Les tables de matières dans les catalogues en ligne

Opportunités, méthodes et coûts

par Majid Ihadjadène*

La nécessité d'enrichir les catalogues en lignes

Depuis une vingtaine d'années, des bibliothécaires et des spécialistes en recherche de l'information ont effectué des études sur l'opportunité d'enrichir les notices bibliographiques par les tables de matières. Du projet novateur SAP (*Subject Access Project*) effectué en 1977 [cf. la bibliographie 1], à l'étude de PALINET 1996 [16], une cinquantaine de travaux ont été publiés sur ce sujet [10]. Nous pensons qu'il y a trois raisons principales qui justifient l'enrichissement des catalogues en lignes par les tables de matières :

1. Le problème de l'accès sujet dans les catalogues

La difficulté de la recherche par sujet dans les catalogues en lignes est bien connue.

L'un des facteurs qui influe d'une façon importante sur la pertinence de l'accès sujet est l'insuffisance du nombre de vedettes matières par notice. Alors que dans les bases de données bibliographiques, le nombre des descripteurs par notice se situe entre 10 et 20 ; dans les OPACs il est inférieur à deux. Sur un échantillon de 900 notices bibliographiques du catalogue de l'ENSSIB, nous avons trouvé 1,47 de vedettes par notices :

type de subdivisions	nombre	pourcentage
a	490	37%
a+1subdivision	627	47.4%
a+2 subdivisions	184	13.9%
a+3 subd. et plus	22	01.7%
Total	1323	100%

- a = tête de vedette

- nombre moyen de vedette par notice : 1.47

- nombre moyen de mots par vedette matière : 3.1

Deux directions sont envisagées pour atténuer ce problème : la première est l'enrichissement des catalogues par des tables de matières, des résumés et des notes. La seconde solution consiste à augmenter le nombre de vedettes matières par notice.

2. Le phénomène de butinage

Le butinage, qui était déjà fort prépondérant quand il y avait les catalogues en papier, l'est toujours avec l'introduction des OPACs [7]. Lipetz [12] a constaté que les utilisateurs se servaient des catalogues imprimés pour savoir ou localiser physiquement les livres susceptibles de les intéresser. C'est souvent en naviguant sur les étagères et rayons des bibliothèques et en ayant le livre dans les mains, que les lecteurs décident si les livres trouvés correspondent à leurs besoins.

Le catalogue n'est plus un outil de recherche, mais un outil de localisation. Selon cet auteur l'une des raisons qui explique cette pratique est le manque d'informations qui donnent un aperçu global du document.

Cette pratique n'a pas évolué avec l'introduction des catalogues en lignes. Les usagers considèrent toujours le butinage dans les étagères comme une stratégie de recherche principale.

3. Les œuvres de collaboration et le problème de l'indexation

Une œuvre de collaboration est un ouvrage qui contient deux ou plusieurs parties écrites par plusieurs auteurs. Ce type d'ouvrages est en général mal indexé.

Les livres contiennent beaucoup de parties qui ne sont pas pleinement représentées à travers les points d'accès des catalogues. Pourtant, souvent ces parties sont succinctement décrites dans les tables des matières. Identifier ces parties devrait four-

* ENSSIB/CERSI.

nir un point d'accès supplémentaire à l'information [15][18].

Sur un corpus de 4 098 de livres, Hoffmann [9] indique qu'approximativement un livre sur cinq est une œuvre de collaboration.

Études	échantillon	pourcentage
Hoffmann	4098	21.3 %
Weintraub	375	12 %
Poulsen	887	24 %
Poulsen	698	18 %

Comme ces études concernent soit des bibliothèques publiques, soit des bibliothèques universitaires, nous avons voulu connaître le pourcentage d'œuvres de collaboration dans une bibliothèque spécialisée comme celle de l'ENSSIB.

Nous avons analysé 420 livres concernant les domaines de l'informatique documentaire et celui de l'économie de l'information.

Nous avons trouvé que 109 livres sont écrits par plusieurs auteurs (soit presque 26% de l'échantillon). Le nombre moyen de parties par document est 18.

Certains auteurs signalent que divers éléments, tels que la langue de publication, la date de publication, la politique d'acquisition suivie, influent sur le nombre d'œuvres de collaboration.

En tenant compte de ces critères, on peut penser que le nombre moyen d'œuvres de collaboration dans une collection varie de 12 à 20 %.

Méthodologie d'enrichissement

Pour initier un projet d'enrichissement, il est nécessaire d'aborder ces trois questions :

— Quelles sont les notices bibliographiques à enrichir ?

— Quelle est la méthode d'enrichissement à suivre ?

— Quel est le coût de cet enrichissement ?

1. Critères de sélection

Pour des raisons économiques, il est nécessaire pour chaque bibliothèque d'établir un ensemble de critères pour le choix des livres à enrichir. Ceux-ci sont fonction du fonds de chaque bibliothèque, de la politique d'acquisition et de catalogage.

Il nous semble intéressant de montrer que suivant les critères choisis pour décider si les ouvrages doivent être enrichis ou non, le taux de livres candidats à l'enrichissement varie considérablement. Ainsi, à l'université de Carnegie Mellon [14], ce taux est de 7,85 % alors que dans le projet ESP de la bibliothèque ADFA [3] [4] et de Weintraub [18], ils sont respectivement de 25 % et 23 %.

Voici un exemple de critères choisis :

— livres écrits par différents auteurs,

— catalogues d'exposition avec au plus 25 artistes cités,

— si les titres de chapitre donnent des informations supplémentaires à celles des vedettes matières,

— anthologie,

— conférences.

On n'enrichira pas les notices en fonction des critères d'exclusion suivants :

— comptes rendus scientifiques, séries, atlas, archives, dictionnaire, répertoires de livres ou de personnes, brochures, guides, annuaires, annales, bibliographies,

— certaines conférences techniques,

— livres pour lesquels les mots clés et vedettes existantes sont suffisants,

— livres très spécialisés souvent bien décrits par leurs titres.

Définir les livres candidats à l'enrichissement dépend essentiellement de la collection étudiée, mais il nous semble raisonnable de penser qu'environ 20% d'une collection peut être enrichie.

2. Méthode d'enrichissement

Deux méthodes sont utilisées, l'une est manuelle, l'autre est semi-automatique.

La première méthode consiste à extraire des mots ou phrases des tables de matières et de les inclure dans les notices bibliographiques correspondantes. Dans le projet ESP (*Enriched Subject Program*) de la bibliothèque de l'ADFA (Australian Defense Force Academy), par exemple, l'enrichissement de 40 000 notices (sur une période de 5 ans) a généré presque 1 500 000 mots clés. Les termes extraits des tables de matières ne remplacent pas les vedettes matières mais les complètent.

Au lieu d'extraire des termes des tables de matières et de les inclure dans les notices, des projets plus récents, notamment RIDDLE (*Rapid Information Display and Dissemination in a Library Environment*) [8] et PALINET/MONO-TOC tentent d'automatiser une partie ou la totalité de la chaîne de traitement des tables de matières.

Bibliographie

- 1. Atherton Cochrane Pauline, 1978: « Books are for use ». *Final report of the Subject Access Project to the Council on Library Ressources*, 1978. Syracuse, N.Y.

- 2. Beatty Sue, 1991. : « ESP at ADFA after five years », Cataloguing Australia 17(3/4), 65-92.

- 3. Beatty Sue, 1992. « Subject Enrichment Using Contents or Index Terms : The Australian Defence Force Academy Experience» , in : *Advances in Online Public Access Catalogs*, vol.1, Wesport, CT : Meckler, 1992, 93-113.

- 4. Byrne Alex, Micco Mary, 1988 : « Improving OPAC subject access : The ADFA experiment », College & Research Library, 1988 sept., 49, 432-441.

- 5. Dillon Martin, Wenzel Patrick, 1989 : « Enhanced bibliographic record retrieval experiments », OCLC newsletter, 181, pp 13-14.

- 6. Dillon Martin, Wenzel Patrick, 1990 : « Retrieval effectiveness of enhanced bibliographic records », Library Hi Tech, 1990, 31(3), 43-46.

Le projet européen RIDDLE étudie la faisabilité d'utiliser des techniques de numérisation pour capturer les sommaires de journaux scientifiques et de les insérer dans un catalogue en ligne. Nous avons identifié quatre étapes importantes :

— numérisation des sommaires en mode image,

— identification du périodique,

— balisage SGML,

— incorporation des sommaires ainsi numérisés dans le catalogue.

Le projet PALINET/MONO-TOC a pour ambition de définir une chaîne de production des tables de matières de monographies. Parmi les aspects étudiés dans ce projet, on trouve les questions relatives au problème du droit d'auteur, le calcul des coûts de production et de vente, le contrôle de qualité ainsi qu'une étude de marché sur l'existence de clients pour ce type de service.

La chaîne de numérisation du projet PALINET/MONO-TOC est composée de quatre étapes principales :

— réception des photocopies des tables de matières et des titres,

— élimination des doublons,

— numérisation en mode image et OCR,

— édition et balisage.

Les tables de matières ainsi numérisées sont envoyées aux bibliothèques participantes dans un format compatible avec ceux de leurs catalogues.

3. Le coût

L 'un des éléments importants d'un projet d'enrichissement de catalogue en ligne avec des tables de matières est l'estimation du coût . Celui-ci dépend essentiellement du facteur temps.

Pour l'extraction des mots des tables de matières (méthode manuelle), la majorité des études que nous avons consultés montre qu'il varie entre 15 à 20 mn.

Ainsi, dans le cas du projet ESP, on aboutit à ces résultats :

photocopie	1,5 mn
choix des termes	8,5 mn
saisie/correction	8,5 mn

En ce qui concerne la numérisation totale des tables de matières, l'étude effectuée dans le projet PALINET/MONO-TOC montre que le temps nécessaire pour le traitement de 1 000 tables de matières est de 332 heures, c'est à dire environ 20 mn par titre.

Une fois les tables de matières numérisées, c'est la fonction édition (correction d'erreurs et mise en page) qui est la plus longue. Ce coût varie d'une bibliothèque à une autre et est fonction de plusieurs facteurs (matériels, personnels,...). On peut estimer ce coût entre 20 F et 30F par table des matières.

Les initiateurs du projet RIDDLE ont établi un ensemble de critères permettant de comparer le coût des méthodes manuelles et celui des méthodes semi-automatiques.

Tables des matières et recherche d'information

1. Tables de matières et recherche d'informations

Les parties d'un document ont chacune plus ou moins de valeur pour la recherche d'informations. Afin d'évaluer la pertinence des tables de matière pour la recherche d'informations, la majorité des études d'évaluation que nous avons

• 7. Hancock-Beaulieu Micheline, 1990 : « Evaluating the impact of an online library catalogue on subject searching behaviour at the catalogue and at the shelves », *Journal of Documentation*, 1990 decembre, 46(4), 318-338.

• 8. Harrisson A.D., Roos F.A., Thomas R.E., 1995 : « (Semi) automatic capturing of biliographic information from journal contents pages for inclusion in online library catalogues : the RIDDLE Project », *The Electronic Library*, vol. 13, n°1, février 1995.

• 9. Hoffman H , 1985 : « Futur outlook: better retrieval through analytic catalogs », *The Journal of Academic Librarianship*, 11, 151-153.

• 10. Ihadjadène M., Porte K, « OPACs et tables de matières », *Rapport de recherche*, ENSSIB, 1996.

• 11. Lin Xian, 1996 « Graphical Table of Contents », in : *Proceedings of ACM DL'96.*

• 12. Lipetz B, A, 1972: « Catalogs use in a large research library », *Library Quarterly*, 42, 129-139

• 13. Knutson Gunnar S., 1991 : « Subject enhancement : Report on an experiment », *College & Research Libraries*, 1991 janvier, 52(1), 65-79.

• 14. Michalak T. J., 1990 : « An experiment in enhacing catalog records at Carnegie Mellon University », *Library Hi Tech*, 31 (3), 33-41.

• 15. Poulsen C, 1996 : « Tables of contents in library catalogs: a quantitative examination of analytic catalogs », Library Ressources and Technical Services, 40 (2), 133-139.

• 16. Rush E. J., 1996 : « PALINET/MONO-TOC pilot project », Rapport, Palinet Headquarters, Pittsburg, August 1996.

• 17. Settel B. et Cochrane P. A., 1982 : « Augmenting subject descriptions for books in online catalogs », *Database*, 5, 29-37.

• 18. Weintraub .T., Shimogushi W., 1993 : « Catalog record contents enhancement », *Library Ressources and Technical Services*, vol. 37, avril 1993.

consultées utilise des critères de performance tels que le rappel[1] et la précision[2]. Seule une étude a utilisé un autre critère qui est celui du taux de circulation des ouvrages dont les notices bibliographiques sont enrichies.

En 1982, Settel & Cochrane [17] conduisent une étude pour déterminer si en ajoutant des mots et des phrases extraits des tables des matières, cela améliore l'accès au sujet et le taux de rappel pour l'utilisateur. Deux types d'enregistrements sont comparés :

1) enregistrements non enrichis,

2) enregistrements enrichis à l'aide de mots ou phrases extraits des tables des matières.

Le deuxième type d'enregistrement double le taux de rappel sur le domaine des sciences sociales et le triple en sciences humaines par rapport au premier type d'enregistrement. Ces auteurs en concluent donc que l'addition de termes extraits des tables des matières augmente significativement le taux de rappel, et donc évite le silence.

Si les tests effectués par Dillon [5] et l'équipe d'ADFA confirment cet accroissement du taux de rappel, ils notent cependant une légère diminution du taux de précision :

	ESP (ADFA)	Dillon
Rappel		
base normale	44%	17%
base enrichie	75%	26%
Précision		
base normale	88.6%	71%
base enrichie	70.16%	59%

2. le rôle des tables de matières dans le choix des documents

Souvent les utilisateurs rencontrent des difficultés pour sélectionner un document. En effet les notices bibliographiques contiennent très peu d'informations sur le contenu du document (auteur, sujet, titre, éditeur, etc.). L'enrichissement des notices bibliographiques par les tables de matières peut aider l'utilisateur à mieux sélectionner les documents. Les tables de matières donnent à la fois une vue d'ensemble et permettent d'identifier les parties du document. L'utilisateur ayant la possibilité de visualiser la table de matières d'un ouvrage, pourrait mieux juger à l'écran la pertinence de ces références sans aller aux rayons. Lorsqu'on sait que presque le quart des collections des bibliothèques universitaires en France ne sont pas en libre accès, on mesure l'importance de cet enrichissement.

Cette difficulté de sélectionner les documents grandit lors d'un accès à distance (par Telnet ou par le Web)

3. Le rôle des tables de matières dans les bibliothèques virtuelles

L'accès direct à de grandes collections numérisées pose le problème de la surabondance d'informations. Nous pensons qu'il est nécessaire d'avoir une étape intermédiaire aussi bien pour effectuer des recherches, que pour le choix d'un document. Les tables de matières peuvent jouer ce rôle. Le lecteur peut ainsi accéder à des parties ou à la totalité du document numérisé. Pour les ouvrages dépourvus de tables de matières, des prototypes récents tentent d'extraire d'une façon automatique des résumés de textes qui peuvent être représentés graphiquement [11].

Conclusion

L'analyse des diverses études nous montre qu'il y a d'une part, une forte amélioration du taux de rappel, donc une meilleure circulation des livres et une meilleure exploitation des fonds des bibliothèques ; d'autre part, les utilisateurs n'ont plus besoin d'avoir le livre en main pour décider de la pertinence d'un livre. Cependant, entreprendre un projet d'enrichissement des notices est une entreprise complexe qui nécessite une meilleure précision des critères de choix, une méthodologie d'enrichissement bien définie, et enfin une évaluation du coût de faisabilité.

A l'image du projet PALINET/MONO-TOC, il serait intéressant que ce travail soit mené par un groupe de bibliothèques pour que le coût de production diminue.

Il est néanmoins indispensable d'effectuer d'autres études pour répondre à la question épineuse de précision.

1. Le taux de rappel est la proportion des documents pertinents retrouvés par rapport à l'ensemble des documents présents dans le système.

2. Le taux de précision est la proportion des documents pertinents par rapport à l'ensemble des documents fournis par la recherche.