L'AVENIR DES FORMATS DE COMMUNICATION

HTML, SGML ET FORMATS BIBLIOGRAPHIQUES : DE L'INFORMATION A LA METAINFORMATION

Adel EL ZAIM
Directeur de l'unité Applications des Inforoutes (ADI)
(aelzaim@crim.ca)
et Sylvie TELLIER
Responsable du Centre de Documentation et du Service de Veille
(tellier@crim.ca)
Centre de Recherches Informatiques de Montréal Canada

INTRODUCTION

Constater le foisonnement de l'information sur l'Internet et la transformation de ce réseau d'un outil de communication réservé à un milieu restreint en un médium de communication grand public est devenu lieu commun. Par conséquent, les besoins de structurer l'information et de la rendre utilisable rapidement et de la façon la plus optimale possible ne se sont jamais autant fait sentir.

Pour répondre à ces besoins, il faut pouvoir représenter l'information de manière synthétique. Dans un milieu traditionnel d'information, le format bibliographique est un moyen répandu de représenter et de diffuser l'information.Quelques initiatives liées à l'Internet se rapprochent de cette façon de faire.Elles touchent la production de métainformation (ou information sur l'information) : les méthadonnées de HTML (HyperText Markup Language), TEI (Text Encoding Initiative), MCF (Meta Content Format), etc.Mais ces expériences semblent ramener le monde Internet vers SGML (Standard Generalized Markup Language), à l'origine de HTML, qui est conçu pour produire de la métainformation. Lequel de HTML ou de SGML l'emportera pour répondre à ces nouvelles exigences?

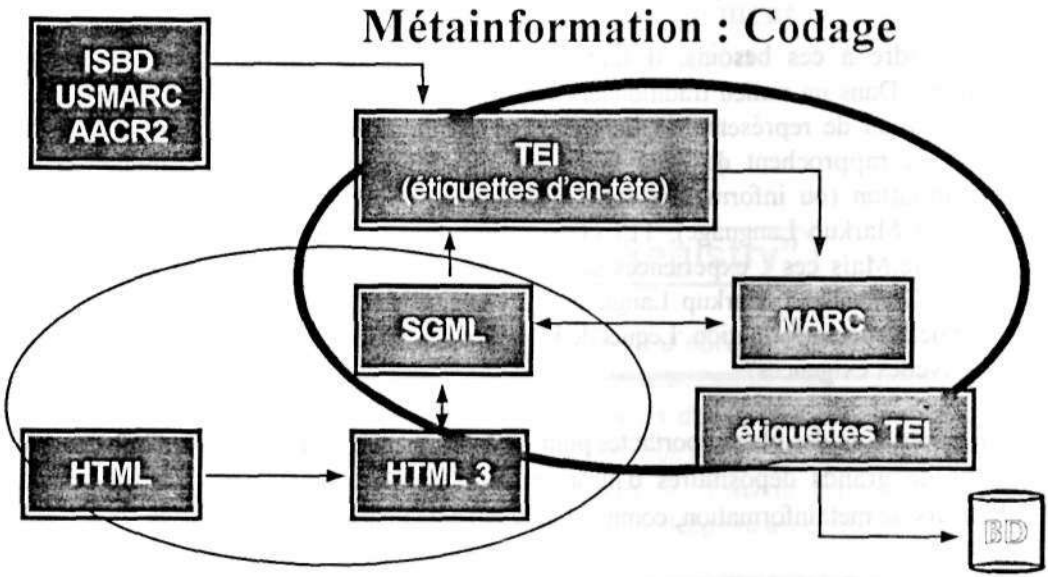
Ces questions sont des plus importantes pour les bibliothèques et les services d'information qui sont de grands dépositaires d'information, comme l'est Internet, et de grands producteurs de métainformation, comme doit le devenir Internet.

1. LES BIBLIOTHEQUES EN TANT QUE PRODUCTRICES DE METAINFORMATION

Les bibliothèques ont de plus en plus recours aux médias électroniques pour leurs produits et services. Le changement rapide des moyens de diffusion, le passage lent mais inéluctable du format papier au format électronique et l'introduction du réseau dans les bibliothèques incitent les professionnels à réorienter leurs services et leurs opérations vers l'information et la diffusion électronique.

Mais cette tendance, au lieu d'atténuer le problème le décuple; les documents électroniques ne sont certainement pas sur le point de remplacer les documents imprimés. S'ensuivent d'énormes problèmes de compilation et d'accès à l'information. D'où l'importance grandissante de la métainformation.

La métainformation est la représentation d'un document.Cette représentation se fait à trois niveaux : signalétique, analytique et référentiel. La description signalétique consiste à extraire des éléments généraux du document, tels l'auteur, le titre, l'éditeur, etc., pour en signaler l'existence.La description analytique consiste à prendre connaissance du contenu du document à l'aide du titre, de la table des matières, des têtes de chapitres, du résumé ou en faisant une lecture en diagonale et en condensant ainsi l'information sous forme de résumé ou d'une liste de mots clés.La fonction référentielle se fait, quant à elle, par le biais d'un autre document qui réfère au document traité.Elle se fait traditionnellement par l'inclusion de la description signalétique dans une bibliographie ou une liste de références. Elle se fait, avec les nouveaux moyens informatiques, de manière beaucoup plus dynamique à l'aide de l'hypertextuel.



Le milieu des bibliothèques est au fait des techniques de production de métainformation.L'Internet l'est de plus en plus puisque, plus encore là qu'ailleurs, le besoin s'en fait sentir.

2. FORMAT INTERNET ET FORMATS BIBLIOGRAPHIQUES

2.1 PEUT-ON PARLER DE FORMAT INTERNET ?

Les données diffusées sur le réseau Internet sont de formats divers qui ne sont pas tous conçus uniquement pour l'Internet. Que l'on pense, par exemple, au format ASCII (texte ou .txt) utilisé sur le réseau Internet entre autres pour le système d'information Gopher (1) et pour la diffusion de documents textuels par FTP (2), mais qui est aussi répandu sur les autres moyens de diffusion et de travail électroniques à commencer par les traitements de texte. Il en est de même des formats multimédias, comme le MIDI, le WAV, le GIF, le JPEG ou le QT, et des formats destinés à l'édition, comme le postscript, le LaTex, ou le PDF et dans une certaine mesure le SGML.Mais tous ces formats retrouvent une vocation nouvelle avec la possibilité de les afficher correctement sur le réseau Internet que ce soit à même le fureteur (browser) ou grâce à des applications indépendantes (3).

Parmi les formats expressément inventés ou adaptés pour le réseau Internet, on retrouve surtout le format HTML, utilisé dans les documents hypertextuels destinés à la diffusion par le WWW (World Wide Web).HTML est issu du SGML, Standard Generalized Markup Language, un système de définition de langages de marquage destiné aux documents textuels électroniques à des fins d'affichage et d'analyse (4).Le HTML n'est que l'un de ces langages; plus précisément, le HTML est une application du SGML.Le HTML est beaucoup plus simple, ce qui en a largement favorisé l'utilisation sur l'Internet.

2.2 Format Internet et formats bibliographiques : possibilités de convergence

L'Internet tend de plus en plus vers l'utilisation des bases de données pour structurer ses données.De ce fait, et à cause du besoin urgent de marquer le contenu de l'Internet pour y permettre une recherche, des groupes travaillent à l'établissement de modes de marquage de l'information Internet qui ressemblent à ceux des notices bibliographiques (auteur, date, mots clés, langue, etc.).

Paramètres/Formats	HTML	Formats bibliographiques
Type d'information	Information	Métainformation
Buts	Présentation de l'information	Compilation de l'information
Finalité	Données non structurées reliées par hyperliens	Données structurées
Diffusion	Electronique	Diverses structurées Diverses formes dont la forme électronique

2.2.1 HTML

Les versions premières de HTML ne présentaient aucune possibilité réelle de concordance avec les formats bibliographiques, aussi bien par rapport à l'information traitée que par rapport aux buts visés.

HTML doit dorénavant répondre à des besoins particuliers et parfois bien définis de catégorisation et de structuration de données à des fins de recherche et de repérage de l'information par des humains ou par des logiciels (ex. : robots) servant à indexer le contenu du réseau.

La dernière version de HTML (3,2) proposée par le W3C est développée conjointement avec de grandes industries: IBM, Microsoft,Netscape Communications Corporation, Novell, SoftQuad, Spyglass et Sun Microsystems. Les éléments de HTML 3.2 sont issus des sources de HTML+ et de HTML 3.0.Le W3C continue ses travaux pour l'ajout et le support d'extensions supportant les objets multimédias, les scripts, les feuillets de style, les modèles (Templates et Layouts), les formulaires et l'amélioration de l'impression des documents HTML ainsi que la reproduction des formules mathématiques (5).

Cette version (3.2) de HTML intègre des fonctions d'identification de métainformation pour l'instant limitées mais en devenir : au titre déjà présent se sont ajoutés la description du document et les mots clés.

2.2.2 TEI

Le TEI (Text Encoding Initiative) est un projet de développement de lignes directrices pour la préparation et l'échange de données informatisées.Les accomplissements du groupe comprennent entre autres la spécification d'une méthode de catégorisation de la documentation compatible avec les règles de catalogage et qui peut être utilisée pour retracer l'historique des documents permettant ainsi l'authentification de leur provenance et des modifications subies.

Le SGML a servi de base syntaxique à TEI

2.2.3 MCF

Le format d'échange de données MCF (Meta-Content Format) est en développement dans les Laboratoires Apple.Le but de MCF est de fournir un langage de représentation du contenu d'un large éventail d'informations.La particularité de ce format réside dans le fait que la métainformation n'est pas codée comme dans HTML ou SGML mais elle est automatiquement extraite et représentée sous le format MCF.Il est prévu que les éléments de métainformation suivants soient pris en compte : description, auteur, affiliation, date de publication hyperliens, langage, sujets, sites miroir, type de média (JPEG, MPEG, Postcript, Java Applet), etc.

2.2.4 Dublin Core

En mars 1995 se tenait à Dublin (Ohio, E-U.) le Metadata Workshop organisé par The Online Computer Library Center et le National Center for Supercomputing Applications. Une première constatation faite par les spécialistes invités est que l'information sur l'Internet est trop volumineuse pour pouvoir en assurer le traitement.Il a donc été décidé de définir des éléments de métainformation qui seraient par la suite mis à la disposition des auteurs et des fournisseurs de documents pour qu'ils puissent eux-mêmes décrire les documents. Ces éléments comprennent le sujet, le titre, l'auteur, l'éditeur, la date, le type, le format, l'identificateur, la langue, la couverture, etc.

Des développements ultérieurs sont prévus, notamment des discussions avec les groupes responsables des formats MARC et SGML.

2.2.5 SGML

Les initiatives d'identification de métainformation sur l'Internet se rapprochent de plus en plus de SGML, technologie stable et complexe (6).Dans la figure qui suit, SGML est au coeur des initiatives de production de métainformation que sont MARC, TEI et HTML. Par ailleurs, depuis peu, le milieu des bibliothèques adopte le format SGML pour l'exploitation des données bibliographiques à la fois pour les catalogues collectifs, la publication électronique en ligne, les documents de commande électronique et la publication pour le WWW sur l'Internet.

Cette adoption de SGML n'étonne pas considérant la liste impressionnante de ses avantages :

- intègre les formats MARC;
- intègre les formats non-MARC;
- permet la recherche en texte intégral;
- permet la navigation hypertextuelle;
- est bien adapté pour les données bibliographiques qui comprennent du texte et sont structurées;
- est indépendant de la plateforme et du logiciel;
- permet de réutiliser les données.

Mais pourquoi cette adoption vient-elle si tardivement?

- SGML est complexe et coûteux à développer.
- Un bon nombre de bibliothèques ont tardé à s'informatiser.
- Le milieu des éditeurs et celui des bibliothèques continuent à fonctionner sur un mode traditionnel.Mais, l'avènement de la bibliothèque numérique est sur le point d'entraîner l'acquisition de documents électroniques codés en SGML directement des éditeurs.

* Le marché des logiciels SGML se développe et des outils de plus en plus puissants sont disponibles.
* Avec le développement du World Wide Web, des solutions de coexistence de HTML et de SGML sont développées. Une première solution consiste à transformer des instances SGML en fichiers HTML.Une seconde solution consiste à stocker les instances SGML sur le serveur et en permettre la visualisation avec un SGML viewer.

2.3 EXEMPLES DE QUELQUES INITIATIVES D'UTILISATION DU SGML DANS LES DOMAINES DE L'INTERNET ET DES BIBLIOTHEQUES

Les cas que nous citons en exemple proviennent de deux mondes, celui de l'Internet qui se structure et celui des bibliothèques qui " s'électronisent ". Le trait commun entre ces deux mondes est l'omniprésence de SGML.

2.3.1 EXEMPLE DE LA RECHERCHE DANS LES SITES WWW DU RCT

Les données du site WWW du Réseau canadien des technologies (RCT) (7) dont les données contiennent des marques (tags) SGML utilisées comme descripteurs de la métainformation et qui sont réutilisées par le logiciel de recherche utilisé sur le serveur WWW de ce site. Les marques, qui décrivent et structurent des zones du texte, étant en SGML, les fureteurs (browsers) du WWW les ignorent tout simplement puisqu'ils ne les reconnaissent pas. Par contre, le logiciel de recherche OpenText les reconnaît et les prend en considération lorsqu'il répond aux requêtes structurées de recherche effectuées par l'usager du site WWW du RCT.

Page HTML (8) contenant des métadonnées et des marques SGML (9)

```
<HTML>

<!-- This HTML code generated by Tycho(tm), a Tranquility Base
 S0ftware Inc. product. gdignard@tranquility.com-->

 <HEAD>
 <TITLE>CNRC - Centre d'information du CNRC, biotechnologie</TITLE>
 <META NAME="CTN-Region" CONTENT="QC">
 <META NAME="CTN-Presence" CONTENT="Regional">
 <META NAME="CTN-Member-Type" CONTENT="Node">
 <META NAME="CTN-Language" CONTENT="French">
 <META NAME="CTN-Build-Date" CONTENT="9/30/96">
 </HEAD>
 <BODY bgcolor="#FFFFFF">
 <FONT SIZE=3>
 <A HREF="http://ctn.nrc.ca/ctn/rct.html">Page d'accueil du RCT</A>
 &#183;
 <A HREF="http://ctn.nrc.ca/ctn/apercu.html">Aper&ccedil;u du RCT</A>
 &#183;
 <A HREF="http://ctn.nrc.ca/ctn/ressourc.html">Ressources du RCT</A>
 &#183;
 <A HREF="http://data.ctn.nrc.ca/recherch.html">Recherche</A> &#183;
 <A HREF="http://ctn.nrc.ca/ctn/autre.html">Autres contacts</A> &#183;
 <A HREF="http://ctn.nrc.ca/ctn/nouvelle.html">Nouvelles</A> &#183;
 <A HREF="http://data.ctn.nrc.ca/comment.html">Commentaires</A>
 </FONT><P>
 <IMG SRC="http://ctn.nrc.ca/ctn/images/mem_actf.gif" VSPACE=20
 ALT="Membre actif">
 <OrgName><H2>CNRC - Centre d'information du CNRC,
 biotechnologie</H2></OrgName>
 <H4>
 Genre d'organisation:
 <OrgTypeID>
 <A HREF="/qc/navigate/bytype/type1/pg1-f.htm">
 Organisme f&eacute;d&eacute;ral</A>
 <OrgMission><DD>La mission de l'ICIST consiste &agrave; offrir de
 l'information scientifique, technique et m&eacute;dicale afin de
 contribuer &agrave; l'atteinte des objectifs &eacute;conomiques et
 sociaux du Canada.</OrgMission>
 <DT><H3>
```

On peut effectuer une recherche sur les sites du RCT à partir de la page http://data.ctn.nrc.ca/recherch.html.

Dans la page des résultats de recherche, dans la figure qui suit, les mêmes marques sont utilisées mais, cette fois, comme paramètres de recherche repris dans les réponses aux requêtes ou tout simplement comme éléments de marquage du texte.

Page de résultats de la recherche dans les sites du RCT (extraits)

```
<HTML>
<HEAD>
<title>The Canadian Technology Network (CTN) / Le Réseau canadien de
technologie (RCT)</title></HEAD>
<BODY bgcolor="#FFFFFF">
<FONT SIZE=3><A HREF="http://ctn.nrc.ca/ctn/rct.html">Page d'accueil
du RCT</A> &#183; <A
HREF="http://ctn.nrc.ca/ctn/apercu.html">Aper&ccedil;u du RCT</A>
&#183; <A HREF="http://ctn.nrc.ca/ctn/ressourc.html">Ressources du
RCT</A> &#183; <A
HREF="http://data.ctn.nrc.ca/recherch.html">Recherche</A> &#183; <A
HREF="http://ctn.nrc.ca/ctn/autre.html">Autres contacts</A> &#183; <A
HREF="http://ctn.nrc.ca/ctn/nouvelle.html">Nouvelles</A> &#183; <A
HREF="http://data.ctn.nrc.ca/comment.html">Commentaires</A> &#183; <A
HREF="http://data.ctn.nrc.ca/search.html">English</A></FONT><P>
<P>
[...]
<P>
Documents <B>1</B> &agrave; <B>1</B> de <B>1</B> documents renfermant:
<CODE><B>metal</B></CODE> <I>And</I> <CODE><B>Biotechnology /
Biotechnologie</B></CODE> in <I>Sector List</I> regions<BR><OL
START=1>
<LI><OTUserMeta><INPUT TYPE="checkbox" NAME="fax" VALUE=1>
<INPUT TYPE="hidden" NAME="OfferTitle" VALUE="Microwave Induced
Catalysis-Destruction of Organic Halides">
<INPUT TYPE="hidden" NAME="AnyOrganization" VALUE="Queen's
University">
<INPUT TYPE="hidden" NAME="AnyDivision" VALUE="PARTEQ Innovations
(R&amp;D)">
<INPUT TYPE="hidden" NAME="OfferDescription" VALUE="&lt;DD&gt;Most
organic chemicals do not interact significantly with microwave
radiation. Under the direction of Dr. Jeffery Wan, [...]
<DL><DT><b>Contact : </b><DD>
<DD>
John P. Molloy
<BR>Executive Director
<AnyTelephone><BR>Phone: (613) 545-2342</AnyTelephone>
<AnyFacsimile><BR>Fax:
```

2.3.2 EXEMPLES D'APPLICATIONS SGML DANS LES BIBLIOTHEQUES DE BELGIQUE

La norme SGML est facilement utilisable pour la description des données bibliographiques parce que ces données contiennent du texte et ont une structure logique (règles de catalogage) qui se prête à la description par une DTD (Document Type Definition ou Définition de type de documents).

Le CCB (10), ou Catalogue collectif de Belgique, est un projet d'échanges bibliographiques entre 40 bibliothèques universitaires et spécialisées; le catalogue collectif comprend 4 000 000 de notices tenant sur deux Cd-Rom.Ce projet a permis à toutes les universités belges d'échanger leurs données en format SGML (11).

Le catalogue Antilope, également en format SGML (12), est une banque de données de périodiques totalisant 75 000 titres avec environ 230 000 localisations dans environ 80 bibliothèques.

Des services d'acquisition et de dissémination sélective de l'information ont également été réalisés à l'aide de SGML.

2.3.3 AUTRES APPORTS DE L'INTERNET AUX BIBLIOTHEQUES

Les protocoles de l'Internet, surtout ceux de repérage de l'information par localisation unique, peuvent aussi être utilisés dans les formats bibliographiques.La conversion du MARC en SGML ou en HTML est un travail nécessaire lorsqu'on veut assurer une diffusion sur plusieurs supports ou tout simplement lorsqu'on veut publier sur Internet. Parmi les expériences et les initiatives à souligner, la Bibliothèque du Congrès américain collabore avec le Network Development et MARC Standards Office pour élaborer les Guidelines for the Use of Field 856. Ce champ utilisé pour l'information sur la localisation de ressources électroniques peut se comparer à l'URL de l'Internet. Le guide détaille les codes et leurs correspondances avec les protocoles de l'Internet. En voici un extrait :

The data in field 856 may be a Uniform Resource Locator (URL), which is recorded in subfield $u, or it may parse the necessary locator information into separate defined subfields. An access method, or protocol used, is given as a value in the first indicator position (if access method is email, ftp, telnet, or dial-up) or in subfield $2 (if access method is anything else, including http). The access method is also the first element of the URL. (http://www.loc.gov/marc/856guide.html)

CONCLUSION

La métainformation est une valeur ajoutée à l'information ; elle en permet la compilation et le repérage. Les bibliothèques produisent depuis toujours ce type d'information. L'Internet, pour sa part, commence à peine et la tâche n'est pas mince : on évalue à plus de 50 000 000 le nombre de pages sur l'Internet qui constituent autant de documents. Indexer cette masse d'information pour en faciliter l'accès aux usagers nécessite avant tout un travail de préparation de l'information qui pour le moment est encore fait après publication. L'ajout de métainformation est un travail inévitable.Et c'est dans ce sens que les mondes de l'Internet et des bibliothèques auront beaucoup à apprendre l'un de l'autre.

Le recours à SGML, ou à des formats similaires, semble de plus en plus aller de soi. Ce format offre une plate-forme conjointe où les besoins plus spécifiques peuvent être comblés tout en assurant l'harmonie et la complémentarité des formats Internet et formats bibliographiques. Mais le volume d'informations exige des techniques hautement automatisées.

Références Bibliographiques

· [CORT95] Jan Corthouts." The use of SGML in the VUBIS-Antwerpen library network ".In 2nd annual conference on the practical use of SGML. 1996.(Antwerp, October 25, 1995.)
http://www.bim.be/BeLuxweb/95/jcorthou.html

· [CORT96] Jan Corthouts et Richard Philips." SGML : A librarian's perception ".The Electronic Library, vol. 14, no. 2 (April 1996) : 101-110.

· [DESA] Bipin C. Desai .Report of the Metadata Workshop Dublin, OH.Montreal : Concordia University, Department of Computer Science.
http://www.cs.conçordia.ca/~faculty/bcdesai/metadata/metadata-work shop-report.html
Montreal, H4B 1R6, CANADA

· [GUHA] R.V. Guha.Meta-content format.Apple Computer.
http://www.atg.apple.com/go/ProjectX/mcf.html

· [IDEN95] Nancy M. Ide et C.M. Sperberg-McQueen."The TEI : History, goals, and future ".Computers and the humanity,vol. 29 (1995) : 5-15.

· [PFAF94] Bryan Pfaffenberger.Internet in plain english. New York : MIS Press, 1994.

· [SPER] C.M. Sperberg-McQueen et Robert F. Golstein. Html to the max a manifesto for adding SGML intelligence to the World-Wide Web.
http://www.uic.edu/~cmsmcq/htmlmax.html

· [WEIB95] Stuart Weibel." Metadata : The foundations of resource description ".D-Lib Magazine, (July 1995) : 1-8.
http://www.dlib.org/dlib/July95/07weibel.html

Notes :

1. Système d'information basé sur la diffusion de documents textuels accessibles par un menu et des sous-menus.

2. FTP (File Transfert Protocol). Un exemple des fichiers textuels diffusés par ce moyen sont les fichiers de type Readme ou apropos contenant habituellement des instructions et des consignes.

3. Helper Applications : logiciels lecteurs de certains formats qui prennent la relève du fureteur lorsque celui-ci reçoit un document formaté dans un format qu'il ne supporte pas.

4. Bryan Pfaffenberger, Internet in plain english, New York, MIS Press, 1994.

5. HTML 3.2 Reference Specification, W3C Working Draft 09-Sep-1996, http://www.w3.org/pub/WWW/TR/WD-html32.html

6. A Lexical Analyzer for HTML and Basic SGML, W3C Working Draft 15-Jun-96, Dan Connolly , http://www.w3.org/pub/WWW/TR/WD-sgml-lex/

7. Réseau canadien des technologies, http://ctn.nrc.ca/

8. Fiche descriptive du Centre d'information du CNRC, biotechnologie, membre du RCT au Québec, http://data.ctn.nrc.ca/qc/content/type1/org521/parent.htm

9. Identifiées en gras dans les sources.

10. Le CCB, ou Catalogue collectif de Belgique : http://www.libis.kuleuven.ac.be/libis/ccb/index.html

11. La DTD du CCB se trouve à l'adresse : ftp://lib.ua.ac.be/pub/ccb/ccb.dtd

12. La DTD du catalogue Antilope est disponible à l'adresse : ftp://lib.ua.ac.be/pub/antilope/atp/atp.dtd