L'INFORMATION BIBLIOGRAPHIQUE DES DOCUMENTS ÉLECTRONIQUES

CATHERINE LUPOVICI

Jouve Digitalisation des informations

clupovici@jouve.fr

L'information bibliographique est créée pour faciliter la sélection et l'accès au document primaire. Cette information électronique ajoutée est codée dans un format qui permet la préparation de différents produits tels que des bibliographies imprimées, des services bibliographiques en ligne et des cédéroms bibliographiques. L'information secondaire est composée de deux types de données : la description bibliographique proprement dite et des points d'accès destinés à faciliter l'accès à la description bibliographique.

La description bibliographique est une sélection de données extraites du document primaire telles que le titre, la mention de responsabilité, les données relatives à l'éditeur, la description physique. Elle est parfois complétée par des commentaires rédigés par le catalogueur. Ces informations sont créées en suivant les principes de règles de catalogage.

Les points d'accès constituent la véritable valeur ajoutée. Ils sont contrôlés par des thésaurus ou des formes d'autorités pour les noms propres d'auteurs ou de collectivités, les noms géographiques ou les descripteurs.

Les bases de données bibliographiques informatisées sont depuis trente ans gérées sous la forme d'une notice bibliographique et les points d'accès sont contrôlés par des dictionnaires externes des formes autorisées, les fichiers d'autorité. Le lien avec le document primaire est fourni par l'intermédiaire d'une cote donnant la localisation de l'unité physique.

La gestion du document électronique dans cet existant nécessite d'étendre l'organisation actuelle, voire de la repenser plus complètement selon l'importance relative des collections électroniques et selon les services que l'on souhaite offrir. La simple extension de l'organisation actuelle à la gestion des documents électroniques consiste à fournir le lien qui permet de passer de la notice bibliographique au document électronique, afin de le consulter à partir du même poste de travail que celui qui a servi à consulter le catalogue. Cependant, dès que l'on pratique la technique des liens, on pense très vite que d'autres documents électroniques peuvent fournir un accès à un document électronique donné, parce qu'ils le citent ou l'analysent, et l'on voudra établir un lien fonctionnel du même type que celui qui va du catalogue au document qui y est décrit.

On peut aussi considérer directement la collection de documents électroniques comme susceptible d'offrir directement accès au profil de l'information descriptive contenue dans les

documents, sans nécessairement la dupliquer dans une notice bibliographique, pour peu que le format utilisé pour le document électronique permette de distinguer cette information. Si tel est le cas, il suffit alors de créer l'information à valeur ajoutée directement au document lui-même et dans le format du document.
Toutes ces approches sont actuellement expérimentées dans différents projets de bibliothèques ou d'archives électroniques et principalement dans le contexte de programmes de numérisation de collections qui permettent aux bibliothèques et archives de choisir les collections, le format des documents électroniques ainsi que l'architecture fonctionnelle du système supportant la gestion et l'accès aux collections.

Extension des formats bibliographiques

L'extension de la philosophie de traitement documentaire du document classique au document électronique doit être examinée selon les deux axes suivants :
– *la création de l'information descriptive et des points d'accès.* En effet, le catalogage des documents électroniques est bien plus lourd que celui du document classique. Il nécessite souvent l'installation puis la désinstallation des logiciels. En revanche, lorsqu'il s'agit d'une reproduction numérique d'un document existant, on peut simplement ajouter une information à la notice bibliographique de l'original si elle existe.
– *la création d'une sorte de cote active* qui va permettre à l'utilisateur de consulter directement le document à partir de la notice bibliographique. Cette formation peut être créée dans un format local ou utiliser des standards pour la localisation tels que les URL (Uniform Resource Locator), si on offre la consultation du catalogue et des documents dans un environnement Web.

On peut utiliser les formats MARC pour cataloguer et localiser les documents électroniques en alimentant un champ particulier à l'aide d'informations relatives à une reproduction électronique d'un document classique.
Le champ 856 a été récemment introduit dans USMARC et dans UNIMARC pour cet usage. Dans UNIMARC[1], le champ 856 est défini pour la publication globale (un périodique et pas un fascicule ni un article par exemple). C'est un champ que l'on peut répéter pour signaler des localisations multiples ou des fichiers informatiques différents. Les indicateurs et sous-champs suivants sont définis :
– indicateur 1 : méthode d'accès (e-mail, ftp, telnet, dial-up, http, spécifique)
– indicateur 2 : blanc (non défini)
– sous-champs :
$a Nom du serveur
$b Numéro d'accès (Adresse IP, numéro de téléphone)
$c Information sur la compression
$d Path
$f Nom électronique (des fichiers)
$g Uniform Resource Name (URN)
$h Responsable des traitements
$i Instruction (si cela est demandé par un serveur pour traiter une requête)
$j Bits par seconde
$k Mot de passe (mots de passe généraux et pas de sécurité)
$l Logon/login (logon/login généraux)
$m Contact pour aide à l'accès
$n Nom de la localisation du serveur spécifié en $a
$o Système d'exploitation du serveur spécifié en $a
$p Port
$q Type de format électronique (ASCII, MIME Internet media types)
$r Paramètres utilisés pour le transfert des données
$s Taille de fichier
$t Émulation de terminal
$u Uniform Resource Locator (URL)
$v Méthode et horaires d'accès
$w Numéro de contrôle de notice
$x Note non publique
$y Méthode d'accès (si ce n'est pas l'un des trois principaux protocoles TCP/IP)
$z Note publique
Exemples :
856 3#$b1-202-7072316$j2400-9600$n Library of Congress, Washington, DC$oUNIX$rE-7-1 $tvt100$zRequires logon and password
200 0#$aBulletin des bibliothèques de France
856 4#$uhttp://www.enssib.fr/Enssib/bbf.htm$zTextes au format PDF

L'INITIATIVE DUBLIN CORE EST UN TRAVAIL DE NORMALISATION INTERNATIONALE POUR LA DÉFINITION DES ÉLÉMENTS DE DONNÉES BIBLIOGRAPHIQUES À INCLURE DANS LES PAGES WEB

L'approche métadonnées du Dublin Core

Dans l'environnement Web, on peut utiliser la syntaxe des métadonnées pour inclure dans une page HTML la description des ressources de la page. Ces données sont utilisées automatiquement par les moteurs de recherche.
Dans le cadre général de la technique des métadonnées, l'initiative Dublin Core est un travail de normalisation internationale pour la définition des éléments de données bibliographiques à inclure dans les pages Web. La liste de base des éléments a été arrêtée en décembre 1996 et ces

1. 856 Electronic Location and Access. URL : http://ifla.inist.fr/vi/3/p1996-1/856.htm

LE DUBLIN CORE PEUT ÉGALEMENT ÊTRE CONSIDÉRÉ COMME UN SOUS-ENSEMBLE D'ÉLÉMENTS DESCRIPTIFS PERMETTANT L'ÉCHANGE ENTRE DES FORMATS HÉTÉROGÈNES PLUS COMPLEXES

données sont en cours de test dans différents projets qui conduisent parfois à des interprétations différentes de ce que sont ces données.

Les éléments de données du Dublin Core sont le résultat d'un consensus sur une description minimale des ressources électroniques à inclure dans la création des pages Web pour permettre la recherche de ces ressources.

Le Dublin Core peut également être considéré comme un sous-ensemble d'éléments descriptifs permettant l'échange entre des formats hétérogènes plus complexes.

Les éléments de données définis dans le Dublin Core (DC) sont :

Métadonnées relatives au contenu

– titre

– sujet et mot-clé : topique de la ressource

– description : une description textuelle du contenu de la ressource

– source : autre ressource à partir de laquelle la ressource est dérivée

– langage

– relation : liens vers d'autres ressources

– couverture : caractéristiques spatiales et temporelles du contenu intellectuel de la ressource.

Métadonnées relatives à la propriété intellectuelle

– auteur ou créateur : responsabilité principale du contenu intellectuel

– éditeur : entité responsable de la mise à disposition de la ressource dans sa forme actuelle

– autre contributeur : personne ou organisme qui a fourni une contribution intellectuelle importante à la réalisation de la ressource

– gestion des droits : lien vers une mention de gestion des droits ou un service donnant ce type d'information.

Métadonnées relatives à l'instance documentaire

– date

– type de ressource : catégorie de la ressource, par exemple page d'accueil, poésie, document de travail

– format : format des données (logiciel et matériel nécessaires pour utiliser la ressource)

– identifiant de la ressource : chaîne de caractères ou numéro utilisé pour identifier de manière unique la ressource (URL, URN, ISBN, etc.).

On peut associer des vocabulaires contrôlés à certains éléments de données. Les éléments DC sont actuellement testés dans un grand nombre de projets dans différents pays – dont des projets européens. Voici quelques exemples significatifs pour l'information bibliographique et qui se déroulent dans des pays européens :

– *En Allemagne*, Metadaten-Projekt (Metadata Project) explore l'utilisation des métadonnées du point de vue des bibliothèques et examine l'impact, sur les règles de catalogage traditionnelles, du développement dans un environnement réseau d'instruments de recherche et de navigation (URL : *http://www2.sub.uni-goettingen.de*).

– *En Scandinavie*, The Nordic Metadata Project est un système coopératif de création de métadonnées. Le Dublin Core a été choisi pour créer l'information qui permettra à l'utilisateur de trouver toutes sortes de documents électroniques sur le Web. URL : *http://linnea.helsinki.fi/meta/*

INDOREG (INternet DOcument REGistration) est un projet du Danish Library Center (DBC) pour l'enregistrement de toutes les publications Internet qui répondent à un profil de documents et pour fournir l'accès à ces documents *via* DanBib (URL : *http://www.purl.dk/rapport/html.uk/*).

– *Au Royaume-Uni*, BIBLINK est un projet européen du programme bibliothèques qui regroupe plusieurs bibliothèques nationales. Il a pour objectif d'établir un lien entre les éditeurs et les agences bibliographiques nationales pour échanger des notices électroniques de métadonnées sur les nouvelles publications électroniques (URL : *http://www.ukoln.ac.uk/metadata/biblink/*).

Dans le prototype ELISE II (Electronic Library Image Service for Europe), les données catalographiques fournies par les institutions participantes sont reformatées en Dublin Core et affichées en même temps que les images au format vignette (URL : *http://severn.dmu.ac.uk/elise/*).

BIBLINK A POUR OBJECTIF D'ÉTABLIR UN LIEN ENTRE LES ÉDITEURS ET LES AGENCES BIBLIOGRAPHIQUES NATIONALES POUR ÉCHANGER DES NOTICES ÉLECTRONIQUES DE MÉTADONNÉES SUR LES NOUVELLES PUBLICATIONS ÉLECTRONIQUES

SGML et l'information bibliographique

SGML (Standard Generalized Markup Language), ISO 8879 est une norme internationale de format logique de documents qui commence à être utilisée dans plusieurs projets de ressources électroniques pour lesquels l'information bibliographique et le document sont gérés avec un niveau de format professionnel équivalent.

Dans ces expériences, la technologie SGML est utilisée de deux manières. La première approche consiste à considérer le document électronique au niveau de granularité le plus fin et à ajouter un « en-tête » SGML au document structuré en SGML en appliquant les principes du TEI (Text Encoding Initiative).

La seconde approche permet d'apprécier une collection de documents électroniques dans son entier et de définir une DTD (Définition de Type de Document) SGML capable de couvrir toute l'information secondaire correspondante allant de la description de toute la collection, depuis la racine de l'arborescence jusqu'au niveau de chaque unité, en suivant l'organisation de la collection. On reproduit ainsi la hiérarchie des instruments de recherche tels que les inventaires, les registres, les index et les guides.

L'approche « en-tête » de texte structuré

La communauté des chercheurs qui s'intéresse à l'utilisation de l'informatique dans les sciences humaines, la littérature et la linguistique, a commencé il y a dix ans à construire un cadre commun de codage pour la création de nouveaux documents ou pour l'échange de documents textuels ou de données d'archives en s'appuyant sur SGML.

Ce projet connu sous le nom de TEI (Text Encoding Initiative) a finalement abouti à une DTD SGML accompagnée de Recommandations pour le codage et l'échange des textes. Une version simplifiée, la DTD TEI Lite[2] a ensuite été créée pour fournir un jeu minimum d'éléments de données de base utilisable pour le cas général.

L'une des caractéristiques importantes de la DTD TEI est de définir la structure d'un en-tête au document

LA COMMUNAUTÉ DES CHERCHEURS QUI S'INTÉRESSE À L'UTILISATION DE L'INFORMATIQUE DANS LES SCIENCES HUMAINES A COMMENCÉ IL Y A DIX ANS À CONSTRUIRE UN CADRE COMMUN DE CODAGE POUR LA CRÉATION DE NOUVEAUX DOCUMENTS OU POUR L'ÉCHANGE DE DOCUMENTS EN S'APPUYANT SUR SGML

fournissant des métadonnées sur le document balisé telles que la source, les principes retenus pour le balisage, des informations sur l'histoire du texte, en particulier ses révisions et modifications. La présence de l'en-tête est obligatoire dans un document TEI. L'en-tête est composé des parties suivantes :

– *description du fichier* (élément obligatoire) : c'est l'équivalent électronique des informations contenues dans la page de titre pour un document papier. La souplesse du cadre TEI permet par exemple de décrire un texte en suivant les AACR2 (Anglo-American Cataloging Rules).

– *description sur l'encodage* : elle donne des informations sur la relation entre le texte encodé et le (ou les) texte(s) sources (dans le cas du traitement simultané dans un même document encodé de plusieurs versions différentes du texte source). Il peut s'agir par exemple d'indiquer le nom du projet d'encodage et les choix de transcription faits.

– *description sur la révision* : elle contient l'histoire des révisions du texte.

L'en-tête peut être très simple ou très complexe selon les besoins de l'application.

Plusieurs DTD TEI ont été développées et sont actuellement testées dans des projets à résonance bibliographique. Deux exemples significatifs sont :

– *Electronic Text Center, University of Virginia*[3]. Ce centre a été créé en 1992 et offre, en ligne, une collection de centaines de textes codés en SGML. Le codage des documents a été effectué en suivant la TEILITE.DTD. Des services de recherche sont offerts aux utilisateurs *via* un service Web pour lequel les documents sont reformatés en HTML à la volée pour l'affichage. Dans cette application, l'en-tête TEI est très complexe : il est composé de la notice bibliographique du document imprimé qui a été reproduit, de la notice relative à la création de la forme électronique et elle fournit en plus quantité d'informations sur l'outil de recherche. L'en-tête une fois créé est utilisé pour générer la notice USMARC qui est intégrée dans le catalogue de la bibliothèque. La notice USMARC est générée à partir de l'en-tête par un programme de conver-

2. Introduction au codage des textes électroniques en vue de leur échange. URL : http://www.uic.edu/orgs/tei/lite/teiu5_fr.html
3. University of Virginia Library. Electronic Text Center. URL : http://etext.lib.virginia.edu/

LA DTD EAD PERMET DE TRAITER LES REGISTRES ET LES INVENTAIRES, QUELLE QUE SOIT LEUR TAILLE, EN DÉCRIVANT L'ENSEMBLE DES COLLECTIONS

sion automatique.

– *Library of Congress American Memory DTD for Historical Documents*[4]. La Bibliothèque du Congrès utilise SGML pour le balisage du texte intégral de livres, de brochures, de manuscrits et d'autres textes historiques. La DTD American Memory (AMMEN.DTD) utilisée dans ce traitement des documents est une application de TEI Lite. Les fichiers sont disponibles sur le serveur Web de la Bibliothèque du Congrès.

Dans cette application, la Bibliothèque du Congrès a décidé de ne pas dupliquer la totalité des informations bibliographiques dans l'en-tête des textes. Seul le titre est copié à partir du champ 245 USMARC de la notice bibliographique complète dans le catalogue. Il est complété d'un commentaire indiquant qu'il s'agit d'une transcription numérique du document. Le numéro LCCN (Library of Congress Card Number), s'il existe pour la notice du document, est aussi donné en mention de responsabilité source.

La DTD Encoded Archival Description (EAD)

Le codage en SGML des instruments de recherche de documents d'archives a commencé à l'université de Californie, Berkeley, en 1993. Une première version de DTD a ensuite été testée par d'autres bibliothèques et améliorée avec la participation du Committee on Archival Information Exchange de la Society of American Archivists (SAA). Les organismes qui ont participé aux travaux du groupe de travail comprenaient la Library of Congress, RLG (Research Library Group) OCLC et le SAA. Une version *beta* a été distribuée en juillet 1996 et est en cours de test dans plusieurs projets. En tant que norme internationale potentielle, cette DTD est maintenue dans le Network Development and MARC Standards Office de la Library of Congress, en partenariat avec la Society of American Archivists.

L'objectif est d'aboutir à une norme non propriétaire pour le format des instruments de recherche lisibles en machine, avec la volonté d'aller au-delà de l'information que l'on peut produire dans une notice MARC traditionnelle. La DTD EAD permet de traiter les registres et les inventaires, quelle que soit leur taille, en décrivant l'ensemble des collections.

Le dispositif permet de décrire des documents de tous types : textuels, vidéo ou enregistrements sonores. L'en-tête est conforme à TEI. L'ensemble du texte descriptif papier que constitue par exemple un inventaire est ainsi converti de manière à pouvoir être traité dans une puissante base de données relationnelle ou orientée objet.

Cette norme offre une passerelle avec les formats et les outils traditionnels par le biais d'attributs d'équivalence MARC. Ceux-ci sont utilisables avec les éléments qui ont une correspondance avec les champs USMARC, y compris la possibilité d'ajouter les formes d'autorité MARC. Enfin, les éléments définis dans la DTD EAD sont conformes aux règles de description internationales des archives ISAD(G) (General International Standard Archival Description).

Quelques exemples de mise en œuvre de la DTD EAD montrent son champ d'application :

– *California Heritage Digital Image Access Project*. Ce projet va jusqu'à inclure des documents numérisés dans la structure électronique de l'instrument de recherche codé avec la DTD EAD. L'instrument de recherche lui-même est reformaté à la volée de SGML (format de stockage et d'indexation) en HTML pour l'affichage dans des pages Web. On peut donc ainsi parcourir classiquement l'instrument de recherche et aboutir aux pièces qui sont numérisées (URL :

L'UTILISATION DE SGML OUVRE LES POSSIBILITÉS D'UTILISATION D'OUTILS ET DE CONCEPTS FACILITANT LES ÉCHANGES AVEC D'AUTRES COMMUNAUTÉS COMME LES MUSÉES, LES CHERCHEURS, L'INFORMATION INSTITUTIONNELLE Y COMPRIS LA LITTÉRATURE GRISE, LA DOCUMENTATION TECHNIQUE

4. American Memory DTD for Historical Documents. URL : http://lcweb2.loc.gov/ammem/amdtd.html

http://sunsite.berkeley.edu/CalHeritage).
– *Library of Congress Finding Aids Project*. La page d'accueil correspondante de la Library of Congress donne accès à tous les instruments de recherche qui ont été codés avec la DTD EAD. Les évolutions prévoient d'offrir une interface de recherche pour tous les instruments de recherche de tous les départements de la bibliothèque, ainsi que des liens allant du catalogue des notices bibliographiques vers les instruments de recherche (URL : *http://www.loc.gov/rr/ead/eadhome.html*).
– *American Heritage Virtual Archive Project*. C'est un projet coopératif entre quatre universités pour analyser les facteurs qui interviennent dans la création et la maintenance d'une base de données collective d'instruments de recherche en format EAD, associés aux documents numérisés décrits dans les instruments de recherche (URL : *http://sunsite.berkeley.edu/ amher*).

Différentes étapes

Les différents formats en cours d'élaboration pour la création et la gestion de l'information bibliographique permettent évidemment de construire des systèmes d'information d'architectures fonctionnelles différentes. Ils représentent également différentes étapes possibles de migration vers les nouveaux environnements d'information numérique.
La simple extension du format bibliographique à la gestion du lien vers la reproduction numérique du document qu'il décrit fournit un accès classique au document *via* le catalogue. Le document est alors manipulé dans un contexte technique séparé et plusieurs fonds documentaires peuvent être supportés dans des environnements techniques différents. Il n'y a pas de changement fondamental pour la bibliothèque : les documents électroniques sont traités dans la chaîne documentaire classique et la gestion des documents électroniques est une nouvelle fonction séparée qui n'a pas d'impact direct sur l'organisation et le système de stockage existant.
Un pas supplémentaire vers une nouvelle architecture est franchi lorsque l'on veut gérer plus complètement des documents HTML en créant des métadonnées en Dublin Core dans les pages elles-mêmes. Les documents sont alors directement accessibles à la recherche dans une application Web grâce aux métadonnées et au texte intégral, sans la médiation du catalogue.
La bibliothèque peut vouloir également gérer l'information bibliographique de ces documents dans le catalogue traditionnel. Le choix nord-américain consiste plutôt actuellement à générer une notice MARC minimum à partir du Dublin Core et à compléter éventuellement cette notice minimum en format complet si des ressources humaines sont disponibles. Une telle approche introduit la notion de catalogage dans le document lui-même et dans le format du document, tout en gardant la possibilité de liens avec les fichiers d'autorité pour les points d'accès.
L'approche la plus élégante est bien entendu la troisième approche utilisant SGML pour englober dans un seul modèle homogène et puissant la collection complète des objets. Cette approche est très intéressante lorsque l'on souhaite décrire une collection du niveau général au niveau de l'unité documentaire, en respectant l'organisation de la collection et en offrant au niveau de la feuille de l'arborescence l'objet numérisé lui-même. Nous avons dans les formats bibliographiques tels qu'UNIMARC avec le mécanisme des liens ou tels que le CCF (Common Communication Format) avec les segments, des fonctionnalités permettant de fournir une information bibliographique hiérarchique en cascade, mais ces formats nous limitent aux systèmes de bibliothèque ou aux systèmes documentaires. L'utilisation de SGML ouvre les possibilités d'utilisation d'outils et de concepts facilitant les échanges avec d'autres communautés comme les musées, les chercheurs, l'information institutionnelle y compris la littérature grise, la documentation technique.
Participer à de telles expérimentations permet d'apprendre et d'ouvrir le monde des bibliothèques (le beau côté), mais c'est aussi investir dans la définition d'une nouvelle conception de l'organisation des collections et dans un nouveau système informatique (le côté sombre).
La seule certitude est que si une bibliothèque a investi dans un format professionnel normalisé et documenté tel qu'UNIMARC, elle pourra à tout moment convertir ce qu'elle a créé, mais seulement ce qu'elle a créé, dans un système de type SGML de gestion des instruments de recherche et des documents numérisés.

Avril 1998

BIBLIOGRAPHIE

« Digital libraries : cataloguing 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.