

Cahiers **GUT** *enberg*

☞ WWW-TED : THESAURUS ÉVOLUTIF ET
DYNAMIQUE POUR BASES DE LIENS HTML

☞ Marcia J. Bossy

Cahiers GUTenberg, n° 28-29 (1998), p. 56-71.

<http://cahiers.gutenberg.eu.org/fitem?id=CG_1998__28-29_56_0>

© Association GUTenberg, 1998, tous droits réservés.

L'accès aux articles des *Cahiers GUTenberg*

(<http://cahiers.gutenberg.eu.org/>),

implique l'accord avec les conditions générales

d'utilisation (<http://cahiers.gutenberg.eu.org/legal.html>).

Toute utilisation commerciale ou impression systématique

est constitutive d'une infraction pénale. Toute copie ou impression

de ce fichier doit contenir la présente mention de copyright.

WWW-TED : thesaurus évolutif et dynamique pour bases de liens HTML

Marcia J. BOSSY

*E.N.S.T. — École Nationale Supérieure des Télécommunications
46 rue Barrault, F-75634 Paris CEDEX 13
bossy@inf.enst.fr*

Résumé. Nous considérons la demande d'un outil de gestion de bases de liens HTML pour la recherche scientifique. Nous proposons ensuite des orientations pour l'implémentation de WWW-TED, un thesaurus évolutif et dynamique pour des collections de taille moyenne de pages HTML.

Mot-clés : document HTML, thesaurus, information scientifique et technique.

Introduction

WWW-TED est un outil de gestion de thesaurus évolutif et dynamique pour l'indexation de bases de documents HTML. Les liens WWW sont indexés — pendant la collecte — par un thesaurus lui aussi élaboré en parallèle avec le développement de la base. Le public visé par cet outil est composé de chercheurs, des groupes de recherche ou des laboratoires demandant une gestion pointue de leurs collections de documents.

WWW-TED se prête à la gestion de collections de liens hypertextes de taille moyenne (300 à 3000 liens), requérant une capacité de recherche documentaire fine. Pour un utilisateur assidu, les outils de gestion de petites collections (*bookmarks*) sont vite dépassés. A l'autre extrême, les moteurs de recherche sont mal adaptés à la gestion documentaire fine et structurée, nécessaire à la recherche scientifique. Les langages structurés tels que le thesaurus offrent aux professionnels de l'information les moyens de gestion minutieuse dans le contexte des documents papier. Notre objectif, en développant WWW-TED, est d'exploiter le potentiel d'application de cette technique traditionnelle à un serveur de documents HTML.

La littérature scientifique et technique sur Internet

Internet est devenu un outil quotidien de communication et de diffusion de l'information parmi les membres des communautés scientifiques : les supports physiques sont en place avec ses capacités de transmission grandissantes, les structures d'échange de données s'installent selon les pratiques propres à chaque domaine. On observe que de plus en plus de chercheurs se tournent vers Internet comme support de diffusion de la littérature scientifique. Les professionnels de l'information doivent re-examiner leur technologie en vue de l'adapter à ces nouvelles formes de production et de circulation de documents scientifiques.

Les communautés scientifiques reconnaissent dans Internet un outil privilégié de travail. Elles y trouvent un large éventail d'information facilement accessible et des possibilités de communication avec ses paires. Les ressources d'information peuvent prendre la forme de bases de données, de répertoires de texte intégral, des sites de codes ou autres. Parmi les ressources de communication nous pouvons classer les listes de diffusion, les conférences électroniques et les tableaux d'affichage (*newsgroups*). Les plus récents développements s'orientent vers la résolution du problème de recherche de l'information (*resource discovery*). Ces modèles de circulation de l'information ont été mis en place par les utilisateurs eux-mêmes, souvent par tâtonnements, au fur et à mesure de leur besoins. Ils définissent le réseau électronique comme le siège d'un nouveau modèle d'auto-organisation par les communautés scientifiques.

Cependant, les utilisateurs doivent apprendre à gérer ce flux ininterrompu d'informations de qualité inégale. Une cause fréquemment d'insatisfaction est l'excès d'information non-pertinente. Les utilisateurs se plaignent souvent d'être submergés par ce qu'on appelle de l'info-pollution (*info-junk*). Parmi les divers types d'utilisateurs, les communautés scientifiques sont particulièrement concernées par ce problème. Le circuit de production de la recherche scientifique se déroule de plus en plus dans les réseaux électroniques. La qualité de cette recherche dépend des outils de circulation, de gestion, de recherche et d'évaluation de la littérature et, par conséquent, du savoir scientifique véhiculé. Le problèmes de gestion de flux d'information est capital pour le futur du réseau en tant qu'outil de la recherche scientifique et sa solution passe par la création d'outils appropriés aux modes de comportement de ses utilisateurs. Il existe une demande pour des outils qui aideront les utilisateurs/producteurs à structurer et à filtrer le flux d'information de façon à organiser leur univers informationnel.

L'utilisation généralisée du réseau électronique en tant que source de documentation scientifique et technique renverse les relations entre deux demandes contradictoires : la validation de l'information et sa vitesse de diffusion. L'information formelle est validée par les pairs et traitée de façon à intégrer, par

le biais des bases de données, un large circuit de diffusion et de capitalisation. Sa production est nécessairement longue, lente et chère. A l'autre bout du spectre, nous trouvons l'information informelle. Volatile, rapide, non contrôlée elle garde néanmoins une grande valeur stratégique pour les utilisateurs. Cette information est par définition non standardisée et difficile à traiter du point de vue documentaire.

La tension entre ces deux demandes est particulièrement aiguë dans l'environnement du réseau. Les réseaux électroniques provoquent un effet de lissage entre ces deux types de documents : l'apparence physique des documents ne dépend plus seulement du producteur mais de l'environnement technique du lecteur, les protocoles d'accès sont uniformes et offrent peu de renseignements sur la qualité du document obtenu. Les modèles de circulation se trouvent ainsi bouleversés. Les pratiques de traitement de la documentation émergeant dans cet environnement peuvent potentiellement affecter la qualité de la production scientifique. Les chercheurs sont d'ailleurs très conscients de ces problèmes et nous notons que, pour le moment, Internet est surtout utilisée soit pour la communication informelle, soit comme support complémentaire de documents validés et traités ailleurs par des méthodes traditionnelles.

Le modèle de traitement documentaire en vigueur a été développé dans un contexte de documents sur support papier, leur organisation en larges collections et de leur subséquente numérisation et automatisation. La documentation de l'information scientifique et technique sert à signaler, à repérer, à rechercher, à diffuser et à mesurer la production scientifique. L'expansion d'Internet a déjà stimulé la création d'outils de visualisation (*browsers*) et de puissants moteurs de recherche pour des grands corps de données.

Des documents scientifiques de qualité circulent bien sur des serveurs WWW. Plusieurs revues scientifiques sont maintenant publiées exclusivement sous format électronique. Ce mouvement migratoire a tendance à s'intensifier. Les documents sont bien là mais tant qu'ils ne feront pas objet d'un traitement documentaire de qualité ils ne pourront pas intégrer les circuits de diffusion usuels parmi les communautés scientifiques. La migration de la littérature scientifique vers Internet pose le problème de filtrer, sélectionner et évaluer des bases de données spécialisées. Pour qu'Internet devienne la principale source de documentation scientifique et technique, il nous manque des outils adaptés aux modèles spécifiques de documents générés par ces nouvelles pratiques. La technologie traditionnelle des sciences de l'information peuvent être transposées avec profit à l'environnement du réseau dans la recherche de solutions à ces problèmes.

WWW comme support privilégié de littérature scientifique et technique

WorldWideWeb et ses interfaces graphiques offrent un environnement idéal pour le support de documents complexes. Le protocole HTTP devient universellement utilisé et les *browsers* les plus récents intègrent des interfaces vers les autres protocoles de base d'Internet (FTP, SMTP, TELNET, Z-39.50). La production de documents y est aisée et riche. Actuellement, HTML reste un standard très simplifié mais des extensions sont en étude, intégrant l'affichage de symboles spécifiques (mathématiques, électriques ...).

Il est aussi facile de collecter des documents avec un *browser* WWW. Dans le cas des bases de données traditionnelles, la recherche des documents correspondant aux références obtenues peut se révéler une tâche longue et frustrante. Avec WorldWideWeb, la distance entre le document final et sa référence « bibliographique » (le lien hypertexte) se mesure par un simple « *click* » de souris. En effet, la situation inverse est la cause de désorientation des utilisateurs : ils se plaignent souvent d'un surplus de liens collectés de façon désordonnée.

Comment organiser la collection de liens sélectionnés ? Comment accéder au bon document ? Et comment peut-on être sûr que le document pertinent ne reste pas « caché », noyé parmi une masse d'information inutile ? Ce sont là des problèmes classiques de la gestion documentaire qui sont rendus plus difficiles à résoudre quand nos documents sont des liens hypertextuels. Dans le contexte du support papier, nous sommes soutenus par notre familiarité avec l'environnement traditionnel et par nos outils et méthodes éprouvés. Dans le cas de collections de liens nous nous trouvons dans un contexte inconnu ; nous procédons par essais et erreurs, en nous efforçant de construire des outils adaptés à nos besoins nouveaux. Notre objectif est d'apporter, avec WWW-TED, une réponse à cette demande.

La taille des collections et l'accès à l'information

En développant WWW-TED nous avons tenu compte de deux points principaux : la taille de la collection et l'indexation des documents en vue d'une recherche documentaire de qualité. Les efforts actuels de gestion de collections de liens concernent deux types de collections. Dans les deux cas, l'effort de classification et/ou d'indexation est inadéquat à la recherche scientifique.

Les « petites collections » personnelles

Il s'agit souvent d'une collection de taille allant jusqu'à 300 liens. Nous disposons de *bookmarks* intégrés aux *browsers* ou de logiciels *bookmarks* indépendants, parfois avec la possibilité d'établir des arborescences. Nous voyons aussi des pages HTML personnelles plus ou moins bien organisées. Dans ces cas, les défauts d'organisation ne constituent pas un handicap grave car l'espace informationnel est restreint. Nous ne courons pas le risque de nous perdre ou de laisser échapper une information importante.

Ces collections sont par ailleurs très pointues et répondent à un besoin individuel. Elles sont, selon la tradition d'Internet, mises à la disposition de la communauté « en l'état » mais les auteurs n'ont pas d'obligation vers un public particulier.

Les moteurs de recherche

A l'autre extrême, nous trouvons les grands moteurs de recherche. Ces grandes machines ont pour objectif de traiter la totalité de l'espace informationnel d'Internet. La masse de documents à traiter impose des choix de gestion qui les rendent peu fiables pour une recherche affinée. Les utilisateurs visés par ces produits sont le grand public, composé forcément par des communautés hétérogènes avec des besoins variés et parfois incompatibles entre eux. Dans cette perspective, les gestionnaires n'opèrent pas de sélection. Les documents récupérés à chaque requête sont donc de qualité très inégale et aucune indication ne permet de contrôler la valeur de l'information obtenue.

Par ailleurs, les gestionnaires de ces moteurs ne pratiquent pas d'indexation fine. Nous ne disposons pas d'informations sur les critères de sélection, les périodicités de mise à jour, les algorithmes d'indexation et de recherche ni les critères de pondération. Ce sont des critères auxquels nous nous référons, dans les bases de données informatisées traditionnelles, pour évaluer la qualité de la recherche effectuée. Le choix a été fait de permettre beaucoup de bruit dans les réponses en laissant à l'utilisateur le soin d'effectuer une sélection ultérieure. Cette procédure mène à une perte de temps considérable pour l'utilisateur. Il n'est pas rare de se retrouver avec des centaines de milliers des liens, ordonnés selon des critères obscurs, en réponse à une requête.

Or, nous nous trouvons souvent confrontés au besoin de traiter des corpus de quelques milliers de documents. Un fonds documentaire scientifique comporte typiquement entre 1000 et 5000 documents. Il s'agit là de la taille usuelle d'un petit centre documentaire pour un projet, pour un laboratoire ou pour une bibliothèque universitaire. Assurément, il est possible d'argumenter que la do-

cumentation de tel ou tel grand projet est beaucoup plus important. Dans la pratique cette documentation n'est par traitée en masse mais sous forme de plusieurs centres attachés à des groupes de recherche.

Par ailleurs, la gestion d'un corpus destiné à la recherche scientifique et technique pose des exigences précises. Par opposition aux critiques exprimées ci-dessus à l'encontre des moteurs de recherche, un fonds documentaire scientifique et technique doit avoir des critères explicites de sélection des documents, de périodicité de mise à jour d'indexation et de couverture du domaine. Ces critères peuvent varier selon le champs de recherche et ils sont sujets à critique. Mais la pratique même de la recherche scientifique exige que ce regard critique puisse être porté en connaissance de cause.

Ceci explique les réserves exprimées par les chercheurs envers les moteurs de recherche présents dans WorldWideWeb en tant qu'outil de recherche documentaire. Les critiques le plus couramment exprimées sont la perte de temps et la qualité inégale des résultats. Nous venons d'analyser les raisons de ces critiques bien fondées.

WWW-TED est un outil d'aide à la gestion de serveurs de liens de taille moyenne (300 à 3000 liens) orientés vers la recherche scientifique et technique. De tels serveurs doivent disposer de moyens de recherche documentaire affinés. Les résultats doivent être précis, évitant les problèmes de bruit et de silence, rencontrés par les utilisateurs des moteurs de recherche. Il n'est pas possible d'atteindre ces objectifs sans l'aide d'une structure d'indexation fine telle que le thesaurus.

Le thesaurus et la technologie hypertexte

Langage de description et qualité de la recherche documentaire

Le document et son image

Tout traitement d'un ensemble de documents commence par une opération de description qui crée une image appelée « substitut du document ». Il peut s'agir d'une notice bibliographique, d'une fiche ou d'un enregistrement informatique ... Dans tous ces cas, les substituts de documents sont naturellement plus petits, plus maniables, plus faciles à combiner entre eux que les documents eux-mêmes. L'ensemble de substituts de documents est donc une représentation de la collection, un catalogue.

Cette représentation est utilisée pour créer des outils de recherche de l'information. La recherche peut s'effectuer sur des caractéristiques physiques du document telles que le titre, le nom des auteurs, la langue, etc. Pour affiner la recherche, des structures de description du contenu est superposée à cette première description simple.

Les possibilités de recherche sont d'autant plus affinées que la structure de description est sophistiquée. Cependant, le prix de ce développement est la distance accrue entre l'ensemble de documents représenté et son image. La structure de description enrichie entraîne la dissociation entre le document physique et sa référence.

La classification et les livres

La structure de description la plus classique est la classification. Il s'agit d'une liste de termes décrivant un domaine et dotée de relations hiérarchiques. Un document ne peut être indexé que par un seul noeud de l'arbre ainsi formé ; chaque document n'a qu'une seule clé d'indexation. Cette structure a été inspirée par l'organisation physique des livres dans des salles, des étagères, des rayonnages et ainsi de suite. Dans ce contexte, l'image des fonds documentaires dans un fichier correspondait très exactement à leur localisation dans des salles de consultation adjacentes. Encore aujourd'hui, il est possible de « consulter le fichier avec ses pieds » dans des bibliothèques en libre accès comme la BPI.

Deux développements rendent peu à peu cette organisation inadaptée à la recherche d'information scientifique et technique :

1. L'importance prise par les revues scientifiques en tant que source de documentation. L'unité d'information intéressante pour l'utilisateur est l'article et non le numéro ou la collection. Le format de description ne correspond plus à la réalité physique des documents.
2. La multiplication des représentations, visant à contourner la contrainte de la clé unique de description. Les fiches intercalaires de renvoi se multiplient pour palier à la contrainte de la clé unique ou pour approcher des concepts complémentaires.

L'instrument de recherche devient plus performant mais sa gestion est complexe. L'image s'éloigne de l'objet qu'elle est censée représenter. Dans ce contexte, la numérisation des références et la création des grandes bases de données informatisées ont été perçues comme une solution à un problème qui risquait de submerger les gestionnaires de documents.

Le thesaurus et les bases de données

L'informatisation des références a permis l'utilisation optimale de nouvelles structures de description du contenu telles que le thesaurus. Outre les relations hiérarchiques, le thesaurus peut intégrer les renvois et les synonymes. La contrainte de la clé unique d'indexation n'a plus lieu d'être ce qui enrichit la description. La recherche est affinée par l'usage des opérateurs booléens.

Toute solution amène avec elle des nouveaux problèmes. Tout chercheur a déjà éprouvé la frustration de consulter une base de données et de se retrouver avec une liste de références de documents très pertinents qui existent bien quelque part, mais pas à proximité du lieu de consultation. En dépit de ses avantages, le défaut de ce système est clair : le lien de proximité entre la structure de représentation et la collection physique est rompu. Le prix de la performance de l'outil de recherche est la difficulté d'accès au document — ce qui est en fin de comptes, l'objectif final de chaque recherche.

Le choix semblait posé entre des outils de recherche rigides et appauvris, mais avec la certitude de retrouver le document correspondant et des outils riches et performants mais éloignés de la collection physique.

Un thesaurus évolutif pour des liens hypertextuels

WWW et la technologie des liens hypertextuels peuvent nous aider à sortir de cet impasse en distribuant la tâche du stockage des documents. Le lien hypertexte réduit la distance entre la référence et le document. Il est possible d'avoir une structure de description riche et accès facile au document dans le même outil. Comme dans le traitement informatique classique, il est possible de multiplier les clés d'indexation car la collection traitée, composée de références hypertextuelles, ne surcharge pas la mémoire. Il nous a donc paru intéressant de transposer la structure de description de thesaurus vers un outil de gestion de liens WWW pour littérature scientifique.

Avec WWW-TED le thesaurus évolue au même temps que la collection. Les utilisateurs peuvent attacher des descripteurs aux liens regroupés dans des pages thématiques ; ils peuvent aussi créer des relations hiérarchiques, de synonymie ou de proximité entre les liens. Le thesaurus est mis à jour à chaque re-indexation de la base. Un outil de visualisation des listes hiérarchiques y est incorporé.

L'élaboration d'un thesaurus est une tâche spécialisée demandant le concours de plusieurs professionnels hautement qualifiés pendant une période de temps conséquent. Leur emploi demande aussi des professionnels qualifiés. Il s'agit donc d'une méthode chère de gestion et d'aide à la recherche. Actuellement, la

pratique courante est de dissocier les étapes de constitution de la collection et celle de la fabrication du langage de description. Il nous semble que ce modèle ne s'accorde pas à notre pratique courante. En réalité, nous ne collectionnons pas des documents « en vrac ». Nous sélectionnons, autour d'un thème qui nous intéresse à un moment donnée, un petit noyau de documents parmi une offre souvent pléthorique.

Cette pratique est particulièrement facile à vérifier dans le cadre de la recherche de documents pertinents dans des serveurs WWW. Typiquement, des liens sont organisés en petit nombre par pages dont les titres représentent les thèmes traités. Il existe déjà des outils simples de recherche dans une collection de pages HTML. La question que nous nous sommes posés est : est il possible de tirer parti de cette pratique en l'enrichissant afin de créer un outil d'indexation fine de liens sélectionnés ?

WWW-TED étend et affine cette pratique courante. Les liens, rassemblés dans des pages HTML, sont en effet indexés par l'utilisateur sous le titre de la page. L'ensemble de ces pages peut être considéré comme un fichier inversé ; les titres comme les termes du thesaurus et les liens comme les documents indexés sous le descripteur représenté par le titre correspondant. L'ensemble des titres constitue la liste initiale des termes du thesaurus évolutif élaboré par WWW-TED. Le serveur de liens WWW est indexé, pendant la collecte, par un thesaurus lui aussi élaboré en parallèle avec le développement de la base.

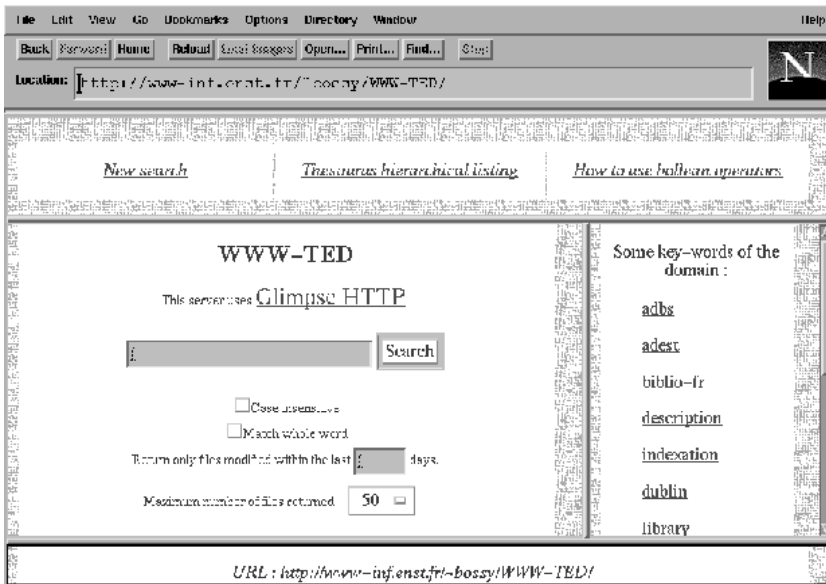
WWW-TED et son implémentation

Dans le cadre de ce travail, WWW-TED a été implémenté sur un système UNIX Solaris v.5.5 et le serveur WWW Apache v.1.2.b6. Les opérations de recherche documentaire utilisent httpGlimpse v.4.0b1 — U of Arizona. Ce dernier est un moteur de recherche pour des pages HTML et comme tel, bien adapté à nos besoins. Des scripts supplémentaires ont été écrits en Perl v.5.003 et Java v.1.1.4. Cependant, il faut noter que les concepts sous-jacents à ce travail sont indépendants d'outils ou plates-formes spécifiques.

WWW-TED est composé de trois couches :

- le **thesaurus** : son implémentation comprend les fichiers de la base de données du thesaurus, des scripts cgi, et divers compteurs (liens par page, mots-clés utilisés, liens accédés). Le thesaurus opère sur la base de données ;
- la **base de données** est composée d'un ensemble de fichiers HTML, contenant chacun un titre et plusieurs liens. La base de données est accédée à travers l'interface utilisateur ;

- l' **interface utilisateur** comprend une page de consultation et plusieurs pages de gestion du serveur. Les actions des utilisateurs déclenchent les scripts de gestion du thesaurus.



Le thesaurus

Un thesaurus est un outil d'indexation, constitué d'un vocabulaire contrôlé enrichi de relations sémantiques entre ses termes. Les composantes d'un thesaurus sont : les descripteurs, les non-descripteurs, les candidats-descripteurs, les notes d'application et les relations sémantiques.

Descripteurs

Les termes d'un thesaurus sont appelés des descripteurs. Il ne faut pas confondre ce concept avec celui de mot-clé, qui est un mot (chaîne de caractères entouré de deux blancs) extrait du texte à indexer. Un descripteur peut être composé (ex. : « RFC 1356 ») et il n'apparaît pas forcément dans le texte (ex. : « indexation automatique » pourrait être un descripteur approprié pour le texte que vous avez en mains bien que ces termes n'y figurent pas).

Dans notre base, les descripteurs sont implémentés comme des titres des pages HTML qui la composent ; chaque page représente donc un descripteur. Les liens

figurant sur chaque page représentent les documents indexés par ces descripteurs. Dans une base traditionnelle, un document peut être indexé par plusieurs descripteurs. Tel est aussi le cas de notre serveur WWW. Un lien peut figurer dans plusieurs pages si cela s'avère utile. La consommation en mémoire est négligeable et cette situation ne pose pas de problème ni pour la recherche de l'information ni pour la gestion du serveur. La liste des descripteurs — les titres des pages — est stockée dans la base de données du thesaurus, suivie des leurs relations, notes d'applications et statistiques d'utilisation respectives.

Non-descripteurs

Un non-descripteur est un terme considéré comme équivalent d'un descripteur (voir détail ci-dessous). Bien que ne faisant pas partie du thesaurus *stricto sensu*, il figure dans la présentation par liste en tant qu'aide à l'indexation et à la recherche.

Les non-descripteurs sont implémentés sous forme d'étiquettes < META > placées dans les pages indexées par leurs termes synonymes. L'inclusion des étiquettes garantit l'extraction de ces pages lors d'une requête employant le non-descripteur. Ces non-descripteurs doivent aussi faire partie de la base de termes du thesaurus.

Candidats-descripteurs

Ce sont des termes employés par des utilisateurs lors d'une requête et qui ne font pas partie du thesaurus. Ces termes doivent être conservés en tant que candidats-descripteurs. Périodiquement, ils seront examinés et une décision doit être prise sur leur intégration dans le thesaurus. Ils ne figurent pas dans la présentation par liste mais ils doivent intégrer la base de termes du thesaurus.

Notes d'application

Certains termes peuvent être suivis d'observations sommaires éclaircissant le contexte de leur utilisation. Ex. :

FTP [sites]

File Transfer Protocol [protocoles]

Les notes d'application figurent explicitement à la suite des descripteurs dans les pages HTML ; ils sont donc un attribut du descripteur dans la base de termes.

Relations sémantiques• *Relation de synonymie*

C'est une relation entre un descripteur et un non-descripteur. Pour éviter les problèmes de silence lors d'une recherche, un seul terme est retenu comme descripteur parmi plusieurs termes trop proches. Les termes écartés sont des non-descripteurs. Il faut noter que ce ne sont pas là nécessairement des termes synonymes au sens grammatical du terme. Ex. :

forum électronique descripteur

liste de diffusion non-descripteurs synonymes de
listproc **forum électronique**
listserv

Les non-descripteurs ne font pas strictement partie du thesaurus mais ils doivent figurer dans la liste de présentation de façon à orienter les indexeurs et les utilisateurs.

• *Relation de voisinage*

C'est une relation entre deux descripteurs. Il s'agit d'une aide à la recherche, signalant des concepts proches pouvant être employés avec profit par les utilisateurs dans leurs requêtes. Ex. :

forum électronique descripteur
newsgroup terme voisin de **forum électronique**

Cette relation n'est pas nécessairement symétrique. Poursuivant l'exemple ci-dessus, nous pouvons imaginer :

forum électronique terme voisin **newsgroup**
 et aussi
newsgroup terme voisin **forum électronique**

Mais :

communication entre chercheurs terme voisin **forum électronique**
 mais non
~~**forum électronique** terme voisin **communication entre chercheurs**~~

La relation de voisinage est implémentée par un attribut des descripteurs dans la base des termes. L'accès aux termes voisins est implémenté par des boutons en bas des pages des descripteurs concernés.

- *Relation hiérarchique*

C'est une relation entre deux descripteurs ; l'un d'eux est le terme général (TG) et l'autre le terme spécifique (TS). Un descripteur peut être terme général de plusieurs termes spécifiques mais chaque descripteur ne possède au plus qu'un terme général.

L'ensemble de ces relations hiérarchiques définissent un ou plusieurs arbres enracinés à l'intérieur du thesaurus. La racine est le terme général le plus haut (celui qui n'est TS d'aucun descripteur). Chaque arbre est appelé un champ ou une facette et est nommé d'après le descripteur racine.

Les relations hiérarchiques sont implémentées par des attributs des descripteurs dans la base des termes. L'accès aux termes généraux ou spécifiques est implémenté par des boutons en bas des pages des descripteurs concernés.

La base de données

Le serveur est constituée de pages HTML contenant chacune un titre, une liste de liens et des boutons d'aide à la recherche. Le titre est le thème général aux liens et il représente le descripteur du thesaurus qui indexe ces liens. Ceux-ci sont en nombre limité (3 à 15). Dès que ce nombre est dépassé, WWW-TED affiche un message qui propose (mais ne force pas) une découpe de la page en plusieurs pages plus spécifiques. Ainsi, les pages restent courtes, facilement visualisées dans un seul écran. Il n'y a pas de surcharge d'information. Les boutons d'aide à la recherche renvoient au terme général, aux termes spécifiques et aux termes voisins du descripteur (titre).

L'interface utilisateur

L'interface utilisateur comprend une page de consultation et plusieurs pages de gestion du serveur. Les pages de visualisation — des résultats, des listes hiérarchiques du thesaurus — sont construites par des scripts cgi.

La page de consultation

La page de consultation permet à l'utilisateur de faire des requêtes par sujet en utilisant des opérateurs booléens. Elle affiche aussi une fenêtre listant des

mots-clés de la base, mais les utilisateurs ne sont pas limités à celles-ci. La liste de mots-clés permet à l'utilisateur éventuel de se faire une idée du domaine traité par le serveur. Les mots-clés affichés sont des liens actifs, simulant des requêtes simples. La page de consultation contient aussi des boutons d'aide vers des pages d'explications sur les opérateurs booléens et vers la description des ressources disponibles dans le serveur.

Les pages de gestion

Il s'agit d'une hiérarchie de pages HTML permettant à l'utilisateur de créer (modifier, supprimer) des pages thématiques, des liens et des relations entre les pages. Il est possible de créer une nouvelle page à partir d'un fichier texte (un *bookmark*, par exemple) ou en spécifiant un à un les liens et le titre.

Afin de conserver l'intégrité de la base, WWW-TED teste la compatibilité des ces actions avec les structures déjà existantes avant de les intégrer. Par exemple, il n'est pas possible de créer une page avec un titre déjà existant. Il n'est pas non plus possible d'éliminer un titre représentant un terme générique à un terme spécifique existant dans la base de gestion du thesaurus. WWW-TED vérifie que les relations hiérarchiques créées respectent les contraintes de l'arborescence (formation de cycles) et contrôle la compatibilité entre les relations existantes et celles en cours de création.

L'utilisateur peut alors forcer la re-indexation. S'il ne le fait pas, la base sera re-indexée automatiquement à intervalles réguliers, mettant à jour les pages de la base et le thesaurus simultanément.

Conclusion

Un nombre croissant de chercheurs se tourne vers Internet comme support de diffusion de la littérature scientifique, créant une demande pour des outils de description, d'indexation et de recherche documentaire adaptés aux besoins spécifiques de la recherche scientifique. WWW-TED adresse les problèmes de la taille des collections et de la recherche documentaire de qualité dans des collections de pages HTML. Par ailleurs, WWW-TED permet aux utilisateurs d'attacher des mots-clés à leurs liens — organisés en pages HTML thématiques — et de créer des relations structurées entre ces pages. Le thesaurus alors, évolue en même temps que la collection. Les développements futurs envisagés sont la création de structures spécifiques à un champ de recherche et le développement d'un modèle multi-utilisateur pour construction collaborative du thesaurus.

Références bibliographiques

- J. Aitchison, A. Gilchrist. *Thesaurus Construction : a Pratical Manual*. 2nd ed., London : Aslib, 1987.
- T. J. Berners-Lee, R. Caillau, N. Pellow, J.-F. Groff, J.-F., B. Pollermann. World Wide Web : the Information Universe. *Electronic Networking : Research, Application and Policy*. 2(1) Spring, pp. 52–58, Westport, USA : Meckler Publishings, 1992.
- M. K. Buckland, M. H. Butler, Y. Kim, B. A. Norgard, C. Plaunt. Partnerships in Navigation : An Information Retrieval Research Agenda, dans *Forging New Partnerships in Information*, Proceedings of the 58th Annual ASIS Meeting, 1995, pp. 84–89, Medford, NJ : Information Today.
- M. K. Buckland, C. Plaunt. On the Construction of Selection Systems. *Library HiTech*, 12(4), 1994, pp. 5–28.
- P. Caplan. Cataloging Internet Resources. *The Public-Access Computer Systems Review* 4(2), 1993, pp. 61–66.
- M. Carroll, W. Scott Downs. *Cyberstrategies : How to Build an Internet-based Information System*. New York : Van Nostrand Reinhold, 1995.
- G. Chartron. IST et réseaux électroniques : les enjeux. *Séminaire Ecrit informatisé et systèmes d'information*, Paris, 4 janvier 1994, URFIST — Unité Régionale de Formation à l'Information Scientifique et Technique.
- J. Chaumier. *Les langages documentaires*. Paris : Entreprise moderne d'édition, 1978.
- J.-P. Courtial. *Introduction à la scientométrie — De la bibliométrie à la veille scientifique*. Paris : Economica, 1990. (Sociologies).
- C. Guinchat, M. Menou. *Sciences et techniques de l'information et de la documentation*. Paris : Unesco, 1990.
- S. Harnad. Implementing Peer Review on the Net : Scientific Quality Control in Scholarly Electronic Journals. *Proceedings of the International Conference on Refereed Electronic Journals*, University of Manitoba, Winnipeg, 1–2 October 1993.
- C. Ollendorf, D. Frochot. L'évolution des méthodes de travail documentaire avec Internet. *Documentaliste*, 32(6) nov.-déc. 1995, pp. 313–318.
- C. Panijel. Repères historiques pour le développement des réseaux électroniques. *Les Bibliothèques et les réseaux électroniques de la recherche* 11 février 1993, URFIST — Unité Régionale de Formation à l'Information Scientifique et Technique de Paris.

J. Price-Wilkin. Using the World-Wide Web to Deliver Complex Electronic Documents : Implications for Libraries. *The Public-Access Computer Systems Review*, 5(3), 1992, pp. 5-21.

B. Stiegler. Machines à lire. *Autrement — La Bibliothèque*, Figuiet Richard (dir.), avril 1991, pp. 143-161 (série Mutations).

B. Vickery, A. Vickery. *Information Science in theory and practice*. London, Butterworths, 1987.

Références WWW

Comment citer un document électronique ?

<http://www.bibl.ulaval.ca:80/doelec/citedoce.html>

Glimpse Working Group — University of Arizona, CS Dept.

<http://glimpse.cs.arizona.edu:1994/index.html>

Innovative Internet Applications in Libraries

<http://frank.mtsu.edu/~kmiddlet/libweb/innovate.html>

World Wide Web searching tools

<http://www.bubl.bath.ac.uk/BUBL/Iwinship.html>

Using Library Classification Schemes for Internet Resources

<http://www.oclc.org/oclc/man/colloq/v-g.htm>

Evaluating World Wide Web Information

<http://thorplus.lib.purdue.edu/libraryinfo/instruction/gs175/3gs175/evaluation.html>