Documentation et bibliothèques

La représentation dans Internet des connaissances d'un domaine
The Knowledge of a Given Field as Represented by the Internet
La representación en Internet de los conocimientos de un campo

Lalthoum Saàdani and Suzanne Bertrand-Gastaldy

Volume 46, Number 1, January–March 2000

URI: https://id.erudit.org/iderudit/1032683ar
DOI: https://doi.org/10.7202/1032683ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Saàdani, L. & Bertrand-Gastaldy, S. (2000). La représentation dans Internet des connaissances d'un domaine. *Documentation et bibliothèques*, *46*(1), 27–42. https://doi.org/10.7202/1032683ar

Article abstract

In order to impart order to the Web and to improve subject access to the information, several organisations have turned toward the more traditional methods that are applied to library and information sciences. Standards, as well as cataloguing projects, indexation and classification abound and web libraries are springing up locally, regionally, nationally and internationally. Aimed at filling the gap left by commercial services, these initiatives are perceived by some as cumbersome, expensive, and outmoded, and by others as a promising solution needed to organise electronic resources and to make them profitable. This research is an inventory of several projects, either terminated or in progress. It also attempts to underline the limits and the strengths of the different approaches used at different levels.

La représentation dans Internet des connaissances d'un domaine

Lalthoum Saàdani
Étudiante au doctorat (Ph.D.)
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Suzanne Bertrand-Gastaldy
Professeure titulaire
École de bibliothéconomie et des sciences de l'information
Université de Montréal

Pour mettre de l'ordre dans le Web et faciliter l'accès à l'information par sujets[1], plusieurs institutions se sont tournées vers les méthodes traditionnellement appliquées en bibliothéconomie et en sciences de l'information. Les normes ainsi que les projets de catalogage, d'indexation et de classification prolifèrent et les web libraries *se multiplient aussi bien au niveau local, régional, national ou international. Ces initiatives, visant à remédier à l'insuffisance des services commerciaux, sont vues par certains comme lourdes, onéreuses et dépassées, et par d'autres comme une solution prometteuse pour organiser les ressources électroniques et rentabiliser leur usage. La présente recherche fait l'inventaire de plusieurs projets proposés, réalisés ou en cours. Elle tente, par la même occasion, de souligner les limites et les forces de ces diverses approches adoptées à différents niveaux.*

The Knowledge of a Given Field as Represented by the Internet
In order to impart order to the Web and to improve subject access[1] to the information, several organisations have turned toward the more traditional methods that are applied to library and information sciences. Standards, as well as cataloguing projects, indexation and classification abound and web libraries are springing up locally, regionally, nationally and internationally. Aimed at filling the gap left by commercial services, these initiatives are perceived by some as cumbersome, expensive, and outmoded, and by others as a promising solution needed to organise electronic resources and to make them profitable. This research is an inventory of several projects, either terminated or in progress. It also attempts to underline the limits and the strengths of the different approaches used at different levels.

La representación en Internet de los conocimientos de un campo
Para poner orden en la Web y facilitar el acceso a la información por temas[1], numerosas instituciones han acudido a los métodos tradicionales que se aplican en biblioteconomía y ciencias de la información. Las normas y los proyectos de catalogación, de indexación y de clasificación proliferan y las bibliotecas de la Web se multiplican tanto a nivel local, como regional, nacional o internacional. Estas iniciativas, destinadas a remediar la insuficiencia de los servicios comerciales, son percibidas a veces como pesadas, onerosas y superadas; otras veces como una solución prometedora para organizar los recursos electrónicos y rentabilizar su uso. Este trabajo de investigación hace un inventario de numerosos proyectos propuestos, realizados o en curso. Asimismo trata de subrayar las restricciones y los puntos fuertes de estos diversos enfoques adoptados a diferentes niveles.

Tout le monde s'accorde sur le fait que les ressources dans Internet, en plus d'être de qualité et de stabilité variables, sont très mal organisées. Ceci rend difficile leur accessibilité, leur conceptualisation, leur visualisation, leur recherche, leur filtrage et les tentatives de les citer en tant que références (Levy cité par Woodward 1996; Zhang et Estabrook 1998). D'ailleurs, nombreux sont ceux qui s'interrogent si, dans son état actuel, Internet peut être organisé. Le caractère d'instabilité des ressources ne favorise pas les tentatives d'une organisation effective. Celle-ci dépend particulièrement de la bonne volonté des divers créateurs de pages et de sites Web (Woodward 1996).

On assiste actuellement à l'émergence d'une multitude de services et à un assortiment de produits (logiciels entre autres), de normes, de directives, de spécifications et de propositions faites à l'intérieur de divers projets locaux, régionaux, nationaux et internationaux, souvent redondants, présentant des chevauchements plus ou moins visibles ou utilisant des approches variées. Le défi, selon Iannella (1997), est de ramener les différentes communautés à se mettre d'accord sur l'utilisation de ces outils et aussi sur le choix d'infrastructures flexibles susceptibles de supporter les diverses normes. Toutes les tentatives visent un seul objectif: mettre de l'ordre dans Internet afin d'améliorer l'accessibilité et le repérage.

En outre, l'intérêt accordé à la représentation des connaissances d'un domaine dans Internet n'est qu'un aspect

1. Permet de faire la recherche en utilisant un vocabulaire contrôlé (thésaurus ou liste des vedettes-matières), un système de classification ou un plan de classement.
Means used to conduct research using a controlled vocabulary (thesaurus or list of subject headings), a classification system or a classification plan.
Permite hacer la investigación utilizando un vocabulario de control (Tesoro o lista de temas actuales) un sistema de clasificación o un plan de ordenamiento.

d'un problème plus global de la description, de la modélisation, de la représentation et de la visualisation des ressources en vue de les rendre accessibles. La notion de représentation couvre une multitude de dimensions. À côté de la représentation iconique et graphique, on parle, toutes définitions confondues, tantôt de représentation de données, tantôt de représentation de l'information et souvent de représentation des connaissances. Le terme représentation tel que défini dans le *Petit Larousse* (1998) fait référence, entre autres, à l'action de rendre sensible quelque chose au moyen d'une figure, d'un symbole, d'un signe, comme l'écriture est la représentation de la langue parlée. En psychologie cognitive, la représentation signifie la perception ou l'image mentale que l'individu a du monde qui l'entoure. On parle alors de représentation interne et de représentation externe. Par ailleurs, le terme représentation est souvent associé au terme visualisation. Ceci s'explique sans doute par le fait que la représentation de connaissances n'est d'aucune utilité sans l'existence de systèmes pour la visualiser et la manipuler (Williams, Sochats et Morse 1995).

Cette recherche s'intéresse à la représentation des connaissances d'un domaine dans Internet, peu importe la forme (linéaire, hiérarchique ou graphique), telle que connue en bibliothéconomie et en sciences de l'information. En particulier, seront inventoriées les différentes approches (manuelles ou automatiques) proposées ou utilisées pour représenter :

- Le document électronique (page ou site Web), et ce, via les métadonnées par les métaétiquettes titre, termes ou concepts ;
- Ce dont traite le document électronique, à savoir les différents concepts exprimant le sujet traduits par des mots clés libres ou contrôlés à l'aide d'instruments tels que les thésaurus ou les listes des vedettes-matières ;
- Le domaine ou sous-domaine dans lequel le document en question est classé (au moyen de plans de classement ou de systèmes de classifications par domaines, encyclopédiques, hiérarchiques ou à facettes).

Tout d'abord sera donné un bref aperçu du contenu d'Internet. Ensuite, un examen de la littérature sur le sujet qui nous intéresse nous permettra d'identifier les différentes approches utilisées pour représenter les ressources disponibles dans Internet, notamment les documents, les sujets, les domaines ou sous-domaines de connaissances. Pour chacune des approches, nous énumérerons quelques normes, propositions, projets ou travaux réalisés ou en cours, qui ont été entrepris pour cataloguer, indexer et classifier les diverses ressources dans Internet en vue d'organiser et de faciliter leur recherche ainsi que leur repérage. Nous tenterons, lorsque c'est possible, de porter un regard critique sur tous les éléments susceptibles de représenter le sujet ou le domaine de connaissances, qu'ils aient été générés automatiquement, intégrés ou proposés en vue de faciliter l'accès ou le repérage. Une brève synthèse nous permettra de mettre en relief les caractéristiques des diverses initiatives ainsi que les aspects marquants des différents travaux visant à faciliter l'accès par sujets. Finalement, après la conclusion, nous présenterons la bibliographie des documents utilisés pour compléter ce travail.

Nous tenterons de répondre, tout au long de cet article, à certaines questions, notamment : Qu'est-ce qui est réellement représenté ? Qui détermine le sujet ou le domaine à représenter ? Qui se charge de sélectionner les termes ou les catégories à classer dans un index ou dans un répertoire ou, plus exactement, sur quels critères se base-t-on pour utiliser l'un ou l'autre moyen pour représenter un domaine de connaissances ou pour générer un répertoire ou un index ? Comment le sujet ou le domaine est-il représenté ?

Notons que cette recherche s'est basée en grande partie sur l'exploration directe de quelques sites pertinents. La consultation des documents électroniques produits et diffusés par quelques organisations ou institutions impliquées dans l'un ou l'autre des projets mentionnés dans ce travail tels Dublin Core, le projet ROADS (Resource Organisation and Discovery in Subject-Based Services), le projet Desire (Development of a European Service for Information on Research and Education) et autres projets s'est avérée nécessaire. En effet, les sites de ces institutions ont été visités afin de voir ce qui a été fait, les travaux à venir et les différentes tentatives d'organiser le contenu du Web. Les autres documents ont été repérés par une recherche dans les bases de données *Current Content*, *LISA* et le réseau Internet en utilisant des mots clés tels organisation, indexation, catalogage, description, format de catalogage, métadonnées, vocabulaire contrôlé (au nom des thésaurus et listes de vedettes-matières), catégorisation, répertoire, classification et noms des classifications, sujet et domaine. Chacun de ces termes a été utilisé en combinaison avec les termes suivants : Internet (exemple : indexation et Internet), Web (organisation et Web, catalogage et Web, etc.), ressources électroniques (indexation et ressources électroniques), collections électroniques, bibliothèque virtuelle, document électronique et d'autres termes pertinents. Les références proposées à la fin d'un document ont également été utilisées pour compléter, approfondir ou avoir plus de précisions sur un aspect du sujet qui nous intéressait.

Le contenu d'Internet

Les ressources accessibles par Internet sont stockées sur de très nombreux serveurs disséminés dans le monde entier. En effet, Internet fournit un lien à plusieurs sources d'information sans qu'il y ait une base de données centralisée pour l'organisation ou la recherche de ces ressources. La problématique de manutention de ces gisements partagés occupe d'ailleurs une place fondamentale dans les recherches entreprises par les bibliothèques virtuelles (Schatz 1997). Les ressources dans Internet sont également très diversifiées couvrant aussi bien des fichiers individuels de textes, des bases de données WAIS (Wide Area Information Server), des systèmes de bases de données, des pages Web contenant des publications électroniques, des métadonnées, des procéduriers et des manuels, des ouvrages de référence, des outils linguistiques et sémantiques, des outils documentaires, des documents et des parties de documents, des données bibliographiques, des données non textuelles telles que des films, des vidéos, de la musique, des graphiques, des photos, des images et une multitude de ressources électroniques. Le type de contenu est variable et peut être de nature scientifique, technique, commerciale, gouvernementale, littéraire, artistique ou autre. Tous les domaines sont représentés. Leur contenu est, toutefois, plus ou moins organisé et l'information disponible sur un sujet ou sur un thème particulier varie largement aussi bien en quanti-

té qu'en qualité. Le Web a permis à ses utilisateurs de publier électroniquement une information accessible à des millions de personnes de façon relativement facile ; cependant repérer une information pertinente relève parfois du défi. En effet, à mesure que le contenu du Web croît, l'efficacité de repérage diminue de façon dramatique. Également, le changement constant qui caractérise les documents électroniques offre un défi différent, car les pages et les sites Web sont éphémères et subissent, selon Koehler (1999), deux formes de métamorphose : la persistance et la volatilité.

La majorité des sites Web présente à partir de la page d'accueil une table des matières ou un index des contenus. Certains utilisent des cartes (*ImageMap*) ou ont un lien avec une carte du site (*SiteMap*). Bon nombre de sites utilisant des cadres fournissent des renvois à une représentation textuelle comme une autre façon de consulter leur site. Les stratégies de navigation les plus connues se font à partir d'une liste des principales sections d'un site ou sont sous forme d'une série de liens textuels ou iconiques regroupés en bloc. Généralement, le bloc navigable apparaît dans une colonne dans la partie gauche de la page Web. D'autres stratégies de navigation se font à partir de liens textuels présentés de façon linéaire horizontalement en haut ou en bas de la page Web. Plusieurs sites Web combinent les deux stratégies et utilisent des liens qui renvoient aux moteurs de recherche utilisés dans le Web ou ont leurs propres moteurs de recherche.

Par ailleurs, les seuls groupements thématiques visibles dans Internet sont les conférences Usenet. Ces conférences sont classées en sept catégories de base : les principales sont l'informatique (comp.), les sujets divers (misc.), les nouvelles ou les informations dans Internet et Usenet (news), les loisirs (rec.), les sujets scientifiques (sci.), les sujets d'intérêt social (soc.), les objets de débat (talk) et onze catégories secondaires dont les plus importantes sont les sujets divers (alt.), le commerce sous toutes ses formes (biz.) et les sujets se rapportant à la biologie (bionet.).

On trouve également les passerelles[2] d'information spécialisées par sujets ou SBIGs (Subject Based Information Gateways). Ces passerelles fournissent des accès par sujets à des ressources d'information d'une certaine qualité. En effet, les contenus sélectionnés sont évalués par des experts du domaine et doivent répondre à des critères bien établis.

Les ressources retenues sont décrites par des experts humains qui, en plus de cataloguer les documents, fournissent un résumé, un ensemble de mots clés et attribuent à chaque document un code de classification. Les codes de classification sont extraits automatiquement à partir des notices et utilisés pour organiser l'espace ou la structure navigable. Les passerelles d'information spécialisées par sujets sont également désignées en tant que services de contrôle de qualité spécialisés par sujets (Brummer 1997). Parmi ces passerelles disponibles dans Internet, on identifie entre autres, ADAM (Art, Design, Architecture and Media Information Gateway), AHDS (Arts and Humanities Data Service), Biz/ed (Business and Economics Information Gateway), EELS (Engineering Electronic Library, Sweden), EEVL (Edinburg Engineering Virtual Library), OMNI (Organising Medical Networked Information), SOGIG (Social Science Information Gateway), BUBL Link (Bulletin Board for Libraries), DutchESS (Dutch Electronic Subject Service), NISS (National Information Services and Systems) Directory of Networked Resources et NOVAGate (Nordic Gateway to Information in Forestry, Veterinary and Agricultural Sciences).

Les diverses approches de représentation des sujets

Plusieurs fournisseurs de services de bases de données Web et les moteurs de recherche commerciaux accessibles via le Web qui leurs sont affiliés sont en compétition pour fournir un accès par sujets à l'information disponible dans Internet (Nicholson 1997). À côté de ces tentatives commerciales, on trouve d'autres initiatives émanant de diverses organisations de normalisation, institutions ou groupes universitaires et gouvernementaux, communautés scientifiques concernées et associations des professionnels de l'information. Ces institutions, qui agissent soit au niveau local, régional, national ou international, visent les mêmes objectifs : organiser les diverses ressources disponibles dans Internet et donner accès à l'information par sujets. La majorité de ces organisations se sont tournées vers des méthodes traditionnellement appliquées dans les bibliothèques. Ainsi, on assiste à l'émergence d'un bon nombre de projets subventionnés soit pour cataloguer, indexer, classifier les contenus électroniques ou pour normaliser les formats de description et d'échange entre systèmes, réseaux et sous-réseaux.

Parmi ces tentatives spécialisées par domaines les plus visibles, JISC (Joint Information System Committee) finance plusieurs projets donnant accès aux ressources thématiques. Les projets peuvent être regroupés selon les deux modèles suivants :

- les passerelles générales qui donnent accès à un ensemble très large de sujets évalués et sélectionnés parmi la totalité des ressources disponibles dans Internet (exemple : BUBL et NISS) ;
- les passerelles spécialisées par domaines qui donnent un accès aux ressources pertinentes se rapportant à un domaine précis ou à un groupe disciplinaire particulier (exemple : SOGIG pour les sciences sociales, OMNI pour la médecine et EEVL pour l'ingénierie).

Il faut noter qu'à l'intérieur de certains projets, il existe parfois quelques chevauchements ou compétitions comme c'est le cas de BUBL et NISS, qui tous les deux ont l'ambition de couvrir l'ensemble des domaines de connaissances. Également, bien que ces services tentent de créer des passerelles spécialisées dans des sujets spécifiques, un bon nombre de domaines restent non couverts (Hyanes 1998).

Par ailleurs, parmi les différentes approches utilisées pour représenter les documents électroniques et leurs contenus, on identifie, entre autres, les métadonnées, les index générés par une indexation automatique ou manuelle en mots clés libres ou contrôlés (usage de thésaurus ou de listes de vedettes-matières), les systèmes de classification et les groupements thématiques.

2. Une passerelle est un équipement d'interconnexion qui relie des réseaux ayant des conventions différentes, leur permettant de communiquer entre eux. Voir le site de l'Office de la langue française : <http://www.olf.gouv.qc.ca/ressources/internet/index/ >

Les métadonnées en tant que forme de représentation des documents Web

Selon plusieurs auteurs dont Woodward (1996), l'usage des métaétiquettes ou *meta-tags* peut apporter une solution à un bon nombre de problèmes, notamment la normalisation de la description d'un site et la fourniture des éléments utiles qui aident les moteurs de recherche et d'autres agents intelligents[3] à déterminer la pertinence du contenu d'une page. La création de métaétiquettes pour les documents électroniques vise également l'organisation des ressources dispersées et disparates afin de permettre une certaine efficacité dans le repérage et faciliter la localisation de l'information. Une métaétiquette est un élément de code comme *HyperText Markup Language* ou HTML qui sert à codifier le contenu d'une page Web en précisant le titre, lorsqu'il est disponible, les mots clés ou les sujets, le créateur et d'autres éléments. Les métaétiquettes fournissent un format électronique lisible par machine favorisant le regroupement et l'indexation de cette information sous forme d'un enregistrement semblable à une fiche de catalogage « cherchable ». Elles permettent aussi d'indiquer au navigateur le format de présentation du document Web, notamment comment afficher certains éléments graphiques.

Les travaux sur les métadonnées, d'après Iannella (1997), couvrent trois aspects majeurs, à savoir la description des ressources, la production des métadonnées et l'utilisation des métadonnées. Le projet Dublin Core, à titre d'exemple, cherche à promouvoir et à développer les éléments de métadonnées requis pour faciliter la découverte des ressources disponibles dans un environnement réseau tel Internet. Il vise également la gestion de « l'interopérabilité » entre les systèmes hétérogènes de métadonnées (Lerinckx 1997). L'importance accordée à la problématique de définition, d'implantation et d'usage des métadonnées est générale. D'ailleurs les participants au groupe de travail Dublin Core proviennent de communautés industrielles, académiques, de chercheurs et de bibliothèques. Les métadonnées telles que finalisées par Dublin Core en décembre 1996 sont composées de 15 éléments dont trois consacrés à la représentation du contenu, à savoir le titre, le sujet ou les mots clés et la couverture. Également, l'OCLC (Online Computer Library Center) et la bibliothèque du Congrès tentent tous les deux de jouer un rôle actif dans la définition des règles et des critères de catalogage des ressources dans Internet, et ce, en développant la documentation et les métadonnées suivant les règles de catalogage Anglo-Américaines (RCAA2). Ainsi, les nouvelles directives incluent une étiquette additionnelle (l'étiquette 856) pour indiquer la localisation des ressources électroniques (Koehler 1999). La bibliothèque du Congrès a aussi joué un rôle important en établissant un système de renvoi entre les éléments des formats de description MARC (Machine Readable Cataloguing), GILS (Government Information Locator Service) et Dublin Core, ce qui favorise l'échange des notices entre ces trois systèmes. Par ailleurs, le projet Intercat, une initiative de l'OCLC, est une tentative expérimentale pour cataloguer le Web en se basant particulièrement sur la participation volontaire des bibliothécaires (Koehler 1999).

D'autres projets de catalogage et de classification ont récemment vu le jour. Bien qu'ils diffèrent, ces projets visent un objectif commun qui consiste à organiser les ressources dans Internet en vue d'améliorer leur repérage. Le projet CATRIONA (Cataloguing and Retrieval of Information Over Networks Applications) se penche sur la problématique d'identification et de localisation des ressources appropriées aux besoins des différents utilisateurs d'Internet. Dans ce sens, les responsables de ce projet cherchent à trouver la solution en adaptant, d'un côté, les nouvelles technologies de l'information et, d'un autre côté, les méthodes établies en bibliothèques. Ainsi, NISS en association avec BUBL encouragent les tentatives de description des ressources électroniques en proposant un formulaire de catalogage très souple. Les bibliothécaires en Angleterre sont invités à décrire et à ajouter leurs ressources au portail[4] de NISS en complétant le bordereau RDT (Resource Description Template) offert en ligne ou sous format électronique dans le site de NISS. Soulignons que le format de RDT permet d'inclure des informations sur le sujet ainsi qu'une description du contenu des documents électroniques.

Un autre projet de catalogage et de classification issu d'un programme des bibliothèques virtuelles britanniques, davantage concerné par la façon de créer, d'organiser, de chercher et de présenter les métadonnées aux utilisateurs, est le projet ROADS (Resource Organisation and Discovery in Subject-Based Services). Ce projet cherche, en outre, à fournir un accès par sujets transparent au contenu d'Internet en créant, acquérant et diffusant les notices contenant la description des diverses ressources. Principalement développé pour répondre aux besoins des services de bibliothèques virtuelles spécialisées, ROADS compte fournir une plate-forme commune pour chercher à travers les multiples services spécialisés par domaines et tente d'impliquer les fournisseurs d'information dans la description de leurs ressources. Dans ce sens, le projet mise sur la disponibilité et la simplicité de la structure d'enregistrement offerte par le bordereau IAFA (Internet Anonymous FTP Archives) proposé par le groupe de travail d'IETF (Internet Engineering Task Force) pour encourager l'implication des différents services d'information. Le format, lisible par l'homme et très simple, permet d'inclure entre autres champs de métadonnées, la description ou le résumé du contenu du document électronique, évite l'usage d'étiquettes numériques ou de sous-zones et a été conçu pour décrire les ressources d'Internet de façon à ne pas appliquer les caractéristiques redondantes associées aux ressources sur papier. La contribution d'une autodescription des ressources de la part des auteurs et des administrateurs de sites Web est, d'après les responsables du projet ROADS, essentielle pour assurer un service durable. D'ailleurs, la tendance qui consiste à permettre

3. Un « agent intelligent appelé également agent mobile est un robot qui quitte l'ordinateur de l'utilisateur pour s'exécuter sur un ou plusieurs ordinateurs distants et ramener éventuellement les renseignements recueillis à l'ordinateur de l'usager. L'utilisation d'un agent mobile permet, par exemple, de confier à un robot le repérage du produit désiré sur le Web et, éventuellement, d'effectuer une transaction d'achat par Internet, sans que l'internaute ne soit en ligne. » Office de la langue française. Terminologie d'Internet. Voir : <http://www.olf.gouv.qc.ca/ressources/internet/index/>
4. « Un portail est un site Web dont la page d'accueil propose, en plus d'un moteur de recherche, des hyperliens avec une foule d'informations et de services attractifs, qui est conçu pour guider les internautes et faciliter leur accès au réseau, mais surtout pour les attirer et fidéliser le plus grand nombre d'entre eux, au point de devenir leur porte d'entrée dans Internet. » Office de la langue française. Terminologie d'Internet. Voir : <http://www.olf.gouv.qc.ca/ressources/internet/index/>

aux utilisateurs de décrire leurs propres documents en utilisant des métadonnées est en émergence (Iannella 1997).

D'autres projets semblables visent particulièrement à faciliter la localisation de l'information gouvernementale. Le projet GILS en est un exemple. L'usage des métadonnées a pour but la description et l'identification des ressources d'information officielles. Pour cela, GILS utilise un robot pour indexer le contenu des pages Web en faisant l'extraction des données à partir des métaétiquettes. Les données ainsi extraites servent à construire une fiche de catalogage électronique reliée aux ressources d'information gouvernementale (Turner 1996). Chaque organisme gouvernemental peut contribuer à enrichir la base de données de GILS en complétant le bordereau disponible dans Internet.

L'indexation par les moteurs de recherche (Indexation automatique)

Au début, les approches pour organiser les ressources d'information dans Internet étaient concentrées autour de la création d'index qui étaient entretenus manuellement. Cependant, la croissance rapide du Web et les changements sans cesse grandissants qui caractérisent l'état des ressources ont rendu les traitements manuels impossibles (Musella et Padula 1996), ce qui a nécessité le recours à des méthodes automatiques utilisant, entre autres, les moteurs de recherche.

Un moteur de recherche est composé de trois parties: le programme informatique appelé *robot, spider, crawler, wanderer* ou *worm* (Cohen 1999), le logiciel de recherche et l'index. L'index est souvent désigné dans la littérature par d'autres appellations, notamment moteur de recherche, catalogue ou base de données Web. Il est semblable à un grand registre permettant de stocker une copie de chaque page Web trouvée par le programme informatique que nous désignerons dans ce travail par le terme robot. La création de l'index se fait donc de façon automatique par le robot. L'usage des robots a été justifié par leur capacité à traiter des quantités d'informations considérables qui, souvent, manquent de structuration (Neuss et Kent 1995). Ainsi, le robot passe au crible le Web, visite chaque page Web identifiée, fait la lecture du contenu, collecte l'information sur un document à partir de l'URL (Uniform Resource Locator), du titre, des mots clés dans le texte, du texte intégral et des métadonnées et suit les liens aux autres pages à l'intérieur d'un site. Le site est ensuite revisité de façon régulière (mensuellement ou tous les deux mois) afin de vérifier si des ajouts ou des modifications ont été apportés aux différentes pages composant le site en question. Tout ce que le robot trouve est copié dans l'index. Lorsque des changements sont apportés à une page, l'index est automatiquement mis à jour. Il arrive parfois qu'un certain délai soit requis avant que les nouvelles pages ou les changements identifiés par le robot en question ne soient ajoutés à l'index. En d'autres termes, la page peut être visitée et en attente d'indexation, ce qui la rend non disponible aux chercheurs utilisant le moteur de recherche ayant servi à l'indexer (Sullivan 1999).

Il existe deux types de moteurs de recherche :

- les moteurs de recherche individuels qui utilisent un robot pour créer leurs propres index ou bases de données « cherchables » tels *Infoseek, Ask Jeeves, Northern Light, AltaVista, HotBot, Excite* et *Lycos*;
- les métamoteurs de recherche qui font la recherche de façon simultanée dans plusieurs index ou dans diverses bases de données créées par les moteurs de recherche individuels (Cohen 1999).

Parmi les métamoteurs de recherche, on identifie, entre autres, le *MetaCrawler* qui cherche simultanément dans plusieurs index créés par des moteurs de recherche et dans des répertoires thématiques, trie les résultats par localisation (URL) et élimine les résultats en double ; le *C4* qui, en plus de chercher dans de multiples index individuels (moteurs de recherche et répertoires), permet de faire une recherche booléenne par proximité et offre des possibilités de personnaliser sa recherche ; le *Inference Find* qui regroupe les résultats par concepts et sites Internet ; le *MetaFind* qui retourne une liste limitée des résultats repérés dans les principaux index de moteurs de recherche et les classe dans un ordre alphabétique par mots clés ou par domaines et finalement *Profusion* qui cherche dans les index de neuf moteurs de recherche, dont les trois index les plus pertinents par rapport à la question de recherche, tout en éliminant les résultats redondants. Les métamoteurs de recherche sont désignés par d'autres noms, notamment les moteurs de recherche parallèles ou les mégamoteurs de recherche.

Les métamoteurs de recherche peuvent à leur tour être classés en deux types :

- ceux qui rapportent tout ce qui a été repéré dans différents index individuels. Souvent, un même document indexé par plusieurs moteurs de recherche est affiché plusieurs fois (exemple : *Dogpile*) ;
- ceux qui retournent une liste des résultats après avoir fait un tri et éliminé les documents en double indexés plusieurs fois par différents moteurs de recherche (exemple : *Inference Find*) (Cohen 1999).

À côté des moteurs de recherche offerts par les fournisseurs de services commerciaux, on assiste à l'émergence d'un nombre croissant de moteurs de recherche limités par secteurs (Limited Area Search Engines ou LASE) utilisant des index basés sur des critères thématiques ou géographiques. Ces moteurs de recherche peuvent inclure toutes les ressources Web concernant un secteur particulier ou un sujet donné ou utilisent des critères de sélection sévères dans le développement de leurs collections (Koehler 1999). Parmi ce type de moteurs de recherche, on identifie *ARGOS*[5] qui est limité à la recherche dans Internet du monde ancien et médiéval et *HIPPIAS*[6] qui se limite à la recherche en philosophie (Koch 1999).

Généralement, les moteurs de recherche revisitent chaque site indexé de façon régulière, sauf en ce qui concerne les sites réputés stables ou modifiant leur contenu une ou deux fois par an. Les responsables ou les concepteurs de sites peuvent également signaler et soumettre au moteur de recherche les modifications significatives apportées à leur contenu en vue d'une réindexation. Il faut noter qu'un changement apporté au contenu d'une page peut, quelquefois, affecter la façon selon laquelle la page en question sera classée lors de l'affichage des résultats. En effet, tout élément peut jouer un rôle dans le classement : le titre de la page, la description du contenu ou le contenu (Sullivan 1999).

5. Voir *Limited Area Search of the Ancient and Medieval Internet* < http://argos.evansville.edu/>
6. Voir *Limited Area Search of Philosophy on the Internet* <http://hippias.evansville.edu >

Les moteurs de recherche utilisent des algorithmes d'indexation simples. Ils font l'extraction des termes à partir des mots du titre ou du texte intégral et ajoutent parfois une pondération aux termes positionnés au début d'un paragraphe ou indiqués à l'aide d'une métaétiquette : titre, description ou résumé. (Weinberg 1996). Le contenu indexé varie d'un moteur de recherche à un autre. Ainsi :

- Tous les moteurs de recherche indexent le plein texte ou l'intégralité du texte visible d'une page Web. Certains moteurs tel *Lycos* indexent les cent premiers mots ou les vingt premières lignes du texte.
- Certains moteurs de recherche excluent l'indexation des mots vides ou *Stop Words*.
- *AltaVista*, *Excite*, *Inktomi*, *Lycos*, *Google* indexent les mots vides.
- Tous les moteurs de recherche, à l'exception de *Google*, *Lycos* et *Nlight*, indexent la description de la page Web, lorsque celle-ci est fournie par le concepteur du site. La description est codée par la métaétiquette « Description ».
- Tous les moteurs de recherche, sauf *Excite*, *Google*, *Lycos* et *Nlight*, indexent les mots clés fournis par le concepteur de la page Web contenus dans la métaétiquette « Mots clés ».
- Seul *Inktomi* indexe les commentaires.

Plusieurs moteurs de recherche indexent les champs : mots clés, résumés ou description, date, auteur, URL, type d'objet, langage et d'autres caractéristiques (Koehler, 1999). Bien qu'une page Web puisse contenir des images ou d'autres informations audiovisuelles, l'usage du texte est essentiel pour le processus de catalogage, d'indexation et de classification. Ainsi, même les informations non textuelles sont traitées en vue d'extraire une information textuelle. Chaque image ou vidéo dans le Web a une adresse Web unique et peut également avoir des étiquettes HTML qui fournissent une information utile pour interpréter l'information visuelle.

En plus de permettre l'indexation du contenu des pages Web, les moteurs de recherche servent également à les repérer. En effet, malgré l'existence d'autres moyens plus directs pour accéder à une page telle l'adresse URL, les moteurs de recherche restent le principal outil que la plupart des utilisateurs d'Internet emploient pour trouver un site ou une page Web. Ceci est dû, en partie, au fait que l'URL ne constitue pas une référence stable. En effet, le fait de renommer ou de déplacer un fichier rend l'adresse URL obsolète. Également, on peut trouver plusieurs versions du même documents localisées dans de multiples URL.

La recherche se fait par mots clés du titre ou du texte intégral à l'intérieur de ce qui a été trouvé et indexé par le moteur de recherche. Malgré le fait que la recherche par sujets puisse s'avérer pertinente, particulièrement pour les non-initiés ou pour les utilisateurs peu familiers avec un domaine de recherche, la majorité des services utilisant les moteurs de recherche privilégient la recherche par mots clés. Par ailleurs, bien que la plupart des moteurs de recherche suivent, jusqu'à un certain degré, les mêmes protocoles ou les mêmes règles d'indexation et de repérage, les résultats d'une même recherche peuvent différer de façon significative d'un moteur de recherche à un autre. Ceci est dû à plusieurs raisons, entre autres, au fait que certains moteurs de recherche :

- règlent différemment les trois parties qui les composent, à savoir le robot, l'index et le logiciel de recherche ;
- indexent plus de pages que d'autres ou indexent les pages Web plus souvent que les autres, ce qui implique qu'ils n'ont pas la même collection ;
- ajoutent des critères d'indexation qui leur sont propres comme la popularité d'un site ;
- ne peuvent pas lire toutes les pages et donc ne sont pas en mesure de les indexer (exemple : les pages qui comportent des cadres, des graphiques ou des tableaux en haut de la page) ;
- ne peuvent pas identifier les liens qui pointent vers les autres pages d'un même site ou ne sont pas en mesure de suivre certains liens à l'intérieur d'un site, ce qui implique que certaines pages ne sont pas indexées.

Lors d'une tâche de recherche, le logiciel du moteur de recherche passe au crible et filtre des millions de pages enregistrées dans l'index afin de trouver les documents correspondant à une question de recherche. Les résultats ou les réponses sont affichés sous forme d'une liste de pages ou de sites Web. Chaque site est en fait un lien hypertexte cliquable qui renvoie à la page concernée. Le classement des résultats suit un ordre de pertinence telle que calculée par le moteur de recherche : les pages les plus pertinentes sont classées au début parmi les dix premiers résultats affichés.

Pour déterminer la pertinence, les moteurs de recherche suivent un ensemble de règles dont les plus importantes sont :

- La localisation d'un mot clé : une page Web est jugée pertinente lorsque les mots clés utilisés dans la question de recherche figurent au titre. Les moteurs de recherche vérifient également si ces mots clés apparaissent en haut de la page Web (ex. en vedette ou dans les premiers paragraphes du texte). Certains moteurs capables de lire les métaétiquettes vérifient l'existence de tels mots clés dans certaines étiquettes tels le titre, les mots clés ou la description.
- La fréquence d'un mot clé dans une page Web est un autre facteur qui permet au moteur de recherche de déterminer la pertinence. Ce dernier analyse également la fréquence d'apparition du mot clé en question en relation avec d'autres mots dans une page Web.

La pertinence telle que calculée par le moteur de recherche n'amène pas nécessairement un résultat exact. En effet, certaines pages non pertinentes peuvent se trouver au début de la liste des résultats affichés ; un réajustement de recherche s'impose donc parfois. Néanmoins, vu la taille des bases de données et la vitesse de réponse, il faut reconnaître que certains moteurs de recherche s'avèrent malgré tout performants. Également, il importe de souligner que dans le cas d'une question de recherche très vague ou mal formulée, le moteur de recherche n'a aucun moyen de la préciser comme c'est souvent le cas dans une entrevue de référence avec un intermédiaire humain (Sullivan 1999).

L'importance d'un site Web ne garantit pas son indexation ni son classement parmi les dix premiers résultats affichés lors d'une recherche. Plusieurs sites Web sont rarement cités ou totalement ignorés ; ceci est dû particulièrement à l'ignorance des concepteurs du fonctionnement des moteurs de recherche et aussi du design facilitant l'indexation et le repérage du site Web.

Par ailleurs, plusieurs moyens peuvent contribuer à propulser une

page Web, lors du repérage et du classement, parmi les premiers résultats affichés. On trouve, entre autres :

- **Le critère de popularité** mesurée par le nombre de liens qui pointent vers une page ou vers le site contenant la page en question. *Excite*, à titre d'exemple, utilise le critère de popularité comme faisant partie de ses conditions d'indexation et de classement, car le nombre de liens pointés vers une page est une mesure de crédibilité du contenu de la page en question auprès de la communauté des internautes.
- **la fréquence d'usage des mots clés utilisés dans la question de recherche** : certains concepteurs de pages Web, conscients de l'importance de mots clés et de leur fréquence dans le repérage pour le positionnement en haut de la liste des résultats ont tendance à répéter les mots clés sans raison. Les moteurs de recherche sont, toutefois, en mesure de détecter ces répétitions et pénalisent ces pages Web en les excluant de leurs index.
- **la présence des mots clés de recherche dans des métaétiquettes** tels le titre, la description ou les mots clés. En effet, certains moteurs de recherche tels *HotBot* et *Infoseek* utilisent les mots clés indiqués dans les métaétiquettes, ce qui augmente la visibilité et le repérage de certaines pages. Cependant, d'autres moteurs de recherche comme *Lycos* ne lisent pas les métaétiquettes, ce qui fait que bien que plusieurs pages n'aient pas de métaétiquettes, elles peuvent être bien positionnées dans le classement.
- **le design de la page Web ainsi que le choix des éléments à indiquer au début** peuvent jouer un rôle important. En effet, indiquer au début de la page Web (titre et premiers paragraphes) les différents mots clés qui reflètent le sujet traité ou fournir une brève description qui traduit le contenu de la page en question aide dans le repérage et le classement.

Le positionnement parmi les premiers résultats, selon Sullivan (1999), a un impact très considérable aussi bien au niveau économique que scientifique et technique. En effet, souvent après avoir consulté le premier et rarement le deuxième écran des résultats affichés, les utilisateurs se lassent très rapidement et abandonnent l'examen des résultats. Beaucoup préfèrent reformuler leur recherche ou changer de moteur de recherche. Ceci implique que plusieurs entreprises ainsi que de nombreux créateurs et auteurs ne seront probablement jamais détectés si le moteur de recherche ne les classe pas parmi les premiers. Pour améliorer leur positionnement dans le classement au moment de la recherche, certains sites Web négocient des liens réciproques avec les sites populaires qui apparaissent souvent parmi les premiers dans une liste (le principe de citant — cité). Notons que tout le Web est principalement bâti sur des liens hypertextes, ce qui implique que ces derniers constituent le moyen le plus efficace pour identifier un site et y accéder. Les liens servent de renvois vers d'autres pages et peuvent être représentés par un texte souligné ou coloré différemment, une image « cliquable » ou une partie d'un graphique « sensitif » (ex. *ImageMap*). Les renvois peuvent concerner d'autres pages Web ou différents services Internet accessibles par l'intermédiaire d'un URL (FTP, Gopher, news, etc.).

L'indexation manuelle

L'indexation dans Internet se fait habituellement de façon automatique. Certains index sont cependant construits de façon manuelle. Les index compilés manuellement sont généralement de couverture très limitée et indexent des masses d'information restreintes (un site Web, une page Web, des passages ou des sections d'un document électronique). Bien que le terme index soit souvent utilisé dans la littérature traitant d'Internet pour désigner un répertoire ou une organisation des sujets par catégories ou sous forme arborescente, il importe de noter qu'un index diffère d'un répertoire dans le sens où il n'offre pas de classes hiérarchiques. En effet, un index est une simple liste de sujets classés par ordre alphabétique ou par thèmes : « *The type of index [that] has a simple structure : direct access to natural-language terms is provided by the alphabetical arrangement* » (Weinberg 1996, 6).

Dans un livre, l'index de type *back-of-the-book index* renvoie aux numéros de pages contenant les sujets. Dans le Web, les sujets figurant dans l'index sont liés à leurs sources d'information par des hyperliens. Ainsi en cliquant sur un sujet, l'utilisateur accède directement à la page ou la partie qui lui est reliée. Ces types d'index sont offerts par certains créateurs de sites désirant guider leurs utilisateurs vers une information précise ou un passage d'une page ou d'un document Web.

L'index dans Internet est rarement utilisé avec la même exhaustivité et le même niveau de granularité qu'offre un index de fin de livre. Il est surtout employé pour organiser alphabétiquement et faciliter l'accès à un document, à une collection ou à une série de ressources se rapportant à un domaine particulier comme, à titre d'exemple, *ResearchIndex* (NEC Research Institute 1999) qui donne accès à la littérature scientifique. Parmi les organisations fournissant ce type d'index, on trouve le *Cornell University Mann Library Gateway Catalog* qui donne accès à des centaines de bases de données dans Internet ; la JISC qui donne accès à une table[7] qui sert aussi bien d'index que de glossaire à l'ensemble des projets dans lesquels JISC est impliqué. Les projets et les activités classés par ordre alphabétique sont des liens cliquables permettant d'accéder à des pages Web à l'intérieur du secteur JISC ou à d'autres sites Web dans Internet. Le fait d'activer dans la liste alphabétique, fournie au début de l'index-glossaire, une lettre permet de se positionner au niveau alphabétique désiré ; l'American Society for Indexers qui donne accès à un ensemble de sites fournissant à leur tour l'accès à un ou à plusieurs index, notamment *Adobe, Association of brewers, BT challenge, California page, Policies and procedures manual at the University of Texas at Austin, Ultime liste, Unixhelp for users, U.S. Census Bureau, U.S. Government Printing Office, Victoria and Albert Museum, IEEE Spectrum* et d'autres. Ces index agissent comme un ensemble de liens ou de points de départ vers divers sites pertinents du même domaine.

À côté des index, on trouve également les tables des matières ou *contents* qui présentent le contenu d'un document Web, généralement une publication électronique (article ou livre). La table des matières agit comme un ensemble de liens

7. <http://www.jisc.ac.uk/glossary/index.html >

hypertextes qui renvoient à différentes parties composant la publication en question (introduction, chapitre ou paragraphe) à l'intérieur de la même page Web ou dans le site Web qui loge les différentes pages du document, et ce, lorsque le document est organisé sous forme de plusieurs pages électroniques.

Par ailleurs, tout chercheur peut facilement constater que les mécanismes d'indexation et de repérage utilisés actuellement dans le Web présentent un certain nombre d'inconvénients dont le plus important est l'absence d'une précision lors de la recherche par mots clés libres (Robertson 1996). De là, l'idée de faire appel à quelques outils traditionnellement employés en bibliothéconomie et en sciences de l'information, notamment les vocabulaires contrôlés se présentent comme une solution prometteuse.

L'usage de vocabulaires contrôlés existants : thésaurus ou listes de vedettes-matières

L'usage de vocabulaires contrôlés dans Internet, notamment les thésaurus, qu'ils soient intégrés ou générés de façon automatique, vise d'un côté l'organisation des contenus des différentes ressources distribuées un peu partout dans le réseau et, d'un autre côté, la recherche d'une solution pour résoudre quelques difficultés d'ordre linguistique tels les problèmes de synonymie, de polysémie, d'homographie ou autres. Le recours aux vocabulaires contrôlés est justifié par le besoin d'améliorer le taux de rappel et de précision lors d'une tâche de recherche et de repérage de l'information. On trouve d'ailleurs dans Internet quelques sites qui tentent de définir ce qu'est un vocabulaire contrôlé et à quoi il sert ou proposent des services, des directives ou des normes pour construire un tel outil. À titre d'exemple, le *Zthes* (Taylor 1999) est un document électronique qui décrit comment représenter et chercher les termes dans un thésaurus ou dans d'autres hiérarchies selon la norme ISO 2788 et spécifie comment un tel modèle peut être implanté en utilisant le protocole Z39.50. Également, le site de Willpower[8] présente une définition du thésaurus, décrit son utilité et donne un aperçu sur l'organisation d'un thésaurus, notamment la présentation des termes et des renvois.

Plusieurs projets ont choisi d'intégrer des thésaurus ou des listes des vedettes-matières existants, généraux ou spécialisés, construits de façon traditionnelle, à des bases de données afin d'indexer les collections disponibles dans Internet et de faciliter leur accessibilité. Parmi ces outils, on identifie, entre autres, la LCSH (Library of Congress Subject Headings) qui a été utilisée pour indexer et donner un accès par sujets aux ressources électroniques universitaires cataloguées dans INFOMINE et par l'OCLC pour gérer le contenu de *NetFirst*[9], le *Women's thesaurus* qui a été proposé par Ward et Olson (1998) en tant que prototype pouvant être utilisé par un groupe particulier, à savoir les femmes. La composante base de données du système contient le thésaurus *Women's Thesaurus* 6000 termes ainsi que des liens vers la classification Dewey. L'interface de recherche permet à l'utilisateur de naviguer à travers le thésaurus, choisir les termes pertinents et visualiser les indices de classification Dewey correspondant au terme sélectionné. Un indice Dewey peut être sélectionné et transmis à l'interface Web d'un catalogue en ligne. Pour créer la base de données, les auteurs ont utilisé une version électronique de *Women's Thesaurus* et ont converti les liens Dewey en une liste de termes qu'ils ont ensuite ajoutés au thésaurus. Dans ce projet, les indices Dewey servent pour une localisation topographique électronique des documents semblable à une localisation sur rayonnages. En d'autres termes, les indices Dewey permettaient de faire le lien entre les termes dans le thésaurus et la collection de la bibliothèque.

D'autres thésaurus spécialisés ont été intégrés au Web pour organiser et donner accès à des collections ou à des sites spécifiques. Parmi d'autres, ADAM utilise le *Art & Architecture Thesaurus*, EELS utilise *Engineering Information Inc's EI Thesaurus*, OMNI utilise *MeSH (Medical Subject Headings)*, SOGIG utilise le thésaurus *HASSET* (*Humanities and Social Science Electronic Thesaurus*) basé sur le thésaurus de l'UNESCO. On trouve également dans Internet le *Macrothésaurus de l'OCDE* spécialisé en sciences économiques et sociales, le *Thésaurus ERIC* spécialisé en éducation, le thésaurus *INFODATA* spécialisé en sciences de l'information et d'autres vocabulaires contrôlés[10]. Ces outils ont été intégrés pour permettre un accès par sujets aux diverses ressources se trouvant sur le Web. Ils agissent en tant que liens et remplissent une fonction semblable à celle des groupements thématiques permettant ainsi de répartir les ressources électroniques par catégories.

La génération automatique de thésaurus

Malgré les nombreuses tentatives visant à améliorer la variété dans les termes utilisés pour faire une recherche automatisée, et ce, en utilisant des thésaurus existants et bien qu'un thésaurus créé humainement offre plus de précision et de richesse sur le plan sémantique, certains auteurs, dont Chen et al. (1998a), pensent que les besoins cognitifs exigés pour créer et maintenir à jour de tels outils sont énormes et dépassent largement la capacité humaine. Ces auteurs voient que la génération automatique des thésaurus[11] est une façon de faire qui, en plus de permettre une couverture plus vaste et plus complète offrant une variété de termes extraits directement des contenus, est plus facile à maintenir et à mettre à jour. Ceci convient parfaitement à un environnement Internet où les contenus se caractérisent par une certaine volatilité. En outre, l'usage des termes extraits automatiquement, selon Schatz et al. (1996), améliore le taux de rappel en exprimant le contexte dans lequel le terme est employé.

C'est dans ce sens que plusieurs expérimentations visant la génération automatique des thésaurus ont été entreprises : Chen et al. (1995) ont testé la création d'un thésaurus pour des biologistes spécialisés en vers et insectes volants, et ce, en utilisant diverses sources disponibles dans Worm Community System (WCS), notamment le *Worm book*, les résumés d'articles de périodiques spécialisés en provenance de *Medline* et de *Biosis*, le *Worm Breeder's Gazette* et les comptes rendus de conférences. La création automatique d'un tel outil dans le cas de Chen et al. (1995) a été justifiée, tout d'abord, par l'absence d'un thésaurus

8. <http://www.willpower.demon.co.uk/thessoft.htm>
9. *NetFirst* est un index qui fournit en plus des résumés.
10. Voir Koch (1996) pour la liste complète des thésaurus spécialisés par domaine.
11. Ces thésaurus ne sont pas constitués d'un vocabulaire contrôlé.

dans le domaine et par la difficulté de créer un thésaurus traditionnel qui exige énormément de temps et d'efforts humains ; ensuite, par la nécessité de pallier les problèmes liés aux différences de vocabulaires utilisés par les divers membres de la communauté scientifique. Ces derniers proviennent aussi bien de l'interne que de l'externe et leur niveau de connaissances varie d'expert à novice. La génération du thésaurus a été faite de façon statistique en se basant sur des techniques de filtrage des termes en les comparant à une liste d'autorité contenant, entre autres les noms des chercheurs dans le domaine, les noms de gènes et les méthodes expérimentales, l'indexation automatique et l'analyse de clusters. L'analyse de clusters permettait d'assigner à chaque terme une pondération selon sa fréquence et d'identifier les termes spécifiques.

De même, Chen et al. (1997) ont entrepris une autre expérience similaire visant à générer automatiquement un thésaurus vu en tant qu'espace concept ou réseau de termes et d'associations pondérées. Le thésaurus en question devrait permettre de faire une recherche interdisciplinaire. Chen et al. (1998a) ont réalisé une autre expérimentation à grande échelle visant la réduction de l'incertitude dans le repérage, particulièrement dans des espaces d'information très larges tel le réseau Internet. L'expérience en question a couvert plus de 400 000 résumés dans une collection spécialisée en ingénierie informatique fournis par INSPEC. Elle s'est basée sur une approche algorithmique utilisant le filtrage par objets, l'indexation automatique et l'analyse de cooccurrences. Les auteurs ont utilisé deux systèmes de génération de thésaurus : 1) la combinaison des méthodes de filtrage par objets en générant une liste des descripteurs, des personnes sujets et des noms des auteurs extraits à partir des étiquettes MARC 650, 651 et 100 et l'indexation automatique permettant d'ajouter d'autres termes (candidats-descripteurs) ; 2) l'indexation automatique seule (extraction des concepts après élimination de mots vides et adjectifs). Ces deux systèmes ont été comparés à un thésaurus construit suivant les méthodes traditionnelles. Les résultats d'une telle expérience ont permis de constater que les thésaurus générés automatiquement selon les deux méthodes étaient supérieurs en ce qui a trait au rappel. Cependant, ils sont égaux avec le thésaurus traditionnel en ce qui concerne la précision des concepts. De même, les auteurs ont pu retenir que les termes proposés par les trois thésaurus (les thésaurus construits automatiquement suivant les deux méthodes et le thésaurus traditionnel) étaient complémentaires et pouvaient donc être exploités pour augmenter la variété des termes de recherche et réduire l'incertitude dans la recherche et le repérage.

D'autres expériences dans le même sens ont été réalisées par Schatz et al. (1996). Ces derniers ont proposé un prototype permettant de combiner l'usage d'un thésaurus construit de façon traditionnelle et une liste générée statistiquement (par cooccurrence de termes présents dans un corpus textuel). Les auteurs cherchaient à développer une interface utilisateur qui devrait servir en tant qu'interface Web pour l'Université de l'Illinois (Digital Library Initiative). Le prototype devrait permettre de faire des recherches en utilisant les termes suggérés par le thésaurus et ceux proposés par la liste produite automatiquement afin de vérifier la complémentarité des deux outils. Les auteurs concluent que la variété dans les termes favorise une meilleure accessibilité aux ressources dans le Web.

Par ailleurs, étant donné l'interactivité qui caractérise l'environnement hypertexte et multimédia du réseau Internet, le recours à une représentation graphique des données est une pratique commune. En plus de permettre la structuration des connaissances, une telle représentation offre des attraits visuels et des schémas plus significatifs pour les utilisateurs finaux. C'est dans cette optique que Chen et al. (1998) ont tenté une autre expérimentation basée sur un algorithme favorisant le groupement des contenus textuels sous forme de clusters. Les auteurs pensent que, vu l'immensité de l'espace d'information couvert par Internet, les problèmes de repérage sont encore plus importants. En effet, les difficultés d'ordre linguistique, notamment les barrières sémantiques (Nadis 1996) sont plus marquantes et génèrent un manque de précision ou de rappel lors d'une recherche par mots clés ou par navigation. Les différences de vocabulaires, particulièrement les problèmes de chevauchements ou de redondance entre les termes, ne permettent pas un repérage d'une certaine efficacité (Chen et al. 1998).

Pour pallier ces problèmes, les auteurs ont testé deux algorithmes qui ont été développés par leur groupe de recherche : 1) *le Kohonen algorithm category map* qui permet d'améliorer la navigation dans l'espace hypertexte et 2) le *Kohonen Self-Organizing Map* ou *SOM* qui permet la génération automatique d'un espace conceptuel facilitant la recherche par concepts. Les résultats de l'expérimentation ont démontré que ce dernier permet de catégoriser efficacement un large espace d'information électronique dans Internet. En effet, l'algorithme a permis de structurer la sous-catégorie « Divertissement » de Yahoo en sous-espaces facilement gérables dans lesquels l'utilisateur peut naviguer avec succès pour localiser une page Web. Chen et al. (1998) concluent que l'algorithme *SOM* convient parfaitement à des tâches de navigation qui couvrent un espace d'information très large et dans lequel l'utilisateur doit se déplacer d'une catégorie à une autre. Par ailleurs, l'expérience a démontré que les aspects visuels et graphiques de la carte ont été très appréciés par les sujets participants. D'autres, par contre, préfèrent l'usage des modèles mentaux familiers telle l'organisation alphabétique linéaire ou hiérarchique et ont trouvé que la carte ne fonctionnait pas convenablement. Dans l'ensemble, les auteurs trouvent que les résultats de cette expérience de recherche dans un environnement vu en tant qu'espace conceptuel sont encourageants. En effet, les résultats indiquent que le fait de combiner les termes des utilisateurs et ceux suggérés par le thésaurus a généré un taux de rappel significatif comparé à un repérage basé uniquement sur les termes proposés par les sujets participants. L'analyse des pages Web obtenues indique très peu de chevauchements et de redondance parmi les pages repérées suivant les deux méthodes (l'usage des termes suggérés par les utilisateurs et ceux proposés par le thésaurus). En effet, dans la majorité des cas, l'ensemble des pages Web repérées étaient variées. Par ailleurs, les sujets participants ont apprécié le niveau de contrôle qu'ils peuvent exercer sur la recherche d'autant plus que les termes suggérés par le thésaurus proviennent directement du contenu des pages Web. Ceci garantissait un certain succès de repérage.

Les systèmes de classifications hiérarchiques

La navigation par les liens hypertextes est souvent utilisée pour visualiser et consulter le contenu des documents Web. La représentation des sujets de façon hiérarchique, selon plusieurs auteurs, notamment Jones (1996), favorise la navigation à travers les structures d'information complexes. On peut, d'ailleurs, facilement constater que ce type de recherche et de consultation des contenus électroniques reste l'un des moyens les plus répandus pour naviguer d'une page Web à une autre, d'une section à une autre ou parmi différents serveurs. Généralement, les informations « navigables » sont regroupées en catégories organisées sous forme d'une table des matières expansible et visualisées par l'utilisateur en tant que données textuelles ou graphiques (Jones 1996). Les informations organisées sous forme arborescente ou *subject tree* sont présentées en tant que réseaux de nœuds. Ainsi, l'activation d'un nœud (parent) permet de visualiser le niveau suivant (nœud enfant) ou l'inverse (enfant — parent) avec toutes les informations se rapportant à ce nœud, permettant ainsi à l'utilisateur de faire davantage de choix. Le déplacement se fait en suivant la hiérarchie du haut en bas, verticalement ou horizontalement.

Bien que le concept de *subject tree* ou d'organisation des sujets sous une forme arborescente date de très longtemps, ses applications pour organiser les ressources dans Internet sont très grandes et trouvent leurs origines dans des expériences plus récentes avec les serveurs Gopher. En effet, bien avant le Web, le serveur Gopher a permis de naviguer à travers les différents types de ressources indexées d'Internet (texte, images, son, vidéo). La navigation dans l'espace Gopher se faisait en passant d'un menu à un autre ou en faisant des recherches par mots clés à l'aide du logiciel *VERONICA* (Very Easy Rodent-Oriented Netwide Index to Computerized Archives). Construites de façon logique sous forme de catalogues, les catégories dans les menus Gopher mènent à des sous-catégories, lesquelles permettent d'accéder à d'autres sous-catégories. De là, il s'est avéré approprié, pour organiser ce type de menus, d'exploiter la structure hiérarchique fournie par les systèmes de classifications traditionnels.

Les systèmes de classifications traditionnels ont été utilisés dans plusieurs projets en tant que schémas de base pour classer hiérarchiquement une table des matières, un domaine de connaissances, un ensemble de ressources électroniques (pages ou sites Web) ou des groupements thématiques. Il importe de souligner que lorsqu'on parle de classification dans Internet, c'est souvent en relation avec une classification traditionnelle existante rendue lisible par machine et intégrée au système, une classification des ressources en catégories (ex. Yahoo) ou une classification générée de façon automatique.

Les classifications numériques universelles

Plusieurs systèmes traditionnellement utilisés dans les pratiques bibliothéconomiques pour cataloguer, indexer, organiser et classifier les collections des bibliothèques ont été repris et réutilisés pour traiter les ressources dans Internet. Selon Woodward (1996, 190), le terme « traditionnel » a une définition arbitraire dans le sens où il est utilisé pour désigner tous les systèmes utilisés exclusivement par les bibliothèques et les services d'indexation ayant été développés vers la fin du XIX^{e} siècle et tout au long du XX^{e} siècle. Ces systèmes incluent, entre autres, la Classification Décimale Universelle (CDU), la classification Dewey (DDC), la classification de la Bibliothèque du Congrès (LCC), les vedettes-matières de la Bibliothèque du Congrès, la *Medical Subject Headings* (*MeSH*) et la *Sears Subject Headings*. Il faut noter, cependant, que les ressources dans Internet ne sont pas cataloguées ou classifiées dans le vrai sens du terme ou du moins tel que ces opérations documentaires sont perçues par les bibliothécaires.

Plusieurs projets ont adopté des classifications numériques existantes pour organiser les ressources électroniques. Ainsi le portail de NISS se présente comme un répertoire où les différents utilisateurs d'Internet sont guidés via une classification par sujets organisés sous une forme arborescente vers le thème de leur choix. Ils peuvent également survoler la liste alphabétique des sujets organisés suivant la classification CDU ou faire une recherche par mots clés dans le titre, la description des ressources ou le vocabulaire fourni par les tables de classification CDU (Haynes 1998, 11).

L'organisation des ressources électroniques dans le Web sous forme de catégories à parcourir a été également testée par bon nombre de projets, notamment CyberDewey, qui a adopté en tant que schéma de base la Classification décimale de Dewey (CDD), WWW Virtual Library et CyberStacks, qui utilisent la classification de la Bibliothèque du Congrès (LCC), et BUBL, qui exploite le système de Classification décimale universelle (CDU). Dans chacun de ces projets, on assigne à chaque ressource dans Internet une notation à partir de la table de classification utilisée. Puisque les ressources dans Internet ne sont pas cataloguées de façon intégrale, l'accès par sujet reste très limité. Les utilisateurs sont contraints de limiter leurs objectifs de recherche à chaque niveau de la hiérarchie correspondant au niveau de la classification en question. Également, bien que certaines ressources soient indexées, la plupart manquent d'organisation (Schatz 1997). C'est en choisissant à partir de plusieurs écrans que l'utilisateur finit par atteindre l'objet de sa recherche. Ceci revient à donner à chaque utilisateur une copie des tables de classification et à lui permettre de suivre les branches se rapportant aux sujets par différents niveaux de spécificité. Malgré la multitude des écrans à parcourir, cette solution paraît plus appropriée que d'explorer les résultats de recherche par mots clés qui, souvent, aboutissent à un cul de sac.

Les notations dans CyberDewey, CyberStacks et WWW Virtual Library se limitent principalement à une notation simplifiée empruntée à des tables des classifications sélectionnées. L'usage de la LCC pour classifier les liens dans Internet permet d'ajouter la structure, le contenu et la description selon McKiernan (1997), concepteur de CyberStacks. Bien que limité aux domaines de science et de technologie, CyberStacks peut éventuellement être un prototype très intéressant pour l'usage d'une classification traditionnelle puisqu'on compte intégrer au système de classification un thésaurus hypertexte avec vocabulaire contrôlé. De leur côté, Ward et Olson (1998) trouvent que la classification Dewey présente certains avantages, notamment le potentiel d'accommoder des changements positifs et la flexibilité organisationnelle. Selon ces auteurs, les responsables du développement de

Dewey travaillent actuellement à développer et à améliorer la représentativité de certains domaines marginalisés telle la classe réservée aux sujets féministes. La classification Dewey est utilisée pour organiser la base de données *NetFirst* de l'OCLC. Elle est appliquée à plus de 40 000 notices décrivant les ressources électroniques dans Internet (Vizine-Goetz et Mitchell 1996). Dans *NetFirst*, on utilise plusieurs indices de classification pour un même document, car la classification n'est pas utilisée pour une localisation topographique, mais plutôt pour assigner un indice sujet à l'intérieur de la structure. Ainsi, lorsque le document traite de plusieurs sujets, on lui attribue plusieurs codes de classification (idem). La CDD a été également adoptée pour organiser les ressources dans certaines passerelles d'information spécialisées par sujets, notamment ADAM et Biz/ed. L'OCLC a également entrepris des expérimentations, et ce, dans le cadre du projet Scorpion, visant à évaluer le coût-efficacité de l'usage de la classification Dewey dans une tentative d'attribuer de façon automatique les indices de classification sujets aux ressources électroniques (Thompson, Shafer et Vizine-Goetz 1997).

Par ailleurs, malgré les progrès réalisés par la classification de la Bibliothèque du Congrès et celle de Dewey, la Classification décimale universelle, en plus d'être rendue la première lisible par machine et prête pour une utilisation électronique, a connu une réorganisation minutieuse mieux adaptée aux besoins des environnements électroniques (Woodward 1996). Le texte complet de la classification, incluant les notations, les descriptions, les annotations, les références et les exemples, se trouve sur le site de la CDU et peut être utilisé par les organisations pour traiter leurs propres collections. Des efforts sont également consacrés à la révision de chaque classe de la CDU et à la mise à jour des tables, particulièrement en ce qui concerne les secteurs scientifiques. Actuellement, au moins cinq services Internet, utilisent la CDU: BUBL, GERHARD (German Harvest Automated Retrieval and Directory)[12] la passerelle d'information NISS, OMNI et SOGIG (Koch 1998).

Plusieurs projets exploitant la CDU ont vu le jour. Ainsi, afin d'améliorer les outils de découverte et de repérage de l'information existants, le projet Nordic (Nordic WAIS/WWW), qui est le résultat des efforts de la Lund University Library en Suède, et de la National technological library du Danemark cherche à fournir un accès par sujets aux bases de données disponibles dans Internet qui sont accessibles via une interface d'accès commun WAIS[13]. L'accès par sujets se fait par le biais du système de notation de la CDU. La classification des ressources se fait d'abord par l'extraction des termes se trouvant dans plusieurs champs, notamment les champs descriptifs, les champs de mots clés, les champs sujets ainsi qu'une liste locale de mots clés construite pour chaque base de données WAIS. Lorsqu'un terme de la base de données correspond au terme utilisé dans le vocabulaire de la CDU, il est ajouté à la liste, appuyé d'une pondération et d'un argument favorable. De telles manipulations permettent d'assigner aux différents éléments composant chaque base de données WAIS une notation et de la relier ensuite à un format de représentation des sujets (WAIS/CDU) organisés sous une forme arborescente. Le projet compte étendre ce genre de traitement à l'ensemble du vocabulaire CDU disponible en format électronique. Le but est de permettre un accès direct à différentes bases de données pertinentes WAIS, et ce, à partir d'un mot clé sans contraindre l'utilisateur à employer et à comprendre la structure de la classification CDU. Bien que les efforts aient été principalement orientés vers l'amélioration d'accès aux bases de données WAIS, des tentatives visant à unifier le WAIS et le Web sont en cours (Ardo et al. 1999).

La Lund University Electronic Library en Suède, la Swedish University of Technology Libraries et la Engineering Electronic Library (EELS) ont développé une sorte de classification combinant la CDU à une liste de sujets organisés sous une forme arborescente nommée UDC-based subject trees . Les concepteurs de ce projet cherchent à faciliter et à améliorer le repérage des ressources d'information électroniques dans Internet en créant des points d'accès pouvant être indexés, consultés et cherchés (Ardo et al. 1999).

Il importe de noter que la majorité des projets d'indexation et de classification automatiques visent avant tout à trouver des solutions permettant de se passer de l'intervention humaine jugée fastidieuse et coûteuse. Parmi ces projets, on identifie le projet UDC-AUTCS (UDC Number Automatic Combination System) de contribution japonaise qui consiste, selon ses concepteurs, en un système interactif personne-machine permettant d'assigner des numéros CDU combinés (Woodward 1996).

Ce système procède d'abord en identifiant les principaux termes susceptibles de traduire le contenu d'un document et en cherchant les numéros de classification (notation de base de chaque terme) correspondants à partir des tables de la classification CDU. A ce niveau, le système attribue à chaque mot identifié une valeur d'importance variable selon que le mot clé a été repéré à partir du champ sujet ou représente d'autres caractéristiques du document à traiter. Les mots clés sont ensuite comparés individuellement au vocabulaire de la classification CDU, et ce, avant qu'on leur assigne une notation. La production d'un indice de classification global et unique se fait en combinant ensemble les notations individuelles de base, et ce, en conformité avec les règles et les procédures de combinaison. Bien que prometteur, le système nécessite quelques réajustements avant de permettre la classification des documents sans l'assistance d'un bibliothécaire. En effet, une comparaison entre la classification d'une vingtaine d'articles choisis au hasard en utilisant le système UDC-AUTCS avec la classification faite suivant les techniques manuelles traditionnelles a permis de relever un certain nombre d'erreurs. L'expérience a permis de constater que dans huit cas, les deux systèmes de classification étaient en désaccord. La divergence est due, selon les concepteurs du système, à un désaccord en ce qui a trait à la méthode d'analyse du contenu employée par le bibliothécaire et l'opérateur du système (Woodward 1996).

Par ailleurs, le laboratoire KBS-Media de la Lund University en Suède travaille sur un projet intitulé KBS-CROSS (Knowledge-Based System for Automatic Cross-Referencing of Classification Systems). Le but de ce projet consiste à produire un outil informatisé qui permet de faire des renvois entre le système de classification de la Bibliothèque du Congrès (LCC) et la Classification décimale univer-

12. <http://gerhard.bis.uni-oldenburg.de/ >
13. WAIS est un immense catalogue de bases de données indexées d'une manière uniforme. Le repérage dans WAIS se base sur la fréquence de mots clés et leur localisation au début du document.

selle (CDU), particulièrement dans le domaine d'architecture et de bâtiments. Le projet vise également à fournir un mécanisme de conversion qui permet aux différentes notices en provenance de sources diverses et variées, notamment en langues étrangères, d'être regroupées ensemble dans un seul index « cherchable ».

Les classifications par domaine ou à facettes

OMNI, qui est un projet supporté par ROADS, vise la construction d'un portail pour la communauté des chercheurs et des enseignants universitaires en vue de leur faciliter l'accès à une information d'un niveau de qualité élevée particulièrement en ce qui touche les aspects cliniques, de recherche et de gestion en santé et en biomédecine. Le projet compte créer un catalogue des ressources d'information disponibles en réseau par un processus de découverte, de filtrage, de description, de classification et d'indexation. Les ressources seront classifiées, entre autres, à l'aide du schéma de classification NLM (National Library of Medicine) largement utilisé en Angleterre par la communauté médicale. D'autres classifications numériques ou alphanumériques spécialisées par domaine ont été intégrées pour organiser diverses collections, notamment la Danish Veterinary and Agricultural Classification, la Engineering Information Classification Codes, la Mathematical Classification, le MathGuide, la Dutch Electronic Codes, ACM Computing Classification System, la Anglo-American Literature Guide, la GeoGuide, l'HistoryGuide et d'autres classifications[14] McKiernan 1999).

Rares sont les projets qui se sont penchés sur l'usage d'une classification à facettes. Pourtant certains auteurs, notamment Richmond cité par Woodward (1996) pensent qu'une classification à facettes permet l'accès aux ressources à partir de plusieurs points d'accès sous forme de liens sujets. Dans ce sens, l'auteur propose un système intitulé Web librarian basé sur le principe d'une organisation à facettes. Pour développer les facettes, il suggère de commencer d'abord par définir le ou les sujets à couvrir, et ce, en examinant les classifications existantes, le ou les thésaurus utilisés, les titres ou les objets dans la base de données à traiter. Les thèmes dérivés de la base de données sont ensuite décomposés pour former des facettes ayant chacune une étiquette distincte. Les facettes ainsi formées sont décomposées en sous-facettes dont les éléments sont classés dans un ordre allant du général au spécifique et des données abstraites aux données concrètes (Woodward 1996).

Les répertoires ou directories

Un répertoire est un service qui offre une série de liens donnant accès à un ensemble de ressources disponibles dans Internet, soumises par les créateurs ou les évaluateurs des sites et organisés en catégories de sujets (Cohen 1999). Ainsi, la classification manuelle consiste à situer chaque site à l'intérieur d'une catégorie prédéterminée et à compiler ensuite des répertoires qui servent également pour la recherche et le repérage du contenu des pages Web. D'ailleurs, le terme moteur de recherche est généralement utilisé dans la littérature pour décrire les logiciels de recherche et les répertoires. Bien que la méthode d'indexation des pages Web (de façon automatique ou manuelle) constitue la différence entre un répertoire et un moteur de recherche, la frontière entre ces deux outils tend à disparaître. En effet, les répertoires sont présents dans certains sites offrant des moteurs de recherche et le contenu des répertoires est parfois « cherchable » à partir du Web. À titre d'exemple, *Alta Vista* offre le répertoire de *LookSmart*; *Infoseek* partage son écran avec le répertoire de *GoNetwork*; *Excite* possède son propre répertoire et *Lycos* offre les contenus de répertoire à partir de *Netscape Open Directory*. Également, la recherche dans un répertoire est limitée aux regroupements des fichiers tels que définis par le serveur opérant le système Koehler 1999) de la même manière que la recherche par moteur de recherche se limite à ce qui est collecté et indexé par ce dernier.

Il faut noter, cependant, que les éléments ou les critères qui, dans un moteur de recherche, peuvent jouer un rôle dans le classement ne garantissent pas qu'un des sites classés parmi les premiers soit retenu dans un répertoire. En effet, pour figurer dans un répertoire, particulièrement ceux conçus par des institutions universitaires, le site doit répondre à un certain nombre de critères, notamment la qualité de son contenu. Les évaluateurs humains font l'inventaire et le tri parmi les différents sites annoncés et sélectionnent ceux identifiés comme prometteurs ou répondant à certains critères de qualité.

Par ailleurs, la création d'un répertoire est une tâche qui consiste à partager un large espace d'information en catégories thématiques distinctes ayant une signification pour les utilisateurs. Le fait de séparer en partitions les sujets permet de créer des bases de données de taille raisonnable, ce qui rend l'exploration plus efficace (Chen et al. 1998). La compilation des répertoires se fait généralement de façon manuelle par des ressources humaines qui surveillent continuellement les sites annoncés et font une sélection parmi les sites les plus pertinents.

Il existe deux types de répertoires :

- Les répertoires universitaires ou professionnels créés et entretenus par des experts du domaine ou des bibliothécaires qui sont généralement affiliés à des institutions universitaires ou à des bibliothèques.

La création de tels répertoires vise, d'un côté, l'amélioration des processus de recherche et, d'un autre côté, l'accès à des ressources Internet de qualité supérieure. En effet, les sites retenus sont soumis à des critères de sélection très sévères et sont annotés par les évaluateurs responsables de créer les répertoires. Généralement, la collection des sites retenus sert aux besoins de l'institution créatrice et à ses membres, mais peut être utile pour n'importe quel chercheur qui s'intéresse au contenu des sites choisis. INFOMINE est un exemple de répertoire universitaire créé par l'Université de Californie. On trouve également d'autres répertoires semblables tels BUBL Link qui donne accès à des ressources sélectionnées et annotées à partir de University of Strathclyde Library in Glasgow; Librarians' Index to Internet qui fournit une collection sélectionnée, bien organisée et mise à jour régulièrement par des indexeurs en Californie; WWW Virtual Library, considéré comme le

14. Pour la liste complète des classifications spécialisées ayant été intégrées pour organiser et donner accès aux diverses ressources disponibles sur le Web, voir McKiernan (1999).

premier répertoire dans le Web, qui fournit un ensemble de collections sur divers sujets annotés et entretenus par des experts à travers le monde.

- Les répertoires créés par des entreprises commerciales appelés également portails commerciaux.

Ces répertoires visent d'abord la génération d'un certain revenu et s'adressent au grand public. Ils sont reliés à un ensemble très large de sujets et se concentrent généralement sur des sites non couverts par les répertoires universitaires, notamment les loisirs, le commerce, les sports, les voyages et d'autres domaines d'intérêt général. Afin d'accroître le trafic et donc d'attirer les annonceurs, ces répertoires sont généralement offerts avec un bon nombre de services et de biens de consommation. Yahoo est un bon exemple de portail commercial. Il a été le premier service de recherche et de navigation Internet qui a pu également offrir un répertoire classé par catégories tels les sciences, les affaires, les loisirs, etc. (Chen et al. 1998). En effet, diverses personnes (professionnelles et non professionnelles de l'information) se chargent de classifier ou de catégoriser les sites ou les pages Web par thèmes. Elles font la description du contenu de chaque site visité et retenu ou utilisent la description fournie par ceux qui soumettent leurs sites à Yahoo afin qu'ils soient répertoriés. La description qui figure dans le répertoire sert lors d'une recherche par mots clés à identifier le ou les pages concernées et également, lors d'un classement des résultats de recherche, à donner un aperçu du contenu du document repéré.

Les répertoires diffèrent de façon significative en ce qui concerne les critères de sélection appliqués et le degré de leur application. D'ailleurs, plusieurs services de répertoires ne sont pas en mesure d'annoncer leur politique de sélection ni les qualifications et les spécialités des évaluateurs de leurs sites (Cohen 1999). Les répertoires compilés par des fournisseurs de services commerciaux tel Yahoo n'évaluent pas vraiment le contenu des sites sélectionnés. En effet, les évaluateurs ne font que classer les sites qui leur sont soumis dans une catégorie ou dans une autre. Par contre, si on prend l'exemple d'un service de répertoire universitaire tel Argus Clearinghouse, les contenus et les sites retenus sont soumis à des critères de sélection très sévères. On trouve d'ailleurs dans le répertoire d'Argus Clearinghouse plusieurs guides spécialisés par sujets de qualité supérieure compilés par des experts. Ces guides sont réévalués par le personnel de Clearinghouse et sont ensuite annotés pour les besoins des chercheurs.

La classification d'un site dans une catégorie ou dans une autre reste un point de divergence entre les concepteurs des sites et les évaluateurs (Woodward 1996). Certains concepteurs voient mal leur site dans une catégorie particulière. En plus, il n'est pas facile d'inclure dans une catégorie des pages d'information personnelles ou institutionnelles qui offrent des informations très diversifiées souvent faisant partie de plusieurs catégories. À titre d'exemple, sur un site universitaire, on peut trouver une variété d'information couvrant aussi bien les programmes, les conditions d'admission, les frais d'inscription, les offres d'emploi, la liste des professeurs avec des renvois à leurs pages Web personnelles, notamment leur curriculum vitæ, les notes de cours, les publications de l'université, les conditions et les frais d'abonnement, les associations spécialisées, les renvois vers d'autres universités du même type, les formations offertes dans Internet et sur le Web, les instructions concernant la création de pages Web, les modalités d'accès aux ressources locales et publiques comme l'accès à la bibliothèque du département et de l'université, les heures d'ouverture, des renseignements sur les services informatiques, des informations sur les activités de chaque département, les groupements et les associations des étudiants, les projets et les recherches en cours, les annonces de congrès et de conférences, etc.

Synthèse

L'examen de la littérature met en lumière la diversité des approches utilisées pour apporter un certain ordre dans Internet et améliorer, par la même occasion, l'accessibilité de l'information. On retient, néanmoins, que les solutions généralement proposées privilégient l'une ou l'autre des approches suivantes :

- le recours aux techniques de traitement, d'analyse et d'organisation de l'information traditionnellement appliquées en bibliothèques pour améliorer l'accessibilité par sujets ;
- l'amélioration des technologies de l'information utilisées dans Internet, entre autres, les moteurs de recherche et d'autres agents intelligents en les dotant de certaines capacités propres aux êtres humains, notamment l'autoapprentissage. Le recours aux méthodes d'intelligence artificielle apparaît comme la solution idéale.

Les tenants de la première approche pensent que l'organisation de l'information dans le Web est nécessaire pour une utilisation effective et efficace des ressources et que le Web fournit des informations utiles qui peuvent très bien être organisées en utilisant les principes bien établis dans les bibliothèques. Ils trouvent également que les bibliothèques ont une grande tradition dans l'organisation des collections et des connaissances. Paradoxalement, la plupart des solutions proposées et appliquées, même si elles s'inspirent des méthodes utilisées en bibliothéconomie et en sciences de l'information, viennent souvent de l'extérieur du cercle des bibliothèques et sont proposées par des spécialistes dans d'autres domaines. Steinberg cité par Woodward (1996, 192) l'a bien souligné en notant que la bibliothéconomie qui, selon toute évidence, devrait fournir les expertises pour organiser Internet ne s'est avérée d'aucune aide.

Ceux qui tiennent à réinventer la roue trouvent que la technologie actuelle employée par les moteurs de recherche ne permet pas l'accès à des ressources d'information de qualité supérieure. En effet, même en ayant recours à la recherche pondérée et à l'application des indices de pertinence, ces outils n'ont pas réussi à régler le problème d'évaluation des ressources et le filtrage des contenus non pertinents ou de qualité médiocre (Haynes et al. 1998). D'ailleurs, plusieurs fournisseurs de services de moteurs de recherche commencent à se tourner vers des ressources humaines afin de développer des approches plus structurées pour localiser et évaluer l'information. Yahoo en est un exemple, bien qu'il n'ait pas encore réussi à résoudre la question de qualité de façon définitive (Haynes et al. 1998).

Le catalogage, l'indexation et la classification sont fréquemment abordés comme des solutions pour organiser les ressources disparates et distribuées dans Internet. L'intégration d'outils documentaires traditionnels existants tels les thésau-

rus ou les systèmes de classifications se présente comme un autre choix pour faire face, d'un côté, aux problèmes de divergence du langage utilisé pour repérer et accéder aux ressources dans Internet, donner accès par sujet et, d'un autre côté, pour organiser et répartir les ressources à l'intérieur de catégories spécifiques. Les projets pour cataloguer, indexer, classifier les ressources dans Internet prolifèrent et les *Web libraries* se multiplient.

Rares sont ceux, cependant, qui font une différence entre les besoins des ressources électroniques et ceux des ressources sur papier. Pourtant, une différence importante existe entre les deux, et ce, sur plusieurs plans, notamment en ce qui concerne les détails de localisation, la structure interne du document, l'absence de stabilité (autant pour la localisation qu'en ce qui concerne les versions), les méthodes de publication ainsi que les caractéristiques des ressources en réseaux. L'administrateur hôte d'un document Web peut différer considérablement de l'organisation qui le publie. Toutes ces particularités propres aux ressources électroniques dictent des processus et des comportements de recherche différents et imposent un design particulier des systèmes de repérage de ces ressources.

Les tenants de la deuxième approche trouvent que les techniques traditionnellement utilisées dans les bibliothèques sont inutiles, particulièrement si les méthodes de repérage fournies par les outils de recherche et d'autres agents intelligents continuent de s'améliorer. Les techniques de catalogage, d'indexation et de classification sont vues comme étant excessivement coûteuses, fastidieuses et impliquant d'innombrables heures de travail.

D'abord, on reproche aux systèmes de classification traditionnelle le fait qu'ils n'ont pas suivi l'évolution de certains domaines de connaissances. La plupart des systèmes de classification en usage, notamment la Dewey et la LCC, développés au XIX^{e} siècle, sont inadéquats pour classifier les connaissances dans les domaines nouvellement établis telle l'ingénierie génétique et électronique (Steinberg cité par Woodward 1996). Selon ce dernier, même les bibliothécaires admettent que les schémas utilisés aujourd'hui sont archaïques et inappropriés. Également, plusieurs pensent que les classifications ne devraient pas être reliées dans l'espace cybernétique à un arrangement ou un classement linéaire qui a été développé pour les livres sur les rayonnages (Rockefeller College Press cité dans Woodward 1996). Weinberg (1996), quant à elle, trouve que le problème avec les systèmes de notations réside dans la difficulté de les maintenir à jour. À ce propos, elle souligne en page 6 que : « *people consult the Internet for the latest information ; they cannot wait for an international committee to decide in which class a new topic should be placed* ».

Ensuite, on reproche aux vocabulaires contrôlés et aux thésaurus traditionnels intégrés leur incapacité d'adaptation. Selon Chen et al. (1998), si ces thésaurus générés par des indexeurs humains sont en mesure de fournir des termes à utiliser dans la recherche, ils n'arrivent pas à régler le problème d'autoacquisition des connaissances. La demande cognitive qu'exige la construction et la manutention de tels outils est excessive. Hjorland (1998) pense aussi que lorsque la signification d'un terme à l'extérieur du système (thésaurus entre autres) est en train de changer et d'évoluer, la signification du terme à l'intérieur du système devient périmée. Aussi, si on tente d'utiliser le descripteur en concordance avec la signification courante, les systèmes perdent leur cohérence et l'idée d'utiliser un vocabulaire contrôlé perd sa raison d'être.

Finalement, on reproche aux divers efforts de catalogage visant l'organisation du Web, bien qu'ils restent (Koehler 1999) des efforts louables, leur inadéquation, et ce, sur deux plans : d'abord, l'inadéquation est stratégique, car bien que les documents électroniques et sur papier partagent des caractéristiques communes, il faut reconnaître que le Web est un médium différent des autres média de publication. En effet, le document cesse d'exister une fois que son propriétaire décide de le supprimer. De même, lorsque le concepteur du document décide de le publier sur le Web, la provenance intellectuelle du document est perdue (Koehler 1998). Ensuite, l'inadéquation est tactique puisque dans un document électronique, il y a la page Web et le site Web qui sont des caractéristiques non partagées avec les documents traditionnels (Koehler 1999, 23-24). D'autres, notamment Jul et al. cités dans Koehler (1999) pensent que les efforts pour organiser le Web sont inutiles pour trois raisons : le contenu du Web est mauvais ; il est éphémère et les techniques de catalogage ont été pensées pour les documents imprimés et ne peuvent donc s'appliquer au Web.

Malgré les divergences, il existe un consensus parmi les bibliothécaires et les utilisateurs d'Internet en ce qui concerne le besoin de passerelles spécialisées par domaines. En effet, même ceux qui sont sceptiques à l'égard de la nécessité de ce type de passerelles et qui sont convaincus que les technologies fourniront les solutions qui rendront ces moyens inutiles reconnaissent que, dans le futur immédiat, notamment dans les trois ou cinq années à venir, les passerelles spécialisées par sujets devraient continuer à fournir leur service (Haynes et al. 1998). En effet, une évaluation comparative des approches en général employées par les passerelles spécialisées par sujets qui donnent accès aux ressources d'information en réseaux a été réalisée par Haynes et al. (1998) à la demande de JISC. Il en résulte que des technologies telles les méthodes d'intelligence artificielle ne remplaceront pas les besoins en ressources humaines, notamment pour évaluer la qualité des sites Internet, et ce, au moins pour les sept prochaines années à venir. Humphreys (1999) pense que, malgré la performance des technologies, notamment les moteurs de recherche et d'autres agents intelligents, ces systèmes ne peuvent remplacer les indexeurs humains. Seul le professionnel de l'information est en mesure d'anticiper les besoins de ses utilisateurs. En effet, lors de l'indexation du contenu, ce dernier ne se contente pas d'extraire les termes dans le texte, mais il enrichit son indexation en ajoutant d'autres concepts et relations jugés utiles pour ses utilisateurs. Or, il est difficile de doter la machine d'un tel raisonnement.

Bien que plusieurs projets de classification et de catalogage concentrent leurs efforts pour trouver des solutions permettant de se passer d'intervention humaine, la plupart des travaux entrepris, jusqu'à récemment, démontrent que le chemin vers l'atteinte d'un tel objectif reste très lointain. De même en ce qui concerne la recherche par sujets, Koehler (1999) pense que, vu la quantité des données qui ne cessent d'augmenter, il est peu probable que la qualité de rappel soit améliorée. Au contraire, ce problème ne fait que s'aggraver à mesure que le Web continue sa croissance. Nombreux sont ceux qui pensent que pour rendre les information d'Internet

utiles, il est nécessaire de combiner les efforts humains avec les technologies de repérage de l'information (Woodward 1996).

Conclusion

Faire l'inventaire des différentes initiatives pour représenter les connaissances d'un domaine dans Internet, tout comme examiner l'organisation des ressources électroniques à travers ces initiatives, ont été le parcours de cette recherche. Dans un premier temps, il a fallu se pencher sur l'état du contenu d'Internet, ce qui nous a permis de répondre à plusieurs questions soulevées dans l'introduction. Nous avons pu constater que le contenu d'Internet est le résultat d'une contribution individuelle de différents créateurs, auteurs et concepteurs de sites Web (individus, organisations de toutes sortes, entreprises, groupes et communautés scientifiques entre autres). Tout en contribuant à alimenter et à enrichir son contenu, les sous-réseaux individuels, souvent mal conçus, n'améliorent pas la qualité de sa structure. En fait, le Web n'est que la somme de ses parties, à savoir les sous-réseaux, et ne peut donc pas avoir une structure de qualité meilleure que ses constituants. Les domaines de connaissances sont représentés de façon inégale aussi bien en ce qui concerne la qualité que la quantité. D'un coté, l'absence d'organisation et d'une bonne structuration du Web aggravée d'un manque de consensus pour la création de pages Web et de mécanismes pour contrôler la qualité du contenu — comme c'est généralement le cas dans la production de documents imprimés — et de l'autre côté, la facilité offerte pour construire un site Web, permettant à chacun d'ajouter au réseau Internet son propre sous-réseau comme bon lui semble, affectent sérieusement son efficacité et sa convivialité, particulièrement lorsqu'il s'agit d'accéder à une information pertinente.

Dans un second temps, l'examen de diverses approches utilisées pour organiser et représenter le domaine ou les contenus thématiques dans Internet nous a permis de constater que la plupart des efforts ne font que proposer des méthodes ayant été appliquées en sciences de l'information avant l'avènement d'Internet ou d'adapter des outils existants conçus initialement pour gérer d'autres types de collections dans d'autres contextes. On retient, par ailleurs, que les multiples initiatives aussi bien manuelles qu'automatiques, malgré le fait qu'elles restent fragiles et insuffisantes, connaissent un développement plus ou moins équilibré. Ces initiatives, qui révèlent une certaine redondance sinon des chevauchements, apportent, chacune à sa façon, une ou plusieurs solutions, bien que partielles, et permettent de résoudre une partie du problème. En effet, ces tentatives, particulièrement celles entreprises localement, contribuent à améliorer l'accès par sujets en développant les passerelles spécialisées par domaines. Elles donnent, par la même occasion, des exemples à suivre et font profiter d'autres communautés en attendant des alternatives plus universelles.

Un aspect de ces diverses tentatives que la recherche met en lumière est l'absence de clivage entre les efforts au niveau local (universitaires ou professionnels entre autres) ou international (organisations de normalisation). En effet, les deux milieux sont sensiblement en accord pour appliquer des méthodes traditionnellement utilisées dans les bibliothèques. On ne peut, cependant, ignorer la fragilité tant des projets mis en place que des espoirs investis dans la recherche de solutions intégralement automatiques. En effet, aucune étude jusqu'à maintenant ne s'est réellement penchée sur l'efficacité des systèmes proposés ou implantés, particulièrement sur leur application à grande échelle ni sur leur utilisation par les différents usagers. La plupart des travaux sont motivés par la disponibilité de subventions et concentrés sur la production de prototypes.

Sources consultées

American Society of Indexers. 1999. *Indexing the Web.* <http://www.ASIndexing.org/webndx.shtml>

Ardo, Anders et al. 1999. *Improving resource discovery and retrieval on the Internet: The Nordic WAIS/ Word Wide Web project — summary report.* <http://www.ub2.lu.se/W4/summary.html >

Brummer, Anna. 1997. *Subject based information gateways.* 1997. <http://www.lub.lu.se/desire/sbigs.html>

Chen et al. 1995. Automatic thesaurus generation for an electronic community system . *Journal of the American Society for Information Science*, 46 (3): 175-193.

Chen et al. 1998a. Alleviating search uncertainty through concept associations: automatic indexing, co-occurrence analysis, and parallel computing. *Journal of the American Society for Information Science*, 49 (3): 206-216.

Chen et al. 1998. Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49 (7): 582-603.

Haynes et al. 1998. *Comparative evaluation of the subject based gateways approach to providing access to network resources: a report to JISC under the eLib supporting studies programme.* London: David Haynes Ass., 53 p. <http://www.ukoln.ac.uk/services/elib/papers/tavistock/subject-gateway-access/ >

Hjorland, Birger. 1998. Information retrieval, text composition, and semantics. *Knowledge Organisation*, 25 (1-2): 16-31.

Humphreys, Nancy K.1999. Mind maps: hot new tools proposed for cyberspace librarians. *Searcher*, 7 (6). <http://www.infotoday.com/searcher/jun/humphreys.htm >

Iannella, Renato et Waugh, Andrew. 1997. *Metadata: enabling the Internet.* <http: //www.dstc.edu.au/RDU/publications/cause97/ >

Koch, Traugott. 1996. *DC subject. Thesauri and classification systems available in the WWW.* <http://www.ub.lu.se/metadata/subject-help.html >

Koch, Traugott, et al. 1998. *Specification for resource description methods. Part 3: The role of classification schemes in Internet resource description and discovery.* <http://www.ub2.lu.se/desire/radar/reports/D3.2.3/class_v10.html >

Koehler, Wallace C. 1998. The Librarianship of the Web: options and opportunities managing transitory materials. In Ching-chih Chen (ed.). *Proceedings of the 10th New Information Technology Conference*, Hanoi, Vietnam, March 1998. West Newton, MA: MicroUse Information, 1998.

_____. 1999. Classifying Web sites and Web pages: the use of metrics and URL characteristics as markers. *Journal of Librarianship and Information Science*, 31 (1): 21-29.

Lerinckx, D.1997. Dublin Core. In *Metadata Workshop*, DG XIII/E-4, Luxembourg, 1-2 December 1997. <http://www.bib.ulb.ac.be/BST/html/meta_dc.htm>

McKiernan, Gerry. 1997. Hand-made in Iowa: organizing the Web along the Lincoln highway. *D-Lib Magazine.* <http://www.dlib.org/dlib/february97/02mcKiernan.html>

McKiernan, Gerry. 1999. *Beyond bookmarks : schemes for organizing the web.* <http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm>

Musella, David et Padula, Marco. 1996. *The Authors catalogue their documents for a light Web indexing.* <http://www.isoc.org/isoc/whatis/connaissances/inet/96/proceedings/a2/a2_4.htm >

Nadis, Steve. 1996. Computation cracks semantic barrier between databases. *Science*, 272 : 1419.

NEC Research Institute. 1999. *ResearchIndex : start using ResearchIndex* (CiteSeer). <http://www.scienceindex.com/ >

Neuss, Christian et Kent, Robert E. 1995. *Conceptual analysis of resource meta-information.* <http://www.igd.fhg.de/www/www95/papers/94/www3.html>

Nicholson, Scott. 1997. Indexing and abstracting on the World Wide Web : an examination of six Web databases. *Information Technology and Libraries* : 73-81.

A Review of metadata : a survey of current resource description formats. 1999. <http://www.ukoln.ac.uk/metadata/desire/overview/>

*ROADS : Resource Organisation and Discovery in Subject-Based Services.*1999. <http://www.ariadne.ac.uk/issue3/roads/intro.html>

Robertson, David W. 1996. Subject indexing on the Web. In *World Wide Web Consortium meeting*, May 28/29. Cambridge (Massachusetts) : Distributed Indexing/Searching Workshop (DISW), 1996. <http://www.ub2.lu.se/desire/radar/ >

Schatz, Bruce R. 1997. Information retrieval in digital libraries : bringing search to the net. *Science*, 275 : 327-334.

Schatz, Bruce R., et al. 1996. *Interactive term suggestion for users of digital libraries : using subject thesauri and co-occurence lists for information retrieval.* Urbana : Digital Library Initiative, 1996. <http://dli.grainger.uiuc.edu/papers/schatzDL96.htm >

Sullivan, Danny. 1999. *Search engine watch.* Internet com. Corp., 1996-99. <http://searchenginewatch.com/>

Taylor, Mike. 1999. *Zthes : A Z39.50 profile for thesaurus navigation.* <http://lcweb.loc.gov/Z3950/agency/profiles/zthes-03.html >

Thompson, Roger ; Shafer, Keith et Vizine-Goetz, Diane. 1997. *Evaluating Dewey concepts as a knowledge base for automatic subject assignment.* Dublin (Ohio) : OCLC, 1997. <http://orc.rsch.oclc.org :6109/eval_dc.html >

Turner, Fay. Le U.S. 1996. *Government Information Locator Service (GILS) : description et situation.* Ottawa : Bibliothèque nationale du Canada, 1996. <http://gils.gc.ca/gils/backg_f.html >

Vizine-Goetz, Diane ; Mitchell, Joan S. 1996a. *Dewey 2000 : cataloguing productivity tools.* <http://www.oclc.org :80/oclc/fp/research/dwy2000/dwy2000.htm>

_____. 1996b. Online classification : implications for classifying and document [-like object] retrieval. In Rebecca Green (ed.). *Knowledge organization and change : proceedings of the 4th international ISKO conference*, Washington, D.C., 15-18 July 1996. Frankfurt/Main : INDEKS Verlag, 1996. <http://orc.rsch.oclc.org :6109/dvgisko.htm >

Ward, Dennis B. et Olson, Hope A. 1998. A Shelf browsing search system for marginalized user groups. In *Proceedings of the 61st ASIS annual meeting*, October 25-29, 1998. 35 : 342-347.

Weinberg, Bella Hass. 1996. Complexity in indexing systems — abandonment and failure : implications for organizing the Internet. In *Proceedings of the 59th ASIS annual meeting*, October 19-24, 1996. 33 : 19-24.

Williams, James G. ; Sochatz, Kenneth M. et Morse, Emile. 1995. Visualization. *Annual Review of Information Science and Technnology (ARIST)*, 30 : 161-207.

Woodward, Jeannette. 1996. Cataloging and classifying information resources on the Internet. *Annual Review of Information Science and Technology (ARIST)*, 31 : 189-220.

Zhang, Yin et Estabrook, Leigh. 1998. Accessibility to Internet-based electronic resources and its implications for electronic scholarship. In *Proceedings of the 61st ASIS annual meeting*, October 25-29, 1998. 35 : 463-473.