

L'information secondaire du document primaire

Format MARC ou SGML ?

par Catherine Lupovici

Responsable des activités bibliothèque chez Jouve¹

L'acquisition et le traitement de documents électroniques conduisent les bibliothèques et organismes documentaires à réexaminer et comparer les méthodes et les outils de traitement de l'information. Cette réflexion sur l'impact des documents électroniques dans la

chaîne de traitement de l'information s'applique nécessairement à un des composants de la chaîne qui est l'information secondaire comme médiation pour accéder à l'information primaire elle-même. Une des conséquences est la comparaison des formats pour la codification de la

structure de l'information primaire et secondaire. Nous examinerons donc successivement les formats professionnels bibliographiques et les formats profes-

1. *clupovici@jouve.fr*
Jouve Systèmes d'Information

sionnels de documents. Quelques exemples de projets pilotes expérimentaux permettront de comprendre les évolutions stratégiques en cours pour une intégration plus forte des fonctions de gestion des documents et de valeur ajoutée documentaire.

Structure de l'information secondaire

L'information secondaire est créée pour permettre un meilleur accès à l'information primaire contenue dans des documents. Elle est composée de grands types de données plus particulièrement destinés à des actions telles que :

- l'identification des documents ;

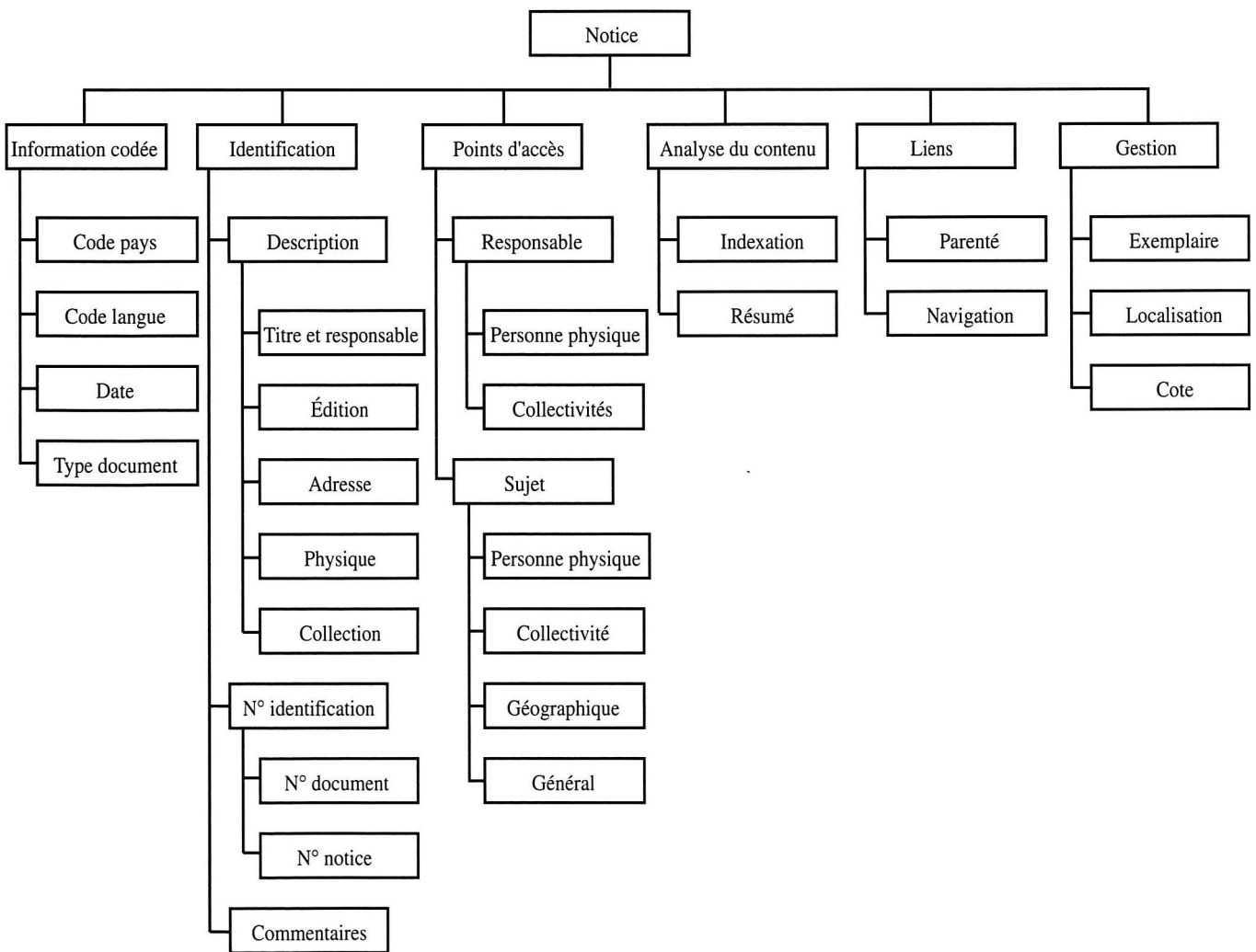
- le signalement de documents, organisé de manière thématique ;
- la recherche d'information sur un sujet ;
- la gestion des documents complets.

L'identification s'effectue à partir des numéros d'identification et/ou de la description bibliographique. Le signalement thématique est construit sur les points d'accès qui permettent de classer les descriptions. La recherche d'information sur un sujet utilise les données de l'analyse de contenu. Elle est complétée, dans les outils les plus récents du WWW, par la navigation s'appuyant sur des liens entre des notices, à l'intérieur d'un même catalogue, ou dans différents catalogues selon le concept actuel d'hypercatalogue. Les données de gestion permettent l'accès au document physique rassemblant l'information primaire décrite par l'information secondaire.

Avec l'informatisation du traitement de l'information secondaire et conformément aux possibilités de l'informatique de gestion des années 1970, on a pris l'habitude d'ajouter également des informations codées pour faciliter les traitements en machine et rajouter des possibilités de recherche et de tri des notices.

L'information secondaire a une structure hiérarchique qui permet de dérouler les grandes classes de données bibliographiques et documentaires correspondant aux différents usages que l'on peut en faire. Il s'agit principalement des grands domaines d'application traditionnels que sont les bibliothèques et les organismes documentaires.

L'arbre de cette classification a au niveau très général la structure suivante :



Cette arborescence peut être dépliée jusqu'aux feuilles qui représentent les données élémentaires constitutives de l'information secondaire et qui correspondent au niveau des sous-zones de formats de type MARC.

Structure du document textuel : la norme SGML

Le secteur de production de l'information primaire est lui aussi très fortement informatisé et pratique une codification en machine du contenu et de la structure des documents. C'est la codification de la structure qui permet de produire la mise en forme typographique et la mise en page du document papier publié. La création de l'information secondaire descriptive est une extraction et un typage des données primaires qui s'appuie encore aujourd'hui sur une interprétation de la présentation typographique et du contenu. Par exemple, la présentation sur la page de titre permet au catalogueur de désigner le titre principal, le complément du titre, les différentes mentions de responsabilité avec leur importance relative (première, deuxième, etc.). C'est d'ailleurs au travers de la normalisation de la présentation des publications que s'est faite pendant plusieurs décennies la normalisation des éléments de données qui doivent servir à l'identification des publications. Les groupes de normalisation internationaux et nationaux en Information et Documentation se partagent encore en deux grands sous-comités : l'informatisation dont les formats et les protocoles d'un côté et la présentation et numérotation de l'autre.

La présentation d'une publication permettant un catalogage descriptif de bon niveau découle donc, *via* la visibilité de la présentation résultante, d'un codage de structure préexistant de niveau au moins équivalent. Il y a corrélation entre le format de catalogage descriptif et le format de publication.

Il existe depuis dix ans une norme internationale d'écriture de formats de documents qui a son origine dans le monde de l'édition. La norme SGML (*Standard Generalized Markup Language*) ISO 8879², publiée en 1986, est à l'origine un langage permettant de décrire la structure logique de classes de documents. Elle correspond à la normalisation de pratiques de codage de fichiers

2. SGML (Langage normalisé de balisage généralisé), ISO 8879, 1986. NF EN 28879, 1990.

de documents pour que les éditeurs puissent faire circuler les informations dans la chaîne de production de manière indépendante des prestataires, des matériels et des logiciels. Cette norme, initialement destinée à permettre une meilleure production du document textuel papier, est rapidement devenue, grâce aux produits commerciaux qui l'ont accompagnée, un outil de saisie du document en base de données structurées. Elle est en train de devenir la norme universelle de production de documents d'autant plus qu'elle sert de base à HTML (*Hypertext Markup Language*), la norme de codage des pages constitutives des documents des services WWW sur Internet ou dans les applications Intranet. La norme SGML, tout comme la forme simplifiée HTML, est potentiellement utilisable pour la distribution de tout ou partie de documents électroniques. Les outils SGML du marché évoluent d'ailleurs d'outils de saisie vers des outils de gestion de documents électroniques. C'est le cas par exemple de Basis SGML Server.

Le format SGML

La structure SGML d'une classe de documents, par exemple les livres, les thèses, les articles dans les publications scientifiques, est définie dans une DTD (Définition de Type de Document), dans laquelle toute la structure générique de la classe de documents est décrite en langage SGML.

Un document codé en SGML contient des balises conformes à la DTD à laquelle il fait référence et dont il est une *instance*. Cela signifie que les balises sont définies dans la DTD mais aussi que leur utilisation en terme de structure hiérarchique arborescente est strictement définie et contrôlable par la DTD.

Les outils de production de documents SGML comportent de façon généralement intégrée :

- un *parseur* ou programme permettant de vérifier qu'une DTD est écrite réellement en SGML et qu'une instance de DTD est bien conforme à cette DTD ;
- un *éditeur* SGML qui est un outil de saisie structurée facilitant la saisie des balises dans un document.

La DTD

La DTD d'une classe de documents permet de définir tous les *éléments* logiques qui peuvent composer les documents de cette classe. Les éléments composites sont décrits par les éléments qui les composent à tous les niveaux de la hiérarchie de la structure. Par exemple l'élément groupe d'auteurs sera composé des auteurs personnes physiques et des auteurs collectivités. L'élément personne physique sera composé du nom, du prénom qui sont eux des éléments terminaux de la structure. En langage SGML ces éléments s'écriront dans la DTD de la manière suivante :

```
<ELEMENT nom
<ELEMENT prénom
```

Des règles d'écriture permettent de donner des règles d'occurrences et de succession entre les différents éléments constitutifs d'un élément composite.

Les éléments peuvent être qualifiés par des *attributs*. On peut par exemple attacher un attribut de code de langue à un élément titre.

Des entités externes peuvent être ancrées de manière codifiée à un endroit précis d'un document SGML. Ces entités peuvent être définies dans la DTD. Elles peuvent par exemple consister dans une chaîne de caractères qui remplacera l'élément du document lors de son utilisation (par exemple développement systématique d'un sigle). Il peut aussi s'agir de l'ancrage d'une illustration attachée à un ou plusieurs endroits d'un texte.

Ce mécanisme d'appel d'entité peut permettre de créer des liens hypertextuels dans un même document ou entre deux documents stockés éventuellement sur des serveurs différents. C'est sur ce principe que s'appuie la technique des URL (*Uniform Resource Locator*) des documents HTML dans les applications WWW.

Les DTD normalisées

Très tôt dans l'histoire du développement de SGML on a songé à normaliser des DTD génériques. Dès 1984 un projet américain regroupant des éditeurs, des imprimeurs, des auteurs, des bibliothèques dont la Bibliothèque du Congrès

et des bases de données ont travaillé à la normalisation de DTD dans le but de les utiliser comme langage d'échange dans la chaîne de création et de distribution de manuscrits électroniques. Cette initiative a conduit à la norme ISO 12083 «Préparation et balisage de manuscrits électroniques» dont les versions anglaise et française ont été publiées fin 1994. Cette norme fixe quatre DTD génériques pour :

- les livres,
- les publications en série,
- les articles,
- les formules de mathématiques.

Après avoir été considérée comme un exercice un peu théorique, cette norme commence à être envisagée comme la base d'échange de documents électroniques dans les domaines de l'édition scientifique, technique et médicale. Les produits commerciaux de saisie de documents SGML l'offrent dans les DTD de base.

Des extensions sont en cours de développement au niveau international, en particulier une DTD des références bibliographiques est un sujet de travail retenu par l'ISO.

HTML (Hypertext Markup Language)

Le format HTML est une DTD SGML simplifiée quant à la structure du document. Il gère des niveaux de titre, des paragraphes de texte et des structures de liste, comme le fait une feuille de style de traitement de texte de base. HTML est par contre très développé sur la gestion des liens hypertextes. Cette DTD correspond au besoin d'affichage et de structuration des documents présentés dans les applications WWW. C'est la base de la construction des « pages Web » et du mode d'affichage des textes correspondants sur le Web.

Pour les publications scientifiques, cette structure n'est pas assez puissante pour gérer correctement les tableaux et les formules de mathématiques et son évolution essaie de prendre en compte les besoins parfois contradictoires des différents groupes d'utilisateurs tout en conservant le principe initial de simplicité sur lequel est fondée l'universalité de l'utilisation de HTML.

Dans les applications WWW les liens mis en œuvre permettent de lier des documents

à des objets externes, éventuellement distants qui peuvent être des objets multimédias, des programmes, d'autres applications Internet (WAIS, FTP, Gopher, etc.).

Les utilisations bibliographiques de SGML

Les informations bibliographiques peuvent être créées par les différents intervenants de la chaîne de création et de distribution des documents. Les auteurs par exemple ont toujours, en suivant les consignes rédactionnelles plus ou moins élaborées de leurs éditeurs, saisi des informations bibliographiques ou potentiellement bibliographiques, telles que les parties d'en-tête de leurs documents et la bibliographie intégrée dans le document. Les parties dites d'en-tête recouvrent le titre, les mentions d'auteur, les mots clés et les résumés pour un article scientifique par exemple. Les éditeurs contribuent tous d'une manière ou d'une autre au signalement commercial de leur publication qui comprend une partie d'identification bibliographique qui pourra également servir à commander le document. Les centres documentaires qui font de l'archivage numérique pour fournir les documents à la demande ont tous besoin d'un minimum d'information bibliographique attachée aux documents stockés pour pouvoir retrouver les documents. Enfin les bibliothèques ont dès le début de l'informatisation de leurs catalogues fixé les formats de l'information bibliographique utilisable en machine. Dans tous ces domaines, des réflexions sont en cours pour définir une utilisation de SGML comme norme d'échange des informations bibliographiques soit entre les différentes catégories d'intervenants d'un processus de publication et de distribution, soit entre les intervenants d'une même catégorie. Les exemples suivants illustrent cette réalité.

Les publications sur Internet

Le format HTML permet de définir et de coder des éléments qui contiennent des informations sur les données et qui pourront être utilisées pour indexer et effectuer des recherches d'information sur le

réseau en utilisant les moteurs de recherche de ressources sur Internet.

Ces *metadata*³ intéressent toutes les communautés d'utilisateurs d'Internet qui souhaitent mettre un peu d'ordre et faciliter l'accès à l'information et qui normalisent certaines *metadata* à l'usage de leur groupe.

C'est ainsi qu'en 1995 un premier atelier de réflexion s'est réuni à Dublin, Ohio, sponsorisé par OCLC et le NCSA (National Center for Supercomputing Applications). Les participants étaient des bibliothécaires, des archivistes, des chercheurs en sciences humaines et en géographie, des experts en normalisation dans les domaines d'Internet, de Z39.50 et de SGML. L'objectif était de définir une liste d'éléments *metadata* permettant une description simple de l'information électronique et qui puisse être utilisée dans différents contextes, en particulier par les auteurs et les éditeurs, c'est-à-dire le plus possible à la source de la publication.

Cette liste de treize éléments porte le nom de Dublin Core (DC) et continue à être examinée. Une seconde réunion internationale a eu lieu en avril 1996 à Warwick⁴. D'autres groupes de travail prolongent cette initiative comme le Nordic Metadata Project qui, entre autres, souhaite étudier la conversion entre le Dublin Core et les formats MARC nordiques. La Bibliothèque du Congrès a déjà également travaillé à la conversion DC vers USMARC. Un champ a d'ailleurs été rajouté pour tenir compte de données DC qui n'avaient pas d'équivalent USMARC.

La démarche des éditeurs

Les éditeurs scientifiques, techniques et médicaux participant au European Workgroup on SGML (Elsevier Science Publishers, Kluwer, Springer, Thieme, etc.) ont dès le début de la décennie travaillé à la mise en œuvre d'une DTD commune pour la structure de l'en-tête des articles scienti-

3. Metadata resources/IFLA. <http://www.nlc-bnc.ca/ifla/II/metadata.htm>

4. Warwick framework and Dublin Core set provide a comprehensive infrastructure for network resource description : report from the Metadata Workshop II, Warwick, UK, April 1-3, 1996 / Juha Hakala, Ole Husby and Traugott Koch. <http://www.bibsys.no/warwick.html>

fiques. L'en-tête couvre les notions de titres, d'auteurs personnes physiques et collectives, d'affiliation des auteurs, de mots clés et de résumé. Cette DTD commune a été publiée par l'éditeur Springer Verlag en 1991 sous le nom de MAJOUR (Modular Applications for Journals). L'objectif initial est, pour ces éditeurs qui transforment tous leur chaîne de production en SGML, d'être en mesure de gérer leur base de données d'articles pendant tout le processus d'acceptation et de publication. Évidemment, ils deviennent aussi potentiellement fournisseurs de ces informations signalétiques pour les bases de données, les bibliothèques et centres de documentation d'autant plus facilement qu'ils utilisent un format normalisé. C'est ainsi qu'Elsevier par exemple propose dans ses services de distribution électronique CAPCAS⁵, service de fourniture de l'en-tête de l'article en SGML, selon la DTD d'Elsevier adaptée pour les besoins de production de la DTD de l'Association of American Publishers (AAP) devenue ISO 12083. Elsevier a commencé une démarche commerciale pour la commercialisation de cette information dont l'intérêt a été mesuré lors de projets de recherche avec des bibliothèques telles que TULIP⁶ ou ELSA⁷.

La gestion du document électronique

Le programme GRISELI⁸ a été initialisé en 1993 par le ministère de l'Enseignement supérieur et de la Recherche, DISTB. Il a pour objectif de mettre en place un dispositif cohérent de collecte, de traitement et de diffusion de la littérature grise française. Il associe des organismes tels que l'INIST, responsable des études préliminaires, l'INRETS, le CEA, l'INRIA. GRISELI se compose de deux axes complémentaires : le circuit des références documentaires et le circuit des documents. SGML a été choisi comme format support pour les échanges entre les partenaires, y compris pour le circuit des références documentaires.

5. <http://www-east.elsevier.com/ees/capcas.htm>

6. TULIP (The University Licensing Program) final report. <http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml>

7. ELSA (Electronic Library SGML Applications). <http://www.elsa.dmu.ac.uk>

8. « L'édition électronique : dossier », *Normatique*, janvier 1995, n° 63, 12 p. (Contient des contributions sur SGML, ISO 12083 et GRISELI.)

La structure d'information étudiée dans GRISELI pour la gestion des documents électroniques se retrouve dans la plupart des projets de gestion de documents, y compris de documents numérisés, comme le National Digital Library Program nord-américain où l'on retrouve aussi SGML pour la structure des textes historiques.

La démarche la plus globale consiste à concevoir une structure SGML enveloppe pour les références documentaires et les documents complets, sachant que le document lui-même peut être du texte balisé en SGML ou une reproduction numérique du document textuel ancré dans un en-tête descriptif à l'aide du codage SGML.

Dans le cadre de GRISELI, les DTD ISO 12083 ont été étudiées pour supporter les échanges bibliographiques et documentaires entre les partenaires.

Les catalogues de bibliothèques

Plusieurs projets nord-américains étudient le reformatage USMARC/SGML dans les deux sens afin de pouvoir échanger des données de type ISO 2709 en SGML.

La bibliothèque de l'université de Californie, Berkeley, a développé et mis à disposition gratuitement sur Internet une DTD du format USMARC en cours de test depuis avril 1994. Cette DTD a pour objectif de permettre de travailler en USMARC sur un système de gestion des notices bibliographiques fondé sur SGML. Elle permet de gérer l'intégralité du format USMARC jusqu'au niveau des sous-champs. À Berkeley, elle permet la conversion d'USMARC en SGML et inversement. Les programmes de conversion développés à Berkeley sur différentes plates-formes Unix sont également disponibles sur le réseau avec la DTD⁹.

La DTD en est à sa huitième version. Elle a été étendue pour gérer les alphabets grecs et cyrilliques ainsi que les jeux de caractères utilisés dans le monde de l'édition puisque l'un des intérêts majeurs d'utiliser SGML pour l'information secondaire est de pouvoir gérer en même temps le document électronique, si pos-

9. <ftp://library.berkeley.edu/pub/sgml/marcdtd/>

sible venant directement de l'éditeur. Elle a également été étendue pour couvrir les versions plus anciennes d'USMARC ainsi que les évolutions importantes en cours pour l'intégration des formats de chaque type de document en un format unifié. Elle a également étendu le format pour la gestion des champs de données locales.

La Bibliothèque du Congrès a effectué un travail monumental qui permet la conversion dans les deux sens de données définies selon le format USMARC complet de ISO 2709 vers ISO 8879 (SGML) en utilisant le jeu de DTD MARC proposées. L'ensemble des cinq formats USMARC est couvert par une DTD bibliographique et une DTD autorités. La DTD bibliographique couvre également le format USMARC état de collection, la DTD autorité couvre également le format classification. L'ensemble des DTD et codages de caractères est disponible sur le réseau¹⁰. Des programmes de conversion sont annoncés.

Conclusion

Ces quatre exemples montrent une convergence des formats d'échanges vers le monde SGML d'autant plus forte que l'on se trouve en situation de gérer les documents électroniques en même temps que l'information secondaire. Des passerelles permettant le reformatage entre le monde MARC et le monde SGML sont en cours de réalisation, actuellement seulement pour le format USMARC. Elles sont directement utilisables par toute bibliothèque qui s'appuie sur un format normalisé et qui dispose d'un système non seulement capable de recevoir un tel format sans particularités locales trop marquées, mais aussi et surtout capable de l'exporter. Au-delà du simple reformatage de structure, la définition des données élémentaires elles-mêmes est à nouveau en cours de discussion entre les producteurs et les utilisateurs des documents électroniques. Le contenu même du catalogue peut être revu pour distinguer ce qui est purement descriptif qui est repris dans le document de ce qui est une création d'information supplémentaire et qui semble davantage l'essence du documentaire et du bibliographique à valeur ajoutée.

10. <gopher://marvel.loc.gov/11/ftppub/marcdtd>