HAL
open science

Aide au référencement d'un site web

Annie Frarin

To cite this version:

Annie Frarin. Aide au référencement d'un site web. Sciences de l'information et de la communication. 2000. dumas-01549149

HAL Id: dumas-01549149

https://dumas.ccsd.cnrs.fr/dumas-01549149v1

Submitted on 28 Jun 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Annie FRARIN

MAITRISE

SCIENCES DE L'INFORMATION ET DE LA DOCUMENTATION

AIDE AU REFERENCEMENT D'UN SITE WEB

Rapport de stage

Stage effectué du 1er février au 30 juin 2000 au service de l'Ingenierie éducative et documentaire du Centre régional de documentation pédagogique du Nord Pas-de-Calais, 3 rue Jean Bart à Lille

Sous la direction de :

Madame C. Van Lancker *responsable universitaire*

Monsieur P. Fernandez *responsable professionnel*

Université Charles de Gaulle Lille 3
UFR IDIST

octobre 2000

B.U.C. LILLE 3
D 021 570474 3

2001 # 1825

Annie FRARIN

MAITRISE

SCIENCES DE L'INFORMATION ET DE LA DOCUMENTATION

AIDE AU REFERENCEMENT D'UN SITE WEB

Rapport de stage

Stage effectué du 1er février au 30 juin 2000 au service de l'Ingenierie éducative et documentaire du Centre régional de documentation pédagogique du Nord Pas-de-Calais, 3 rue Jean Bart à Lille

Sous la direction de :

Madame C. Van Lancker *responsable universitaire*

Monsieur P. Fernandez *responsable professionnel*

Université Charles de Gaulle Lille 3
UFR IDIST

octobre 2000

Remerciements à

Mme Christine Van Lancker, pour ses conseils et sa disponibilité

Mme Sylvie Prieux et Mr Pierre Fernandez, pour leur accueil et leur aide

mes collègues, qui m'ont supportée…

TABLE DES MATIERES

I. LE STAGE, BESOINS, RAISONS, OBJET.....p. 1

I.1 Le lieu..... p. 1

I.1.1 L'institution, ses missions, son public..... p. 1

I.1.2 Le service concerné, ses missions spécifiques.....p. 2

I.2 Le moment et la demande..... p. 4

II. ETAT DE LA QUESTION DU REFERENCEMENT : DES PISTES POUR FAVORISER L'ACCES A L'INFORMATION..... p. 5

II.1 Outils de recherche..... p. 6

II.1.1 Les outils généraux..... p. 6
Moteurs et annuaires.....p. 6
Evolutions de ces outils.....p. 9

II.1.2 Le catalogage des ressources..... p. 11

II.1.3 L'identification des ressources..... p. 16

II.1.4 Les microstructures.....p. 17

II.1.5 Le filtrage d'informations.....p. 18

II.2 La recherche par l'utilisateur.....p. 20

II.2.1 Utilisation de mots-clés.....p. 20

II.2.2 La norme Z 39.50..... p. 21

II.3 Le référencement des sites.....p. 22

III. APPLICATIONS DE CES RECHERCHES AU SITE..... p. 23

III.1 Mise en place d'une recherche-type.....p. 23

III.2 Lancement des recherches, évaluation des résultats.....p. 29

III.3 Propositions de modifications..... p. 30

III.4 Mise à jour du référencement et résultats.....p. 31

IV. CONCLUSION..... p. 37

BIBLIOGRAPHIE.....p. 39

ANNEXES..... p. 40

I. LE STAGE, BESOINS, RAISONS, OBJET

I.1 Le lieu

I.1.1 L'institution, ses missions, son public

Le lieu de mon stage, le CRDP (Centre régional de documentation pédagogique) du Nord Pas-de-Calais, est un des vingt-huit établissements publics nationaux à caractère administratif faisant partie du réseau des Centres de documentation pédagogique, le CNDP (Centre national de documentation pédagogique, organisme à caractère administratif, sous la tutelle du ministère de l'Education nationale) étant le pilote de cet ensemble. La mission des CRDP est la mise à disposition auprès de son public prioritaire, les enseignants, d'une part, de documentation administrative (textes réglementaires, programmes, etc.), et d'autre part, de ressources pédagogiques et didactiques multimédias, et ce à travers des services d'édition, de documentation (vente, prêt, consultation de documents, aide à la recherche), d'ingénierie éducative (formation à la recherche et à l'outil Internet, présentations et analyses de logiciels didactiques).

Une enquête sur les activités des CRDP en 1998, impulsée par le ministère de l'éducation nationale, a été analysée et a fait apparaître le manque d'utilisation, par les enseignants, des ressources documentaires offertes par le réseau des centres de documentation pédagogique ; sur l'académie de Lille, qui est la plus importante académie de France en nombre d'enseignants, l'ensemble des cinq médiathèques académiques a le plus fort taux de prêts de documents sur l'ensemble des médiathèques pédagogiques nationales (ratio de prêts par inscrits : 25 par an, 187497 prêts/an), et cependant, elles ne touchent « que » 6% de l'ensemble des enseignants du Nord Pas-de-Calais ; même si l'on double ce nombre pour prendre en compte ceux qui ne font que consulter ces mêmes ressources (qu'on ne peut donc comptabiliser à travers une inscription), on est encore bien loin du compte, surtout lorsque l'on sait qu'à peu près la moitié du public qui fréquente ces centres est constitué d'étudiants - certes pour la plupart en

sciences de l'éducation ou en formation à l'IUFM (Institut Universitaire de Formation des Maîtres), donc potentiellement futurs enseignants -.

Il est certain que l'étendue des deux départements constituant cette académie explique une partie des difficultés qu'ont les enseignants à venir sur des lieux de documentation proportionnellement peu nombreux (trois sur le Nord, deux sur le Pas-de-Calais) : la distance, mais aussi les horaires et jours d'ouverture sont une raison de la défection de ce public.
Aussi l'accès distant aux ressources ou à leur signalement pourrait être une réponse apportée à ces problèmes de distance, de temps et de disponibilité qui se posent à nombre d'entre eux, ainsi qu'un moyen d'améliorer la connaissance des outils pédagogiques disponibles, puisqu'il permet (au moins en théorie) de toucher d'une part tous les établissements scolaires via Internet et la messagerie électronique, et d'autre part les enseignants eux-mêmes en tant que pratiquants individuels d'Internet.
Cependant, mettre à disposition ne suffit pas, loin de là, surtout au milieu de la profusion d'informations disponibles sur Internet ; il faut signaler cette information, la rendre repérable, il faut être sûr que le public concerné arrive à la retrouver.
C'est dans ce questionnement qu'il m'a semblé intéressant d'entrer : la mise à disposition à distance de ressources résout-elle pour autant le problème d'accès à ces mêmes ressources ?
De là découlait une recherche sur la structuration de l'information sur Internet : comment accède-t-on à l'information, par quels moyens, comment est-elle mise à disposition ?

I.1.2 Le service concerné, ses missions spécifiques

Le CRDP, avec ses cinq centres implantés dans l'académie (Dunkerque, Lille et Valenciennes pour le département du Nord, Arras et Outreau pour le Pas-de-Calais), a mis en place des outils pour répondre à (ou anticiper) toute recherche documentaire de son public :

- des services proposés localement, sur place, comme les catalogues informatisés des médiathèques et les catalogues (papier cette fois) des

productions en vente, proposant ainsi aux enseignants le prêt et/ou l'achat de documents ; mais aussi la consultation ou la vente de textes officiels, de programmes disciplinaires, de rapports de jury ; et aussi des documentalistes pouvant aider tout usager à faire une recherche à travers différents outils (allant du catalogue interne au centre concerné à des bases bibliographiques externes, en passant par la consultation du site web du CRDP, qui offre différentes pistes pour trouver de l'information, et/ou de sites distants pour compléter celle-ci)

- des services proposés à l'extérieur (accessibles depuis n'importe quel ordinateur individuel, et plus particulièrement dans les centres de documentation et d'information des établissements scolaires) à partir du site web du CRDP, qui va pointer les différents services fournis par le centre pouvant être utiles à son public (dont l'accès aux catalogues des médiathèques afin de localiser des documents et des pages « thématiques » par disciplines d'enseignement), et des sites pertinents pour les enseignants (ministériels, académiques, CNDP, ...).

A l'intérieur du CRDP, un de ses services, celui de l'ingénierie éducative et documentaire (IED) a pour mission plus spécifique d'assurer un rôle de service conseil et d'expertise dans le domaine de l'informatique pédagogique, en passant par des animations autour de thèmes et d'outils liés à l'information électronique, en participant à la formation des enseignants-documentalistes de l'académie ; il gère également la mise à jour du site web.

Ce site existe depuis 1996. Son adresse a d'abord été http://www.univ-lille1.fr/~puel/CRDP/homecrdp.html, puis http:// nordnet.fr/crdp, ensuite http://www.ac-lille.fr/crdp, sous-catégorie du site de l'académie de Lille, et dernièrement http://crdp.ac-lille.fr.

La création des pages web du site, sa maintenance et sa mise à jour ont été assurées jusqu'à présent par deux personnes, dont une à mi-temps.

Le site a été envisagé au départ comme un moyen de présenter les différents services internes du CRDP, susceptibles d'intéresser un public enseignant ; on y trouve donc beaucoup de pages descriptives. Des pages spécifiques pour l'enseignement de certaines matières y ont été créées afin d'apporter des pistes pédagogiques et des aides aux enseignants concernés.

Les pages web ont été créées à partir du logiciel FrontPage par des personnes « techniciennes », ayant la connaissance du logiciel en question et des impératifs liés aux signalements de sites et à leurs changements (voir le chapitre sur le référencement de sites).

I.2 Le moment et la demande

Cependant aucun travail sur la façon dont le site peut répondre à une demande d'usager dans le cadre de services documentaires n'avait pu être mis en place, et le positionnement des pages du site dans une liste de réponses pouvait aider à évaluer la pertinence du signalement qui en avait été fait. Ma proposition d'y consacrer du temps lors d'un stage a de ce fait fortement intéressé les « webmasters » qui n'avaient jamais pu aborder ce sujet.

Par ailleurs, je m'étais moi-même intéressée, à travers la recherche documentaire sur Internet (moyens de recherche des usagers et types de réponses reçues), aux signalements des sites Web et des documents contenus dans ceux-ci, et en particulier à l'utilisation possible d'une indexation documentaire dans ce cadre.

Arrivant à un moment où le site du CRDP était en pleine restructuration, la période de stage pouvait permettre d'observer la « fabrication » des pages du site, puis de faire des recherches sur le web quant à la possibilité de retrouver le contenu de celles-ci, en se plaçant en tant que public particulier ; à partir des observations faites, on pouvait alors proposer des ajouts, des modifications, afin d'améliorer le positionnement des pages du site lors de recherches documentaires, en s'appuyant sur l'évolution des outils d'aide à la recherche informatisée.

Pour ce faire, on fera donc un rapide tour d'horizon de ces outils afin d'en connaître les possibilités.

II. ETAT DE LA QUESTION DE LA RECHERCHE SUR INTERNET : DES PISTES POUR FAVORISER L'ACCES A L'INFORMATION

La recherche d'informations ne date certes pas d'aujourd'hui. On trouvera un exemple intéressant sur les pratiques de Voltaire (1694-1778) en la matière dans un article de Jean-François Cerisier (voir bibliographie), où, à travers le travail d'une étudiante de maîtrise de sciences de l'information, on découvre que l'écrivain, maîtrisant les outils d'accès à l'information et aux documents (tels que répertoires, catalogues, etc.), avait une démarche d'appropriation de l'information et un réseau d '« informateurs » lui permettant de compléter ses connaissances.

Cette démarche reste toujours valable. Seulement, avec l'arrivée de l'informatique, la quantité d'informations créées, stockées et pouvant être interrogées, a connu une croissance énorme, et les outils d'accès à l'information ont dû tenir compte de cette masse.

Cependant tant que l'on accédait à des catalogues ou des banques de données structurées, on restait dans la rigueur des systèmes documentaires. Le contenu du document était décrit sous une forme synthétique (la notice) par son auteur ou un spécialiste du traitement de l'information, garantissant ainsi une homogénéité de la présentation des documents à travers l'utilisation de règles particulières (normes) et le choix de mots-clés appropriés. La recherche dans un système informatisé pouvait alors se faire sur certains critères prédéterminés de ces notices, choisis en fonction de leur pertinence (tout comme les fiches de catalogage des catalogues de bibliothèques) permettant d'accéder au document lui-même ou non. En complément il était possible de rechercher sur le texte du document lui-même , permettant alors d'enrichir la technique booléenne de recherche sur les notices (ET, OU, SAUF) par des fonctions de recherche sur le texte lui-même (opérateurs de proximité, recherches dans un paragraphe, une phrase, ...), parfois couplées à des techniques, statistiques et linguistiques essentiellement, permettant de réduire le bruit ou le silence liés à la recherche en texte intégral.

Sur Internet, plus d'accès « normalisé » au document, l'information est censée être disponible directement. Simplement, le volume de documents

accessibles en ligne et la vitesse de croissance de ce même volume ont fait qu'il est impossible pour un usager d'accéder à l'ensemble des informations l'intéressant sans l'aide d'outils appropriés à cette recherche.

II.1 Outils de recherche

II.1.1 Les outils généraux

Moteurs et annuaires

Il existe à la base deux types d'outils fonctionnant soit en interrogation directe sur le contenu du document, soit en interrogation par mots-clés. Il y a donc deux types d'approche :

- la collecte automatique suivie de l'indexation également automatique de ressources à l'aide de robots de recherche (« crawler » ou « spider ») : ce sont les moteurs de recherche.
- la constitution de répertoires organisés avec une validation, une classification humaine des ressources et une indexation automatique sur des zones particulières : ce sont les annuaires ou répertoires.

Dans le premier cas, l'usager pose une requête sous forme d'une liste de mots séparés par des opérateurs implicites ou choisis et le moteur cherche en texte intégral sur le contenu des pages HTML (HTML ou HyperText Markup Language est le langage de description des pages web, qui permet de structurer leur présentation et d'en décrire un minimum le contenu). Dans le second cas, l'usager peut poser une requête sur un index limité qui interrogera en fait sur les descriptions des sites et les titres de catégories, il peut aussi naviguer dans l'arborescence de la classification.

Les annuaires signalent les sites dans leur ensemble (en pointant vers leur page d'accueil), alors que les moteurs traitent indistinctement toutes les pages qu'ils rencontrent, même si elles proviennent du même site.

Quelques précisions :

Les robots :

Les services comme Alta Vista ou Excite se décomposent en *3 modules majeurs* : le module de collecte automatique des ressources (c'est un logiciel

qui passe son temps à demander des copies de pages html), le module d'indexation (au sens constitution de fichiers inverses), le module de recherche d'information gérant aussi l'interface avec l'usager.

Choix des sites explorés : cette stratégie est variable mais très souvent on utilise les pages " What's new ", des classements de sites très demandés, les déclarations spontanées des auteurs mais également une exploration aléatoire en privilégiant les URL courts.

Contenus indexés : très variable. On peut indexer les pages entières (ex : Altavista), les titres html, les premiers paragraphes et les liens (ex : Lycos). L'exploration des liens en profondeur et en largeur peut varier également. Cette indexation est associée à une pondération des mots qui tient compte à la fois de critères de fréquence mais aussi de l'encodage HTML, ainsi les mots des titres (balise ou tag *title* en langage HTML) reçoivent des poids plus importants.

Il faut aussi souligner l'utilisation des métadonnées d'indexation (dans les balises HTML) fournis directement par l'utilisateur sous forme d'une liste de mots-clé et d'un résumé. Certains robots vont utiliser ces données pour indexer et surtout pour afficher le descriptif du site. Cette indexation laissée à l'auteur n'est absolument pas normalisée et peut subir des dérives facilement identifiables (du genre répéter x fois le même mot clé pour donner un poids très fort à la ressource en fonction de ce mot clé). Certains robots vérifient un minimum ce type de dérive.

La production et l'analyse de ces champs métadonnées ne sont cependant pas généralisées.

Utilisation : pour utiliser les robots, il faut connaître la syntaxe du langage d'interrogation. Un guide en ligne est toujours disponible. Les interfaces présentent généralement 2 modes d'interactions:

* un mode simple visant l'utilisateur final : on donne une liste de mots et le système retrouve des documents répondant "au mieux" à cette liste selon des algorithmes peu explicités mais qui tiennent compte de la localisation, de la proximité et de l'occurrence des termes. La troncature à droite est souvent implicite.

* un mode "avancé" où l'utilisateur peut utiliser des opérateurs différenciés

entre les termes de sa question et éventuellement orienter la pondération des documents retrouvés.

L'utilisateur doit se rappeler quelques règles de base qui tendent à se généraliser :

- utilisation des guillemets (") pour spécifier une expression
- utilisation d'opérateurs de présence (+) ou l'exclusion (-) absolue de termes
- tester la sensibilité ou non aux caractères accentués et à la casse majuscule/minuscule des caractères.

Avantages pour l'usager : compte tenu de l'automatisation du processus, les robots sont les outils ayant la couverture la plus exhaustive de l'espace public Internet, la mise à jour est rapide. L'interaction est simplifiée. Mais les réponses sont hétérogènes, les doublons ne sont pas détectés, le bruit peut être considérable.

Les répertoires validés ou annuaires :

Ce sont des inventaires thématiques de ressources organisées selon une classification. Ces ressources sont aussi indexées sur certaines zones. Ces répertoires couvrent de nombreuses thématiques.

Choix des sites explorés : la majorité des ressources recensées sont des ressources déclarées par les auteurs et validées par les responsables du répertoire. Des collectes automatiques peuvent assister éventuellement ce travail de recensement.

Contenus indexés : contrairement à l'approche robots, on indexe ici des zones limitées (URL, titres, résumés éventuels fournis par l'auteur, commentaires des administrateurs, titres des rubriques thématiques de la classification).

Utilisation : l'utilisateur peut soit naviguer en empruntant la classification prévue, soit formuler des équations de recherche sur une partie de la classification. Il faut aussi comprendre la syntaxe des requêtes pour contrôler sa recherche.

Avantages pour l'usager : l'intérêt majeur de cette approche est la valeur ajoutée humaine (contrôle, validation, organisation) qui limite ainsi le bruit des réponses. La classification des ressources permet de répondre à des requêtes de type exploratoire sur une thématique. Par contre, la couverture est moins exhaustive que celle des outils automatiques de type robots.

Cependant, il devient de plus en plus difficile de faire la différence entre ces outils, d'autant plus que les annuaires sont souvent alliés à des moteurs de recherche pour leur permettre de répondre à des requêtes trop spécifiques pour être dans leur propre base de données, tandis que les moteurs de recherche peuvent être associés à des annuaires pour leur permettre de localiser des informations commerciales par exemple, très difficiles à localiser par mots-clés dans des documents indexés en texte intégral. Cette mixité peut être illustrée par celle de Voilà, où la recherche est dégrossie dans l'annuaire puis sont fournies les réponses du moteur.
Pour approfondir la question, on trouvera des informations entre autres sur le site d'Olivier Andrieu et dans le document de Jean-Pierre Lardy cités dans la bibliographie.

Evolutions de ces outils

Avec la pression concurrentielle, ces outils ont évolué vers la « portalisation », transformant les sites des outils de navigation en « galeries marchandes », en essayant de proposer à l'utilisateur les produits convenant le mieux à son profil et en lui offrant de multiples services pratiques afin de le fidéliser. Le concept de « métaportail » vient même s'ajouter à cela (venant des Etats-Unis, l'un d'eux, « consacré aux professionnels de l'information », vient d'apparaître en France : http://www.drapo.com/indexfr.html), proposant à l'usager d'interroger différentes bases de données accessibles directement sur la barre de menus (sites des professionnels de l'information, listes de diffusion, mais aussi moteurs de recherche, sites de vente d'ouvrages) tout en gardant sa recherche initiale.

Certains moteurs cependant restent non accessibles au public (comme Inktomi qui loue ses fonctionnalités et ses bases à des clients comme Yahoo!, AOL, Hotbot ou Nomade par exemple) ou « purs », en contrepoint des « portails », mettant en avant la recherche (ainsi Google, AllTheWeb, mais aussi Raging lancé par Altavista, qui a le même moteur et la même base, mais dont l'interface, dépouillée, est beaucoup plus rapide), s'adressant ainsi aux chercheurs et non plus au grand public. Parmi eux, une nouvelle génération met en avant de nouvelles fonctionnalités, reprises

peu à peu par les autres outils ; ainsi certains essaient d'analyser le comportement de leurs visiteurs à travers « l'indice ou l'analyse de clic(s) » : si l'utilisateur revient directement sur la page des résultats après avoir cliqué sur un lien, le moteur en conclut que le contenu du site ne correspondait pas à son attente et que l'association entre les mots-clés utilisés dans la requête et ceux du site n'est pas bonne, le site est alors rétrogradé dans le classement des résultats. Ce type d'étude peut être doublé d'une analyse des liens : plus le nombre de sites traitant d'un sujet similaire et pointant vers une page est grand, plus la page a des chances d'être pertinente sur un sujet (« indice de popularité ») ; pour contourner le problème du « spamming » (créer de multiples fausses pages qui pointent vers une page pour optimiser son classement), certains moteurs comme Google calculent la popularité à deux niveaux : seront mieux classées les pages vers lesquelles pointent des pages elles-mêmes populaires.

Une autre direction est celle du « moteur de questions » (tel Ask Jeeves) : le moteur va chercher dans sa base de données de questions préalablement définies celle qui se rapprochera le plus de celle de l'usager, et lui ramènera les réponses déjà trouvées (Ask Jeeves a d'ailleurs été intégré dans Altavista en complémentarité de ses résultats).

D'autres encore, tout en proposant des listes de liens vers des pages Web susceptibles de correspondre à la demande de l'usager, rangent également ces pages dans des dossiers, qu'il est alors possible d'ouvrir et d'organiser (NorthernLight).

Les métamoteurs interrogent en parallèle plusieurs moteurs de recherche et annuaires, et fusionnent leurs résultats : cela peut être intéressant, puisque chaque moteur n'a dans sa base de données qu'une petite partie des documents disponibles sur le Web (30% d'après les études faites) ; en les regroupant, on aura sûrement plus de références de pages, mais reste le problème de l'interrogation avec la même syntaxe alors que l'on sait que la façon d'interroger n'est pas la même sur tous les moteurs, et de la synthèse de résultats de moteurs qui n'ont pas les mêmes critères de pertinence.

Les outils appelés « agents intelligents » peuvent désigner les métamoteurs les plus évolués, les moteurs de questions et ceux qui tiennent

compte de l'indice de popularité, mais aussi d'autres logiciels qui eux aussi profitent de l'apport de la recherche en ingénierie linguistique et statistique : ainsi Altavista utilise la technologie Cow9 dans sa fonction *Refine* (un tri statistique des termes posés est fait parmi les pages issues de la recherche et un regroupement par paquets ou *clusters*, définis par des séries de mots, est fait ; cette fonction révèle autour des termes de la recherche les réseaux des mots utilisés, il est donc possible d'affiner sa question avec des termes plus précis. A l'heure actuelle, des outils, qualifiés d'agents de recherche d'informations ou d'agents de filtrage, sont utilisables dans un contexte hors ligne et permettent d'assurer globalement de façon plus ou moins automatique ces quatre grandes fonctions :

- l'analyse de la question de l'utilisateur
- l'analyse des résultats de recherche
- le rapatriement des informations sélectionnées pour en dégager l'information pertinente
- le filtrage et la mise à jour des résultats.

Ainsi DigOut4U(société Arisem) exploitant les traitements linguistiques (http://www.arisem.com), InformationMiner4U, qui le complète, travaillant sur les profils de recherche, et Class4U permettant de constituer une classification conceptuelle, tout comme UmapWeb (société Trivium) et Sampler TM (société Cisi) (http://www.trivium.fr ; http://www.cisi.fr).

II.1.2 Le catalogage des ressources

Les professionnels de l'information, les chercheurs apprécient sûrement l'évolution de tous ces outils ; néanmoins, l'utilisateur non averti ne sait pas obligatoirement les utiliser et obtient des résultats souvent insatisfaisants, dûs à plusieurs raisons :

- beaucoup de « bruit », généré par une indexation informatique en texte intégral sur du texte non structuré sémantiquement. En effet, les données présentes sur le réseau sont décrites à l'aide d'un langage très simple, le langage HTML, qui précise la structure et la mise en forme de l'écrit et permet d'inclure des éléments extérieurs (images, sons, vidéos...), et ce à l'aide de balises ou tags nommés *title, link, url...* Cependant, ce langage ne

rend pas compte de la structure du texte lui-même ni de son « intelligence », et mélange les descriptions des structures logiques et de simples éléments de mise en page. On a vu que, néanmoins, les outils de recherche cherchent à limiter le bruit à travers des algorithmes de pondération en présentant d'abord à l'usager les ressources où les termes de sa question apparaissent fréquemment et dans des zones telles que les champs *title* HTML.
- l'instabilité des ressources : les adresses changent sans que les index des outils soient mis à jour en conséquence.
- le manque d'identification des ressources : pas de date de création, pas de nom d'auteurs, pas d'information générique sur le contenu... Lors du référencement de son site, le concepteur peut soumettre un résumé auprès des annuaires, mais ce ne sont en général que quelques lignes succinctes, et qui n'apparaissent pas toujours en entier lorsque l'adresse du site apparaît.
- l'hétérogénéité des ressources trouvées : l'élément de base est un fichier HTML et l'utilisateur n'a aucun moyen de spécifier qu'il ne s'intéresse qu'à un certain type d'objet, par exemple uniquement à des cours ou à des articles.

Pour pallier à ces types de problèmes, plusieurs solutions ont été envisagées :
- comme on l'a vu, certains robots proposent des algorithmes statistiques de regroupement des ressources pour fournir à l'utilisateur une certaine lecture des réponses : on présente des agrégats de ressources partageant de nombreux termes et l'utilisateur peut reformuler sa question en sélectionnant ou en éliminant des termes proposés dans ces agrégats. Alta Vista propose ce type de fonctionnalité automatique appelée " Live Topics ". Ceci ne résout qu'en partie le problème du bruit et n'apporte pas de solutions aux autres points soulignés précédemment. La fonction est cependant appréciable comme aide à l'exploration d'une thématique, elle ne nécessite pas d'intervention humaine.
- une autre approche est l'élaboration par un être humain de descriptions des ressources qui sont appelées des **metadata**. Ces metadata sont plus complètes que les tags que nous avons évoqués pour l'indexation effectuée par les robots et plus détaillées que les commentaires des administrateurs

des répertoires. On rejoint la fonction classique de catalogage effectuée par les bibliothèques.

Les metadata ou métadonnées sont des données structurées sur les données, qui vont aider à l'identification, à la description et à la localisation des ressources sur le réseau. Elles sont véhiculées par un langage de description de pages web qui est en train d'évoluer : le HTML , qui était lui-même une version « simplifiée » du SGML (Standard Generalized Markup Language), commence à être remplacé par le XML (eXtended Markup Language), avec à ses côtés le XHTML, normalisé fin janvier 2000 par le W3C (World Wide Web Consortium : regroupement de compagnies qui préside au développement du World Wide Web), couplant la puissance du XML et la simplicité du HTML, et permettant de conserver une compatibilité entre la production éditoriale existant sous HTML et celle qui naît sous XML.

Un cadre, reposant sur une structure XML, va recevoir les métadonnées des documents : c'est le RDF (Resource Description Framework) : le RDF est un ensemble de conventions qui supportera l'interopérabilité entre des applications qui échangeront des métadonnées. La syntaxe sera exprimée en langage XML mais la sémantique sera définie par les besoins des usagers. C'est un cadre : il peut donc recevoir toutes sortes de métadonnées (même celles à venir !) et même emboîter plusieurs standards de métadonnées.

Dans le modèle RDF, les métadonnées peuvent être associées à la ressource de quatre manières différents :

- encapsulées dans la ressource (« *embedded* ») ; dans ce cadre, deux standards principaux ont été définis :

  le projet TEI (Text encoding initiative) qui s'est concrétisé par une DTD (Définition de type de document) SGML, et qui s'applique à chaque document électronique encodé selon la DTD et définit pour chaque unité documentaire électronique un en-tête obligatoire comportant les métadonnées. Un des chapitres de la TEI P3(Public proposal number 3) est consacré aux citations et références bibliographiques, et précise quelles sont les balises à retenir pour la description bibliographique d'un document, quels intitulés leur donner et comment les présenter.

la DTD HTML, définie pour la structuration des pages Web de l'Internet, sur laquelle ont travaillé des communautés d'utilisateurs afin de standardiser les types de métadonnées. De là a été créé le Dublin Core (DC), qui prévoit l'utilisation de 15 balises pour décrire « bibliographiquement » une ressource électronique (**10**) ; c'est une norme servant à la description des ressources sur Internet, ses éléments sont optionnels, répétables, elle est interdisciplinaire, internationale et extensible grâce à l'ajout possible de sous-éléments dans chaque élément. *Exemple :* dans l'élément « Creator » on peut ajouter « Given Name », « Surname », « Affiliation », « Contact Info », etc. Différents projets importants commencent à utiliser cette standardisation (voir les liens rassemblés par l'IFLA sur le sujet :http://ifla.inist.fr/II/metadata.htm) et des sites ont mis en place des métadonnées selon le Dublin Core : ainsi celui de la Bibliothèque nationale française (http://www.bnf.fr) ou celui de coopération entre sites Web en bibliothéconomie et sciences de l'information Sitebib (http://www.abf.asso.fr/sitebib).

- externes à la ressource mais fournies avec elle dans le mode de transfert de la réponse à une recherche d'information (« *along-with* »)
- utilisées séparément de la ressource pour une recherche d'information, éventuellement dans des bases de données différentes (« *service bureau* »). Ceux-ci concernent la description bibliographique ou documentaire, et recouvrent les règles de description (définition des éléments de description et règles à observer pour créer ces descriptifs à partir des informations figurant dans les ressources que l'on décrit) et les formats de description (codage en machine permettant de structurer les données en vue de leur exploitation dans des systèmes d'interrogation). Certains se sont déjà adaptés au contexte des ressources électroniques en étendant les règles de description et les formats aux besoins de recherche et d'accès direct aux ressources électroniques. Ainsi, le format MARC (Machine readable cataloging), format de gestion et d'échange de données bibliographiques informatisées, offrent désormais la possibilité d'indiquer la localisation de la ressource électronique, ses caractéristiques techniques, ainsi que le lien à partir duquel on pourra y

accéder : les formats USMARC et UNIMARC qui en sont dérivés ont défini le champ 856 à cet effet.

Et en décembre 1999, un logiciel libre XMLMARC a été lancé ; ce programme convertit le format MARC en XML en se basant sur une DTD détaillée pour les formats bibliographiques et d'autorité (http://www.dlib.org/dlib/march00/03inbrief.html#MILLER)

- la ressource est encapsulée dans les métadonnées qui la décrivent (« *wrapped* »). La DTD EAD (Encoding archival description) a été développée pour la structuration et l'encodage en XML des instruments de recherche décrivant des collections spécialisées telles que les archives ou les collections de papiers, et permet une structuration de l'instrument de recherche en douze niveaux hiérarchiques et l'attachement de la ressource électronique à sa description.

D'autres métadonnées existent ou sont en cours d'élaboration : le PICS (Platform for Internet content selection), permettant de classer le niveau du document, son contenu moral, le P3P codant son niveau de difficulté ou de secret, MPEG-7 (*Moving picture experts group*) pour le catalogage de documents numériques comme les vidéos, etc.

Les métadonnées doivent permettre :

- de meilleures performances dans la recherche d'informations (les moteurs ayant une structure où naviguer, comme les champs dans une base de données)
- une meilleure description du contenu et des relations entre les différents fichiers d'un site web
- la description de collections de documents qui représentent un document logique
- le classement du contenu suivant un degré de difficulté ou un public cible
- la description des droits de propriété intellectuelle ou d'accès à des pages web
- l'encodage d'une signature numérique pour valider les textes circulant sur Internet

On peut se demander aussi si tout ceci sera réellement respecté et utilisé. Pour le moment, ces standards ne sont pas généralisés mais il est fort

probable qu'ils puissent le faire assez vite. Les outils de recherche tels que les robots ne proposent pas à l'utilisateur d'exploiter ce type de données pour le moment mais ils le feront certainement dans un proche avenir. Il est intéressant néanmoins de savoir quelles sont les balises interrogées par les moteurs, et à quoi ceux-ci attribuent le plus d'importance pour ramener l'adresse d'une page web ; pour ce faire, on se reportera au site web d'Olivier Andrieu (cf. bibliographie).

II.1.3 L'identification des ressources

Des numéros ou codes d'identification des documents existent depuis longtemps : l'ISBN (International standard book number) pour le livre et l'ISSN (International standard serial number) pour les publications en série. Ces identifiants sont utilisés à la fois par les éditeurs et par les bibliothèques et centres de documentation, d'un côté pour la gestion de la commercialisation, de l'autre pour la gestion des commandes et l'identification bibliographique et documentaire des ouvrages et périodiques. Envisagée dans le contexte du Web, cette identification doit non seulement identifier et localiser une ressource, mais aussi y donner accès. C'est le système URI (*Uniform resource identifier*) qui va proposer un cadre et des règles, s'appuyant sur le dispositif suivant :

l'**URN** (*Uniform resource name*) de la ressource est son nom unique et persistant : la ressource pourra être présente sur plusieurs sites comme autant d'exemplaires. Ce numéro est résolu (c'est à dire qu'il indique le ou les URL du ou des sites où on peut consulter la ressource) par une agence (agence de résolution du nom) où il est enregistré et qui maintient la validité des URL.

l'**URL** (*Uniform resource locator*) est l'adresse du site où se trouve la ressource (ce lien n'est pas persistant et sa validité n'est pas garantie).

l'**URC** (*Uniform resource characteristic*) contient des métadonnées sur la ressource, en particulier les modalités d'accès et la gestion des droits.

Le système DOI (*Digital object identifier*) est une application du système URI : le numéro DOI est un URN au sens Internet, composé d'un préfixe attribué à l'éditeur et d'un suffixe qui est le numéro de la ressource chez l'éditeur ou le détenteur des droits. Il peut être un identifiant issu des systèmes existants, tel l'ISBN ou l'ISSN.

II.1.4 Les microstructures : clôture de l'espace

On observe également de nombreuses initiatives pour délimiter l'espace à un sous-ensemble de ressources utiles à une discipline, à une communauté d'usagers partageant les mêmes centres d'intérêt. Chaque micro-structure construit ses clés d'accès correspondant aux demandes des utilisateurs et aux ressources disponibles.

Par exemple, dans le domaine des ressources éducatives, différents acteurs ont initié des recensements spécialisés de ressources utiles à l'enseignant et à l'élève, l'objectif majeur étant d'accroître l'efficacité et la richesse des recherches de l'utilisateur dans ce domaine. Très souvent, les ressources ont été évaluées et organisées par discipline et/ou par niveau de classe. Ces répertoires constituent des filtres permettant de guider au mieux l'utilisateur intéressé par ce type de ressources. Par exemple :

Cyberscol : c'est un projet visant à favoriser l'exploitation des ressources éducatives d'Internet et le développement de ressources francophones. Le projet est d'origine québécoise et implique différents partenaires éducatifs. Le répertoire propose à la fois des sélections de sites et des scénarios pédagogiques utilisant ces ressources.
URL: http://www.cyberscol.qc.ca/Accueil.html

Les ressources du réseau CNDP, CRDP : sur les sites web de ces organismes, on trouve déjà certains documents pédagogiques utilisables par les enseignants. Un projet national a permis de mettre en place une base de données appelée **Educasource,** pilotée au départ par la Direction de l'information scientifique, des technologies nouvelles et des bibliothèques et le CNDP. Cette base est le produit d'une collecte nationale des supports pédagogiques disponibles, l'objectif est de fournir aux enseignants une aide pour leurs recherches d'informations, leurs préparations de cours, l'animation

des cours et les échanges d'expériences.
URL : http://www.educasource.education.fr.
Un vaste projet américain dans ce domaine est le programme **K-12** fédérant un ensemble de sites américains fournissant des ressources éducatives.
On retrouve aussi le concept de « portail », cette fois non plus sur des sites généraux comme ceux des outils de recherche, mais sur des sites déjà spécialisés, qui vont rassembler toute l'information concernant un domaine particulier.
Ainsi *Educlic*, « portail des professionnels de l'éducation » (accessible à l'URL http://educlic.education.fr ou à partir du site du CNDP (http://www.cndp.fr), mis en place le 30 juin de cette année par le CNDP à la demande du ministère de l'éducation nationale, dont l'objectif principal est de valoriser la production éducative des acteurs publics ou associatifs sur le Web français, qui couvre plus de 200 serveurs éducatifs, et permet de faire une recherche par thèmes, ou directe par mot ou expression dans l'ensemble des documents hébergés par ces mêmes serveurs ; il va également donner accès à des informations utiles aux enseignants (liste des vidéos exemptes de droits pour pouvoir être utilisées en classe, par exemple).

II.1.5 Filtrage d'informations et info-agents

Face à la quantité, à la diversité et au renouvellement continuel des informations disponibles sur Internet, le marché des services d'information personnalisée se développe également dans différents domaines, correspondant à un repérage automatique. Le principe est de filtrer les informations selon un profil personnalisé et de diffuser régulièrement et automatiquement les nouvelles informations collectées à l'utilisateur (DSI diffusion sélective de l'information). Le filtre va être réalisé par un agent, c'est à dire un logiciel qui agit pour accomplir des tâches fixées par son utilisateur.
Un agent va en fait être caractérisé par :

- son architecture technique
- son domaine d'application (courrier électronique, news, dépêches, ...)

- son " intelligence "(pour la définition du profil : mots fournis par l'utilisateur ou fonction d'auto-apprentissage, pour la stratégie de recherche : choix explicite des outils par l'usager ou auto-sélection).

Il existe encore très peu d'agents intelligents, c'est à dire capable de prendre des décisions pour l'usager avec des fonctions d'auto-apprentissage. On trouve, par contre, de nombreux programmes paramétrés par l'usager et effectuant ce travail de filtres dans des domaines variés : filtrage des pages web en fonction d'un profil, filtrage d'articles de revues et magazines (comme les services d'alerte des éditeurs scientifiques), filtrage des dépêches de presse, mais aussi fonction de DSI dans certains logiciels de gestion de bibliothèques ou de centres de documentation, grâce à laquelle le lecteur de la bibliothèque ou l'usager (interne ou externe) du centre peut être mis au courant de ce qui existe dans le domaine qui l'intéresse et informé régulièrement des nouveaux documents correspondants.
Concernant le domaine éducatif, un filtrage important concerne la non-visualisation de sites problématiques : pornographie, violence, racisme... Rien ne garantit aujourd'hui qu'un élève utilisant un robot de recherche par exemple, ne puisse visualiser ce type de ressources. Des logiciels particuliers ont été développés pour permettre de bloquer certaines requêtes ou empêcher la visualisation de certains sites.
Des mesures internationales sont en cours de développement comme celles définies par le groupe PICS (voir plus haut) créé à l'initiative du W3C. L'objectif est de développer des moyens de contrôle sur le contenu en vue de sélectionner notamment les ressources accessibles par des enfants.
Le principe repose sur l'établissement d'un " label " mesurant en particulier les niveaux de violence, nudité, sexe et correction de langage, qui serait codé dans les ressources. Les visualiseurs seraient capables d'interpréter ce label et de filtrer les ressources à la demande (notamment celle des parents et des enseignants). Plusieurs labels pourraient être associés à une ressource car en fait plusieurs acteurs pourraient attribuer ces labels : le fournisseur de la ressource mais aussi des services extérieurs ayant un rôle d'évaluation des sites.

II.2 La recherche par l'utilisateur

Les experts de la recherche sont peu nombreux à savoir utiliser toutes les possibilités des outils de recherche, et même s'il existe des « outils cachés » qui font du tri, du classement, ..., dans les moteurs, il reste toujours que les mots posés par l'usager sont l'élément premier et donc essentiel.
Je citerai ici un extrait d'un texte de François Bourdoncle (**3**), qui souligne les limites du grand public face à la recherche sur Internet : «... *l'effort que le grand public est prêt à faire pour obtenir une information pertinente est extrêmement faible. Ainsi, d'après une étude portant sur plus de 1 milliard de requêtes du moteur Altavista, les requêtes consistent en moyenne en 2.35 termes, le plus souvent imprécis et mal orthographiés, et 80% des requêtes ne comportent aucun opérateur. Qui plus est, 85% des utilisateurs se contentent des 10 premiers résultats fournis sur la première page (alors qu'ils ont potentiellement accès aux 200 premiers résultats) et 78% des requêtes ne sont pas modifiées dans le but de les améliorer.* »

II.2.1 Utilisation de mots-clés

Pourquoi, malgré toutes les avancées vues précédemment, celui qui cherche de l'information sur Internet est-il souvent déçu ? (si encore une fois, il n'est pas « expert » dans cette recherche de par sa profession ou sa pratique).
C'est qu'au départ le choix des mots n'est pas innocent. Cette notion de « mot-clé » (utilisée d'abord en documentation et dans l'enseignement du français) nécessite un apprentissage, et l'usager lambda l'ignore, ne réalisant pas toujours qu'un moteur de recherche n'est qu'un programme informatique qui va « comparer » des chaînes de caractères pour ramener celles qui sont identiques aux mots de la recherche, et contenues dans la base de données du logiciel pour lequel il travaille. Il va chercher une <u>identité de forme</u>. Ce qui explique les résultats parfois ahurissants que l'on peut obtenir en réponse à une requête. En outre le moteur va chercher chacun des mots de la recherche, entraînant un « bruit » documentaire (nombre important de réponses obtenues non pertinentes).
On ne peut pas lui préciser : « non, ce que j'ai voulu dire, c'est... » Il faut trouver le ou les mots-clés qui vont apparaître dans le document qui va

correspondre à la recherche (ou qui seront intégrés dans les métadonnées correspondant à la description du contenu intellectuel du document), et qui seront suffisamment peu fréquents pour que le moteur ait indexé peu de pages avec ceux-ci, mais des pages adéquates. Il est important aussi de connaître le réseau sémantique des mots utilisés et cela suppose un vocabulaire riche (on peut même se trouver dans la situation de trouver des mots-clés spécifiques dans un domaine qu'on ne connaît pas encore – puisque le but de la recherche serait justement de découvrir ce domaine -).
Chercher de l'information sur les « voitures » ne doit pas laisser de côté les mots « automobile(s) » et « auto(s) ». A l'inverse, une recherche sur l'animal « jaguar » devra exclure le domaine automobile et le football (équipe du même nom).
L'utilisateur du moteur doit donc indiquer ce qu'il veut associer ou exclure, en tenant compte des options proposées par les moteurs de recherche (celles-ci sont accessibles en général par un bouton « Recherche avancée » sur la page d'accueil), et écrire par défaut ses mots-clés en minuscules et sans accents (à moins de vouloir restreindre la portée de la recherche).

II.2.2 La norme Z39.50

Dans un cadre plus restreint d'interrogations de bases de données bibliographiques (catalogues de bibliothèques, mais aussi toute base d'information documentaire), un outil existe qui permet à l'usager d'interroger des bases de présentations, natures et formats différents sans avoir à tenir compte de ceux-ci lors de son interrogation, et même d'en récupérer les informations bibliographiques obtenues pour les intégrer dans une base de données locale. Il s'agit de la norme Z39.50 qui, s'appuyant sur des normes de communication adaptées à l'environnement Internet, permet d'interroger toute base bibliographique reliée à Internet, à travers un logiciel client Z39.50 (qui traduit les requêtes utilisateurs en Z39.50, lance l'interrogation et décode les résultats reçus en Z39.50) et un logiciel serveur Z39.50 (qui décode les requêtes reçues, les interprète et les exécute sur sa base de données, code en Z39.50 les réponses obtenues et les envoie au client).

L'usager n'a donc pas à s'inquiéter de la façon d'interroger ces bases de données, il lui suffit de poser sa requête dans son format local habituel ; seule limite, et non des moindres, c'est qu'il ne peut obtenir de l'information qu'à partir de bases sur lesquelles sont implantés des serveurs Z39.50. En revanche, la possibilité de récupérer des notices bibliographiques pour les intégrer dans son système local, est très importante, permettant un partage du catalogage sans que ce soit pour autant véritablement du catalogage partagé, puisqu'il n'y a pas obligation de normes communes ni d'échanges de données entre celui qui interroge et celui qui est interrogé.

II.3 Le référencement de sites

Au vu de tout ce qui précède, on peut comprendre le gros souci du mainteneur de site (« webmaster ») qui se trouve face à des outils de recherche qui ont chacun leur mode de fonctionnement, et une catégorie d'internautes (le public visé par son site) qui aura son mode d'interrogation, selon ses connaissances dans la façon d'interroger les outils de recherche et ses centres d'intérêt dans le domaine concerné. Comment signaler son site et comment optimiser ce signalement ?

Il existe heureusement des aides apportant des réponses à ces questions (voir bibliographie **(1)** et **(2)**, et surtout le site d'O. Andrieu http://www.abondance.com, qui est une source d'informations très intéressante sur les outils de recherche, leur fonctionnement, le référencement, etc., qui présente des listes de questions (avec leurs réponses) à propos du positionnement et de la promotion de sites et des outils d'analyse de sites ; on y propose même un audit des pages du site proposé par le visiteur (http://www.abondance.com/audit/code.html). Il existe aussi des grilles d'évaluation de sites proposées aux lecteurs de sites mais qui peuvent tout à fait servir au créateur de pages web, par exemple celui de l'INSA : http://csidoc.insa-lyon.fr/sapristi/fristi36.html) [consulté le 03.03.2000]

Le mainteneur doit donc penser à remplir avec les mots appropriés les « champs » de la page HTML qui seront « visités » par les moteurs de recherche (« title », « description », « keywords »...), en les choisissant judicieusement : vocabulaire utilisé bien sûr, mais aussi singulier et/ou

pluriel, voyelles accentuées ou non, majuscules, ...Il devra faire un résumé de présentation de son site (surtout pour les annuaires), penser à établir des liens vers des sites connus et très visités (les moteurs indexent d'abord les sites les plus visités, s'il existe un lien, le robot l'utilisera pour arriver à la page qui a fait ce lien).
Lorsque les pages sont créées, le mainteneur va les inscrire et donc les référencer dans les moteurs et annuaires les plus connus en suivant la procédure indiquée sur chacun d'entre eux (pour connaître le classement des outils de recherche, là encore, le site d'O. Andrieu donne des tableaux de statistiques actualisés régulièrement : fréquentation, mais aussi pertinence des résultats obtenus (http://www.abondance.com/outils/barometres.html).

III. APPLICATIONS DE CES RECHERCHES AU SITE

Le stage avait pour objectif d'apporter un point de vue documentaire en complément au travail technique fait auparavant lors de la création des pages web du site du CRDP. C'était d'autant plus intéressant que l'ensemble du site étant en cours de mise à jour, les pages de présentation concernant les services documentaires étaient modifiées et ces modifications pouvaient être prises en compte dans leur signalement.
La première étape a donc été celle de la reprise de l'existant (voir Annexe A : « Promotion du site CRDP ») afin de savoir quels types de données avaient été utilisées, quels mots avaient été choisis pour la description globale du site et/ou de chacune des pages, quels moteurs et annuaires avaient servi au référencement. Les observations ont donc été faites sur cet existant, ainsi que les ajouts et/ou modifications ultérieurs.

III.1 Mise en place d'une recherche-type

Les questions à poser dans le cadre d'une simulation de recherche documentaire par le public enseignant devaient prendre en compte deux impératifs :

- on était « dans la peau » d'un enseignant à la recherche de documents concernant son niveau d'enseignement et/ou sa discipline, parfois de textes administratifs au sujet de sa carrière ou des programmes scolaires
- néanmoins il fallait tenir compte du type d'informations signalées sur les sites des centres de documentation pédagogique, et en particulier sur celui du Nord Pas-de-Calais, pour espérer trouver des références à ces sites.

Un enseignant en exercice a besoin de documentation pédagogique utilisable en classe telle quelle ou intégrable dans une préparation de cours. Les sites des centres de documentation pédagogique peuvent proposer des contenus de cours, mais le plus souvent signalent des sites-ressources et/ou des sites-portails ; ils présentent leurs propres services et leurs produits éditoriaux, voulant mettre en avant leurs spécificités auprès de leur public .En revanche, ce dernier cherche des documents pédagogiques, ou des sites spécialisés dans sa discipline ou sur un thème bien précis, mais ne s'intéresse pas directement aux services des centres.

Il fallait donc poser des questions à champ relativement large pour espérer cibler des sites de centres (en ayant fait l'expérience, je peux affirmer qu'une question très précise relative à un programme de cours particulier va ramener des informations récupérables sur des sites créés par des enseignants ou des sites d'académie – rectoraux - , presque jamais sur des sites de centres de documentation pédagogique). Etant en outre moi-même responsable de médiathèque pédagogique dépendant du CRDP du Nord Pas-de-Calais, je voulais également savoir comment les ressources pédagogiques mises à disposition en prêt ou en consultation dans celle-ci pouvaient apparaître en réponse à une recherche documentaire d'enseignant, sachant qu'un catalogue avait été mis en consultation à partir d'une page de présentation du site.

Le travail de référencement qui avait été fait par les mainteneurs du site portait sur une dizaine de moteurs de recherche et d'annuaires ; j'ai donc repris ces mêmes outils afin là aussi de partir de l'existant et du travail fait précédemment (voir leur liste dans l'annexe A : « Promotion du site CRDP »). Le choix des champs à remplir lors de la création des pages et des mots-clés à utiliser comme « descripteurs » avait été fait par S. Prieux et P.

Fernandez selon leurs propres critères (ces mots sont également dans le document cité précédemment).
Je n'ai bien sûr pas tenu compte dans les recherches à effectuer, des termes choisis dans ce cadre, puisque je me plaçais en tant qu'usager lambda et que celui-ci n'est pas censé savoir comment ces pages ont été créées et signalées.
Les mots-clés que j'ai choisis ont été les suivants : CRDP, documentation pédagogique, logiciel(s) éducatif(s) ou pédagogique(s). Le choix a été limité en nombre, en raison du temps passé pour chaque recherche sur une dizaine d'outils de recherche et de l'urgence qu'il y avait à faire ces tests avant que n'ait lieu le changement des pages du site (prévu au départ pour septembre 2000), qui nécessitera de nouvelles créations ou modifications de métadonnées.
Pour le terme « CRDP », j'ai supposé la recherche d'une personne cherchant à localiser ce lieu dans sa région, à en connaître les missions, les services. Cela me permettait également de tester la place que pouvait occuper le site du CRDP du Nord Pas-de-Calais dans une liste de réponses donnée par annuaires et moteurs.
Le terme « documentation pédagogique » permettait de prendre également dans certains outils de recherche le terme « document » sous la forme avec troncature « document* peda* » englobant aussi le terme « pédagogie » (on pourrait ainsi avoir des « documents sur la pédagogie de... ») et regroupant singuliers et pluriels. Cet ensemble de termes pouvait faire partie de ceux employés par un enseignant cherchant un ou des sites contenant ce type d'informations, espérant ensuite cibler sur une discipline, un niveau, etc.
« Logiciel éducatif ou pédagogique » cible un support sur lequel de nombreux enseignants cherchent de l'aide quant à l'utilisation : comment évaluer la valeur pédagogique ou didactique d'un logiciel avant de l'acheter pour l'établissement scolaire ? Ils sont donc à la recherche de fiches descriptives, évaluatives, de ces produits, d'adresses où les consulter, les emprunter, les acheter, de lieux où se font des présentations de certains d'entre eux. J'ai voulu interroger sur les deux adjectifs, pédagogique et éducatif, ceux-ci faisant partie du langage « courant » concernant ce type de support. Ces termes, à l'inverse des précédents, n'avaient pas été introduits

dans les métadonnées du site initial ; cependant, les tester me permettait de savoir si des sites d'autres CRDP pouvaient apparaître et comment.
Les interrogations ont été faites sans accents, sans majuscules pour éviter d'éventuelles restrictions de recherche. Sur chaque annuaire et moteur, j'ai vérifié la prise ou non de la troncature, l'espace entre les mots considéré ou non comme une expression d'addition, la prise implicite du pluriel, et j'ai tenu compte pour chaque recherche de ces limites.
Le tableau suivant reprend les requêtes posées, avec d'éventuelles adaptations, et les résultats obtenus (le site est toujours localisé à l'époque à l'adresse suivante : http://www2.ac-lille.fr/crdp).

	crdp	document* pedago*	logiciel* educ* pedago*
Altavista.fr Altavista.com	crdp : *rien* url:crdp : *pas celui du Nord Pas-de-Calais* crdp : *rien* url :crdp : *n°169 (17e page), une page du site*	+document* +pedago* langue = français *bien ciblé, le CNDP, des sites de CRDP entre autres sortent, mais pas celui du Nord Pas-de-Calais*	logiciel* AND (educat* OR pedago*) *1 CRDP apparaît en 91e position, mais pas celui du NPdC*
Ecila	crdp : *n° 23 (3e page), donne accès à la toute 1ère adresse du site, d'où renvoi à la seconde, puis à la 3e actuelle*	documentation pedagogique *rien*	logiciels educatifs *rien* logiciels pedagogiques *rien*
Voila	crdp : *n°3 (1ère page), page d'accueil du site*	*n°3 (2e page), cité par l'intermédiaire d'un site portail* *n°7 et 8 (même page), deux pages du site (« Guide des sources... » et accueil)*	*une page de CRDP est donnée, pour les liens qu'elle fait sur d'autres pages ; pas le NPdC*
HotBot	crdp : *n°8 (1ère page), page d'accueil du site* *n°19 (2e page), donne accès à la 1ère adresse... (voir Ecila)* *n°23 (3e page), adresse de redirection à partir de la 2e adresse*	*le CNDP et des CRDP sont cités dès la 1e page, mais pas celui du Nord Pas de Calais*	*même remarque que pour Voila*
Infoseek	crdp : *n°3 (1ère page), page de redirection (2e adresse)* *n°4 (1ère page), une page du site [sommaire Nouvelles technologies]*	*même remarque que pour HotBot, puis le site du NPdC est cité à travers les liens du site Educasup en n° 23* *n° 28, page des nouvelles technologies* *n° 32, page d'accueil du site*	*des pages de CRDP et CDDP apparaissent, pas pour le NPdC*

Lokace	crdp : *rien*	*n°11, sur un site personnel, le site du CRDP NPdC est cité dans une liste de CRDP*	*rien*
Yahoo	crdp : *n°6 (2e page)* *n°11 (3e page)* *rangé dans des catégories*	*rangement en catégories, même classement que pour* crdp	*rien*
Lycos	crdp : *20e page, une page du site [russe]*	*des sites de CRDP sont cités, pas celui du NPdC*	*le CNDP apparaît, 2 sites de CRDP, des sites académiques apparaissent, pas le NPdC*
Nomade	crdp : *n°2 (1ère page), adresse du 1er site, d'où 2 redirections pour arriver sur l'actuel, le descriptif est à réactualiser*	*même remarque que pour HotBot*	*dans une catégorie, 60 sites apparaissent, le CRDP du NPdC apparaît sur la 2e page*
Francité	crdp : *n°1, adresse précédente* *n°32, page accueil du site*	*même remarque que précédemment*	*des pages de CRDP, du CNDP, pas le NPdC*
WebCrawler		*rien*	*rien*

Interrogations faites entre le 15 et le 25 juin 2000

III.2 Lancement des recherches, évaluation des résultats

Pour la lecture des réponses, je suis allée souvent très loin dans celle du nombre de pages proposées (souvent même jusqu'à la fin) pour savoir si le site apparaissait, tout en sachant que nombre d'utilisateurs n'auraient pas cette démarche puisque la plupart ne lisent que les premières pages de résultats (voir page 20 à propos du comportement des utilisateurs), parfois la dernière en plus.

Le seul mot « crdp », qui est celui sous lequel on s'attend certainement à voir apparaître le site du Nord Pas-de-Calais, montre beaucoup de manques, de nombreuses redirections d'anciennes adresses en plus récentes, sans toujours tomber sur la plus actuelle. Il faut savoir que peu de ces outils proposent au « référenceur » d'un site web de faire l'annulation, ou simplement la modification d'une adresse antérieure ; ce qui oblige à faire un nouveau référencement, en laissant de côté ces corrections. Situation toujours délicate car rien ne dit qu'un usager tombant sur une première adresse « à renvoi » ira jusqu'à la troisième, comme c'est le cas ici.

Globalement, l'ensemble des mots choisis fait apparaître des réponses pertinentes, et les sites de centres de documentation pédagogique sont inclus (CNDP, CRDP et CDDP) ; néanmoins, sur une recherche plus précise (les logiciels), les centres n'apparaissent presque plus, on voit en revanche s'afficher beaucoup de sites d'éditeurs (les centres le sont pourtant aussi) et des sites académiques, sur lesquels des enseignants présentent des produits pédagogiques et des fiches-ressources aidant à la préparation des cours.

Sur les annuaires, tels Yahoo ou Nomade, le site apparaît dans une liste alphabétique, dans une catégorie dans laquelle il a été rangé (les catégories sont proposées au référenceur du site, au nombre de deux maximum): en fonction de la catégorie dans laquelle il se trouve, il apparaîtra donc ou non selon la recherche.

Au vu des réponses, il me semblait donc indispensable de faire un nouveau référencement du site sur tous les outils de recherche choisis, d'ajouter des métadonnées à celles d'origine afin d'essayer de couvrir l'ensemble des informations accessibles sur le site. A la lecture d'Olivier Andrieu (**2**), il aurait

été intéressant également de modifier l'URL afin que les moteurs qui interrogent sur celle-ci trouvent le mot accessible immédiatement (et non pas en deuxième position après la barre oblique comme à ce moment).

III.3 Propositions de modifications

Les pages sur lesquelles ont porté les modifications ont été en nombre restreint afin de pouvoir aller vite dans la relance du référencement (nous étions en juin, les vacances du personnel approchaient, et on avait parlé de bascule vers le nouveau site avant celles-ci...). Celles qui ont été choisies : la page générale d'accueil du site, la page d'accueil des librairies, la page d'accès aux catalogues des médiathèques et la page de présentation des TPE (travaux personnels encadrés, mis en place cette année dans les classes de première des lycées d'enseignement général) [voir Annexe B], l'ont été par moi en fonction de la proximité qu'elles avaient avec le public fréquentant les médiathèques et leurs demandes, la page d'accueil étant, elle, la porte d'entrée pour tous, sur laquelle il fallait mettre le maximum d'informations . Les modifications n'ont concerné que les métadonnées, absolument pas la partie concernant la mise en forme des pages ; c'est donc la partie comprise entre les balises <head> et </head> de chaque page concernée qui a été retravaillée.

Pour la balise <title> et la métadonnée « description », j'ai voulu qu'apparaissent à la fois la localisation et le descriptif général ciblant l'offre du site (pour cela, j'ai tenu compte de la façon dont d'autres centres de documentation se présentaient lorsqu'ils étaient répertoriés par les moteurs de recherche), tout en employant peu de mots afin que la lecture puisse être rapide et explicite. En mots-clés (« keywords »), j'ai utilisé des mots décrivant les ressources du réseau du CRDP Nord Pas-de-Calais, relevant à la fois du vocabulaire courant utilisé par les enseignants et de véritables mots-clés d'indexation inclus dans un thésaurus (ici, il s'agit de Motbis, thésaurus utilisé par les centres de documentation de l'éducation nationale : « didactique des disciplines », « documentation administrative », « CDI » en font partie ; « TPE », « RIP », « TICE » sont des abréviations utilisées fréquemment dans le monde enseignant) ; les mots utilisés (singulier ou pluriel) permettent des

croisements dans les interrogations : la recherche « ressources multimédias » touchera « ressources pédagogiques » et « documents multimédias » ou « animations multimédias »... Certaines expressions sont indiquées au singulier puis au pluriel, tenant compte en cela des moteurs qui ne prennent qu'une des formes ; cependant, seuls certains mots ont été concernés afin d'éviter un excès de mots-clés qui pouvait devenir pénalisant par rapport aux moteurs de recherche.

J'ai tenu compte également de la norme du Dublin Core en ajoutant des métadonnées DC (à l'heure actuelle, ce sont des sites de bibliothèques qui ont commencé à se référencer ainsi) pour ajouter des possibilités supplémentaires de repérage par les moteurs qui en tiendront compte.

Pour les pages de présentation intérieures au site (comme celle des librairies ou des médiathèques), j'ai repris certains mots-clés utilisés (comme « didactique des disciplines » ou « documents pédagogiques »), mais j'ai aussi ciblé plus finement le public concerné par la présentation : ainsi apparaissent des intitulés de disciplines, vente et achat pour les librairies, des natures de documents et des services pour les médiathèques.

Les moteurs et annuaires sur lesquels porte le nouveau référencement sont restés les mêmes, j'ai simplement fait le choix des versions françaises lorsqu'il y en avait (Altavista, Yahoo). De ce fait, j'ai également indiqué dans une balise « language » des pages HTML la langue utilisée (cette balise est utilisée par des moteurs comme Altavista lorsque la requête de l'utilisateur demande des réponses dans une langue)

III.4 Mise à jour du référencement et résultats

La mise à jour du référencement a été commencée mi-juin 2000, en envoyant une page par jour aux moteurs de recherche (donc sur quatre jours). Pour les annuaires, un nouveau signalement a été fait. Sur Yahoo, il a été possible de faire une modification de description du site pour pouvoir être réenregistré (sans cela, il n'y avait pas de possibilité de changement de catégorie). Sur Nomade, le site était déclaré comme déjà référencé, pas de possibilité de changement.

En consultant les pages d'enregistrement de sites sur chacun des moteurs, on trouve des informations intéressantes sur le référencement ; c'est en lisant celles de Lycos que j'ai appris qu'une balise « robots » pouvait être ajoutée dans la page HTML, signalant aux moteurs que la page en question devait ou non être indexée par eux et qu'il y avait ou non des liens à suivre à l'intérieur de cette page (voir également O. Andrieu, op. cit.). Il est possible de le faire également par l'intermédiaire d'un fichier *robots.txt* créé sur le serveur ; mais cela n'étant pas dans mes attributions, j'ai utilisé la méthode précédente.

J'ai donc modifié les pages en conséquence et demandé à ce que ces pages modifiées soient envoyées sur le site. Sur ces entrefaites, l'hébergeur du site du CRDP (l'académie de Lille) a proposé d'installer celui-ci sur une machine dédiée à lui seul, ce qui voulait dire une adresse propre. Après acceptation, celui-ci a été mis en place et la bascule s'est effectuée le 27 juin 2000, avec pour adresse désormais http://crdp.ac-lille.fr.

Il n'y avait plus qu'à recommencer le référencement, ce qui a été repris dès le lendemain. Il est à noter que certains moteurs indiquent un délai de 48 h pour ajouter la page dans leurs références (Altavista) tout en précisant que leur moteur viendra ensuite indexer les pages (ce qui peut prendre « un certain temps »...), d'autres deux semaines (WebCrawler, Nomade), six semaines (Ecila) pour cette indexation, soixante jours pour HotBot, certains localisent immédiatement la page dès son signalement (Lycos). Le référencement du nouveau site et de certaines de ses pages a été terminé au début du mois de juillet, les recherches ont été reprises au début du mois de septembre en raison des vacances d'été et des délais de mise à jour des moteurs et annuaires.

Les résultats sont rassemblés sur le tableau des pages suivantes.

	crdp	document* pedago*	logiciel* educ* pedago*
Altavista.fr Altavista.com	url:crdp : *rien* crdp : *n° 44 (5e page), seconde ancienne adresse, pas de lien* *n° 51 (6e page), présentation par un autre site (IUFM)* *n° 55 (même page), avant-dernière adresse amenant sur une page ciblée du site, mais message d'erreur* url :crdp : *rien* crdp : *rien*	*pas le site du Nord Pas-de-Calais*	même requête que lors de la 1ère interrogation, langue = français: *pas le site du NPdC (un site de CR apparaît à la 4e page, puis un autre à la 12e, ...)* *pas de sites du réseau, mais la recherche amène des réponses ciblées correctement*
Ecila	crdp : *rien* recherche dans l'URL : *n° 9 (1ère page), une page du site [sur le Kent]* description:crdp : *n°1, 2, 3, 4 et 5 (1ère page), page d'accueil du site*	*non*	*non*
Voila	crdp *ou* url:crdp : *n°3 (1ère page)*	documentation pedagogique : *n° 10 (1ère page), page d'accueil, bonne adresse*	logiciel pedagogique, logiciels pedagogiques, logiciel educatif, logiciels educatifs: *non*
HotBot	crdp : *n°4 (1ère page), avec adresse précédente n°12 (2e page), page de redirection vers l'adresse précédente*	documentation pedagogique, document pedagogique (prend le pluriel par défaut) : *le CNDP, des CR, des CD, pas celui-ci*	logiciel pedagogique, logiciels pedagogiques, logiciel educatif, logiciels educatifs: *rien*

Infoseek	crdp : *n°3 (1ère page), page de redirection [comme la précédente]* *n°4 (1ère page), avec adresse précédente* *n°27 (3e page), page d'accueil du site actuel*	documentation pedagogique, document pedagogique, documents pedagogiques : *rien sur les sites du réseau*	logiciel educatif, logiciels educatifs, logiciel pedagogique, logiciels pedagogiques : *rien*
Lokace	crdp : *rien [arrêt à la 13e p., ce sont presque toujours les pages des mêmes sites qui reviennent]*	documentation pedagogique, document pedagogique (prend le pluriel par défaut) : *pas le site du N PdC*	logiciel pedagogique, logiciel educatif (pluriel par défaut) : *rien*
Yahoo	crdp : *n° 3 (2e page)* *n°6 (3e page)*	document* pedago* : *dans une catégorie, sur une liste alphabétique, n° 6 (1ère page)*	logiciel* educ*, logiciel* peda* : *non*
Lycos	crdp (avec le choix : web français) : *n° 2 (1ère page), une page du site [italien], lui-même indiqué comme recommandé* *n° 21 (3e page), une page du site [anglais]*	documentation pedagogique, document pedagogique (pluriel par défaut) : *non*	logiciel pedagogique, logiciel educatif (pluriel par défaut) : *rien*
Nomade	crdp : *n°1 (1ère page) avec la bonne adresse*	document pedagogique (espace = ET implicite) : *n°42 (5e page), on trouve beaucoup de sites CN/CR/CD)* document +pedagogique : *n°35 (4e page), peu de sites de CR cette fois*	logiciel pedagogique, logiciel educatif (pluriel par défaut) : *rien*
Francité	crdp : *n° 1 (1ère page), page d'accueil adresse précédente* *n°3 (même page), même chose* *n°4 (même page), page d'accueil avec la bonne adresse*	documentation pedagogique (web français) : *n° 7 (1ère page)*	logiciels educatifs : *rien* logiciels pedagogiques : *n° 18 (2e page), accueil adresse précédente* *n° 19 (même page), accueil, bonne adresse*
WebCrawler	crdp : *rien [uniquement des pages de sites canadiens sur le droit : Centre Ressources de Droit Public]*	documentation pedagogique, document pedagogique, documents pedagogiques : *rien*	logiciels educatifs, logiciels pedagogiques, logiciel educatif, logiciel pedagogique : *rien*

Interrogations faites entre le 30 août et le 15 septembre 2000

On se rend compte que pour la recherche sur les logiciels, on obtient toujours des résultats négatifs concernant en particulier le site (on s'aperçoit de plus qu'avec certains moteurs, des sites académiques, celui du CNDP et certains sites de CRDP qui apparaissaient au mois de juin n'y sont plus maintenant. On peut se demander ce qui a pu modifier leur repérage...) ; ceci s'explique par le fait que peu de moteurs utilisent les mots-clés « keywords » pour répondre aux requêtes, ceux concernant les logiciels sont en outre enfouis assez loin dans la page HTML d'accueil du site. Il faudrait en fait créer un document spécifique concernant les logiciels pédagogiques mis en vente par le CRDP par exemple ou présentés en animations, derrière lequel on placerait une page HTML avec des métadonnées particulières à ce document. Seul Francité répond à la demande sur « logiciels pédagogiques » en citant le site du CRDP dans sa deuxième page de références.

Les adresses fausses sont maintenant en nombre limité : HotBot, Francité ont encore l'adresse précédente (qui correspond au référencement de juin) mais ont néanmoins une référence ensuite à la nouvelle adresse ; Infoseek a gardé les références d'adresses précédentes, mais y a ajouté la nouvelle ; seul Altavista amène deux mauvaises adresses, dont une correspondant à celle du mois de juin, ce qui pourrait se comprendre en raison du référencement de cette période, malheureusement, la page citée amène un message d'erreur si on clique dessus.

Les annuaires ont pris en compte le changement dans les catégories, la place du site dans les réponses s'est améliorée et l'adresse a bien été modifiée.

Si la recherche par « crdp » donne plus de réponses que précédemment avec le site du Nord Pas-de-Calais et en meilleure place, celle avec « document* pedago* » reste encore faible pour ce même site ; en fait seuls les annuaires le placent bien dans leur liste, alors que d'autres sites de CRDP sont en plus référencés par les moteurs pour cette requête.

Il est évident que vérifier régulièrement le référencement du site est un élément essentiel ; pour les annuaires, cela doit être en principe juste « par acquis de conscience » ou pour mettre à jour un descriptif qui s'avèrerait inadapté ou incomplet. Les moteurs, de par leurs fonctionnements différents,

ne permettent pas de mettre en place une stratégie claire. On voit par Ecila que la recherche dans l'URL et la balise « description » est très positif pour la position du site dans la liste des réponses ; il faut simplement que l'usager clique sur la proposition qui est faite en haut de la page des réponses : chercher « ... » dans les URL, dans la description. Y en a-t-il beaucoup qui chercheront à faire cette démarche alors qu'ils ont déjà une liste devant les yeux avec des sites répondant à leur demande ?

On s'aperçoit également qu'il faut essayer de contourner le problème des adresses modifiées, ce qui n'est pas évident avec des outils de recherche qui n'acceptent que les soumissions de nouveaux sites.

En faisant analyser la page d'accueil du site par l'outil proposé sur le site abondance.com, j'ai pu constater que le titre pourrait être repris en incluant d'autres mots significatifs, puisque le nombre de mots utilisés n'atteint pas le maximum préconisé pour une bonne lecture ; même chose pour la balise « description ». Le début du texte du corps de la page (ce qui correspond au « texte visible ») n'est absolument pas significatif, or c'est un endroit particulièrement important pour certains moteurs qui l'utilisent pour indexer la page ; il serait donc à revoir pour en soigner le choix des mots significatifs.

J'ai proposé aux concepteurs du site de référencer d'autres pages de celui-ci, en ajoutant auparavant des métadonnées également. Je pense qu'il faudrait reprendre celles qui ont été écrites avec des caractères accentués pour les écrire également sans accents, et transformer les majuscules en minuscules, supprimer peut-être les virgules pour ne laisser que les espaces de séparation entre les mots, rajouter des mots-clés ciblant des éléments plus précis ou omis précédemment (audiovisuel, malvoyants, ...). Il serait également possible de mettre un moteur de recherche sur le site : le CNDP en a installé un sur son propre site (Spinoo) et a proposé aux sites du réseau intéressés de faire un lien entre le site et le moteur pour qu'il puisse être utilisé en local. Ce moteur serait un moyen supplémentaire donné à l'usager lorsqu'il est sur le site, mais n'aurait évidemment pas d'influence sur la recherche initiale.

IV. CONCLUSION

Il est certain que l'on reste avec de nombreuses questions à l'issue de ce stage ; il est en effet difficile de savoir à quel point les modifications apportées aux pages HTML du site ont été ou non prises en compte par les moteurs (le positionnement dans les annuaires restant sensiblement le même puisqu'il s'agit d'un classement thématique), ce d'autant plus que ces mêmes moteurs mettent beaucoup de temps à prendre en compte les données apportées (on nous a parlé de plusieurs mois pour Altavista !), et si oui, ce qui a pu entrer en ligne de compte dans la modification du classement du site. On se rend compte à quel point il est important à la fois de connaître le fonctionnement des outils de recherche et la façon dont l'information doit être agencée pour pouvoir être « ramassée » par eux.

Peut-être émerge là une « nouvelle » profession, celle de « référenceur », dans le monde de l'information et de la documentation. Profession à multiples facettes, où l'on trouve à la fois une face commerciale (être le mieux placé dans une liste de résultats de recherche), et une face documentaire (Comment mon public va-t-il chercher l'information ? Quels termes ont pu être choisis par les créateurs d'un document pour que j'arrive à le retrouver ?). Le fait de référencer n'est pas totalement nouveau : les documentalistes, les bibliothécaires ont toujours utilisé ou créé des références, pour signaler ou savoir que quelque chose existe quelque part. Ce qui est plus nouveau, c'est l'outil sur lequel il se réalise (Internet) et l'abondance de documents existants. Se référencer devient alors un moyen de se situer autrement que les autres. Il est intéressant de consulter des discussions qui ont eu lieu à propos de ce métier sur la liste de diffusion de l'ADBS (Association des documentalistes et bibliothécaires spécialisés) : adbs-info@cru.fr, messages des 5 juin, 30 juin et 2 juillet 2000. On voit que savoir référencer est extrêmement important sur le réseau Internet. Et on passe du mythe du savoir universel, de l'information à portée de tous, de la « liberté » totale à une certaine prise en charge par des professionnels, parce que le grand public ignore toutes les subtilités de la recherche de documents et se retrouve souvent confronté au « tout ou rien », sans la connaissance de certaines des règles qui la sous-tendent.

Ce travail sur le référencement a également mis l'accent sur les métadonnées : des normes se sont mises en place autour de leur construction (Dublin Core, XML, ...). Cependant reste un problème, celui du langage utilisé : à l'heure actuelle, chaque concepteur de page web, lorsqu'il crée les métadonnées « keywords » en particulier, utilise un langage « naturel », le sien, celui de l'environnement qu'il connaît, celui de son public-cible (on retrouve là le souci du référenceur déjà évoqué). Mais ce langage dit naturel est-il vraiment commun à tous, chacun utilise-t-il le même vocabulaire pour décrire la même chose ? Tout catalogueur de notice bibliographique s'est retrouvé un jour confronté à la difficulté de trouver les bons mots-clés (d'indexation documentaire cette fois) pour décrire le contenu d'un document, en fonction du public à qui il est destiné ; et l'on retrouve cette même difficulté lorsqu'il faut trouver les mots décrivant le contenu d'une page ou d'un site web : les mots choisis rencontreront-ils la recherche d'un usager ? Faudrait-il aller jusqu'à la mise en place d'un vocabulaire normalisé, l'utilisation de thésaurus ?

L'utilisation d'Internet, nouvel outil, laissant ouvertes encore de nombreuses questions, ramène toujours cependant à une notion essentielle, familière à nombre de bibliothécaires et de documentalistes : le souci du public, de l'usager, utilisateur final du document qui lui est destiné...

BIBLIOGRAPHIE

1. ANDRIEU, Olivier. *Trouver l'info sur l'Internet.* Paris : Eyrolles, 1999. 425 p. ISBN 2-212-08992-9.

2. ANDRIEU, Olivier. *Créer du trafic sur son site Web : conception, référencement et promotion.* Paris : Eyrolles, 1999. 464 p. ISBN 2-212-09062-5.

3. La bibliothèque en ligne: actes de la journée d'étude du 21 juin 1996. *Bulletin d'informations de l'Association des bibliothécaires français*, 1er trimestre 1997, n° 174, p. 7-174.

4. BOURDONCLE, François. *Panorama et perspectives des outils de recherche d'information textuelle sur Internet = Searching the Internet : state-of-the-art and perspectives.* [en ligne]. Paris: Ecole nationale supérieure des mines, 1999.[Consulté le 10.05.2000] Disponible sur:
<URL : http://www.cma.ensmp.fr/François.Bourdoncle/idt99.htm>

5. CERISIER, Jean-François. Pour une pratique en réseau de la recherche de l'information...: accès au document et recherche collaborative. *Dossiers de l'ingénierie éducative*, octobre 1999, n° 29, p. 10-13.

6. CZAPLINSKI, Jean-Marc, MORET, Yves. Z39.50. *Bulletin d'informations de l'Association des bibliothécaires français*, 1er trimestre 1997, n° 174, p. 77-81.

7. LAHARY, Dominique. Le jeu de puzzle de l'accès aux catalogues: World Wide Web et/ou Z39.50. *Bulletin d'informations de l'Association des bibliothécaires français*, 1er trimestre 1997, n° 174, p. 82-85.

8. LARDY, Jean Pierre. RISI, recherche d'information sur l'Internet : outils et méthodes [en ligne]. Paris: ADBS, 2000. [Mise à jour le 05.02.2000] Disponible sur:
<URL : http://www.adbs.fr/adbs/sitespro/lardy/risi.htm>

9. LARHER, Tanguy, MILBEAU, Katia. Les métadonnées *Dossiers de l'ingénierie éducative*, octobre 1999, n° 29, p. 37-39.

10. LUPOVICI, Catherine. Identification des ressources sur Internet et métadonnées: diversité des standards. *Documentaliste – Sciences de l'information,* 1999, vol. 36, n° 6, p. 321-325.

11. Optimisez vos recherches sur Internet [dossier]. *Archimag,* juin 2000, n° 135, p. 28-30.

12. Recherche intelligente sur l'Internet. *Pour la science*, août 1999, n° 262, 6 p. et [en ligne]
URL : http://www2.pourlascience.com

13. ROBERT, Florence. Cataloguer le Web. *Arabesques,* juillet-août-sept. 2000, n° 19, p. 6-7.

14. SAMIER, Henry, SANDOVAL, Victor. *La recherche intelligente sur l'internet et l'intranet : outils et méthodes.* 2e éd. revue et augm. Paris : Hermès Science Publications, 1999. ISBN 2-7462-0070-8.

15. VERCOUSTRE, Anne-Marie (trad.). *Eléments de métadonnées du Dublin Core, version 1.1 : description de référence* [en ligne]. Le Chesnay: INRIA Rocquencourt, 2000. [consulté le 08/06/2000] Disponible sur: <URL : http://www-rocq.inria.fr/~vercoust/METADATA/DC-fr.1.1.htm>

Et quelques adresses intéressantes à consulter en ligne

Le site d'Olivier Andrieu
URL : http://www.abondance.com

La Lettre du bibliothécaire québécois, mine d'informations pour les bibliothécaires et documentalistes, à laquelle il est possible de s'abonner gratuitement
URL : http://www.sciencepresse.qc.ca/lbq/lbq.html

ANNEXE A

« PROMOTION DU SITE CRDP »

Promotion du site CRDP

Il existe trois types de référencement possibles : le référencement par les moteurs de recherche, le référencement sur les répertoires ou bases de données hiérarchiques, et le référencement sur les sites de nouveautés.

Comment signaler le changement d'adresse de mon site ?

Enregistrez vous à nouveau tout simplement. Les robots détecteront automatiquement que votre ancien site n'est plus accessible et supprimeront les anciennes références.

Promotion sur les moteurs de recherche

Alta Vista

Apprécié mondialement - Très utilisé

Il détermine le placement des pages en fonction des mots et des phrases : les mots-clés doivent apparaître dans les premiers mots du document - ils doivent être proches les uns des autres dans le document - le document doit contenir le maximum des mots demandés dans la recherche.

Pour référencer le site auprès d'Alta Vista, il suffit de communiquer l'URL de la première page, le moteur se chargeant de l'exploration des autres pages liées. Le moteur fait référence à des pages sur Nordnet. Ces pages sont à supprimer du serveur Nordnet. Référencement fait le **18 Mai 1998.**

Ecila

Moteur de recherche francophone.Il faut lui communiquer l'URL du site et l'adresse E-mail du webmaster pour une visite rapide et régulière par le moteur.

Le référencement a été effectué le **18 Mai 1998** pour les trois pages du site CRDP : **page d'accueil générale, page sommaire du site Italien, page d'accueil du guide des sources d'information**. Le texte qui sera normalement affiché est celui des méta-tags

Echo

Moteur de recherche dans l'univers francophone

Il faut lui communiquer l'URL du site et l'adresse E-mail du webmaster pour une visite rapide et régulière par le moteur. La recherche s'effectue par Les pages Web et QuiQuoiOu comme pour Infospace.

Hot-bot

C'est une importante base de données. La sélection se fait par la répétition de mots clés - leur utilisation dans la page-titre, leur utilisation dans les méta-tags. La seule possibilité offerte par le moteur est la soumission d'URL, sans commentaire particulier. Les URL des 3 pages principales ont donc été soumises le **19 Mai 1998**. Elles devraient être indexées dans la base de données 48 hs après. Vérifier dans quelle catégorie.

Info-seek

Moteur connu. Les trois pages principales ont été référencées le **19 Mai 1998** sans pouvoir choisir de catégorie. Dans ses résultats de recherche, le moteur affiche cependant encore des liens vers des pages www.ac-lille.fr...., pages qui ne sont plus valides. Il semble donc qu'une mise à jour des adresses mortes soit peu fréquente. On peut par contre communiquer les adresses détruites. A essayer quand le serveur sera plus rapide

Lokace

Il faut lui communiquer l'URL du site et l'adresse E-mail du webmaster pour une visite rapide et régulière par le moteur, mais aussi la catégorie où il doit figurer ainsi qu'un commentaire.

Lycos

Connu et largement utilisé. Il faut inclure les mots-clés dans le titre puis les répéter tôt dans le début de la page. Le référencement de la page Sommaire général a été effectué le **19 Mai 1998** dans la catégorie Education. Il reste à inclure les 2 autres pages.

Webcrawler

Son classement est issu du nombre de mots-clés par page sur le nombre total de mots du document.

Infospace

Moteur de recherche doublé d'un répertoire thématique : QuiQuoiOu. Il faut lui communiquer l'URL du site et l'adresse E-mail du webmaster pour une visite rapide et régulière par le moteur, mais aussi la catégorie où il doit figurer ainsi qu'un commentaire.

Le référencement général a été effectué le **18Mai 1998**.

- Le site CRDP a été de plus proposé dans la catégorie **Enseignement > Services pour l'enseignement > Bibliothèques, centres de documentation**
- Le site Italien a été proposé dans la catégorie **Education > Formateurs, professeurs** et **la catégorie Enseignement > Séjours linguistiques**
- Le guide de ressources a été proposé dans la catégorie **Education > Services pour l'enseignement > Outils pédagogiques** et **Sciences, technologies et recherche > Sciences de l'éducation**

Les descriptifs ci-dessous ont été fournis.

Les répertoires ou bases de données hiérarchiques

Nomade

Il est considéré comme le must des outils de recherche francophone. Deux modes de navigation sont proposés : par mot clé et par navigation dans une arborescence thématique.

Les trois pages principales ont été référencées le **18 Mai 1998**, chacune dans trois catégories différentes (maximum autorisé) (cf copies d'écran). La saisie des mots clés ayant échoué techniquement, des modifications devront être apportées ultérieurement.

Yahoo et Yahoo France

Francité

Moteur de recherche de la francophonie. Pour inclure une référence, on parcourt la base de données à la recherche d'une sous-catégorie adaptée. La catégorie retenue est : **Education > Outils pédagogiques** pour le référencement de la page Sommaire général, fait le 19 Mai 1998. Référencer les autres pages

Méta-tags

Méta-tags utilisés pour le site Italien

```
<title>Le site italien du CRDP Nord Pas-de-Calais</title>

<meta Name="description" Content="Un site conçu par des
professeurs d'italien pour les enseignants, les élèves,

les parents et tous ceux qui s'intéressent à l'Italie">

<meta Name="keywords" Content="CRDP,Centre Régional de
Documentation Pédagogique,italie,italien,langues,enseignants
italien,

professeurs italien"
```

Méta-tags utilisés pour le site CRDP

```
<title> Centre Régional de Documentation Pédagogique - CRDP Nord
Pas-de-Calais </title>

<meta Name="description" Content="Un site institutionnel :
présentation du CRDP Nord Pas-de-Calais, ses services, ses
missions">

<meta Name="keywords" Content="CRDP,Centre Régional de
Documentation Pédagogique,éducation,enseignement,

documentation
pédagogique,médiathèques,enseignants,documentalistes,pédagogie">
```

Méta-tags utilisés pour le Guide des sources d'information

```
<meta Name="description" Content="Un guide documentaire pratique
des sources en ligne d'information en éducation pour l'enseignant,
le documentaliste, l'étudiant, le chercheur en éducation">

<meta Name="keywords" Content="CRDP,Centre Régional de
Documentation Pédagogique,éducation,sciences de l éducation,

documentation
pédagogique,médiathèques,enseignants,documentalistes,pédagogie,doc
umentation,guide,étudiant">
```

Descriptifs de sites - commentaires

Le site global CRDP

http://www2.ac-lille.fr/crdp/default.htm

Site institutionnel du CRDP Nord Pas-de-Calais. Présentation du CRDP, ses missions, ses services, ses animations. Accès en ligne au catalogue des médiathèques du réseau. Serveur d'informations pédagogiques pour la communauté éducative (enseignants et documentalistes)

Le site italien

http://www2.ac-lille.fr/crdp/site-italien/sommaire-italien.htm

Site conçu par des professeurs d'italien pour les enseignants, les élèves, les parents, et tous ceux qui s'intéressent à l'Italie. Photos, adresses, programmes, sujets d'examen, bibliographies.

Le site Guide des ressources en éducation

http://www2.ac-lille.fr/crdpguide/default.htm

Guide documentaire pratique des sources d'information en éducation pour les enseignants, les documentalistes, les étudiants, les chercheurs en éducation. Edité par le CRDP Nord Pas-de-Calais, ce guide en ligne est mis à jour en permanence.

Mots-clés

Mots-clés CRDP

Education, enseignement, documentation pédagogique, documentation, médiathèques, enseignants, documentalistes, BCDI, pédagogie, braille, nouvelles technologies, nte, ntic, nord, pas-de-calais, audiovisuel

Mots-clés Italien

Italie, italien, langues, enseignants italien, professeurs italien, apprendre italien, CRDP, Nord, Pas-de-Calais

Mots-clés Guide

Guide, documentation, information, éducation, enseignants, pédagogie, documentalistes, chercheurs éducation, adresses sites, sciences de l'éducation, annuaire de sites, répertoire de sites, CRDP, Nord, Pas-de-Calais

Promotion d'une page (exemple : la page Braille)

Pour permettre à une page donnée du site d'apparaître en haut des résultats de recherche on peut la référencer directement auprès de quelques moteurs de recherche connus. C'est ce qui a été fait pour les 3 pages **Sommaire général**, **Sommaire du site italien** et **Sommaire du guide de ressources**.

On peut aussi utiliser d'autres astuces directement dans la conception de la page. C'est ce qui a été retenu pour la page de présentation du service Braille . Par exemple :

- Placer des mots clés dans le titre de la page (balise Title)
- Placer le maximum de mots clés dans le corps du texte (le mot braille a été rajouté à plusieurs endroits)
- On ajoute des méta-tags dans le code source HTML :

```
<meta Name="description"

Content="Le CRDP Nord Pas-de-Calais assure depuis 1982, au niveau
national, une fonction spécifique : la conception et l'édition de
livres et de documents pour les élèves aveugles et malvoyants.">

<meta Name="keywords"

Content="braille,aveugles,malvoyants,CRDP,Centre Régional de
Documentation Pédagogique,éducation,enseignement,documentation
pédagogique,pédagogie,">
```

Il reste à vérifier dans quelque temps l'effet produit dans une recherche par mot-clé.

Marqueurs de fréquentation

Les moteurs de recherche proposent des outils de mesure de fréquentation à inclure sur les pages du site. Nous en avons utilisé deux le plus discrètement possible.

E-Stat

E-Stat (proposée par la société Echo) est un outil de mesure del'audience d'un site Web. Un contrat gratuit avec E-Stat permet de placer 10 marqueurs personnalisés sur 10 pages différentes du site. Le choix s'est porté sur les pages :

- Sommaire général
- Sommaire du site italien
- Sommaire de la branche Réseau
- Sommaire de la branche Animations
- Sommaire de la branche Services
- Sommaire de la branche Catalogue des médiathèques
- Sommaire de la branche Ressources pour documentalistes
- Sommaire de la branche NTE
- Sommaire des sujets d'examen
- Page d'accueil

Les statistiques sont consultables directement depuis le site.

Francité

Ce marqueur peut être utilisé sur autant de pages qu'on le souhaite mais les statistiques proposées indiquent le n° IP d'une machine et nom son identifiant explicite. (Il faut trouver le logiciel WSPING32 pour effectuer les conversions)

Il a été inclus sur certaines pages du guide :

- La page d'accueil
- La partie 1
- La partie 2
- La partie 3
- La page de commande

Les statistiques sont consultables à l'adresse : `http://francite.com/compteur.html` (référence 18300 sans mot de passe)

Interdire des pages aux robots

Si **vous ne voulez pas** que les robots des **Pages Web** ou d'autres robots visitent votre site, ou bien si votre site a déjà été visité et si vous désirez qu'il soit retiré de la base d'URLs, **vous devez placer** un fichier `/robots.txt` à la **racine** de votre site.

Le fichier `/robots.txt` vous permet de spécifier par qui et comment vous autorisez l'indexation de votre site.

Sa syntaxe peut être assez complexe, c'est pourquoi nous ne vous donneront que les principes de base et vous invitons à consulter un document plus complèt à l'adresse : info.webcrawler.com/mak/projects/robots/robots.html.

SYNTAXE DE /ROBOT.TXT

Ce fichier est constitué de plusieurs lignes de commande séparées par des "retour chariot". Les deux commandes principales sont:

- **User-agent :** puis le nom d'un robot. En général vous aurez du mal à le connaître, à moins que vous puissiez accéder à vos fichiers de log **Agent_Log**. Pour les interdire tous, mettez une *.
- **Disallow :** puis les répertoires où vous désirez interdire l'accès. Pour les interdire tous, mettez une *.

Si par exemple vous désirez interdire l'accès à vos répertoires de statistique (`/Statistiques`), votre fichier `/robots.txt`devra ressembler à ceci:

```
# robots.txt pour le site http://www.monsite.fr

User-agent: *
Disallow: /Statistiques
```

Si par contre vous désirez interdire l'accès à tout votre site, votre fichier `/robots.txt` devra ressembler à ceci:

```
# robots.txt pour le site http://www.monsite.fr

User-agent: *
Disallow: /
```

Attention!, malgré un effort de standardisation de la syntaxe du fichier `/robots.txt`, la majorité des robots ignoreront tout simplement l'ensemble de votre site s'ils le découvrent à la racine du serveur. Si vous désirez rendre certaines parties de votre site privées, mieux vaut exploiter un système de mots de passe (`.htpasswd`).

ANNEXE B

Extraits des pages HTML modifiées du site du CRDP Nord Pas-de-Calais

Page d'accueil

```
<head>
<meta name= "robots" content="index, follow">
<meta name="language" content="french">
<title>CRDP du Nord Pas-de-Calais : documents pédagogiques multimédias pour
l'enseignement</title>
<meta name="description" content="Le CRDP, centre régional de documentation pédagogique,
du Nord Pas-de-Calais vous propose des documents pédagogiques, des animations multimédias et
dela maintenance technique">
<meta name="author" content="CRDP, centre régional de documentation pédagogique du Nord
Pas-de-Calais (Lille)">
<meta name="keywords" content="CRDP, CDDP, documents pédagogiques, TPE, travaux
personnels encadrés, travaux croisés, TICE, enseignement, didactique des disciplines, ressources
pédagogiques, pédagogie, éducation, édition pédagogique, édition en braille, documentation
administrative, sujets d'examens, enseignants, CDI, maintenance BCDI, documentalistes, prêt,
consultation, réseau de médiathèques, catalogue, catalogues, achat, vente, animations
multimédias, ateliers multimédias, RIP, produits reconnus d'intérêt pédagogique, logiciels
éducatifs">
<meta name="DC.title" content="CRDP du Nord Pas-de-Calais : documents pédagogiques
multimédias pour l'enseignement>
<meta name="DC.creator" content="CRDP,centre régional de documentation pédagogique du
Nord Pas-de-Calais (Lille)">
<meta name="DC.subject.keywords" content="CRDP,CDDP, documents pédagogiques, TPE,
travaux personnels encadrés, travaux croisés, TICE, enseignement, didactique des disciplines,
ressources pédagogiques, pédagogie, éducation, édition pédagogique, édition en braille,
documentation administrative, sujets d'examens, enseignants, CDI, maintenance BCDI,
documentalistes, prêt, consultation, réseau de médiathèques, catalogue, catalogues, achat, vente,
animations multimédias,.ateliers multimédias, RIP, produits reconnus d'intérêt pédagogique,
logiciels éducatifs">
<meta name="DC.description" content="Le CRDP, centre régional de documentation
pédagogique, du Nord Pas-de-Calais vous propose des documents pédagogiques, des animations
multimédias et de la maintenance technique">
<meta name="DC.publisher" content="CRDP du Nord Pas-de-Calais">
<meta name="DC.language" content="french">
</head>
```

Page de présentation des librairies

```
<head>
<meta name="language" content="french">
<meta name="robots" content="index, follow">
<title>Librairies des CDDP du Nord Pas-de-Calais : vente, achat de documents pédagogiques
multimédias"</title>
<meta name="author" content="CDDP du Nord (Lille, Dunkerque, Valenciennes)">
<meta name="author" content="CDDP du Pas-de-Calais (Arras, Outreau, Calais)">
<meta name="description" content="Pour l'achat de documents pédagogiques du réseau CNDP-
CRDP, les librairies et les délégués pédagogiques des CDDP du Nord et du Pas-de-Calais sont à
votre disposition">
<meta name="keywords" content="vente, achat, documents pédagogiques, didactique des
disciplines, école, collège, lycée, primaire, secondaire, supérieur, instituteurs, PE, professeurs des
écoles, concours, PLC, PLP, programmes, rapports de jury, français, mathématiques, histoire,
géographie, SVT, STI, STT, SMS, physique, chimie, technologie, art, EPS, musique, arts
plastiques, langues vivantes, anglais, allemand, néerlandais, italien, espagnol, portugais, russe,
grec, latin, rapport de jury, programme">
<meta name="DC.title" content="Librairies des CDDP du Nord Pas-de-Calais : vente, achat de
documents pédagogiques multimédias">
<meta name="DC.creator" content="CDDP du Nord (Lille, Dunkerque, Valenciennes)">
<meta name="DC.creator" content="CDDP du Pas-de-Calais (Arras, Outreau, Calais)">
<meta name="DC.subject.keywords" content="vente, achat, documents pédagogiques, didactique
des disciplines, école, collège, lycée, primaire, secondaire, supérieur, instituteurs, PE, professeurs
des écoles, concours, PLC, PLP, programmes, rapports de jury, français, mathématiques, histoire,
géographie, SVT, STI, STT, SMS, physique, chimie, technologie, art, EPS, musique, arts
plastiques, langues vivantes, anglais, allemand, néerlandais, italien, espagnol, portugais, russe,
grec, latin, rapport de jury, programme">
<meta name="DC.publisher" content="CRDP du Nord Pas-de-Calais">
<meta name="DC.description" content="Pour l'achat de documents pédagogiques du réseau
CNDP-CRDP, les librairies et les délégués pédagogiques des CDDP du Nord et du Pas-de-Calais
sont à votre disposition">
</head>
```

Page de présentation des catalogues des médiathèques

```
<head>
<meta name="language" content="french">
<meta name="robots" content="index, follow">
<title>Les médiathèques pédagogiques du Nord Pas-de-Calais : catalogues de documents
multimédias"</title>
<meta name="description" content="Les catalogues des médiathèques pédagogiques des CDDP
du Nord et du Pas-de-Calais, consultables en ligne, localisent des documents multimédias en
consultation ou en prêt">
<meta name="keywords" content="documents pédagogiques, TPE, didactique des disciplines,
pédagogie, enseignants, documentalistes, étudiants IUFM, PE, professeurs des écoles, PLC, PLP,
prêt, consultation, documentation administrative, sujets d'examens, bibliographies, recherche
documentaire, ouvrages, revues, périodiques, imprimés, vidéos, VHS, cassettes,audiocassettes,
vidéocassettes, diapositives, films, logiciels éducatifs, CDROMS,cédéroms, disques, CD">
<meta name="DC.title" content="Les médiathèques pédagogiques du Nord Pas-de-Calais :
catalogues de documents multimédias">
<meta name="DC.description" content="Les catalogues des médiathèques pédagogiques des
CDDP du Nord et du Pas-de-Calais, consultables en ligne, localisent des documents multimédias
en consultation ou en prêt">
<meta name="DC.subject.keywords" content="documents pédagogiques, TPE, didactique des
disciplines, pédagogie, enseignants, documentalistes, étudiants IUFM, PE, professeurs des écoles,
PLC, PLP, prêt, consultation, documentation administrative, sujets d'examens, bibliographies,
recherche documentaire, ouvrages, revues, périodiques, imprimés, vidéos, VHS, cassettes,
audiocassettes, vidéocassettes, diapositives, films, logiciels éducatifs, CDROMS,cédéroms,
disques, CD">
<meta name="DC.coverage" content="Lille, Dunkerque, Valenciennes, Arras,Outreau, Calais">
</head>
```

Page présentant les travaux personnels encadrés (TPE)

```
<head>
<meta name="language" content="french">
<meta name="robots" content="index, follow">
<title>Les TPE, travaux personnels encadrés</title>
<meta name="description" content="Le CRDP et les CDDP du Nord Pas-de-Calais vous
proposent des ressources documentaires multimédias pour les TPE"
<meta name="keywords" content="TPE, travaux personnels encadrés, ressources documentaires,
documents pédagogiques, Nord, Pas-de-Calais, documentation pédagogique, médiathèques,
académie de Lille">
<meta name="DC.title" content="Les TPE, travaux personnels encadrés">
<meta name="DC.description" content="Le CRDP et les CDDP du Nord Pas-de-Calais vous
proposent des ressources documentaires pour les TPE"
<meta name="DC.subject.keywords" content="TPE, travaux personnels encadrés, ressources
documentaires, documents pédagogiques, Nord, Pas-de-Calais, documentation pédagogique,
médiathèques, académie de Lille">
</head>
```