

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE

La recherche d'information sur les réseaux

**Cours INRIA, 30 septembre - 4 octobre 2002,
Le Bono (Morbihan)**

*Ouvrage coordonné
par Jean-Claude Le Moal, Bernard Hidoine
et Lisette Calderan*

ADBS Éditions

Ce cours INRIA a été organisé avec le patronage de l'Association des professionnels de l'information et de la documentation (ADBS) et celui du ministère de la Jeunesse, de l'Éducation nationale et de la Recherche, direction de l'enseignement supérieur et du ministère de la Recherche, direction de la technologie.

Les ouvrages publiés à l'occasion des précédents cours Inria sont épuisés, à l'exception de : Bibliothèques numériques, cours INRIA, 9-13 octobre 2000, La Bresse, ouvrage coordonné par Jean-Claude Le Moal et Bernard Hidoine. 2000. 246 p. ISBN 2-84365-044-5

N. B. Dans cet ouvrage, les notes de bas de page, ainsi que les références bibliographiques indiquées par des chiffres entre crochets, figurent à la fin de chaque chapitre.

Révision du manuscrit et mise en page : Isabelle Kersimon

Sommaire

Avant-propos	5
<i>par Jean-Claude Le Moal,</i> INRIA, Unité de communication et information scientifique	
1 - Instruments de recherche sur le Web	11
<i>par Sylvie Dalbin,</i> ATD - DECYBEL	
2 - XML et la documentation structurée : des principes aux techniques	71
<i>par François Role,</i> Ministère de la Recherche	
3 - Les métadonnées : accès aux ressources électroniques	99
<i>par Marie-Élise Fréon,</i> Jouve	
4 - Traitement automatique des langues et recherche d'information	137
<i>par Pascale Sébillot,</i> IRISA	
5 - Des bibliothèques traditionnelles aux « bibliothèques virtuelles »	169
<i>par Dominique Lahary,</i> BDP Val d'Oise	
6 - De la sémantique des contenus à la sémantique des structures	203
<i>par Laurent Romary,</i> INRIA/Loria	

7 - Recherche interactive dans les documents multimédias	231
<i>par Nozha Boujemaa, INRIA Rocquencourt</i>	
8 - Veille stratégique sur les réseaux	257
<i>par Armelle Thomas, Inforizon</i>	
<i>Répertoire des sigles utilisés</i>	<i>301</i>
<i>Table des matières</i>	<i>309</i>
<i>Adresses des auteurs</i>	<i>321</i>

Avant-propos

par Jean-Claude LE MOAL

Depuis 1982, l'INRIA organise, tous les deux ans, un cours sur l'IST et l'informatique, en variant, session après session, les thèmes traités. Pour la première fois cette année, un titre déjà utilisé en 1996 a été repris : « La recherche d'information sur les réseaux ¹ ». Il y a six ans, l'information disponible sur Internet, certes déjà importante, était encore peu considérée ; l'objectif principal était alors de sensibiliser les professionnels de l'information à ces nouvelles potentialités.

Aujourd'hui, tout le monde est conscient de la quantité d'informations disponibles sur le Web. Il est nécessaire, par contre, de savoir comment fonctionnent les systèmes de recherche, comment organiser les documents de son centre de ressources numérique ou virtuel afin qu'ils soient aisément accessibles en intranet ou sur Internet.

Les textes que vous trouverez dans cet ouvrage reprennent les interventions de ce cours qui s'est déroulé au Bono (Morbihan), du 30 septembre au 4 octobre 2002.

INTERNET ET CONFIANCE

En ayant la possibilité de disposer de l'information depuis leur lieu de travail ou leur domicile, les utilisateurs, au fil des années, ont rapidement privilégié le document numérique aux dépens des sources imprimées. Pourtant, des réticences existent encore. Les sources sur Internet sont certes nombreuses, mais la quantité disponible est si importante

1. La recherche d'information sur les réseaux. Internet : pour en savoir plus : cours INRIA, 30 septembre - 4 octobre 1996, Trégastel. Paris, ADBS Éditions, 1996 (épuisé).

que l'on peut craindre de ne pas savoir comment aborder sa recherche ; lorsque des documents pertinents ont été trouvés, il est souvent possible de s'interroger sur leur fiabilité ou leur crédibilité. Et, comble de l'inquiétude, n'est-on pas passé à côté du document qui fait référence sur le sujet ? Au final, l'internaute prend souvent conscience d'une perte de temps importante.

Dans la recherche d'information comme dans diverses activités – de plus en plus présentes en nombre – sur Internet, on constate un manque de confiance dû, au moins partiellement, à l'absence de repères reconnus comme « dans le monde réel ». Dans un premier temps de développement du Web, certains ont cru à l'utopie de la désintermédiation mais, pour retrouver la confiance, aujourd'hui, la médiation apparaît plus que jamais nécessaire.

Celle-ci doit être menée par les spécialistes de l'activité, et non par les spécialistes de l'outil informatique (une étude de notaire dressera des actes authentiques dématérialisés, et la Poste sera tiers de confiance pour les envois recommandés électroniques). Les outils changent mais les métiers restent ; il ne faut pas confondre technique et métier.

Dans la recherche d'information, ce qui manque ce sont des repères, des classements, des catalogues, des bibliothèques : une organisation connue pour les documents imprimés. Les services d'édition et les bibliothèques-centres de documentation doivent assurer, aujourd'hui comme hier, leur rôle de médiation, rôle d'interfaces entre les documents et leurs lecteurs. Il y a toujours des métiers, même si les frontières sont moins rigides.

Pour exercer leur métier, il est indispensable que les professionnels sachent acquérir les connaissances leur permettant de s'emparer des nouveaux outils disponibles. Une fausse route serait qu'ils soient tentés d'exercer d'autres fonctions voisines, au lieu de faire preuve d'imagination dans l'accomplissement de leurs missions en profitant astucieusement des opportunités nouvelles offertes.

CENTRES DE RESSOURCES VIRTUELS

L'utilisateur final appréciera la fréquentation d'une bibliothèque virtuelle de qualité, lui donnant un accès aisé à des sources sélectionnées, analysées, classées, de la même manière qu'il appréciait hier de venir consulter dans une bibliothèque traditionnelle.

Dans la mise en place d'une bibliothèque ou d'un centre de documentation virtuel, il faut être attentif à quelques points. Une première difficulté est de savoir donner des limites au domaine pris en compte par son centre de ressources, en privilégiant évidemment les utilisateurs que l'on a mission de servir. Le risque est grand, sinon, d'être submergé par le traitement de la masse de documents disponibles. Une bonne maîtrise de la recherche d'information facilitera une collecte efficace. Une autre difficulté concerne la mise en place d'un accès aisé aux sources très hétérogènes (pages, ouvrages, sites, documents primaires et secondaires) qui constituent la collection. Celle-ci n'est plus physique, elle peut n'être qu'une simple base de liens mais dont la valeur ajoutée est constituée par la sélection, l'intégration, le classement, le catalogue de notices descriptives, l'index, etc.

Le fait de parler de métadonnées ou de portails ne modifie en rien les fonctionnalités attendues d'une bibliothèque.

En présentant des sources de valeur accompagnées de signalements descriptifs et de métadonnées, en créant des ontologies, des taxonomies qui structurent le vocabulaire de l'organisme, les professionnels sensibilisent leurs utilisateurs à une information externe de qualité. Ces derniers découvrent rapidement que, dans un tel cadre, leur recherche est beaucoup moins aléatoire qu'en utilisant les moteurs généralistes. L'intégration de toutes les sources relatives à un même thème, qu'elles soient ou non structurées, est évidemment une forte valeur ajoutée.

Enfin, si l'objectif est bien de réaliser un centre de ressources virtuel où l'utilisateur pourra aisément trouver seul l'information recherchée, il faut veiller à ne pas oublier de mettre en évidence, sur le site, la possibilité d'un contact humain par messagerie, *chat* ou téléphone. À l'instar du visiteur qui, auparavant, n'ayant pas trouvé le document souhaité, s'adressait au bureau d'accueil ou au bureau de référence, il faut aujourd'hui permettre à l'internaute de s'adresser à un spécialiste pour lui soumettre sa demande. Ceci suppose bien évidemment, pour être efficace, une organisation et une grande réactivité des personnes assurant ce rôle. Dans de nombreuses bibliothèques étrangères, ce service est déjà mis en place en réseau, durant la plus grande plage horaire possible. Les moteurs de recherche sont disponibles 24 heures sur 24 ; la valeur ajoutée du « moteur humain » ne peut se satisfaire d'un fonctionnement de quelques heures par jour.

MISE À JOUR DES CONNAISSANCES

Suivant « les bonnes pratiques » de ce cours, cet ouvrage désire apporter aux lecteurs un point de vue sur la situation existante, et des perspectives sur les évolutions à venir. C'est pourquoi les contributeurs sont d'origines diverses : bibliothécaire, consultant en IST ou informaticien, professionnel de l'industrie de l'information, chercheur ou enseignant. Son objectif est de fournir au lecteur l'occasion de mettre à jour ses connaissances sur la recherche d'information ; connaissances utiles dans la recherche au service d'utilisateurs ou pour alimenter sa propre collection de documents ; connaissances utiles également pour organiser l'accessibilité aux ressources d'une bibliothèque ou aux éléments d'un système d'information.

Un panorama des sources et des outils sur le Web

Dans le premier chapitre, Sylvie Dalbin offre un panorama des sources d'information et des instruments de recherche sur le Web. Elle aborde successivement les moteurs, annuaires, outils collaboratifs, fonctionnalités et technologies, avant de traiter de la manière de les évaluer.

Armelle Thomas complète ce paysage par la veille stratégique au chapitre 8. C'est le besoin d'un accès rapide à l'information, notamment dans les entreprises, qui a conduit à rationaliser les activités naturelles de veille et à bénéficier de la modernisation des processus.

Parmi les sources disponibles sur le Web, il n'est pas question de passer sous silence les bibliothèques auxquelles est consacrée la totalité du chapitre 5. Dominique Lahary n'y dresse pas seulement une typologie des bibliothèques virtuelles, mais montre comment celles-ci continuent de fournir les services assurés par les bibliothèques traditionnelles. Il évoque notamment les perspectives apportées par XML dans la structuration des données.

Le Web sémantique

En effet, pour que les machines puissent réellement traiter l'information et donc la rechercher, il faut d'une part que la représentation de celle-ci soit rigoureuse, structurée et, d'autre part, que des métadonnées, description formelle du document, l'accompagnent².

2. Vincent Quint. Documents structurés sur le Web. In : L'information numérique : actes du 19^e congrès IDT/Net, Paris, 4-6 juin 2002. Paris, Ditinfo, 2002. - P. 83-87.

François Role présente XML et la documentation structurée au chapitre 2, en soulignant le rôle normalisateur de ce langage utilisé dans la représentation des structures, en décrivant les principales techniques de vérification et de manipulation, et en terminant par quelques exemples d'applications.

Les métadonnées font l'objet du chapitre 3. Marie-Élise Fréon en donne une typologie illustrée d'exemples détaillés. Ces données sur les données décrivent le contenu d'une ressource mais peuvent également être le support de bien d'autres informations utiles pour la gestion des documents.

Enfin Laurent Romary, dans le chapitre 6, « De la sémantique des contenus à la sémantique des structures », montre la nécessité de la normalisation pour permettre des échanges d'information efficaces. Pour parvenir à un véritable Web sémantique, il faut déjà, par exemple, s'appuyer sur une base conceptuelle unifiée pour décrire la variété de métadonnées susceptibles d'être associées à des bases documentaires.

Information textuelle ou multimédia

Si la mise en place de métadonnées nécessite généralement le concours ou au moins le contrôle d'humains, il est une autre voie explorée depuis de nombreuses années pour les recherches sur des textes : la recherche en texte intégral. Le traitement automatique des langues laisse envisager des perspectives intéressantes pour accéder aux ressources d'Internet ou d'un Intranet. Pascale Sébillot nous détaille, au chapitre 4, ces techniques qui ne sont encore que partiellement utilisées sur le Web. L'objectif général reste de parvenir à une réelle amélioration de la qualité des systèmes de recherche d'information, c'est-à-dire d'en faire croître parallèlement les taux de rappel et de précision.

L'information disponible est de moins en moins exclusivement textuelle et il est nécessaire de prendre également conscience de la complexité dans la recherche des documents multimédias. Nous ne pouvons traiter, dans les limites de ce cours, de toutes les spécificités. Dans le chapitre 7, Nozha Boujemaa présente, à titre d'exemple, l'indexation et la recherche d'images par le contenu visuel.

DE LA NÉCESSITÉ DE COOPÉRER

À la lecture de cet ouvrage, que vous soyez archiviste, bibliothécaire ou documentaliste, vous serez à même de mieux apprécier les

immenses potentialités de la recherche d'information sur les réseaux... mais aussi les vastes chantiers qui restent à ouvrir ou à développer. On oublie encore trop souvent – volontairement ou non – qu'à l'ère du document numérique la localisation de ce dernier comme celle de qui le traite importent peu. La coordination, la collaboration sont indispensables pour ne pas gaspiller des moyens nécessairement limités. Ceci passe par le respect des normes et l'adoption de standards ; tout au long de ce livre les auteurs en auront souligné l'importance.