – c'est-à-dire qu'il indique le ou les URL du ou des sites où l'on peut consulter la ressource – par une agence où il est enregistré et qui maintient la validité des URL. Cette agence s'appelle une agence de résolution du nom.
- L'URL (*Uniform Resource Locator*) est l'adresse du site où se trouve la ressource. On se connecte à celle-ci en cliquant sur ce lien qui n'est pas persistant et dont la validité n'est pas garantie.
- L'URC (*Uniform Resource Characteristic*) contient des métadonnées sur la ressource et en particulier les modalités d'accès et la gestion des droits. Ces métadonnées peuvent être stockées dans une banque de données d'enregistrement des URN maintenue par l'agence d'enregistrement et de résolution. Elles peuvent également se trouver sur un site différent et liées aux URN.

Le système DOI

Le système DOI (*Digital Object Identifier*) est une application du système URI mise en œuvre par une fédération d'éditeurs commerciaux par l'intermédiaire de la Fondation DOI.

Cette Fondation s'est constituée comme agence d'enregistrement des identifiants DOI des publications. Le numéro DOI est un URN au sens Internet et il en a la structure, composée d'un préfixe attribué à l'éditeur et d'un suffixe qui est le numéro de la ressource chez l'éditeur ou le détenteur des droits. L'attribution de ce numéro relève de la responsabilité de l'éditeur ou du détenteur des droits. Il peut être un identifiant issu des systèmes existants, tels l'ISBN ou l'ISSN. L'attribution d'un préfixe est conditionnée au paiement d'un droit par l'éditeur et elle comporte une clause sur le respect du *copyright* et l'engagement de mettre à jour, en cas de modification, les URL servant à la résolution des DOI.

Les éditeurs entretiennent des bases d'information sur les ressources qu'ils gèrent : elles contiennent des URC, et c'est sur celles-ci que pointent les numéros DOI de la base d'enregistrement. Un utilisateur effectuant une recherche par numéro URN aura ainsi accès à la base URC

associée à ce numéro et aura connaissance des conditions d'accès à la ressource (accès gratuit, paiement à l'acte, conditions d'abonnement, etc.).

Les publications électroniques sur Internet utilisent et utiliseront de plus en plus souvent des données multimédias avec des droits complexes et entrelacés. De plus l'évolution programmée de la structure des pages Web, qui verra le passage du codage HTML au codage XML, conduit à imaginer d'attribuer des identifiants à des portions de contenus à l'intérieur même des ressources ; cela conduira à un découpage en unités plus fines que la page Web. Les éditeurs et la Fondation DOI travaillent donc, en suivant les évolutions de la standardisation Internet, à des extensions de ces notions d'identification des ressources et aux moyens de mettre en œuvre les métadonnées associées aux identifiants dans le contexte de l'édition électronique sur le Web.

Les tendances présentes sont l'identification d'un document ou d'une ressource et de ses inscriptions dans des publications papier, électronique hors ligne, électronique en ligne, une des difficultés étant que cette diversification des types de manifestations d'une œuvre fait éclater les typologies classiques et, en particulier, la notion de publication en série, avec tout l'appareil d'identification qui l'accompagnait. À cet égard, l'un des thèmes de réflexion actuellement le plus important – et directement lié à l'identification et aux métadonnées – est la mise en œuvre de la gestion des droits d'accès aux ressources par navigation à partir des citations dans les articles.

De nouveaux partenaires

Les publications sur Internet ne peuvent être gérées à l'aide des standards des publications classiques. On assiste actuellement à un grand mouvement de réflexion et de modélisation conceptuelle qui englobe les publications classiques et les ressources électroniques. Cette modélisation s'accompagne de standards plus génériques que ceux que nous avons connus pour les publications traditionnelles. Leur élaboration se fait en collaboration avec de nouveaux partenaires pour lesquels l'apport des professionnels de l'information peut être très important. Un exemple frappant en est le cas du Dublin Core.

Sources consultées

Digital libraries : cataloging and indexing of electronic resources. Bibliography. In : IFLA electronic collections. <URL : http://ifla.inist.fr/II/catalog.htm>

Digital libraries : metadata resources. In : IFLA electronic collections. <URL : http://ifla.inist.fr/II/metadata.htm>

Lupovici, Catherine. 1998. Le Digital Object Identifier : le système DOI. *Bulletin des bibliothèques de France* 43 (3) : 49-54. <URL : http://www.enssib.fr/Enssib/bbf/bbf.htm>

_____. 1998. L'information bibliographique des documents électroniques. *Bulletin des bibliothèques de France* 43 (4) : 42-47. <URL : http://www.enssib.fr/Enssib/bbf/bbf.htm>