Construire le classement d'un annuaire Internet

Pour gérer au mieux l'importante masse d'information disponible sur Internet, les bibliothèques et les organismes de documentation ont remis au goût du jour les techniques de classification généralistes et universelles, ainsi que les thésaurus. L'emploi d'une classification numérique universelle ou d'une classification hiérarchique permet d'organiser le savoir du Web en catégories (les classes), en allant du général au particulier. L'utilisateur peut ainsi consulter les ressources dans une arborescence thématique particulièrement adaptée à ces différents modes de recherche d'information. La « navigation », dans ces systèmes hypertextuels d'information[1], se fait directement par l'information.

Géraldine Gourbin

www.cyber-documentaliste.com

Aux États-Unis, the WWW Virtual Library[2] et CyberStacks utilisent la classification de la bibliothèque du Congrès (LCC), BUBL le système de Classification décimale universelle (CDU), tandis que CyberDewey exploite la Dewey (CDD). *The Clearinghouse for Subject-Oriented Internet Resource Guides*[3] a été mis au point par des professionnels de la documentation de la bibliothèque de l'université du Michigan. En France, l'Association des directeurs de bibliothèques départementales de prêt (ADBDP) a utilisé la Dewey pour proposer une sélection encyclopédique de sites Web.

L'accès à l'information passe par un travail intellectuel qui optimise son repérage. L'indexation des ressources doit être de qualité et le classement doit proposer un accès aisé aux utilisateurs. Comment organiser cette information ? Comment construire le classement d'un annuaire de sites Web ? Y a-t-il des modes de classification propres aux annuaires de recherche ? La mise en place de ces annuaires thématiques[4] préfigure-t-elle d'une révolution en marche pour les pratiques bibliothéconomiques ?

Les fondements d'un classement du Web

Yet Another Hierarchical Officious Oracle[5] (Yahoo) a été développé en avril 1994 par David Filo et Jerry Yang, deux étudiants de l'université de Stanford (Californie, États-Unis), afin de repérer les sources d'information relatives à leurs centres d'intérêt. Yahoo constitue aujourd'hui le répertoire le plus connu et le plus utilisé à travers le monde. Pionniers

1. Lalthoum Saadani, « La représentation dans Internet des connaissances d'un domaine », *Documentation et bibliothèques*, janvier-mars 2000, volume 46, n° 1, p. 27-42.

2. Jean Michel, « Les professionnels de l'information-documentation à l'heure du document numérique et des réseaux électroniques », *Document numérique*, juin 1997, volume 1, n° 2, p. 224.
3. http://www.clearinghouse.net

4. Bien souvent gérés par des professionnels de la documentation et bibliothécaires, des cyber-documentalistes (Nomade), net-surfeurs (Yahoo), ou cyber-writers, responsable de la nomenclature (QuiQuoiOù intégré depuis peu à Voilà).
5. Traduction littérale de Yahoo : « Encore un autre oracle à classement hiérarchique officieux ».

Géraldine Gourbin *est titulaire d'une maîtrise de la documentation et de l'information. Responsable de la documentation chez Nomade de 1996 à 1999, elle a ensuite encadré la conception de l'infothèque de l'université virtuelle francophone à l'Agence universitaire de la francophonie. Elle était, jusqu'à une date récente, responsable de la gestion de l'information pour un « incubateur de start-up ».*

dans ce domaine, les créateurs de Yahoo ont inspiré de nouvelles initiatives de classement thématique des ressources du Web. D'autres annuaires thématiques, tels que Open Public Directory, Snap ou LookSmart, sont fondés sur la même approche de catégorisation des ressources, et ont depuis vu le jour sur Internet.

La mise en ligne de l'annuaire de recherche Nomade date quant à elle de juillet 1996. Le premier objectif était de concevoir un guide d'orientation sur le Web, avec un point de vue francophone sur les ressources d'Internet en français ; le second, d'identifier les ressources en français représentatives du Web.

Nomade

La mise en place de l'arborescence thématique de Nomade a nécessité deux mois de travail à temps plein, durant lesquelles les classifications universelles de type Dewey ou CDU, mais également des thésaurus ou des annuaires de pages jaunes, ont été passés au crible. Le choix d'utiliser les plans de classement déjà existants plutôt que d'en créer un de toutes pièces est lié avant tout à des raisons économiques : l'élaboration d'un plan de classement est difficile et coûteuse en temps. Il en existe de nombreux dont on peut s'inspirer, en les adaptant simplement à ses besoins.

Le plan de classement, très largement surdimensionné par rapport aux besoins du guide dans un premier temps, a été maintes fois retravaillé, épuré puis enrichi à nouveau pour répondre aux préoccupations des utilisateurs (en juillet 1996, Nomade comprenait 2800 sites pour environ 3 000 catégories. Aujourd'hui, 100 000 sites sont répertoriés dans près de 9 000 catégories). La définition des grands thèmes, les « méta-catégories », a été conforme à l'esprit d'Internet et au contenu d'un guide de recherche généraliste, avec notamment des catégories liées aux médias, au divertissement et bien sûr aux pages personnelles, classées dans *Loisirs et tourisme/Pages personnelles.* Les rubriques se devaient d'être aussi homogènes que possible, sans décalage ou fossé entre les différentes catégories ou les différents niveaux. Un plan de classement doit être clair, complet, simple et souple pour pouvoir s'adapter aux évolutions du nombre de sites indexés. Peu à peu, les contours d'un plan de classement propre aux spécificités d'Internet se sont dessinés.

> Mieux vaut ne pas mélanger classification, liée au contenu et typologie, liée à la nature de l'information

Sur Internet, les fonctions documentaires ne sont en effet employées que dans un cadre virtuel : elles n'impliquent pas le rangement physique des documents, mais simplement une classification adaptée. Ainsi, si chaque annuaire a ses propres critères de catégorisation, certaines règles de construction et de maintenance des plans de classement sont communes.

Distinguer les types de sites

En 1996, la différence était très nette entre les sites personnels et les sites officiels d'entreprises, mais, depuis lors, l'information commerciale a envahi Internet. La classification de Nomade marque une distinction très nette entre l'information gratuite et l'information commerciale (ex. : la catégorie *Loisirs, tourisme* et la catégorie *Entreprises, économie/Loisirs, tourisme, restauration).* Cette distinction est également faite dans d'autres annuaires, comme Yahoo, notamment ; les concepteurs de QuiQuoiOù ont pris, quant à eux, le parti de ne pas marquer cette distinction dans les premiers niveaux de l'arborescence, mais seulement dans les sous-niveaux. Toutefois, ce classement aurait, comme le nôtre, ses inconvénients.
Pour certains types de sites commerciaux, un double classement est effectué. Par exemple, le site Web d'un journal quotidien se retrouve dans les deux catégories *Actualité, presse* et *Entreprises, économie/Presse, radio, TV.* Certains types de sites associatifs, qui font payer leurs prestations comme les entreprises, ne sont pourtant pas classés dans *Entreprises, économie,* mais dans leur catégorie principale. Par exemple, une association axée sur la formation continue sera classée dans *Enseignement, Emploi/Formation,* tandis qu'une entité considérée comme une entreprise sera incluse dans *Entreprises, Économie/la discipline/formation continue.*

D'autres critères sont également pris en compte pour permettre une identification de la provenance de l'information :
- la nature du site : est-il géré par une association, une entreprise, un service public ? Est-ce un site personnel, présente-t-il un caractère éducatif ?
- le public visé : tout public, enfants, ados/adultes, adultes, professionnels ?

L'indexation des sites dans les catégories est complexe, car la typologie des sites s'exprime à deux niveaux : les catégories et la nature du site. Mieux vaut donc ne pas mélanger *classification,* liée au contenu et *typologie,* liée à la nature de l'information.

Pour les sites personnels, deux choix sont communément admis. Si le site traite d'un sujet particulier, il est classé dans la catégorie en rapport avec le sujet. Si c'est un site personnel quelconque, il est alors classé dans les *Pages personnelles* de l'annuaire. Les sites commerciaux sont répertoriés dans la catégorie *Entreprises, économie*. Certains annuaires proposent un classement géographique en complément du classement thématique. Ce classement doit alors être bien conçu et n'indexer que les sites ayant un intérêt certain pour le pays, la région ou la ville concernée.

Des atouts et des contraintes spécifiques

La page Web permet toutes les extensions. L'annuaire peut contenir autant de catégories que nécessaire. Plus de limites alors ? Au contraire : si les possibilités de l'outil sont infinies, celles de l'utilisateur ne le sont pas. L'aisance dans la navigation à travers les catégories et la visibilité des catégories sans déplacement du curseur sur l'écran sont nécessaires. Et si l'accès à l'information est facilité par la navigation hypertextuelle, la construction d'un répertoire requiert en revanche nombre de précautions.

Des entrées multiples

À la différence d'un document matériel inclus dans un plan de classement, le Web permet d'inclure un même site dans plusieurs catégories. Généralement, les annuaires choisissent de classer les sites dans deux ou trois catégories maximum. En cas de modification sur le formulaire d'un site, les informations sont mises à jour dans toutes les catégories où ce site apparaît. Il doit y avoir une parfaite concordance entre la description du site, son contenu et les catégories retenues.

Il est possible de lier les catégories les unes aux autres, afin d'indiquer aux utilisateurs les catégories voisines en terme de contenu. C'est le système des *crosslinks* « @ » qui a fait le succès de Yahoo et a été largement repris par tous les autres annuaires. On navigue de thème en thème en multipliant les possibilités d'atteindre la catégorie désirée, grâce aux liens transversaux qui renvoient directement vers la catégorie d'origine (renvois directs). La position dans la classification est toujours indiquée en haut de la page. Le classement des sites se fait par ordre alphabétique des titres.

De nouveaux types d'annuaires fleurissent, à l'image de l'Open Directory Project, un projet fondé sur le modèle de la collaboration bénévole entre éditeurs

Un risque d'éparpillement

Pour permettre une bonne mise en valeur des ressources, il ne faut pas dépasser 12 à 14 catégories par niveau. La plupart des répertoires l'ont bien compris, même si, dans les niveaux inférieurs, cette convention n'est pas toujours respectée. Si vous choisissez, par exemple, un secteur d'activité dans la méta-catégorie *Sociétés* de Yahoo, vous risquez de perdre plusieurs minutes à parcourir la cinquantaine de catégories proposées.

Un annuaire vit et évolue au fur et à mesure que le nombre de sites indexés augmente. Le développement de l'arborescence peut s'opérer en largeur ou en profondeur, selon que l'on souhaite élargir le nombre de domaines ou préciser les sous-domaines. Toutefois, il est bon de limiter la profondeur du nombre de niveaux. En règle générale, il ne faut pas excéder cinq ou six niveaux pour un annuaire généraliste, deux ou trois niveaux pour un annuaire sélectif (*cf.* Weborama, Hachette.net, etc.). S'il y a plus d'une cinquantaine de sites dans une catégorie, un reclassement plus pointu doit être effectué et de nouvelles catégories spécifiques créées pour permettre à l'internaute d'effectuer une recherche plus ciblée et correspondant à ses besoins.

Prolifération des ressources

L'annuaire de l'Urec (Unité réseaux du CNRS) a été le premier annuaire de sites Web français. Créé au début de 1994, il a d'abord recensé les sites académiques, puis son contenu est devenu plus généraliste. À l'automne 1997, ses responsables prennent conscience que la gestion de l'annuaire devient très difficile, du fait de l'accroissement constant du nombre de sites, et notamment de sites commerciaux. L'Urec cesse donc d'assurer la mise à jour de cet annuaire généraliste, qui devient un annuaire spécialisé consacré à l'enseignement supérieur et à la recherche. C'est un exemple parmi beaucoup d'autres des contraintes liées à l'augmentation incessante du nombre de ressources sur Internet.

De gros moyens financiers sont-ils les seuls garants de la maintenance d'un annuaire généraliste ? Peut-être pas. De nouveaux types d'annuaires fleurissent, à l'image de l'*Open Directory Project*[6], un projet fondé sur le modèle de la collaboration bénévole entre éditeurs. Ce dernier postule qu'avec l'accroissement du Web, les moteurs de recherche automatisés et les répertoires gérés par de petites équipes d'édition ne pourront plus prendre en charge tous les sites. La conception d'annuaires sous forme de collaboration entre organismes de

6. http://www.demog.com

documentation et bibliothèques se développe (Sitebib par exemple).

Bien que le nombre de sites indexés dans un annuaire soit aujourd'hui infinitésimal en comparaison des prouesses technologiques des moteurs de recherche, l'orientation et le repérage n'en sont que plus nécessaires. La généralisation d'une catégorisation en masse sur les différents moteurs de recherche majeurs[7] est un signe de bonne santé pour les techniques de classification. Il faut considérer les classifications comme un outil d'accès complémentaire à la recherche par mots clés, et non comme un outil de substitution. Si l'hypertexte bouleverse notre manière de penser le langage documentaire, il est bon de rappeler que la fonction documentaire est omniprésente sur Internet et que les techniques documentaires ont de beaux jours devant elles.

Octobre 2000

7. Géraldine Gourbin, « Les moteurs de recherche deviennent des annuaires », *Lettre du bibliothécaire québécois*, avril/mai 1999, n° 17. http://www.sciencepresse.qc.ca/lbq/lbq.html

Bibliographie

Gourbin, Géraldine, « Une nouvelle profession cyber-documentaliste, l'exemple de Nomade », *Documentaliste-Sciences de l'information,* 1998, vol. 35, n° 3, p. 175-177.

Accart, Jean-Philippe ; **Rethy,** Marie-Pierre, *Le métier de documentaliste,* Paris, Éditions du Cercle de la librairie, mars 1999.

Balpe, Jean-Pierre ; **Lelu,** Alain ; **Papy,** Fabrice ; **Saleh,** Imad, *Techniques avancées pour l'hypertexte,* Paris, Hermes, février 1996.

Maniez, Jacques ; **Mustafa el Hadi,** Widad (textes réunis par), *Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information,* Villeneuve d'Ascq, Édition du Conseil scientifique de l'université Charles-de-Gaulle, Lille 3, 1999, coll. « UL3 : Travaux et recherche ».