THESE

Présentée

devant l' 'UNIVERSITE CLAUDE BERNARD - LYON1

pour l'obtention

du DIPLOME DE DOCTORAT

en Sciences de l'Information et de la Communication

présentée et soutenue publiquement
en mai 2001

Par

Youcef AMEROUALI

Metadonnées basées sur l'association d'éléments de description de ressources et d'éléments de profil d'utilisateur

Directeur de thèse : Professeur Richard Bouché
JURY:

Richard Bouché	**professeur**	**ENSSIB**	
Hassoun Mohamed	**Professeur**	**ENSSIB**	
Hubert Fondin	**Professeur**	**Bordeaux3**	**Rapporteur**
David Amos	**HDR**	**Nancy2**	**Rapporteur**
Stéphane Chaudiron	**Mdc**	**Paris10**	

REMERCIEMENTS

J’adresse mes plus vifs remerciements au professeur Richard Bouché pour m’avoir accepté dans son équipe de recherche et pour m’avoir soutenu et encouragé tout au long de ce travail. Je souhaiterai qu’il trouve ici l’expression de ma profonde reconnaissance.

Je remercie vivement messieurs Hubert Fondin et David Amos pour l’honneur qu’ils me font d’être les rapporteurs de cette thèse et pour le temps qu’ils ont consacré à sa lecture.

Je remercie également messieurs Hassoun Mohamed et Stéphane Chaudiron pour avoir bien voulu participer à ce jury.

Je tiens aussi à remercier toutes les personnes qui m’ont encouragé et soutenu jusqu’à l’aboutissement de ce travail, particulièrement mon ami Madjid.

A tous les membres de l’équipe SII et du laboratoire RECODOC ;

A mes enfants….

SOMMAIRE

TABLE DES ILLUSTRATIONS

A- FIGURES :

B- TABLEAUX :

C- IMAGES :

L’identité des individus n’est pas une donnée à priori ; chaque individu appartient à plusieurs scènes et peut donc avoir plusieurs rôles.

Erving Goffman

INTRODUCTION

La recherche de l'information est un terme vague, approximativement défini si l'on se réfère uniquement aux systèmes automatiques. Evidemment, je précise bien automatique par opposition au manuel et information par opposition aux données. Malheureusement le mot « information » peut induire en erreur. Dans le contexte de la recherche d'information, cette dernière, selon la définition technique qu'en donne la théorie de la communication de Shannon, n'est pas quantifiable (Shannon et Weaver) [93]. En effet, dans la plupart des cas, ou peut d'une manière convenable décrire le genre de recherche simplement en remplaçant le mot document par le mot information. Pourtant des auteurs et non des moindres (Cleverdon [68], Salton [72], Sparck Jones [70]) ont mené des études approfondies sur la recherche de l'information. Selon Lancaster [79], bien que l'expression recherche de l'information ne soit pas vraiment exacte, son utilisation est admise par convention. En général le système de recherche de l'information ne renvoi pas directement à l'utilisateur l'information ou le document recherché, mais il l'informe simplement de son existence (ou de sa non-existence) et de sa localisation. Ceci exclue spécifiquement les systèmes questions-réponses présentés par Winograd [66] et ceux présentés par Minsky [80]; et exclue aussi les systèmes de recherche de données (tel celui qui donne les cotations en ligne de la bourse par exemple). Il est clair aussi qu'il y a une nuance certaine entre recherche d'information et recherche de données.

Dans la recherche de données, nous cherchons un appariement exact, c'est- à- dire, nous voulons retrouver un élément bien précis. Dans la recherche de l'information, ceci peut aussi nous intéresser, mais ce que nous voulons trouver en général, ce sont des éléments qui apparient en partie la requête, et en sélectionner quelques-uns uns parmi ceux qui nous semblent les plus pertinents….

Maintenant la plupart de ces notions ont considérablement évolué.
Face à l'évolution rapide des moyens d'information et de communication, une masse importante d'informations est produite chaque jour à travers des milliers de livres, des dizaines de milliers d'articles de périodiques et d'innombrables pages web. Par conséquent, la nécessité de structurer l'information et de la rendre utilisable et accessible de la façon la plus optimale possible ne s'est jamais autant fait sentir. Pour cela, il faut pouvoir décrire cette information d'une façon synthétique.
Dans le milieu des bibliothèques, la notice bibliographique constitue un moyen de description et de représentation de l'information, très utilisé. Pour l'information numérique stockée dans des bases de données, ou en réseau, ce besoin s'est fait sentir aussi très tôt. Ce fut le cas pour les bases de données géographiques : Leurs gestionnaires ont ainsi été amenés à développer des standards de metadonnées, pour décrire leur information.

Devant le foisonnement d'informations sur Internet, des initiatives pareilles se mirent en branle ; et d'autres standards de metadonnées furent développés. L'un d'eux émergea du lot par sa simplicité d'abord et ses qualités ensuite ; et fut très vite adopté par une large communauté, pour des besoins aussi divers que spécifiques : c'est le standard de metadonnées du Dublin Core. (Partie A-II ; page 51)

Devant cette multitude de réflexions et de projets qui se sont construits sur le thème de la recherche et du filtrage de l'information (Partie B), toujours dans le sens d'une continuelle recherche de l'efficacité des systèmes qui en résultent, nous avons essayé de développer, nous aussi, un modèle sémantique, que nous avons appelé DREPU, basé sur une association d'éléments de description de ressources et d'éléments de profil d'utilisateur.
Evidemment, nous situant sur un axe de recherche d'actualité et donc investi par plusieurs équipes de recherche, nous nous sommes donnés pour missions, au préalable, de passer en revue les principaux projets ayant des points de contingence appréciable avec notre théorie ; afin d'en voir la genèse et d'identifier la nature, la sémantique et le codage des informations qu'ils prennent en compte. (Partie C)

Après cela, nous avons entrepris de développer un système de recherche et de filtrage de l'information basé sur notre modèle DREPU (Partie D)
Dans cette dernière partie, nous nous sommes donnés comme missions de construire notre modèle, de le structurer en un standard de metadonnées dérivé du Dublin Core ; et de définir la notion d'interrogation à relance qui va permettre d'activer la fonction de filtrage de notre système. Nous y présentons aussi les outils logiciels (DREPUZ et G-MET) développés pour permettre de tester et d'évaluer le modèle DREPU….

Notre pari est ambitieux et simple à la fois : mettre en place un système de recherche et de filtrage de l'information efficace, avec un mode d'indexation particulier et un principe d'interrogation spécifique : l'interrogation à relance.

Actuellement tous les systèmes de recherche d'informations en réseau évoluent de plus en plus vers une meilleure appréciation de la notion de documents pertinents. Evidemment pour cela, il faut d'abord améliorer les modes d'indexation. Mais jusqu'à maintenant, les quelques tentatives qui existent se limitent à une timide introduction des éléments de description des documents, tirés souvent des titres de ces derniers et de quelques mots-clés ou description très sommaire. Dans notre travail, nous avons introduit tous les éléments classiques de description de ressources pour une véritable indexation du document et nous y avons associé des éléments de profil d'utilisateur pour une fonction de filtrage. Nous pensons par cette association, améliorer grandement la pertinence des documents retrouvés par un tel système. Nous avons parcouru les différentes voies des principales équipes de recherche qui travaillent sur des projets ayant le même objectif (Partie C) ; cela nous a évidemment beaucoup aidé à tracer notre voie. Nous pensons que certains travaux sont bridés par un formalisme purement informatique qui ne permet pas d'entrevoir avec le même entendement que celui des véritables détenteurs du savoir-faire (pour ne pas dire de la science) de la « recherche d"information » que sont les bibliothécaires et les documentalistes, des notions très importantes telles l'indexation et la pertinence.

D’autres travaux que nous n’avons pas cités, abondent dans le même sens et affluent vers le même objectif.

Intégrer des éléments de profil d’utilisateur dans un processus de recherche d’information, avec des définitions parfois aux antipodes d’un projet à un autre ; est devenu un fait d’actualité. Les chercheurs du CNET (Centre de recherche de France Télécom) qui ont développé le moteur de recherche « Voilà » travaillent sur un outil de recherche intégrant le profil utilisateur selon leur entendement bien-sûr ; le Groupe Amazone.com prévoit un système de captation du profil des utilisateurs qui se connectent sur ses serveurs….

Tout cela nous a orientés vers la construction d’un modèle plus proche des schémas de définition des bibliothécaires et documentalistes. Les résultats des tests et évaluations nous ont donné satisfaction ; et nous pensons que nous avons ainsi introduit une autre option dans le domaine de la recherche d’information qui est appelée à se développer et à se perfectionner….

Partie A :

LA DESCRIPTION DE RESSOURCES :

I-Description classique de ressources bibliographiques : le catalogage

II- Description de ressources électroniques : les metadonnées

Introduction :

La description de ressources d'information est une opération qui permet de sélectionner des éléments descriptifs pertinents de ces dernières qui aideront à les retrouver lors d'un processus de recherche d'information ultérieur.

Pour les documents sur supports classiques, les bibliothécaires ont recours depuis très longtemps à des opérations de catalogage pour les décrire.

Pour l'information en réseau, des standards de metadonnées ont été développés pour la synthétiser et la décrire.

I- Description classique de ressources bibliographiques :

Le catalogage :

Définitions :

- Le catalogage classique consiste en la rédaction de toutes les fiches nécessaires aux différents répertoires (catalogues) d'ouvrages, à partir d'une description normalisée des éléments permettant l'identification d'un ouvrage donné.

- L'indexation consiste en la recherche d'un symbole numérique ou nominal pour un ouvrage, à partir de l'analyse de son contenu. Ce symbole peut être évidemment, soit :
  -un chiffre tiré d'une classification (Dewey ou Universelle) ;
  -un mot clé (indexation alphabétique matière)

- La cotation est une opération qui complète l'indexation numérique généralement par les trois premières lettres du nom de l'auteur ou du titre d'anonyme afin de faciliter le classement de l'ouvrage sur les rayons de la bibliothèque.

Rédaction de la fiche de base :

Les fiches des différents catalogues, à part celles des catalogues *titres* et *collections,* sont établies à partir d'un même modèle : La fiche de base ; qui se compose des éléments suivants :

- La cote ;
- La vedette-auteur ;
- Le corps de la notice.

Les renseignements portés sur cette fiche sont tirés :

- de la page de titre essentiellement ;
- des parties qui précèdent la page de titre et son verso ;
- de l'achevé d'imprimer ;
- du livre entier ;
- des sources externes au livre tels des ouvrages de références pour préciser une donnée.

La cote :

La cote est un signe conventionnel qui indique la place du livre sur les rayons. Elle est donc établie en fonction du système de classement choisi.

La vedette-auteur :

La vedette-auteur est constituée du nom de l'auteur de l'ouvrage. Le nom de famille est inscrit en lettres capitales, suivi du prénom, en lettres minuscules et entre parenthèses.

Le corps de la notice :

Il se compose de différentes zones, elles-mêmes composées de différents éléments qui sont introduits par des signes de ponctuation bien déterminés.

Différents types de catalogues :

Ils existent plusieurs types de catalogues qui permettent de retrouver des ouvrages dans un fonds de bibliothèque :

- ✓ *Le catalogue auteurs et titres d'anonymes* qui permet d'identifier tous les auteurs, principaux ou secondaires, ainsi que les titres des anonymes, par manque ou excès d'auteurs ;
- ✓ *Le catalogue par titre* qui permet de retrouver le nom de l'auteur et la cote d'un ouvrage dont seul le titre est connu (ouvrages de fiction) ;
- ✓ *Le catalogue matière* permet de recenser tous les ouvrages de la bibliothèque, traitant d'un sujet donné. Il se subdivise en deux types :
  - Le catalogue alphabétique ;
  - Le catalogue systématique.

Ces deux derniers catalogues abordent l'ensemble des connaissances de façon très différente.

Le classement alphabétique est celui du langage naturel (à la manière d'un dictionnaire). Il est le plus utilisé.

Le classement systématique permet de regrouper les connaissances par disciplines. L'utilisation d'un langage conventionnel et abstrait rend son accès difficile.

Le type de catalogue choisi est déterminé en fonction des besoins des usagers de la bibliothèque. Généralement la précision des termes et la facilité de la recherche militent en faveur du catalogue alphabétique.

- ✓ *Le catalogue des collections* : Les collections regroupent sous un même titre, un certain nombre d'ouvrages, écrits à différents moments et par différents auteurs. Généralement les ouvrages d'une même collection, comportent des traits communs qui sont des caractéristiques communes et qui peuvent être soit :
  - intellectuelles (idéologie, contenu) ;
  - matérielles (format, présentation).

- ✓ *Le catalogue topographique* : appelé aussi catalogue-inventaire, il reproduit exactement l'ordre des ouvrages sur les rayons ; et permet de vérifier en principe chaque année, l'état du fonds livres de la

bibliothèque. Les fiches catalographiques sont classées dans l'ordre des cotes des ouvrages.

Fiche de base et fiches secondaires :

Dans un catalogue, la fiche de base est la fiche principale sur laquelle sont enregistrées toutes les données catalographiques et qui a pour vedette le nom du premier auteur. Elle sert de base à la duplication des fiches secondaires ; auxquelles sont attribuées d'autres vedettes (auteur, matière), pour alimenter différents fichiers catalographiques.

LES FICHIERS INFORMATISES :

Le catalogage est une opération automatisable. Un certain nombre de bibliothèques pratiquent depuis les années 1970 un catalogage informatisé pour permettre un échange standardisé de données bibliographiques ; et ce dans le cadre d'un contrôle bibliographique universel (CBU).

En fait, l'automatisation du catalogage consiste à faire remplir par le bibliothécaire des bordereaux de saisie, selon des règles normalisées. C'est donc ce dernier qui sélectionne les données à traiter et qui détermine les termes définissant le contenu du document à décrire.

DESCRIPTION DE PARTIES DE DOCUMENTS :

Actuellement les techniques de description catalographique privilégient de plus en plus les niveaux bibliographiques élémentaires; c'est à dire qu'on arrive à décrire aujourd'hui des parties de documents, tels des chapitres d'un livre ou d'un rapport, une communication extraite des proceedings d'un congrès, une illustration dans un album, une carte dans un atlas, etc…. Ce sont des composantes d'un document global qui sont appelés unités documentaires ou dans le cadre de l'information en réseau plus proprement unités d'information.

II- DESCRIPTION DE RESSOURCES ELECTRONIQUES :

LES METADONNEES :

1)Définition :

Les metadonnées (ou metadata) peuvent être définies comme étant des données relatives à d'autres données (*data about data* : données sur des données). Par conséquent, une notice catalographique classique peut être considérée comme une métadonnée. Le terme « metadata » est surtout utilisé pour désigner l'information « lisible par machine » concernant des fichiers de données « lisibles par machine» : donc ce terme désigne en quelque sorte une information référentielle sur des données électroniques.

Utilisées dans le contexte de l'information numérique géospatiale, les metadonnées sont perçues comme l'information de fond qui décrit le contenu, la qualité, les conditions et autres propriétés et caractéristiques des données.

Tim BARNERS L. du **W3C** définit les metadonnées comme suit :

« Metadata est une information compréhensible (interprétable) par machine (ordinateur) sur des ressources d'information du Web ou d'autres choses (autres sources de données). L'expression « compréhensible par machine » est la clef. Nous parlons ici de l'information que des logiciels traitent pour nous faciliter la vie, nous assurer que nous obéissons à nos principes, à la loi, vérifier que nous pouvons croire en ce que nous faisons, et exécuter tous les travaux régulièrement et rapidement. Metadata définit bien la sémantique et la structure. Metadata a été appelée ainsi à sa création; et est encore actuellement une information sur des ressources du Web, ou des données sur des données. Plus tard, lorsque les metadonnées, les langages, les moteurs de recherches seront plus évolués; ils formeront une base plus forte pour l'information « compréhensible par machine » du Web, sur n'importe quoi : les gens, les choses, les concepts et les idées. Gardons dans notre esprit ce modèle (préconisé), bien que la première étape soit de réaliser un système d'indexation de l'information en réseau. »

2)Historique :

2.1)Première époque :

Le premier standard de metadonnées, le format MARC (Machine Readable Cataloguing) fût développé en 1965 par la bibliothèque du Congrès américain à des fins d'échanges de notices bibliographiques sur bandes magnétiques. A côté du format USMARK vers lequel évolua le format MARC, apparurent rapidement d'autres formats dérivés de ce dernier. Ce fut le cas d'abord en 1968 de UKMARC qui fut établi par la British National Bibliography en Grande-Bretagne. La bibliothèque Universitaire de GRENOBLE (France)définit sur la même lancée, un format analogue en 1970 : Monocle. A partir de là, ce fut le déferlement : AUSMARC en AUSTRALIE CANMARC au Canada, DANMARC au DANEMARK, et finalement INTERMARC pour l'EUROPE. Ce foisonnement de formats nationaux risquait de devenir un sérieux obstacle aux échanges d'information bibliographique entre pays. Ce qui amena l'IFLA[1] à développer un format d'échange international : UNIMARC. La première version de ce format apparut en 1977. Il fut révisé et complété en 1987 et en 1994. Actuellement une instance de l'IFLA, le *Permanent Unimarc Commitee* est en charge de son évolution et de la promotion de son utilisation dans le monde...

Quant au terme *metadata*, il fut employé pour la première fois par un informaticien américain **Jack E. Meyers** en définissant des architectures informatiques appelées plus tard metamodèles, au début de l'été 1969. Il fit d'ailleurs des recherches dans des publications et des bases de données pour s'assurer que ce mot n'a pas été utilisé avant lui. Après confirmation, il décida en 1986 d'enregistrer le mot « metadata » aux USA comme une marque d'une compagnie de services informatiques. Cela n'a évidement aucun lien avec le sens que nous connaissons au mot metadata, aujourd'hui.

En novembre 1987, un projet de norme d'échange de documents électroniques (TEI pour Text Encoding Initiative) fût lancé, lors d'une conférence préparatoire qui s'est tenue au VASSAR Collège de New York.

[1] International Federation of Libraries Associations

Les objectifs généraux de la TEI ont été définis par la résolution finale, d'ailleurs connu sous le nom de 'Principes de Poughkeepsie' (du nom de l'endroit où s'est tenue la conférence). La TEI connue deux cycles de développement :

- ✓ Le premier cycle (1988\1990) déboucha sur la publication d'une ébauche des recommandations de la TEI (document TEI P1).
- ✓ Le deuxième cycle (1990\1994) permit la publication de la version officielle des recommandations de la TEI, en mai 1994.

Par leur structure, les entêtes TEI peuvent être perçues comme des metadonnées.

2.2)Deuxième époque :

Nous situons le début de cette deuxième époque au moment où des agences gouvernementales Américaines, grandes utilisatrices des systèmes d'informations géographiques ont ressenti le besoin d'organiser et de cataloguer l'information géospatiale ; ce qui était devenue une nécessité pour faciliter l'accès à cette dernière.

Dés le 19 octobre 1990,une circulaire présidentielle (A16) américaine établissait des missions pour une structure fédérale le *FGDC(Fédéral Geographic Data Commitee*) de développement d'un standard de catalogage de l'information géospatiale. Cela permit le lancement d'un standard de metadonnées : le ***CSDGM*** (Content Standard Data for Geospatial Metadata) qui fut basé sur le *SDTS* (Spatial Data Transfert standard) et l'*USMARC*.

Le CSDGM fut rapidement développé et implanté au niveau des agences gouvernementales entre 1993 et 1994.

En parallèle, d'autres standards de metadonnées se développaient aux USA, dont les plus connus :

- ✓ ***IAFA*** (Internet Anonymous Ftp Archives) standard conçu par l'IETF (Internet Engineering Task Force) pour cataloguer les archives des sites ftp.

✓ Le ***GILS*** (Governement Information Locator Service) élaboré pour cataloguer et ainsi faciliter l'accès à l'information détenue par les organismes gouvernementaux. En décembre 1994, le GILS fût adopté comme norme fédérale à laquelle devait se conformer tous les départements d'état.
Ce même standard a été adopté aussi par le Canada en 1995 pour cataloguer toute l'information des organismes gouvernementaux.

✓ Les standards développés par les agences américaines de l'environnement :

- ❖ ***SMSP*** (Scientific Metadata Standards Projects) en 1994 par l'EPA (US Environment Protection Agency)
- ❖ ***CERES(***California Environmental Ressources Evaluation System) en 1995.

✓***MRLC*** (Multiple Résolution Landscape Characteristics) développé par le NBS (National Biological Survey) structure dépendant du ministère américain de l'intérieur.

2.3)Troisième époque :

Nous situons le début de cette dernière époque en 1995, c'est à dire à l'année où, sous la houlette d'OCLC (On-line Computer Library Center) un colloque international fût organisé à DUBLIN, dans l'état de l'OHIO (USA). Il permit d'aboutir au lancement d'un standard plus simple et ouvert : le Dublin Core.
Celui-ci est supporté par une série de workshops qui lui assurent un état d'évolution permanent. Au workshop, qui s'est tenu a l'université de Warwick, en avril 1996 au Royaume-Uni, une syntaxe particulière est proposée pour le Dublin Core, qui est approuvée par le W3C. Cela le raccroche à des développements majeurs du W3C, que sont la RDF(Ressource Description Framework)et XML. Une autre conférence d'OCLC sur les metadata, relative aux images accessibles en réseau, qui s'est tenue en septembre 1996, a rendu possible l'utilisation du Dublin Core dans le domaine des images. Enfin le quatrième workshop qui a eu

lieu a Camberra en Australie en mars 1997 fit l'unanimité sur une proposition d'élever le Dublin Core à un standard international.

Plus d'une quarantaine de projets, se basant sur ce standard commencent à s'implanter un peu partout dans le monde. Des standards dérivés ont vu aussi le jour tel le standard de l'EDNA (Educationel Network Australy)ou le standard du secteur de l'éducation américaine : GEM metadata (**G**ateway to **E**ducational **M**aterials).

Il faut signaler qu'en octobre 1995 fut fondée aussi la '***Metadata Coalition'***, association regroupant alors 53 compagnies d'informatique (dont IBM et Microsoft)et d'autres organismes, dans le but de définir un ensemble de spécifications standards pour l'interchangeabilité et la prise en charge des metadonnées par des outils logiciels. Cela aboutit à un projet dénommé : ***MDIS*** (Metadata Interchange Spécification Initiative).

Devant ce début de prolifération de standards de metadonnées, et sur les recommandations de l'ISO (International Standardisation Organisation), s'est tenue une conférence appelée "Metadata Registry" du 8 au 11 juillet 1997, à BERKELEY (CALIFORNIE\USA). Celle-ci, sponsorisée par l'*Agence de Protection de l'Environnement* U.S., On-*line Computer Library Center* et la *Metadata Coalition*, avait pour but de dégager des recommandations ou des normes pour la mise en place ou le développement de standards de metadonnées afin d'assurer pour ces derniers :

- ✓ l'interopérabilité ;
- ✓ le mapping ;
- ✓ l'évolution (caractère évolutif) ;
- ✓ la flexibilité .

3)Les différents standards de metadonnées :

3.1)Content Standard Digital Geospatial Metadata :*(CSDGM)*

Le CSDGM est un standard de metadonnées développé par le FGDC (Federal Geographic Data Commitee). La première version fut approuvée le 8 juin 1994.

Ce standard comporte 334 éléments, dont 119 sont prévus uniquement pour référencer d'autres éléments. C'est un système ouvert, c'est à dire qu'il est possible de lui rajouter d'autres éléments (ou d'en enlever). Ses utilisations majeures sont :

- ✓ aider à organiser et à maintenir les données géospatiales ;
- ✓ fournir de l'information sur des données pratiques et théoriques, fournir de l'information pour interpréter et traiter des données reçues par un transfert de sources externes.

Les éléments de données du CSDGM fournissent de l'information sur :

- l'identification(information de base sur le lot de données) ;
- la qualité des données;
- l'organisation des données géospatiales ;
- les références géospatiales ;
- les attributs et les entités(information sur le contenu du lot de données) ;
- la distribution ;
- référence metadata (information sur l'auteur des metadonnées).

Les éléments du CSDGM peuvent être crées sous SGML (qui est son format d'échange conformément à la DTD développée par le FGDC) en utilisant un simple éditeur de texte, mais pour éviter les risques d'erreur, il est préférable d'utiliser des outils tels que XtME (Xt Metadata Editor)qui fonctionne sous Unix.

Il existe évidement d'autres outils :

- •ARC/INFO GIS Metadata Generator AMLs
- •ASCII template
- •ASCII template (autre version)

•blmdoc : pour ARC/INFO version 7
•CORPSMET: outil de création de metadata de l'Armée US
(Corps des Ingénieurs)
•Document.aml : ARC/INFO
•Document aml : version de l' USGS pour ARC/INFO 7.0.4.
•Geolineus 3.0: organise les données, visualise la lignée et crée des modèles de flux de données pour ARC/INFO
•Informix SQL template
•mp : Compilateur pour des metadata formalisées (USGS)
•NOAA's FGDC Metadata Toolkit: logiciel
•Word Perfect template

- **Standards de metadonnées dérivés du CSDGM** :

a)Le DENVER CORE :

C'est un standard développé à l'Université de DENVER (USA). Les champs ou éléments suggérés par le DENVER CORE SONT :

- Thème-Mots clés;
- Position-Mots clés;
- Coordonnées extrêmes;
- Résumé;
- But;
- Temps-période-contenu;
- références;
- Données géospatiales-Présentation-Forme;
- Créateur;
- Titre;
- Langage;
- Description des ressources.

Tous ces champs se retrouvent dans le CSDGM, sauf l'élément '*Langage'.*
Il est à noter que ce standard de metadonnées est quasiment inconnu. Il n'est cité que par ceux qui critiquent le CSDGM pour le temps que ce dernier

nécessite à sa mise en place et sa relative complexité (description trop technique).

b) The Australia New Zealand Land Information Council (ANZLIC):

Le développement de ce standard entre dans le cadre de la mise en place d'un plan stratégique (ANZLIC's Strategic Plan pour 1994/1997) d'implantation d'un système d'information géographique global pour la Nouvelle-Zélande et l'Australie.

Dés le mois de Décembre 1995,une ébauche de recommandations pour le standard de metadonnées ***ANZLIC*** fut remise à 180 représentants d'organismes potentiellement utilisateurs ou en relation avec ce type d'information. Les auteurs de ces recommandations se sont inspirés du CSDGM, tout en reconnaissant être moins ambitieux pour leur standard, dont ils limitent le nombre d'éléments au strict minimum. Le groupe de travail chargé de la définition des champs du standard ANZLIC mit 18 mois de consultations pour proposer un condensé de 220 éléments du CSDGM :

Catégorie	**Elément**
Dataset.....	-Titre
	-Conservateur
	-Juridiction
Description	-Résumé
	-Mots de recherche
	-Nom de la position géographique
	Ou
	-Coordonnées géographiques
Data Currency.....	- date de début
	- date finale
Dataset Status.....	-Progression
	-Maintenance et fréquence de mise à jour

Access.....- Format de stockage de données
- Type de format valable
- Contraintes d'accès

Data Quality.....-Lignée
-Position Actuelle
-Attribut Actuel
-Consistance logique

Contact Information.....-Point (ou personne) de contact de l'organisation
-Point(ou personne)de contact de la position
-Mail Address 1 (adresse postale)
-Mail Address 2 (extension de l'adresse postale)
-Place ou Localité
-Etat ou Localité 2
-Pays
-Code postal
-Téléphone
-Fax
-Adresse électronique

Metadata Date.....-Date de création de la metadata

Additional Metadata.....-Metadata additionnelle (références d'autres répertoires ou systèmes contenant davantage d'informations sur les données traitées)

⇒ Les catégories d'éléments *Dataset* et *description* fournissent essentiellement de l'information sur le contenu des données décrites, de l'agence responsable de leur collecte et organisation et du terrain géographique couvert.

⇒ Les catégories *Data Currency* et Data *Status* établissent le cadre temporel des données décrites.

⇒ La catégorie *Access* informe l'utilisateur sur les formats des données et les éventuels formats d'échange.

⇒ L'inclusion de la catégorie *Data Quality,* constituée des éléments *lineage, positional accuracy, attribute accuracy, logical consistency* (lignée, précision de position, précision d'attribut, consistance logique), a entraîné un grand débat chez la communauté de l'information géographique dans la mesure où cela risque d'être source d'incompréhension de la part des utilisateurs non habitués à ce genre d'information.

⇒ La catégorie *Contact Information* fournit les détails d'adresse de l'emplacement des contacts (personne contact) au sein de l'organisation responsable de la distribution des données aux autres utilisateurs.

⇒ *Metadata Date* donne la date de création des éléments de metadata relatifs aux données décrites.

⇒ *Additional Metadata* fournit un lien vers éventuellement d'autres sources d'informations sur les données décrites.

3.2) Government Information Locator Service: (*GILS*)

Le GILS a été élaboré pour appuyer les politiques gouvernementales américaines en vertu desquelles les départements et organismes du gouvernement sont tenus de rendre l'information qu'ils détiennent accessible au public et de mettre au point des systèmes d'information utilisables dans un environnement de systèmes ouverts.

En décembre 1994, *le Département of Commerce* a approuvé le profil GILS à titre de norme fédérale de traitement de l'information (FIPS 192) à laquelle tous les départements devaient se conformer. La création et l'exploitation du GILS sont devenues obligatoires aux Etats-Unis lorsque la loi dite *Paper Reduction Act* de 1995 (article 3511) a été adoptée. Dans le contexte du projet de société d'information mondiale du G7, le GILS a été proposé à titre de modèle de localisateur d'information mondiale. En février 1995, les participants à la conférence ministérielle du G7 ont approuvé un projet pour

la gestion de l'environnement et des ressources naturelles, qui comprendra un localisateur d'information mondiale.

Profil GILS :

Le profil est un guide à l'intention des préposés à la mise en application, qui augmente les probabilités que les systèmes GILS élaborés par divers préposés à la mise en application et fournisseurs soient interconnectables et interopérables. Le profil traite de l'intégration des systèmes et des données devant être échangées aux fins du GILS. Par exemple, il précise que la norme d'extraction de l'information ANSI/NISO (Z39.50) est la norme d'échange de l'information. Toutefois, le GILS ne précise pas l'apparence de l'interface d'utilisateur ni la structure interne de la base de données renfermant les enregistrements de localisation GILS. La première version du profil d'application GILS a été approuvée en 1994, à titre de norme fédérale américaine de traitement de l'information (FIPS 192). La deuxième version, rédigée en 1996, tient compte de l'expérience des fournisseurs de services GILS ainsi que des commentaires du sous-groupe GILS du gouvernement canadien.

Enregistrements GILS :

Les enregistrements GILS peuvent être utilisés pour décrire différents genres de ressources en information. Toutefois, la priorité doit toujours être accordée aux genres de ressources en information suivantes :

a- Produits de diffusion de l'information :

Les organismes devraient créer des enregistrements de localisation GILS pour décrire les produits de diffusion de l'information, comme les livres, les CD-ROM, les publications, les études, les rapports et les brevets, sans égard au support. Ces localisateurs (qu'il ne faut pas confondre avec le GILS lui-même) cataloguent ou décrivent les produits de diffusion de l'information. Par exemple, un enregistrement GILS pourrait exister pour un catalogue de publications ministérielles qui, lui, sert de localisateur pour ces publications.

b- Systèmes d'information automatisés :

Les enregistrements GILS doivent être créés pour décrire les systèmes d'information automatisés, surtout ceux auxquels le public a accès directement ou indirectement.

c- Ressources sur Internet:

Les enregistrements GILS doivent également servir à identifier et à décrire les ressources en information gouvernementale sur Internet, qu'il s'agisse de sites Web ou de documents particuliers. Les enregistrements GILS peuvent appuyer la recherche précise d'information gouvernementale sur Internet et aider les utilisateurs à savoir si les renseignements repérés sont à jour, exacts et authentiques.

GILS ET USMARC :

Le profil d'application GILS fournit des renvois entre les éléments de base du GILS et le format de catalogage lisible par machine (USMARC).

LISTE DES ÉLÉMENTS DE BASE DU GILS :

Tous les éléments sont facultatifs et non répétitifs par défaut, sauf mentions contraires :

TITRE (Obligatoire)
CRÉATEUR (Obligatoire, répétitif)
AUTEUR (répétitif)
DATE DE PUBLICATION(Obligatoire pour les publications ou les ressources possédant des dates de création ou de mise à jour distinctes)
DATE DE PUBLICATION -STRUCTURÉE
LIEU DE PUBLICATION
LANGUE DE LA RESSOURCE(Obligatoire, s'il y a lieu, répétitif)
RÉSUMÉ
INDEX IDÉOLOGIQUE NORMALISÉ (répétitif)
THÉSAURUS IDÉOLOGIQUE
TERMES NORMALISÉS
TERME NORMALISÉ (Répétitif)
TERMES NON NORMALISÉS
DOMAINE SPATIAL

COORDONNÉES DE DÉLIMITATION

THÉSAURUS DE MOTS CLÉS DE LIEU

MOT CLÉ DU LIEU(répétitif)

DURÉE (répétitif)

DATE DU DÉBUT

DATE DE LA FIN

DISPONIBILITÉ (Obligatoire, répétitif)

SUPPORT (Obligatoire)

DISTRIBUTEUR (Obligatoire)

DESCRIPTION DE LA RESSOURCE (répétitif)

TRAITEMENT DE LA COMMANDE (Obligatoire)

INFORMATION SUR LA COMMANDE (Obligatoire)

COUT

INFORMATION SUR LES COÛTS

PRÉ REQUIS TECHNIQUES

DURÉE DE DISPONIBILITÉ(répétitif)

DATE DU DÉBUT

DATE DE LA FIN

LIEN DISPONIBLE (répétitif)

TYPE DE LIEN

LIEN

SOURCES DE DONNÉES

MÉTHODOLOGIE

CONTRAINTES D'ACCÈS

CONTRAINTES D'ACCÈS GÉNÉRALES

CONTROLE DE CRÉATION OU DE DISSÉMINATION

CONTROLE DE LA COTE DE SÉCURITÉ

CONTRAINTES D'UTILISATION

POINT DE CONTACT

INFORMATION SUPPLÉMENTAIRE

BUT

PROGRAMME DE L'ORGANISME

RENVOI (répétitif)

RAPPORT DE RENVOI (répétitif)

LIEN DE RENVOI (répétitif)
TYPE DE LIEN
LIEN (répétitif)
NUMÉRO DE CALENDRIER
IDENTIFICATEUR DE CONTRÔLE (Obligatoire)
IDENTIFICATEUR DE CONTRÔLE INITIAL
SOURCE DE L'ENREGISTREMENT (Obligatoire)
LANGUE DE L'ENREGISTREMENT (Obligatoire)
DATE DE RÉVISION DE L'ENREGISTREMENT

- **Application du GILS dans l'administration canadienne** :

En novembre 1995, le Conseil du Trésor du Canada a créé le sous-groupe GILS SGG) du Groupe de travail sur les normes des documents électroniques auquel il a confié la responsabilité d'évaluer la possibilité de faire du GILS une *Norme du Conseil du Trésor sur la Technologie de l'Information* (NCTTI). Après avoir comparé le GILS à un certain nombre de standards de metadonnées servant à décrire l'information administrative et les ressources en information en général, le groupe a conclu que le GILS satisfaisait aux exigences des ministères fédéraux qui sont tenus d'identifier clairement leurs ressources en information et d'offrir un bon accès aux utilisateurs. Le Sous-Groupe GILS s'est réuni périodiquement en 1996 pour passer en revue l'actuel profil d'application GILS, formuler des suggestions pour l'adaptation de ce dernier, rédiger les lignes directrices relatives au GILS canadien, planifier un projet pilote sur le GILS et préparer l'ébauche de la NCTTI sur le GILS.

3.3) TEXT ENCODING INITIATIVE:

La TEI (Text Encoding Initiative) permet l'échange des données textuelles et d'autres types de données comme les images et les sons. Elle tire son origine d'une part de l'anarchie qui règne dans la communauté scientifique en matière de format, et d'autre part du nombre croissant de traitements que les chercheurs opèrent sur les textes sous forme électronique.

Les recommandations de la TEI fournissent le moyen de rendre explicites certaines caractéristiques d'un texte, de façon à faciliter son traitement par des programmes informatiques pouvant s'exécuter sur des plates-formes différentes. Cette tâche est appelée balisage ou codage.

Les recommandations s'appuient sur le langage SGML (Standard Generalized Markup Language) pour définir leurs règles de codage . Tous les outils SGML généralistes sont capables de traiter des textes conformes à la TEI.

Les recommandations de la TEI peuvent être appliquées aussi bien pour créer de nouvelles informations que pour échanger des informations existantes.

La TEI est soutenue par *l'Association for Computers and the Humanities, l'Association for Computational Linguistics et l'Association for Literary and Linguistic Computing*. Le projet a été en partie financé par *le National Endowment for the Humanities américain*, la DG XIII de la CEE, la fondation Andrew W. Mellon et le *Social Science and Humanities Research Council* du Canada.

Les recommandations ont été publiées en mai 1994, après six ans de travaux auxquels ont participé des chercheurs de toute nationalité et de toute discipline.

Au début de cette entreprise, les objectifs généraux de la TEI ont été définis par la résolution finale de la conférence préparatoire tenue au Vassar Collège de New York en novembre 1987. Cette résolution connue sous le nom de « Principes de Poughkeepsie » fut peu à peu précisée à travers une série de documents de travail. D'après ces documents les recommandations devaient :

- ✓ être suffisamment précises pour représenter les propriétés des textes intéressants pour les chercheurs;
- ✓ être simples, claires et concrètes;
- ✓ être utilisables facilement par les chercheurs et ne pas nécessiter l'utilisation de logiciels spécifiques;
- ✓ permettre une définition rigoureuse des textes en vue de traitements efficaces;
- ✓ être modifiables par l'utilisateur;

✓ respecter les normes en vigueur ou sur le point d'être adoptées.

Le monde de la recherche est large et divers. Pour que ces recommandations aient une large audience, il était important de s'assurer que :

1- les descriptions des caractéristiques fondamentales d'un texte puissent être facilement échangées;
2- les descriptions spécialisées puissent être facilement ajoutées (ou supprimées) d'un texte;
3- la même caractéristique puisse être encodée, en parallèle, de plusieurs façons;
4- la richesse du balisage puisse être déterminée par l'utilisateur de la façon la plus simple possible;
5- une documentation relative au texte et à la façon dont il a été codé soit fournie.

La TEI prévoit des mécanismes pour paramétrer le nom des balises et donc, si on le souhaite, utiliser des balises de son choix (par exemple des balises en français).

L'En-tête TEI :

Tout texte conforme à la TEI comporte :

1- une en-tête TEI (balisé comme un élément <teiHeader>)
2- la transcription du texte lui-même (balisé comme un élément <text>).

L'en-tête TEI contient des informations analogues à celles que l'on trouve sur la page de titre d'un texte imprimé. Elle contient quatre parties :

1- une description bibliographique du texte électronique;
2- une description de la manière dont il a été codé;
3- une description non-bibliographique du texte (le « profil » du texte);
4- un historique de révision.

Par conséquent, une entête TEI peut être assimilée à des metadonnées. Sa fonction est d'assurer que l'information nécessaire pour créer une notice catalographique soit facilement repérable. Les recommandations 'TEI' expliquent que les en-têtes TEI n'ont pas le même rôle que les notices MARC. Alors que les enregistrements MARC sont fondamentalement une version électronique d'une fiche de catalogue qui fait référence à un objet physique, les en-têtes TEI fournissent non seulement toute l'information catalographique, mais en plus toute l'information non bibliographique, déterminante dans le traitement du texte électronique.

L'en-tête TEI, avec ses zones descriptives, peut être facilement repérée et analysée par machine et assure un lien direct avec le texte décrit. Cependant les recommandations relatives à ces en-têtes n'ont pas le statut de normes, ce qui limite la généralisation de ces dernières.

3.4)Internet Anonymous Ftp Archives Templates : (IAFA)

IAFA Templates ont été crées pour décrire le contenu et les services d'archives Ftp (File Transfert Protocol), et ainsi, en faciliter l'accès.

A l'origine, IAFA Templates fut développé pour être utilisé avec le protocole Whois++, par le groupe de travail IAFA de l'IETF (Internet Engineering Task Force)et l'ébauche des directives de IAFA fut publiée en juillet 1995.

Des compagnies privées (BUNYIP, NEXOR,...) s'y sont investies et ont développé des outils de navigation et des répertoires de services. Le but visé par la conception d'IAFA Templates était de fournir un moyen que pourraient utiliser les administrateurs des archives Ftp pour décrire les diverses ressources qu'ils détiennent. Ces dernières peuvent être sous différents formats : images, textes, sons, services tels des listes de diffusion ou des bases de données, aussi bien que des archives de listes de diffusion ou de groupes Usenet et des logiciels. L'intention à l'origine était que les administrateurs de sites Ftp soient responsables de l'implantation de IAFA Templates pour chaque fichier contenu dans leurs archives. Ainsi, l'information serait repérable par tout individu, visitant le site. Il existe actuellement plusieurs applications utilisant IAFA Templates. Le système de recherche ALIWEB développé par NEXOR, est le premier outil de recherche

qui ‘pointe’ les archives Ftp décrites avec IAFA Templates. Un outil logiciel de recherche ROADS (Ressource Organisation And Discovery in Subject-based services)est venu utiliser IAFA Templates pour la description de ressources. Les auteurs ont incorporé dans la version de mai 1996 de ce logiciel, le protocole Whois++.

IAFA Templates est un standard de metadonnées d’implantation facile, avec des éléments de description très simples. Il y a différents types de gabarits définis dans les directives de l’IETF pour décrire les diverses ressources accessibles par le réseau :

- ✓ Documents;
- ✓ Lot de données ;
- ✓ Archives de listes de diffusion ;
- ✓ Archives de groupes de discussion ;
- ✓ Packages de logiciels ;
- ✓ Images.

D’autres types de gabarits ont été conçus dans le contexte des archives Ftp pour fournir de l’information sur des sites Ftp particuliers :

- ✓ Information sur des configurations de sites,
- ✓ Configuration logique des ‘archives’,
- ✓ Services (catalogue en ligne, informations sur serveurs),
- ✓ Miroir (détails des sites ‘miroirs’ incluant des informations sur la fréquence des mises à jour du site ‘source’).

Définition des gabarits :

a)Informations sur le Site :

Champs de ce gabarit :

- Nom du gabarit;
- Nom de l’hôte;
- Alias de l’hôte;
- Admin-(Utilisateur);
- Autre-(Organisation) ;
- Sponsor-(Organisation);
- Ville;
- Etat;

- Pays;
- Latitude-Longitude ;
- Fuseau horaire;
- Fréquence de mise à jour;
- Temps d'accès;
- Règles d'accès;
- Description;
- Mots clés ;
- Notes pour ce gabarit.

b) Information logique d'archive :

- Type de gabarit;
- Admin-(Utilisateur);
- Nom de l'hôte;
- Alias de l'hôte;
- Autre-(Organisation);
- Sponsor-(Organisation);
- Règles d'accès;
- Description;
- Fréquences de mises à jour;
- Mots clés.

c)Fichiers Automatiques de mise à jour de l'Information :

Un certain nombre de ces fichiers doit exister au niveau de l'archive.

d) Information contenue :

Pour cette catégorie, l'information dont il est question, est celle contenue dans l'archive, plutôt que celle disponible sur le serveur Ftp Anonyme.

e) Information Utilisateur :

Le gabarit, qui décrit cette information peut être stocké dans une place assurant un seul usage. Ce type de gabarit est désigné: 'USER'.

f)Information d'organisation :

D'une façon similaire que la précédente, le gabarit 'organisation' fournit une information commune qui doit orienter vers d'autres gabarits qui décrivent la source centrale d'information.

g) Information de service :

Les champs pour ce gabarit sont :

- Type de gabarit;
- Titre ;
- URI ;
- Admin-(Utilisateur) ;
- Autre-(Organisation) ;
- Sponsor-(Organisation);
- Description;
- Authentification;
- Enregistrement;
- Règles de chargement;
- Règles d'accès;
- Temps d'accès;
- Mots clés;
- Sujet-Descripteurs-Schéma;
- Sujet-Descripteurs;

h)Autres gabarits :

Les gabarits relatifs aux Documents, aux lots de données, aux archives de listes de diffusion, aux archives USENET, aux packages de logiciels, aux images et autres objets contiennent les mêmes champs mais des valeurs différentes pour le champs « type de gabarit » :

Type d'objet:	***Type de gabarit:***
Document:	DOCUMENT
Image:	IMAGE
Package de logiciels:	SOFTWARE
Archive des listes de diffusion:	ARCHIVE de MAILS
Archive de groupes de discussion:	USENET
Fichier son:	SON
Fichier Vidéo:	VIDEO
Fichier des FAQ:	FAQ

Donc nous aurons pour chaque gabarit les champs suivants :

Type de gabarit: (voir liste ci-dessus)

Catégorie: **Type d'objet :**

Titre: titre complet de l'objet

URI-V: description de l'accès à l'objet.

Titre court : partie du titre (s'il est long)

Auteur: description/contact information sur l'auteur/créateur de l'objet.

Admin-(Utilisateur): description/contact information sur l'administrateur de l'objet.

Source: information sur la source de l'objet.

Conditions: conditions d'utilisation de l'objet.

Description: description (résumé) de l'objet.

Bibliographie: bibliographie

Citation: citation de l'objet quand il est utilisé dans un autre travail.

Status-Publication: statuts de la publication courante (draft, pré-publication, etc.).

Editeur: éditeur(information)

Copyright: droit de copyright .

Date de création: date de création

Discussions: description (si possible) des forums de discussions relatifs à cet objet.

Mots clés: mots clés

Version-v: version désignée pour cet objet.

Format-v: formats dans lequel cet objet est donné.

Taille-v: taille(en octets).

Language-v: langue dans lequel, l'objet est écrit si c'est un document, ou langage de programmation si c'est un logiciel.

Caractères-v: jeu de caractères(ASCII ou "ISO Latin-1").

ISBN-v: *International Standard Book Number* de l'objet.

ISSN-v: *International Standard Serial Number* de l'objet.

Dernière date de révision-v: dernière date de révision.

Sujet-Descripteurs-Schéma-v: nom de la classification utilisée.

Sujet-Descripteurs-v: classification pour la ressource.

3.5) Summary Object Interchange Format: (SOIF)

Le standard ***SOIF*** (Summary Object Interchange Format) a été défini dans le cadre du Projet HARVEST,en Janvier 1994. Il est dérivé du standard ***IAFA Templates*** et du format bibliographique ***BIBTEX.***

Les enregistrements SOIF ont été conçus pour être générés par les outils du Projet HARVEST et permettre la recherche d'information auprès des brokers(fournisseurs d'information) HARVEST. Depuis le mois de Mars 1996, la campagnie Netscape Communications a annoncé qu'elle allait utiliser les enregistrements SOIF, dans ses produits.

Les gabarits SOIF peuvent être générées à la main par les auteurs ou ceux qui maintiennent les archives, bien que la majorité de ceux qui sont utilisés actuellement, soient générés automatiquement par des robots. SOIF est réellement un format d'enregistrement interne au projet HARVEST. Il a été conçu dans un but très spécifique (Indexation sommaire de ressources)mais le format de base peut-être étendu à d'autres besoins.

Les éléments descriptifs de base d'un gabarit SOIF sont :

- Résumé;
- Auteur;
- Description;
- Mots clés;
- Titre.

3.6) Metadata IMS: (Instructional Management System)

C'est un standard de metadata développé à l'université de l'Etat de Californie, en 1997, pour les besoins des personnes et organismes qui travaillent pour, ou entretiennent des relations avec le monde de l'éducation. Le standard IMS a un système de contenants (objets) types avec des ensembles de metadonnées bien définis. Les membres de chaque ensemble sont tirés d'un dictionnaire de champ de metadonnées commun. L'ensemble minimal de metadonnées, dont tout contenant IMS doit s'accommoder, est le core (IMS core) de 12 champs. Le module 'contenant' comporte 25 champs,et est d'un intérêt particulier pour la

création des ressources éducatives . Des outils de création de metadonnées ont été développés, pour faciliter la création des modules.

Le standard de metadonnées IMS comprend 35 champs tirés du Dictionnaire des Champs IMS :

- Résumé;
- Auteur;
- Catalogue ID;
- Concepts;
- Type de contenant;
- Crédits;
- Date d’expiration;
- Forme;
- Format;
- Guide;
- Niveau d’intéractivité;
- Mots clés;
- Language;
- Niveau d’apprentissage;
- Localisation;
- Version de la metadata;
- Objectifs;
- Pedagogie;
- Plate-forme;
- Pré-requis;
- Presentation;
- Code des prix;
- Relation;
- Rôle;
- Taille;
- Source;
- Distributeur;

- Structure;
- Sujet;
- Titre;
- Droits d'utilisation;
- Support de l'utilisateur;
- Temps d'utilisation;
- Date de la version;
- Version.

La plupart des champs de Metadonnées sont structurés, permettant des termes multiples, et des hiérarchies. Ceux-ci ont été définis en utilisant le Format de Définition de Ressource du W3C (RDF).

Les types de 'contenant' de metadonnées IMS :

Le système de metadonnées IMS possède différents types de contenants (objets) et définit un minimum pour chacun.

Tous les types incorporent un ensemble de cores de metadonnées, ***l'IMS core***. Ce dernier comprend quatre (4) types de contenants:

- L'article
- Le module
- Le profil
- L'outil

Les metadonnées IMS sont dérivées du Dublin Core. Nous y retrouvons d'ailleurs tous les éléments du Dublin Core, en plus d'autres éléments plus spécifiques.

3.7)Meta Content Framework : (*MCF*)

Le format d'échange MCF(Meta Content Framework)a été développé dans les laboratoires de la campagnie APPLE.

Le but de MCF est de fournir un langage de représenation du contenu d'un large éventail de ressources d'information. Sa particularité réside dans le fait

que les metadonnées ne sont pas placées dans des balises HTML ou SGML, mais elles sont extraites automatiquement et representées sous un format MCF.

Pour comprendre MCF, il est nécessaire de connaître :

1- les objets,les catégories, et les propriétés qui forment un bloc conceptuel dans MCF;
2- la syntaxe XML avec laquelle MCF peut-être stockée;
3- le modèle mathématique DLG (Directed Linked Graph), qui peut-être employé par des programmeurs informatiques pour développer efficacement des mise-en-oeuvres de MCF.

Propriétés, objets et Catégories :

MCF fournit l'information sur l'information en attachant des propriétés aux objets. Depuis toujours, dans un ordinateur, les objets sont structurés en mémoire, mais ils sont normalement employés pour représenter des choses telles que des pages Web, des entreprises, des peuples, des pays et des évènements. Par exemple, une page Web pourrait avoir une propriété qui donne sa taille, une autre qui donne son URL, et une autre qui identifie la personne qui l'entretient . Les propriétés sont utilisées pour donner de l'information sur ces objets. Nous distinguons entre des types de propriété et les propriétés. Un exemple d'un type de propriété est « sizeInBytes », qui pourrait s'appliquer à toute page Web. Quand il est appliqué à un objet particulier, il devient une propriété, un exemple du type de propriété, qui a une valeur: par exemple la page Web :http://www.textuality.com/ actuellement, a un « sizeInBytes » de 5,676 (propriété dont la valeur est : 5,676).

Il faut signaler que les types de propriété sont aussi des objets; cela signifie qu'ils peuvent eux aussi avoir des propriétés. Par exemple, le type de propriété qui donne la taille d'une page Web pourrait être nommée « sizeOfPage », il pourrait avoir une propriété qui s'appliquera aux pages Web, une deuxième fixera sa valeur (nombre), une troisième donnera le nombre en octets, et une quatrième fournira un texte explicatif .

Un objet peut avoir plus d'un type; par exemple, un objet représentant une personne pourrait être du type Docteur et du type Chercheur. La clé finale du concept est la categorie - dans l'exemple précédent, "Docteur" et "Chercheur" sont des catégories.

Utilisation de la syntaxe XML *:*

Les objets, propriétés, et catégories de MCF, nécessitent une syntaxe pour les stocker, avec une facilité de traitement manuelle ou par ordinateur. Cette syntaxe est basée sur XML (Extensible Markup Language). Dans XML, les documents contiennent des éléments, lesquels ont des types et sont l'un ou l'autre vides ou sont délimités par des start-tags et des end-tags (début de balise et fin de balise), et ont des attributs avec des noms et des valeurs, par exemple:

<p secret ="faux">. Cette phrase est dans le contenu d'un élément dont le type est 'p'; le contenu est placé entre le start-tag et le end-tag (début de balise et fn de balise). Le paragraphe a un attribut nommé "secret" dont la valeur est "faux".

Le Formalisme *Direct Linked Graphic* *:*

MCF est bâti sur un modèle mathématique. Mathématiquement, MCF est un graphique relié direct (DLG). Les objets sont représentés par des noeuds; les propriétés sont représentées par des arcs, qui sont des flèches reliant deux noeuds; les arcs ont des étiquettes, qui sont des types de propriété.

Les types d'éléments de metadonnées pris en compte dans MCF sont :Description, Auteur, Affiliation, Date de publication, Hyperliens, Langage, Sujets, Sites miroirs, Types de media(JPEG, MPEG, Poscript, Java Applet).

3.8- Metadata for Interchange of Files on Sequential Storage Media between File Storage Management Systems :(FSMS)

Ce standard, développé par *l'Association for Information and Image Management International (AIIM*), précise le format et le contenu de metadonnées pour l'échange de fichiers sur des médias amovibles à stockage

séquentiel. Ces derniers comprennent les bandes optiques et magnétiques et autres médias .

Le standard de metadonnées de FSMS permet, entre autres :

- l'échange de média à stockage séquentiel entre deux Systèmes de gestion de stockage de fichiers diffèrents;
- l'échange positionnel de média à stockage séquentiel,
- l'échange de média à stockage séquentiel entre des environnements manuels et automatisés,
- l'échange de média à stockage séquentiel entre des plates-formes materielles différentes,
- l'échange de média à stockage séquentiel entre des systèmes d'exploitation différents.

Le but de ce standard est de spécifier une voie de déscription de l'information supplémentaire ajoutée par un FSMS, avec d'autres informations, afin de permettre à un autre FSMS de lire ce que le premier a écrit et de reconstruire le fichier original comme il a été généré par l'application logicielle de départ.

Le standard de metadonnées du FSMS comporte une collection d'enregistrements. Chaque enregistrement consiste en une séquence de champs. L'enregistrement prend le nom du premier champ et les autres champs sont nommés champs d'informations ou simplement champs. Les champs d'information suivent aussitôt leurs champs de nom. Chaque enregistrement est soit nécessaire (obligatoire) soit sélectable (optionnel).

3.9) Climate and Environmental Data Retrieval and Archive:(*CERA*)

Le standard de metadonnées CERA a été conçu pour décrire des données climatologiques et écologiques, des modèles de données numériques, ainsi que des données d'observations. Il peut prendre en charge d'autres types de données spatiales (satellitaires).

Le standard de metadonnées CERA est basé sur le modèle de référence de IEEE. Le développement était fondamentalement guidé par l'intention de

garder ce standard de metadonnées aussi simple que possible, mais aussi flexible que nécessaire pour prendre en charge les éxigences des utilisateurs pour le système de base de données de climatologie et à incorporer des normes internationales de description de données comme DIF(Directory Interchange Format) et INFOCLIMA(World Climate Data Information Retrieval Service). INfOCLIMA est un thésaurus pour l'environnement et l'écologie réalisé par l'UNESCO.
L'usage de normes internationales dans les systèmes de base de données climatiques est nécessaire pour la communication future avec d'autres bases de données ou avec des systèmes d'information qui sont basés sur ces normes.

Il faut signaler que le standard de metadonnées CERA a été développé par l'Institut de Climatologie Allemand (DKRZ), et qu'il fait partie d'un ensemble d'outils et systèmes entrant dans la mise en place de la base de données climatiques et écologiques CERA.

3.10) The California Environmental Resources Evaluation System: *(CERES)*

Le standard de metadonnées ***CERES*** est constitué du même contenu de base que le ***CSDGM*** (standard de metadonnées du FGDC), mais le format varie dans les schémas de description.
Basé sur le langage SGML, sa DTD ne correspond pas en tout point non plus à la DTD du standard ***CSDGM***, dont pourtant il a pris ses sources en partie.
Les auteurs du standard de metadonnées CERES se sont inspirés aussi du draft(ébauche) du ***Content Standards for Non-Geospatial Metadata*** du National Biological Survey (NBS) et ont adopté une bonne partie des recommandations de l'American Institut of Biological Sciences relatives à ce standard.

Les rares modifications apportées ont été faites dans le but d'améliorer l'utilité, spécialement dans le contexte de la diffusion et de la recherche des données éléctroniques. Nous en énumérons les plus importantes :

- Modifications des éléments 'Citation Information', 'Contact Information'et des sous-éléments qui font référence au contenu des metadonnées;
- L'élément 'Citation Information' est remplacé par 'Citation Definition';
- Attribuer les noms des entrées des élèments définis par le FGDC (CSDGM)pour inclure des données aussi bien que pour limiter le nombre de balises requises pour la représentation (tag HTML);
- Réduction du rôle de quelques noms d'entrée d'élèments du FGDC vers SGML;
- Quelques entrées du FGDC ont été implémentées comme attributs de spécifications d'éléments dans des groupements de données;
- La méthode de création de site est spécifiée, mais un unique identificateur de metadonnées est publié;
- Introduction des pointeurs Nameloc de Hytime pour simplifier la réutilisation et l'inclusion de metadonnées;
- La notation des références pour permettre des enregistrements de données cohérentes avec les spécifications du standard CSDGM;
- Utilisation de courtes références pour simplifier la création de documents de metadonnées individuels;
- Définition d'éléments additionnels non spécifiés dans le CSDGM, et suggestion d'éléments dérivés du Draft Content Standard for Non-Geospatial Metadata;
- Réarrangement de quelques entrées type FGDC pour implémentation par compilateur SGML;
- Enlèvement de quelques entrées types FGDC(exemple :***Cloud_Cover***).

3.11) Multiple Resolution Landscape Characteristics: (*MRLC*)

Le MRLC Consortium a cherché a développer une approche unifiée sur les metadonnées, qui est conséquente des éxigences de metadonnées de chacun des programmes participants, et conforme au standard de metadonnées

développé par le ***Federal Geographic Data Commitee***. Le projet GAP(Gap Analysis Programm) a permis de développer un standard de contenu de metadonnées qui a servi de base pour les metadonnées du MRLC(Organisme rattaché au National Biologic Survey, lui-même rattaché au Ministère de l'intérieur US). Il s'en est suivi un document de base, redigé le 5 septembre 1994, et intitulé : 'Standard de metadonnées pour Gap Analysis'....

Avec des centaines de chercheurs à travers les USA contribuant à Gap Analysis Programm, le rôle pour les metadonnées devient capital: Elles doivent fournir un moyen d'accès sélectif aux données. Par exemple, il peut y avoir un besoin de chercher les enregistrements de GAP pour l'information sur la propriété terrestre dans un emplacement géographique spécifique. Un type d'information disponible dans les metadonnées est la description de l'étendue spatiale d'une base de données. Une question basée sur la latitude et la longitude peut produire une liste de tous les produits cartographiques de la région décrite.

Le standard de metadonnées GAP(du MRLC) est derivé du CSDGM(standard du FGDC) dont il maintient l'axe d'évolution afin de toujours assurer une certaine compatibilité entre eux.

Le format du standard GAP consiste en huit sections majeures de documentation contenant un ou plusieurs éléments de metadonnées. Chaque élément porte un nom. Le "Type" d'entrée (texte, entier, date, temps) et le champ de l'entrée sont aussi définis. Actuellement 276 éléments du CSDGM ne sont pas décrit dans le document de base qui définit le standard de GAP(standard de metadata MRLC).

Onze nouveaux éléments considérés nécessaires pour le projet GAP ont été ajoutés, sans qu'ils figurent dans le CSDGM.

Les données complètes, incluant des metadonnées et le dictionnaire des données, doivent être conçues pour le transfert numérique qui permettront aux metadonnées d'être distribuées isolément, cependant la base de données réelle doit toujours contenir des metadonnées.

3.12) Content Standard for Non-Geospatial Metadata:

Ce standard de metadonnées a été developpé par le National Biological Survey (NBS) pour le compte du National Biological Information Infrastructure (NBII), pour décrire l'information non géospatiale. Le NBS a utilisé l'approche suivante, dans la consultation avec ses partenaires, pour développer un standard de metadonnées :

- Le but est d'avoir un standard de metadonnées NBII, qui est essentiellement dérivé du CSDGM/USMARC, tout en étant le plus étendu possible;
- NBS a sollicité L'Institut Américain de Sciences Biologiques (AIBS) pour organiser un atelier des experts nationaux dans les sciences biologiques pour étudier et recommander des modifications à son projet de standard de metadonnées;
- NBS a procédé à une revue interne de la version beta (draft recommandé par l'atelier des experts) afin de déterminer l'utilité, la faisabilité et éventuellement apporter des modifications, si nécessaire;
- NBS a sollicité L'Académie Nationale des Sciences/ Conseil National de la Recherche pour un avis sur ce projet de standard de metadonnées.
- NBS a repassé en revue le standard renvoyé par l'Academie Nationale des Sciences pour d'autres modifications possibles. Le standard final a été présentée au Federal Geographic Data Commitee pour étude et adoption formelle comme standard FGDC.

Cette version finale est basée sur :

- Les définitions non spatiales par opposition aux données biologiques géospatiales;
- L'utilisation du CSDGM pour les données géospatiales;
- L'utilisation de USMARC pour cataloguer la portion des données non-spatiales.

3.13) Inter-university Consortium for Political and Social Research: (ICPSR)

Etabli en 1962, le consortium inter-universtaire pour la recherche politique et sociale(ICPSR) est une organisation du type associative, permettant d'accéder aux plus grandes archives du monde, dans le domaine des sciences sociales.
ICPSR fournit des équipements et des services pour la communaute internationale de ses membres.
Les données d'exploitation couvrent un large intervalle de disciplines, telles les sciences politiques, la sociologie, la démographie, les sciences économiques, l'histoire, l'éducation, la gérontologie, le droit, et la santé publique.
ICPSR encourage les sociologues dans tous les domaines à contribuer à l'enrichissement de ses bases de données et à utiliser ses ressources de données.
ICPSR inclut parmi ses membres plus de 325 universités en Amérique du Nord et plusieurs centaines d'établissements répartis en Europe, en Australie, en Asie et en Amérique Latine. Le siege social est situé dans l'Institut de Recherche Sociale de l'Université de l'Etat du Michigan (USA).

ICPSR ne dispose pas à proprement parler d'un standard de metadonnées, mais uniquement d'une sorte de formulaire, que tout dépositaire d'informations dans leurs bases de données ou d'archives, se doit de remplir pour décrire son document ou ce qui le constitue. Ces formulaires appelées : *Data Deposit Form,* exige de l'auteur ou du producteur de données de fournir des information sur les caractéristiques techniques et référentielles d'une collection de données. En complétant ce formulaire de 24 champs (simple description de texte a remplir manuellement), l'auteur ou le producteur assure que la collection de données sera exactement et entièrement décrite pour des analystes secondaires potentiels dans la communauté des sciences sociales.
Le *Data Deposit Form* permet au personnel d'ICPSR de préparer une description d'étude (souvent considérée comme une metadonnée) pour divers publications d'ICPSR ainsi que pour ses bases de données en réseau. Cette information permet aussi la création de citations bibliographiques d'autorité pour la collection.

3.14)National Environmental Data Referral Service : (NEDRES)

Le National Environmental Data Referral Service, NEDRES, est un programme de la *National Oceanic and Atmospheric Administration (NOAA).*
Les utilisateurs peuvent accéder à une large gamme d'informations écologiques par des repertoires NEDRES, offerts en réseau(en ligne). NEDRES est un réseau auquel coopèrent des organisations fédérales, étatiques, académiques et privées, pour améliorer l'accès aux données écologiques. La base de données NEDRES est un catalogue informatisé de données écologiques. Ces données concernent le rayonnement du soleil à travers l'atmosphère, sur les terres et les océans, La physique solaire et de l'atmosphère supérieure , l'océanographie, la climatologie, la météorologie, la pollution, les substances toxiques, la géophysique et la géologie, la géochimie.
Le répértoire *NEDRES* contient seulement les descriptions, et non pas les données réelles, et renvoit l'utilisateur à la source pour une information intégrale.
La base de données fournit des documents sur plusieurs types de description d'information écologique. Elle inclut:

1- les données des centres, programmes et organisations;
2- les fichiers de données non imprimés ;
3- les données des publications séries ;
4- les données publiées ;
5- les atlas ou les données publiées sous la forme graphique;

6- les publications contenant des compilations extensives, des analyses extensives ou des applications de données ;
7- les manuels, les guides de l'utilisateur, ou la documentation des lots de données ;
8- les catalogues de données, les inventaires, ou les bibliographies ;

Une recherche dans la base de données NEDRES fournit une description complète des sources de données disponibles qui satisfont les spécifications

de recherche. Les informations obtenues décrivent les données dans le détail. Les champs de définition sont:

AN Accession Number

TI Title

AB Abstract

DC Data Collection

DD Data Processing and Quality Control

PE Period of Record (earliest date and latest date)

LR Length of Record

GE Geographic Place Name

GC Geographic Codes (FIPS, Ecoregion,Water,Resource)

GL Geographic Grid Locators

PA Parameter Description

DE Descriptors of the data

CO Contact address to obtain the data

AV Availability Characteristics

PI Principal Investigators

PR Program or Project

PO Processing Organization

PR Related Publications

RR Related Records

DT Date Entered and Date Reviewed

CC Category Codes

Les chercheurs et scientifiques qui utilisent la base de données NEDRES sont de tous les horizons, incluant des institutions académiques, privées, associatives, et gouvernementales.

Le format du catalogue NEGRES est crée selon le standard CSDGM du FGDC, dont il est finalement une application particulière. D'ailleurs un manuel : 'Content Standard for Digital Geospatial Metadata Workbook (version 1.0)', est prévu à cet effet.

3.15)DUBLIN CORE :

3.15.1)Introduction :

En Mars 1995 s'est tenu un workshop sur les Metadonnées, parrainé par ***Online Computer Library Center*** (***OCLC***) et ***le National Center for Supercomputing Applications (NCSA),*** rassemblant 52 chercheurs et professionnels des bibliothèques, de l'informatique, et des spécialités connexes, pour faire avancer l'état de l'art dans le développement des descriptions de ressources électroniques. Les buts de ce workshop incluaient une compréhension commune des besoins, des points forts, des défauts, et des solutions à proposer; et l'atteinte d'un consensus sur un ensemble d'éléments de metadonnées pour décrire des ressources d'informations. La complexité du problème des descriptions de ressources a fait limiter l'étendue des discussions. Sachant que la majorité des informations diffusées sont sous forme de documents, et que les enregistrements de metadonnées sont requis pour faciliter la recherche de ressources sur Internet, l'ensemble proposé d'éléments de metadonnées du Dublin Core est destiné à décrire les caractéristiques essentielles de documents électroniques. D'autres enregistrements de metadonnées tels ceux décrivant des informations de coût ou d'archives n'ont pas été pris en considération. Il fut reconnu, néanmoins que ces éléments pourraient être inclus dans une future version du Dublin Core.

Le Dublin Core n'est pas destiné à supplanter les autres descriptions de ressources, mais plutôt à les compléter. Il y a actuellement deux types de descriptions de ressource pour des documents éléctroniques diffusés: les index automatiquement générés tels ceux employés par Lycos et WebCrawler; et le catalogage des enregistrements (USMARC). Générés automatiquement, les enregistrements contiennent souvent un minimum d'information pour être utiles, tandis que générés manuellement, ces enregistrements sont trop coûteux à créer et à entretenir pour le grand nombre de documents électroniques actuellement disponibles sur Internet. Les enregistrements du Dublin Core sont destinés à remplacer ces extrêmes, fournissant un enregistrement structuré simple et qui peut être amélioré.

Le travail du workshop de 1995 est une premiere étape qui devait être suivi d'autres pour améliorer au fur et à mesure, la description des informations. Un comité, issu du premier workshop, a été formé pour suivre le travail par une série d'activités similaires pour faire évoluer le Dublin Core.

Depuis qu'Internet contient plus d'information complète que de simples résumés professionnels, les indexeurs et catalogueurs tentent de gérer cela en utilisant les méthodes et systèmes existants: Il était évident qu'une alternative pour obtenir des metadonnées utilisables pour des ressources électroniques doit donner aux auteurs et aux fournisseurs d'information un moyen permettant de décrire les ressources eux-mêmes, sans formation intensive et spécifique préalable.

C'est pour atteindre ce but, que la tâche majeure du workshop sur les metadonnées, était d'identifier et de définir un ensemble simple d'éléments pour décrire des ressources électroniques diffusées(en réseau).

La discussion, lors du workshop, était davantage limitée aux éléments de metadonnées nécessaires pour la découverte de ce qui était appelé 'Document-Like Objets'(objets documents) ou DLOs.

DLOS n'étaient pas rigoureusement définis, mais étaient compris par l'exemple. Par exemple, une version électronique d'un article de journal ou un dictionnaire est un DLO, tandis qu'une collection de diapositives ne l'est pas. Bien sûr, le coeur du problème est que dans un environnement réseau, les DLOs peuvent être arbitrairement complexes parce qu'ils peuvent consister en un texte avec des images, des clips de vidéo, du son, ou des documents hypertextuels. Les participants au workshop n'ont pas essayé de limiter la complexité des DLOs, sauf de remarquer que le contenu intellectuel d'un DLO est principalement du texte, et que les metadonnées susceptible de décrire des DLOs ont une forte ressemblance avec celles qui décrivent des textes imprimés traditionnels.

Donc, finalement les participants devaient définir un ensemble d'éléments de metadonnées qui permettraient aux auteurs et fournisseurs d'information de décrire leur travail et à faciliter l'intéropérabilité parmi les outils de recherche d'information.

La première version de la synthèse des travaux du workshop établissait un ensemble minimal de treize éléments de metadonnées qui fut nommé : ensemble d'éléments de core de metadonnées de Dublin (ou plus simplement Dublin Core).

La syntaxe fut laissée vague délibérément et assimilée à un détail de mise en oeuvre. La sémantique de ces éléments était assez claire pour être comprise par une large gamme d'utilisateurs.

Les discussions entre les participants au workshop ont permis de révéler plusieurs principes qui devaient guider davantage le développement de l'ensemble des éléments. Ces principes sont: propriété intrinsèque, extensibilité, indépendance de syntaxe, optionalité, répétabilité et modifiabilité.

Propriété intrinsèque:

Le Dublin Core a été dirigé, dés le départ pour décrire des propriétés intrinsèques de l'objet. Par exemple, l'élément "Sujet" est une donnée intrinsèque, tandis que des informations de transaction telle que coût et droit d'accès sont des données extrinsèques.

Extensibilité :

En plus de son emploi en traitant de la propriété intrinsèque des données, le mécanisme d'extension permettra l'inclusion de données intrinsèques pour des objets qui ne peuvent pas être décrits suffisamment par un petit ensemble d'éléments.

L'extensibilité est importante parce que les utilisateurs peuvent vouloir ajouter des descriptions supplémentaires à des champs, même ayant des buts spécifiques. En outre, la spécification du Dublin Core lui-même peut changer dans le temps, et le mécanisme d'extension permet des révisions.

Indépendance de Syntaxe :

Les constructions syntaxiques sont évitées parce qu'il est trop tôt de proposer des définitions formelles et parce que le Dublin Core est destiné à être employé tôt ou tard dans une large gamme de programmes d'application et de disciplines.

Optionalité :

Tous les éléments sont optionnels, pour deux raisons :

La première est que le Dublin Core peut tôt ou tard être appliqué aux objets pour lesquels certains éléments n'ont pas de signification : Qui est l'auteur d'une image satellitaire?

La seconde est qu'il semble futile de donner des descriptions complexes lorsque les auteurs du contenu prévoient de fournir la matière descriptive.

Répétabilité :

Tous les éléments du Dublin Core sont répétables. Par exemple, plusieurs éléments 'auteur' seraient employés quand une ressource a plusieurs auteurs.

Modifiabilité :

Chaque élément dans le Dublin Core a une définition qui est sensée être évidente. Cependant, il est aussi nécessaire que les définitions des éléments satisfassent aux besoins de communautés différentes. Ce but est accompli en permettant à chaque élément d'être modifié par un qualificateur optionnel. Si aucun qualificateur n'est proposé, l'élément prend son sens commun.

Les qualificateurs seront typiquement dérivés de conventions connues dans la communauté des bibliothèques ou du domaine des connaissances propres à la ressource. Les qualificateurs sont importants parce qu'ils donnent au Dublin core, un mécanisme pour supprimer ou amoindrir l'écart entre des spécialistes et des utilisateurs quelconques.

Par exemple, les données dans l'élément 'sujet' consistent en mots ou expressions qui décrivent le contenu de l'objet. Cependant, un professionnel du catalogage peut vouloir se référer à la source d'où les termes du sujet sont pris. Dans un tel cas, l'élément peut être écrit ainsi :

Sujet(scheme=LCSH), indiquant que les termes de sujet sont pris de la liste: ***Library of Congress Subject Headings***.

3.15.2)Liste des éléments du Dublin Core :

La liste actuelle des éléments, ainsi que leur définition générale, a été établie en décembre 1996. Elle comporte 15 éléments, alors que le premier workshop avait définit 13 éléments seulement.

Description des éléments *:*

1.Titre :

Etiquette: titre

Le nom donné à la ressource par le créateur ou l' auteur.

2.Auteur ou Créateur :

Etiquette: créateur

La personne ou l'organisation principalement responsable de la création du contenu intellectuel de la ressource.

3.Sujet et mots-clef :

Etiquette: sujet

Le sujet de la ressource. Typiquement, le sujet sera décrit par un ensemble de mots-clefs ou de phrases qui précisent le sujet ou le contenu de la ressource.

4.Description :

Etiquette: description

Une description textuelle du contenu de la ressource, y compris un résumé, dans le cas d'objets tels que des documents,ou une description du contenu dans le cas de ressources visuelles.

5.Editeur :

Etiquette: editeur

L'entité responsable de la diffusion de la ressource dans sa forme actuelle, telle qu'une maison d'édition, un département universitaire, une entreprise.

6.Autre contributeur :

Etiquette: contributeur

Une personne ou une organisation, non mentionnée dans un élément créateur, qui a fait une contribution intellectuelle significative à la ressource mais dont la contribution est secondaire comparée à celle de toute personne ou organisation spécifiée dans un élément créateur (par exemple un rédacteur, un traducteur, un illustrateur).

7.Date :

Etiquette: date

La date à laquelle la ressource a été publiée dans sa forme actuelle. L'usage recommandé est sous la forme d'un nombre de 8 chiffres tel que YYYY-MM-DD, comme défini par la norme ISO8601. Dans ce schéma, l'élément date 1994-11-05 correspond au 5 Novembre 1994. Beaucoup d'autres schémas sont possibles, mais si un autre schéma est utilisé, il devrait être précisé de façon non ambigue.

8.Type de ressource :

Etiquette: type

La catégorie de la ressource, telle qu'une page personnelle, un roman, un poème, un document de travail, un rapport technique, une dissertation ou un dictionaire.

9.Format :

Etiquette: format

Le format de la ressource, utilisé pour identifier le logiciel et, éventuellement, le matériel qui peuvent être nécessaires pour afficher ou traiter la ressource.

10.Identificateur de la ressource :

Etiquette: identificateur

Chaîne de caractère ou nombre utilisé pour identifier de façon unique la ressource. Exemples pour des ressources réseau incluent URLs et URNs (si implementé). D'autres identificateurs globaux et uniques, tels que ISBN (International Standard Book Numbers), ou d'autres noms formellement définis, sont des candidats potentiels pour cet élément,

dans le cas de ressources privées.

11.Source :

Etiquette: source

Une chaîne de caractère ou un nombre, utilisé pour identifier de façon unique le travail d'où la ressource est dérivée, si applicable. Par exemple une version PDF d'un roman peut avoir un élément source contenant un numéro ISBN correspondant à la version physique du livre à partir de laquelle la version PDF a été réalisée.

12.Langage :

Etiquette: langage

Langage(s) du contenu intellectuel de la ressource. Si approprié, le contenu de ce champ devrait correspondre à la norme RFC 1766.

13.Relation :

Etiquette: relation

Les relations de cette ressource avec d'autres ressources. Le but de cet élément est de fournir un moyen d'exprimer les relations formelles entre des ressources qui existent aussi en temps que ressources indépendantes. Par exemple des images dans un document, les chapitres d'un livre, ou les élément d'une collection.

14.Couverture :

Etiquette: couverture

Definit les caractéristiques spatiales et/ou temporelles de la ressource.

15.Gestion des droits :

Etiquette: droits

Un lien sur les droits de reproduction, les droits d'utilisation, ou renvoi à un service capable de fournir l'information sur les conditions d'accès à la ressource.

3.15.3)Les qualificateurs du Dublin Core :

Nous reprenons ici un ensemble de qualificateurs pour le Dublin Core, qui ont été proposés par REBECCA GUENTER (spécialiste d'USMARC à la Bibliothèque du Congrès US). Le document a été mis à jour à la suite de discussions qui ont eu lieu lors du cinquième workshop qui s'est tenu a Helsinki en octobre 1997. Il traite des qualificateurs "shéma" et "type". Un qualificateur "shéma" est employé pour interpréter la valeur du contenu et est généralement basé sur des normes externes. Un qualificateur "type" affine la définition de l'élément de données lui-même.

La présente liste de qualificateurs est une approche intermédiaire entre les 'Minimalistes' (partisans du principe de n'utiliser qu'un minimum de qualificateurs et de laisser une certaine liberte aux auteurs)et l'approche des 'Structuralistes', qui voudraient fixer toutes les règles au départ (définition de tous les qualificateurs possibles et obligatoires). Il faut remarquer que malgré cela,tous les qualificateurs restent encore optionnels.

Liste des qualificateurs du Dublin Core :

1. Titre :

Shéma: non nécessaire

TyPE:

- DC.TITRE
- DC.TITRE.ALTERNATIF (employé pour tout titre autre que le titre principal incluant le sous-titre, etc.)

2. L'auteur ou Créateur :

Shéma:

- LCNAF (Library of Congress Name Authority File)

TyPE:

- DC.CREATOR
- DC.CREATOR.Nom Personne

- DC.CREATOR.Nom Campagnie
- DC.CREATOR.Nom Personne(inclut tout type d'adresse, ou email)
- DC.CREATOR.Nom Compagnie(inclut adresse)

3. Sujet et Mots-clés :

Shema:

- (Non qualifie:Mots-clés pris par défault)
- LCSH (Library of Congress Subject Headings)
- MeSH (Medical Subject Headings)
- AAT (Art and Architechture Thesaurus)
- LCNAF (Library of Congress Name Authority File)
- DDC (Dewey Decimal Classification)
- LCC (Library of Congress Classification)
- NLM (National Library of Medicine Classification)
- UDC (Universal Decimal Classification)

TyPE: non nécessaire

4. La description :

Shéma:

- Le résume est pris par défaut
- URL

TYPE: non nécessaire

5. L'éditeur :

Shéma: non nécessaire

TYPE:

- DC.PUBLISHER (non qualifié)
- DC.PUBLISHER.NOM PERSONNE
- DC.PUBLISHER.NOM COMPAGNIE
- DC.PUBLISHER.NOM PERSONNE+ADRESSE
- DC.PUBLISHER.NOM COMPAGNIE+ADRESSE

6. Autre Contributeur :

Shéma:

LCNAF (Library of Congress Name Authority File)

TYPE:

DC.CONTRIBUTOR (non qualifié)

DC.CONTRIBUTOR.NOM PERSONNE

DC.CONTRIBUTOR.NOM COMPAGNIE

DC.CONTRIBUTOR.NOM PERSONNE+ADRESSE

DC.CONTRIBUTOR.NOM COMPAGNIE+ADRESSE

7. Date :

Shéma:

- ISO 8601 par defaut
- ANSI X3.30
- IETF RFC 822
- Autres?

TYPE:

- DC.DATE.CREATION_du_contenu intellectuel
- DC.DATE.CREATION/Modification_de_la presente forme
- DC.DATE.FORMAL_PUBLI
- DC.DATE.ACCESSIBLE
- DC.DATE.VALIDE (inclut la vérification)
- DC.DATE.ACQUISITION/Accession
- DC.DATE.ACCEPTEE
- DC.DATE.DATAGATHERIN(Collecte de Données)

8. Type de Ressource :

Shéma: non nécessaire

Type :La liste de type de ressources est en voie de développement.

9. Format :

Shéma:

- IMT (i.e. MIME)

- DCPMT (Dublin Core Physical Medium Type))

Type: non nécessaire

10. Identificateur De ressource :

Shéma:

- URL est pris par défaut
- URN (Uniform Resource Name)
- ISBN (International Standard Book Number)
- ISSN (International Standard Serial Number)
- SICI (Serial Item and Contribution Identifier)
- FPI (Formal Public Identifier)

Type: Non nécessaire

11. La source :

Shéma:

- Le texte libre est pris par défaut
- URL
- URN
- ISBN
- ISSN

Type: non nécessaire

12. La langue :

Shéma:

IETF RFC 1766

Z39.53

ISO 639-1

ISO 639-2/B (après la publication finale)

Type: Non nécessaire

13. La relation :

Shéma :

- texte libre est pris par défaut
- URL
- URN
- ISBN

Type:

- Créatif (e.g. traduction, annotation)
- Mécanique (copier,changement de format etc...)
- Version (edition, draft)
- Inclusion (collection, part)
- Référence (citation)

14. La couverture :

Shéma:

Sera déterminé par le Groupe de travail sur Coverage (Couverture)

Type:

La liste suivante a été détérminée par le Groupe de travail sur Coverage :

DC.COVERAGE.PERIODNAME

DC.COVERAGE.PLACENAME

DC.COVERAGE.T

DC.COVERAGE.X

DC.COVERAGE.Y

DC.COVERAGE.Z

DC.COVERAGE.POLYGON

DC.COVERAGE.LINE

DC.COVERAGE.3D

15. gestion des droits :

Shéma:

Texte libre pris par défaut

URL

URN

Type: Non nécessaire.

3.15.4)Syntaxe du Dublin Core :

Pour transcrire les éléments du Dublin Core en un document électronique, il est évident qu'il est nécessaire de disposer d'une syntaxe spécifique.

Actuellement, le web constitue l'outil stratégique d'Internet. La communauté du Dublin Core a commencé donc par préconiser l'utilisation du langage HTML

(voir le site: http://www.oclc.org:5046/~weibel/html-meta.html)

La convention proposée est la suivante:

<META NAME = "schema_identifier.element_name.qualifier" CONTENT = "string data">

ou :

NAME: Cet attribut précise un nom de propriété.

CONTENT: Cet attribut précise la valeur d'une propriété.

SHEMA: Cet attribut designe un shema de description à employer pour interpréter la valeur de la propriété.

HTTP=: Cet attribut peut être employé à la place de l'attribut de nom.

L'élément META peut être employé pour décrire des propriétés d'un document (auteur, date d'expiration, une liste de mots-clés, etc.)et attribuer des valeurs à ces propriétés. Cette spécification ne définit pas un ensemble normatif de propriétés.

Exemples :

1-Titre:

<META NAME="DC.title" CONTENT="Standards de metadonnées">

<LINK REL=SCHEMA.dc HREF=

"http://purl.org/metadata/dublin_core_elements#title">

<META NAME="DC.title.subtitle" CONTENT="Dublin Core">

2-Auteur or Créateur :

<META NAME="DC.creator" CONTENT="Amerouali Y">

<META NAME="DC.creator.email"

CONTENT="ameroual@enssib.fr">

3-Sujet et mots-clés :

<META NAME="DC.subject" CONTENT="Dublin Core, metadata, ressource électronique">

<META NAME="DC.subject" CONTENT="(scheme=LCSH) Catalogage de ressources éléctroniques">

<META NAME="DC.subject" CONTENT="(scheme=UDC) 518.118">

4-Description:

<META NAME="DC.description" CONTENT="Description textuelle du contenu de la ressource, incluant éventuellement un résumé.">

5-Editeur:

<META NAME="DC.publisher" CONTENT="NORDINFO">

6-Autres Contributeurs:

<META NAME="DC.contributors" CONTENT="Monia Lind">

7-Date:

<META NAME="DC.date" CONTENT=
"(scheme=ANSI.X3.30-1985) 19980220">

<META NAME="DC.date.current"
CONTENT="(scheme=IETF.RFC-822) Thu, 11 Mars 1998 21:12:34 +0100">

8-Type de ressource:

<META NAME="DC.type" CONTENT="Note de synthèse">

9-Format:

<META NAME="DC.format" CONTENT="(SCHEME=imt) text/html">

<LINK REL=SCHEMA.imt HREF=
"http://sunsite.auc.dk/RFC/rfc/rfc2046.html">

10-Identificateur de la ressource :

<META NAME="DC.identifier"CONTENT="http://www.ncl.ac.uk/~napm1/dublin_core/index.html">

<META NAME="DC.identifier.ISBN" CONTENT="1-56884-452-2">

11-Source :

<META NAME="DC.source" CONTENT="Levine and Baroudi: Internet secrets">

<META NAME="DC.source.ISBN" CONTENT="1-56884-452-2">

12-Language :

<META NAME="DC.language" CONTENT="(SCHEME=NISOZ39.53) FRE">

13-Relation :

<META NAME="DC.relation.IsDerivedFrom" CONTENT="http://www.oclc.org:5046/oclc/research/conferences/metadata2/">

14-Couverture :

<META NAME="DC.coverage.local" CONTENT="Scandinavie">

15-Gestion des droits :

<META NAME="DC.rights" CONTENT="Domaine public">

3.15.5)Structure de Description de Ressources : (RDF)

La Structure de Description de Ressources est un cadre pour les metadonnées ; qui permet l'intéropérabilité entre des applications qui échangent l'information compréhensible par machine sur le Web. Elle améliore les facilités de traitement automatisé des ressources d'internet. La RDF peut être employée dans des applications diverses; par exemple: dans la recherche d'information en permettant de meilleures capacités aux outils de recherche; dans le catalogage pour décrire le contenu et les rapports avec les contenus disponibles dans un site Web

particulier, une page Web, ou une bibliothèque numérique; par des agents logiciels intelligents, pour faciliter le partage et l'échange de connaissances.

Le groupe de travail qui suit les développements de la RDF est sous les auspices du W3C. La RDF a commencé comme une extension des éléments de description définis par le W3C pour PICS(Plateform for Internet Content Selection). Elle intervient maintenant aussi dans le langage XML(produit XML de Microsoft et XML/MCF de Netscape). Les orientations du Dublin Core discutées à WARWICK ont eu une certaine influence aussi sur les développements de la RDF(version du 16 Fevrier 1998).

3.15.6)Les projets[2] utilisant le Dublin Core :

Projets Australiens :

1-DSTC :

Page d'accueil : http://www.dstc.edu.au/RDU/

Le DSTC participe dans le Groupe de travail du W3C sur la RDF. Le DSTC compte développer une spécification et des outils pour l'utilisateur pour créer des éléments de Dublin Core compatibles avec la RDF et fournir des interfaces de recherche.

2-AGCRC : (Australian Geodynamics Cooperative Research Center)

Page d'accueil : http://www.agcrc.csiro.au/

L'AGCRC, qui est une collaboration entre deux organismes publiques de recherche et deux universités, utilise le Web comme un premier système de publication des résultats de ses recherches. Le projet utilise deux standards de metadonnées différents pour le texte et les données numériques,avec un lien entre eux.

3-Projet de Metadonnées de la Bibliothèque d'état de Queensland :

Page d'accueil: http://www.slq.qld.gov.au/meta/overview.htm

Le but du Projet est la prise en compte de metadonnées dans la page

[2] Je cite sommairement les projets les plus connus, avec des URL de renvoi pour plus d'informations.

Web de la Bibliothèque D'état de Queensland. C'est aussi une tentative initiale pour établir des normes de déploiement de metadonnées dans les bibliothèques du Queensland.

4-Le projet PANDORA :

Page d'accueil: http://www.nla.gov.au/politique/pandje97.html

La Bibliothèque Nationale d'Australie (NLA) a engagé un projet de développement d'un système de gestion d'archives éléctroniques appelé PANDORA, pour fournir l'accès en ligne, à long terme, à de significatives publications Australiennes. Le projet intégrera un système de déscription des documents archives, basé sur le Dublin Core.

5-EdNA (Education Network Australia) :

Page d'accueil: http://www.otfe.vic.gov.au/edna/dc5edna.htm

EdNA est un projet collaboratif entre tous les territoires et Etats Australiens et tous les secteurs de l'éducation et de la fomation. EdNA utilise un standard de metadonnées basé sur le Dublin Core.

6-Le secteur de l'environnement Australien :

Page d'accueil: http://www.environment.gov./

Le secteur de l'environnement en Australie utilise le Dublin Core pour les informations qu'il met sur le Web et dans son intranet. Le service d'information en réseau du secteur de l'environnement Australien propose plus de 8000 documents en ligne, répartis sur plusieurs bases de données.

Projet Canadien :

7-SearchBC: Vancouver Webpages

page d'accueil: http://vancouver-webpages.com/VWBOT/ searchBC.html

Ce projet consiste en un robot pour parcourir le web, un moteur de recherche de base de données, et un générateur de scripts de metadonnées.

Projet Français :

8-MedExplore & Metadata :

Page d'accueil:

http://www.loria.fr/%Educloy/PUB/Publi97/DC5/DC5.HTM

L'objectif de MedExplore est de permettre à un utilisateur de naviguer grâce à un concept de graphes et à manipuler divers morceaux d'information de grandes bases de données internationales, des documents de sources locales et des informations tirées d'Internet.

Projets Allemands :

9-Metadaten-Projekt : (Projet de metadata)

Page d'accueil : http://www2.sub.uni-goettingen.de

Ce projet explore l'emploi des metadonnées d'un point de vue des bibliothèques. Il vise la transmission du savoir-faire accumulé dans la recherche de source d'information en réseau dans le monde, vers les bibliothèques Allemandes. Il constitue une partie d'un plus grand projet impliquant plusieurs bibliothèques Allemandes.

10-SSG-FACHINFORMATION (SSG-FI) Mathematick :

Page d'accueil: http://www.sub.uni-goettingen.de/ssgfi

Les metadonnées sont générées pour l'évaluation de l'information relative aux mathématiques. Les sources incluent des serveurs Internet, des CD-ROMS, et des référence de livres.

11-SSG-FACHINFORMATION (SSG-FI)Geowissenschaften

Page d'accueil : http://www.sub.uni-goettingen.de/ssgfi/

Ce service d'information est hébérgé par le même serveur que le précédent, mais il est orienté vers les sciences de la terre. Le projet de metadonnées leur est commun.

12-Deutscher Bildungs - Serveur : (secteur de l'éducation)

Page d'accueil: http://dbs.schule.de/indexe.html

Ce site contient actuellement près de 2000 documents sur l'éducation.

13-Math-Net :

Page d'accueil: http://elib.zib.de/math-net/

Ce projet concerne les metadonnées crées par les auteurs eux-même, pour des documents relatifs aux mathématiques.

14-Gestion de l'information électronique et Metadata en physique :

Page d'accueil:

http://www.physik.uni-oldenburg.de/EPS/EurophysNet/PhysDep/dep-links.html

Le but de ce projet est de développer un système distribué d'information électronique en physique. La première version de ce système est basée sur Harvest (modèle US).

15-Bibliothèque Electronique De visualisation :

Page d'accueil: http://visinfo.zib.de/Bibliothèque/

La Bibliothèque Electronique de visualisation (EVlib) est un service distribué de publications électroniques.

1***6-Bibliotheksservice-Zentrum(BSZ)Baden-Wuerttemberg (Sudwestdeutscher Bibliotheksverbund-SWB-VERBUND)*** :

(Réseau de Bibliothèques dans le Sud-ouest Allemand)

Page d'accueil:

http://www.swbv.uni-konstanz.de/wwwroot/s71800_d.html

Ce réseau de bibliothèque fonctionne avec une base de données bibliographiques pour la région, de 16000 000 de références, et incluant des documents électroniques avec l'accès en ligne.

Les Pays-Bas :

17-Koninklijke Bibliotheek :Bibliothèque Nationale des Pays-Bas

Page d'accueil: http://www.konbib.nl:8000

La Bibliothèque Nationale des Pays-Bas est en train de développer une

nouvelle version de son service d'information sur le Web. C'est une volonté de changement avec des nouvelles caractéristiques de fonctionnement et l'incorporation des éléments de metadonnées du Dublin Core dans les pages HTML.

La Scandinavie :

18-Le Projet de Metadata Nordique :

Page d'accueil: http://linnea.helsinki.fi/meta/

Dans les pays Nordiques, il y a un besoin spécial pour un système de création de metadonnées, pour une meilleure exploitation des documents en commun. Le Dublin Core est employé pour fournir et améliorer les services pour l'utilisateur final, en permettant une recherche plus efficace et un meilleur accès aux documents numériques.

La Suéde :

19-Le projet EnviroNet Suédois :

Page d'accueil: http://smn.environ.se/smnproj/proj/summary.htm

EnviroNet est un projet du gouvernement Suédois qui est perçu comme une voie d'accès à l'information et aux données électroniques sur l'environnement en Suéde. Il devra fournir des liens , des descriptions de données par le Dublin Core et autres services aux sites web des grandes agences publiques et campagnies privées travaillant dans le domaine de l'écologie.

Danemark :

20-Netpublikationer :

Page d'accueil: http://www.fsk.dk/fsk/publ/+connecté-pub/

Dans le but de rendre l'information publique plus efficiente et plus accessible sur le web, toutes les nouvelles publications des ministères Danois ont commencé a être mises sur le web, en parallèle avec les éditions imprimées, à partir de 1997.

21-INDOREG: (INternet Document REGistration)

Page d'accueil: http://www.purl.dk/rapport/html.uk/

Projet du Centre Danois des bibliothèques (DBC) pour fournir des enregistrements de toutes les publications d'internet et

l'accès à ces documents par DanBib (un système conjoint de superstructure pour le système de bibliothèque Danois)

Le Royaume-Uni :

22-ADAMIGVADS : (Art,Design,Architecture & Media Information Gateway and the Visual Arts Data Service)

Page d'accueil: http://adam.ac.uk/

'Art,Design, Architecture & Media Information Gateway' et 'Arts Data Service' sont deux services dont l'objectif est de fournir à la communauté anglaise de l'éducation un accès fiable et rapide, par réseau, à des ressources d'information dans les arts visuels.

23-AHDS Arts & Humanities Data Service :

Page d'accueil: http://ahds.ac.uk/

AHDS est une organisation fédérale , consistant en un Exécutif central et cinq fournisseurs de service incluant l'archéologie, l'histoire, l'étude de textes et les arts visuels. L'objectif de cette organisation est de développer un système intégré capable de fournir à l'utilisateur des ressources électroniques disponibles chez chaque fournisseur de service.

24-Projet BIBLINK :

Page d'accueil: http://www.ukoln.ac.uk/metadata/BIBLINK/

Le projet est commun à plusieurs bibliothèques nationales de l'Union Européenne. Il vise à établir un lien entre les éditeurs et les Agences Bibliographiques Nationales (NBA'S), pour l'échange des enregistrements de metadonnées des articles nouvellement publiés. La phase de démonstration de ce projet a été prévue entre Novembre 1997 et Mars 1999.

25-Projet 'DESIRE' :

Page d'accueil:

http://www.nic.surfnet.nl/surfnet/projette/désir/desire.html

Le projet DESIRE aborde deux approches pour la recherche de ressources d'information: Un service basé sur la sélection manuelle et la description de ressources de haute qualité, et un service régional de recherche basé sur les metadonnées générées par des agents de recherche du Web. Les objectifs du projet sont de contrôler et d'incorporer de nouveaux développements dans la gestion des metadonnées.

26-SCRAN :(Scottish Cultural Ressources Access Network)

Page d'accueil: http://www.scran.ac.uk

SCRAN est un projet pour créer une base de données en réseau, de ressources multimédia, pour les études, l'enseignement et les travaux d'histoire en Ecosse. Les partenaires sont les Musées Nationaux d'Ecosse, le Comité Royal sur les monuments Historiques et anciens d'Ecosse, et le Conseil Ecossais des musées. Il est prévu de permettre un accès facile à 1.5 millions d'enregistrements de texte d'objets et monuments historiques et 100,000 ressources multimédia connexes seront disponibles pour l'an 2001.

27-NewsAgent : (pour les bibliothèques)

Page d'accueil: http://www.sbu.ac.uk/litc/newsagent/

L'objectif du projet NewsAgent est de créer un service de news et un service actualisé de sensibilisation pour le personnel de bibliothèque avec un mélange de contenus, fournissant jusqu'à des descriptions de date de documents pour les utilisateurs finaux (basé sur les préférences de l'utilisateur).

ELISE II : (Electronic Library Image Service for Europe)

Page d'accueil: http://severn.dmu.ac.uk/elise/

Le service ELISE opérera sur le modèle client/serveur, en integrant l'emploi de la norme Z39.50 et le Dublin Core.

Dans le prototype ELISE II, les données de catalogue fournies par les institutions participantes, sont reprises selon la syntaxe du Dublin Core.

Les Etats-Unis :

29- Bibliothèque Electronique de Monticello :

Page d'accueil: http://www.solinet.net/monticello/monticel.htm

La fonction fondamentale de la Bibliothèque Electronique de Monticello est de relier et distribuer des ressources régionales indépendamment de la source ou du type d'information. Le Dublin core est employé pour fournir l'intéropérabilité sémantique entre plusieurs bases de données de medias électroniques et des types d'enregistrements incluant SGML, EAD(Encoded Archival Description), USMARC et des collections GILS.

30-Médical Metadata Project :

Page: http://medir.ohsu.edu/~maletg/MedMetadata.HTM

L'Université des sciences de la santé d'Oregon,l'Association Américaine d'informatique Médicale (Groupe de travail sur Internet) et l'institut National du Cancer fournissent un ensemble de tests de la base nationale de données génétiques.

31-FIUDL : (Florida International Universty Digital Library)

Page d'accueil: http://www.fiu.edu/~diglib/

Ce projet de bibliothèque numérique portera sur des images, du son et de la vidéo, incluant des modules multimédias de cours et de présentations. Il soutient les sujets des chercheurs et enseignants de l'Université de l'état de Floride.

32- Bibliothèque Numérique de l'Université de Washington :

Page d'accueil: http://content.engr.washing/

La Bibliothèque Numérique de l'université de Washington contribue au développement et à l'adoption de descripteurs de ressource d'information; ces efforts initiaux portent sur des collections d'image. Le projet utilise maintenant les éléments du Dublin Core pour la description

des collections d'images.

33-Everglades Information Network & Digital Library :

Page d'accueil: http://everglades.fiu.edu/

La collection de documents disponibles sur ce réseau inclue des rapports techniques, des articles scientifiques, des résumes et des programmes de conférences, des données sur la qualité de l'eau, des cartes, des diapositives et des photos, des documents légaux et des documents d'archives couvrant les sujets de recherches en écologie, en hydrologie, en écologie marine, et en restauration de marécages.

34-UMDLRD : (Universty of Michigan Digital Library Registry Database)

Page d'accueil: http://dns.hti.umich.edu/enregistrement/

Ce projet vise à créer une base de données à partir des ressources du Web, choisies pour leurs valeurs académiques et institutionnelles. Il fournira des éléments de metadonnées, assez simples d'utilisation pour des non spécialistes.

35-Digital Library Catalog :

Page d'accueil: http://sunsite.berkeley.edu/Cataloguer

Ce projet concerne la description d'un fonds comprenant des livres, des dissertations, des discours, et autres textes, écrits en HTML, des rapports techniques (dans divers formats), des photographies, des gravures et autres images, des clips vidéo et des clips audio.

36-GEM : (Gateway to Educational Materials)

Page d'accueil: http://gem.syr.edu

Le projet GEM est une initiative du département de l'éducaton des USA et de la Bibliothèque Nationale de l'éducation.

Son but est d'améliorer l'organisation et l'accessibilité des importantes collections de documents de l'éducation, non catalogués et qui sont déjà disponibles sur divers sites Internet.

37-SRS : (Scout Report Signpost)

Page d'acueil: http://www.signpost.org/signpost/index.html

Développée par 'Internet Scout Project', avec le soutien de la NSF(National Sciences Fondation), Signpost est une base de données permettant la 'navigation' et la recherche d'information. C'est un projet de recherche utilisant une structure de Dublin Core modifiée dans le but de décrire des ressources Internet appropriées.

38- SiteSearch : (Universite de l'Arizona)

Page d'accueil:

http://dizzy.library.arizona.edu/sitesearch/welcome.html

La Bibliothèque de l'université de l'Arizona a créé ce site pour fournir l'accès à diverses bases de données multimedia sur Internet.

39-CIMI : (Consortium for the computer Interchange of Museum Information)

Page d'accueil: http://www.cimi.org/documents/metafina/PD.html

CIMI est un groupe d'institutions et d'organisations qui encourage une approche de gestion de l'information électronique des musées, par des standards simples et ouverts. C'est dans ce sens que ce consortium a adopté le Dublin Core pour cataloguer ses ressources d'information.

40- Plusieurs autres projets impliquant le développement de bibliothèques numériques ont adopté le Dublin Core :

Exemple: ***La bibliothèque numérique de la sécurité aerienne US***

Cette bibliotheque a mis en œuvre un site web contenant principalement de l'information choisie pour l'aviation en générale.

4)Metadonnées et outils de recherche :

Les metadonnées ne sont pas prises en compte encore par tous les moteurs de recherche. Certains ont commencé à en tenir compte succinctement dans l'indexation des documents mis sur Internet. Nous pouvons citer :

-Infoseek :Les balises Meta sont repérées dans le document

HTML, mais sans aucune priorité.

-Altavista :L'indexation se fait d'abord sur le titre et les premiers mots du document, puis les champs « description » et « keyword » définis par les balises Meta.

-Hotbot :Les balises Meta sont prises en compte dans l'indexation, mais avec une moindre importance que les mots du titre.

-Ecila :Les balises Meta « keywords » sont prises en compte avant les mots du titre et du document, en l'absence de mots clés.

-HotMeta :Moteur de recherche(encore en période d'évaluation) orienté entièrement vers les documents décrits par des metadonnées (Dublin Core et SOIF).

5) METADONNEES ET BIBLIOTHEQUE NUMERIQUE:

INTRODUCTION *:*

Souvent, les expressions bibliothèques virtuelle, électronique et numérique sont perçues comme un seul et même concept. Pourtant , elles diffèrent de manière sensible.

- ✓ La bibliothèque virtuelle peut être considérée comme le vocable générique.

La première bibliothèque virtuelle sur le web , fondée par Tim Barnners, est :

http://vlib.org

http://vlib.org/AboutVL.html

Le concept de bibliothèque virtuelle repose sur la notion d'une collection de liens basés sur des pages HTML ou des bases de données et qui sont maintenus et mis à jour par plusieurs contributeurs (éditeurs).

Les catalogues virtuels sont des meta – catalogues où la recherche d'information se fait en ligne, simultanément sur plusieurs catalogues.

- ✓ La bibliothèque électronique est une très vieille expression, qui date de plus de vingt ans. A l'origine, cette notion incluait tout moyens rentrant

dans l'automatisation de bibliothèques utilisant des dispositifs électroniques tels des unités centrales de traitement (mainframe) ,des terminaux, des PDP-11 etc.….

Dans ce sens-là , le terme « électronique » désigne la manière selon laquelle la machine ou le dispositif fonctionne ; plutôt qu'une caractérisation des données . Aujourd'hui la bibliothèque électronique est comprise comme une composante de la bibliothèque virtuelle, représentant des collections électroniques et les services s'y rattachant .

- ✓ La bibliothèque numérique est constituée de tout ce qui fut analogique et qui maintenant est devenue numérique : texte, image, son , vidéo, photographie, donnée factuelle, diagramme, etc.….

Ces données numérisées ont besoin d'être indexées et d'avoir des points d'accès autant qu'une collection de bibliothèque classique a besoin d'un catalogue. Ces données décrivant d'autres données sont appelées metadonnées et peuvent être de deux types :

- metadonnées décrivant l'entité électronique (document);
- metadonnées décrivant le contenu de l'entité électronique.

En réalité, le terme metadonnées est utilisée le plus souvent dans le premier sens soit dans le sens des données utiles à l'identification, à la description et à la localisation de ressources électroniques en réseau.

Le choix du format de description est souvent difficile à faire pour le moment. Certains tendent vers l'utilisation et l'adaptation de formats traditionnels de description comme le format MARC (MAchine Readable Cataloging). De nouveaux formats de description sont aussi pris en compte (TEI, EAD, Dublin Core). Mais, jusqu'à présent, aucun de ces formats n'a priorité sur

les autres même si de par l'engouement qu'il a suscité auprès d'une plus large communauté, le Dublin Core évolue vers une plus large utilisation.

Toutefois, une chose est certaine: les metadonnées ne peuvent être ignorées tant pour organiser l'information que pour en faciliter le repérage et la description de ressources. Le monde de l'information a besoin d'un format de description normalisé permettant de faire référence à des entités électroniques. De par ses rapides développements et son caractère évolutif, le Dublin Core est en passe d'être le meilleur candidat en matière de description de ressources, pour les bibliothèques numériques.

Types de bibliothèques numériques *:*

En partant de la définition donnée précédemment, nous pouvons rencontrer deux types de bibliothèque numérique:

- Type 1: Ensemble de metadonnées :

La bibliothèque numérique est constituée exclusivement de metadonnées d'une collection qui n'est pas disponible à même la bibliothèque virtuelle.

Par exemple, on peut penser au catalogue automatisé d'une bibliothèque accessible en ligne. Nous pouvons imaginer que dans ce cas , ces metadonnées, peuvent aussi être accompagnées de certaines fonctionnalités particulières comme un service de prêt entre bibliothèques par courrier électronique. Nous nous retrouvons ainsi dans le premier état qu'une bibliothèque physique adopte pour aller vers le monde de la virtualité.

-Type 2: Ensemble de metadonnées et de données :

La bibliothèque numérique comporte dans ce cas-là des metadonnées et les données décrites par ces dernières. Dans cette catégorie, nous pouvons retrouver, par exemple des sites Web comportant des bibliographies thématiques de ressources accessibles en réseau.

Cet ensemble de metadonnées et de données peut ici aussi, dans le cas d'une bibliothèque virtuelle, s'accompagner de certains services externes comme un service de prêt entre bibliothèques.

Ce type de bibliothèque numérique peut être utilisé tant par les bibliothèques physiques (bibliothèques classiques), afin de compléter leur fonds ou collection, que par une communauté ne possédant pas de bibliothèque physique comme telle mais voulant présenter un ensemble de documents

numérisés ou d'informations organisées portant sur des thèmes spécifiques ou spécialisés.

Problématique linguistique :

Dans le contexte international de l'information en réseau (Internet), la question linguistique est plus que d'actualité. Actuellement des équipes de recherches tel le *Digital library research Group* du *College of library and Information Services* de l'Université du Maryland , travaillent sur des systèmes de recherche d'information permettant le lancement d'une requête dans une langue pour retrouver un document dans une autre langue (Cross-language Information Retrieval Resources).

Les metadonnées étant des clés d'accès à la collection ou au fonds d'une bibliothèque numérique, il serait plus que souhaitable de les avoir en plusieurs langues. Ainsi les usagers seraient à même de juger de la pertinence de consulter une ressource dans une langue étrangère grâce à une bonne compréhension de la description y étant rattachée dans une langue mieux maîtrisée.

Préférences d'utilisateurs :

Tout projet de bibliothèque numérique doit s'adapter à une catégorie d'utilisateurs.

Le fonds numérisé, les services et les outils utilisés sont fonction de leurs besoins. Il est donc important, avant toute chose, d'identifier clairement leurs préférences.

Une bonne connaissance des préférences des utilisateurs potentiels d'une bibliothèque numérique peut permettre de mieux cibler son contenu ou tout au moins de mieux l'organiser.

Lors de l'élaboration de tout modèle, la prise en compte des caractéristiques propres des utilisateurs et ce dès le début de la conception, constitue un aspect fondamental. Mais le gage majeur permettant l'adéquation entre les besoins des utilisateurs et les fonctionnalités de la bibliothèque numérique, ne peut être obtenu que par la présentation, au fur et à mesure de la conception, de prototypes aux utilisateurs potentiels afin d'observer et d'analyser leurs réactions et leurs performances pour ainsi apporter les correctifs nécessaires (Cycle de conception itératif de John D. Gould et Clayton Lewis).

6-CONCLUSION :

Les metadonnées sont , somme toute , une valeur ajoutée à l'information. Elles en permettent la compilation et le répérage. Pour leur part, les bibliothèques produisent ce type d'information depuis longtemps. Par contre Internet en est à ses débuts. Indexer la masse d'information qui y circule, pour en faciliter l'accès aux utilisateurs nécessite un certain travail de préparation de l'information. L'ajout d'élément de metadonnées devient une action nécessaire. Nous venons de passer en revue les principaux standards de metadonnées existants, même s'ils diffèrent souvent dans leurs structures, il n'en demeure pas moins qu'ils poursuivent un objectif principal commun :offrir des éléments de description de l'information pour en faciliter l'accès.

Il faut signaler que les standards de metadonnées du domaine géospatial avait pour but de fournir toute l'information sur les données décrites, et pas seulement l'information bibliographique. De plus, des éléments de description trop nombreux et trop technique dont l'emploi peut échapper au simple utilisateur, les éloignent de tout parallèle avec des fiches catalographiques classiques.

C'est peut-etre cette relative complexité qui a fait que le Dublin Core, en offrant une structure simple, ait fédéré autour de lui beaucoup d'organismes et d'institutions. Au bout de cinq ans, le Dublin Core est déjà opérationnel dans plus de 60 projets, de par le monde. Il est en constante évolution.

Une série de workshops lui assurent un suivi constant dans les développements. Des propriétés d'ouverture, d'extensibilité et d'intéropérabilité, le rendent applicable à tous les domaines.

Des schémas de conversion des autres standards vers le Dublin Core, ou inversemnt, ont été développés, pour montrer sa portabilité. Il faut rappeler qu'au départ, les promoteurs du Dublin Core voulaient relever un défi important : offrir une aide améliorée pour le repérage de ressources d'information. Du moment qu'Internet renferme plus d'information que tout ce que les professionnels des bibliothèques peuvent prendre en charge par les méthodes classiques, il semblait logique de donner des outils aux auteurs et fournisseurs d'information électronique pour leur permettre de décrire eux-même leurs documents. Cela a certe suscité quelques critiques, mais il

n’empêche que depuis qu’une syntaxe spécifique, basée sur SGML a été approuvée par le W3C en 1996, le Dublin Core s’est mis sur la voie d’une norme internationale. C’est aussi en 1996, qu’a été adopté le *Warwick Framework,* au troisième workshop qui s’est tenu à la Warwick University, au Royaume-Uni. Ce dernier est une architecture ayant le potentiel de mettre ensemble des metadonnées à structures syntaxiques diverses, qui sont accessibles et maintenues séparément.

En se conformant aux développements du W3C (RDF et XML), le Dublin Core s’affirme directement comme un futur standard de metadonnées international.

Partie B :

LA RECHERCHE D'INFORMATION :

I- Introduction

II- Les outils de recherche d'information

III- Le filtrage d'information

IV- Le cross-language

V- Conclusion

I- INTRODUCTION:

La recherche d'information est en fait un procédé complexe qui inclut des phases multiples, itératives, et très dynamiques. Dans la bibliothèque traditionnelle, le catalogue classique, constitué de notices ou sous forme de listing , permet de retrouver des documents et donc de l'information.

L'importance des catalogues dans la recherche d'information traditionnelle est bien connue .

Une perspective intéressante sur le rôle des catalogues émane de David Levy [124] (Centre de recherche Xeros, Palo Alto). Levy présente le catalogage comme une méthode de création d'une illusion d'ordre (comme un schéma) de l'univers d'informations chaotiques. Selon ses termes, " Le catalogage est un ensemble de pratiques qui permet l'ordre d'une collection littérale et fournit un accès à travers un ensemble de substituts bien organisés ." Le catalogage et les substituts nous permettent , en tant que chercheurs, de supposer que les ressources ont des attributs communs tels que le titre, l'auteur, et le sujet. En effet , les attributs ne peuvent pas exister actuellement , mais sont dérivés et associés à des objets d'informations par suite de catalogage professionnel. Nous utilisons ces attributs comme base de formulation des critères de recherche et comme un moyen de conceptualiser et examiner les résultats des recherches.

Dans la bibliothèque traditionnelle, le bibliothécaire joue un rôle vital dans la transmission des besoins courants de l'utilisateur vers les substituts statiques. Terry Smith [16], décrit cela comme une cartographie itérative des modèles de connaissances, dans ce qu'il appelle " l'environnement de meta-information " des bibliothèques. Le modèle de connaissances de l'utilisateur est basé sur son environnement, l'explication détaillée du sujet, de la bibliothèque, et des besoins courants d'informations (représentée dans la "question" pour laquelle il voudrait une "réponse"). Le modèle de connaissances de la bibliothèque se base sur les ressources d'informations disponibles actuellement, et les meta - informations (par exemple., enregistrement du catalogage descriptif et des index) et comme substituts à ces ressources et aux ressources relatives extérieures à la bibliothèque. Réunir ces modèles parfois différents, est la raison d'être d'un bon bibliothécaire.

Nous sommes en plein milieu d'une transition rapide d'une structure de l'‘information classique (bibliothèque traditionnelle) au domaine numérique (bibliothèque virtuelle). La combinaison d'un accès facile et d'une quantité pure et simple d'informations courantes a permis l'accès sur Internet, et a révélé la source d'information préférée de la plupart des gens dans le Web. Toutefois, toute proposition qui viserait à supplanter le rôle de la bibliothèque par le Web, est imprudente. Une bibliothèque n'est pas seulement un entrepôt de documents, mais c'est une collection riche de services avec une sélection, une conservation, une catégorisation, une localisation….

Dans le nouvel univers de l'information numérique l'OPAC constitue l'analogue du catalogue traditionnel ; et permet de faire de la recherche d'information en ligne, dans des réseaux d'information. Ces réservoirs de références bibliographiques, appelés par les Anglo-saxons, substituts statiques, sont des aides précieuses dans la recherche d'information, avec ou sans la médiation des bibliothécaires. Par extension, les grands réservoirs de descriptifs de ressources d'information qu'offrent les réseaux d'informations tel qu'Internet seront appelés substituts dynamiques. Nous pensons que des améliorations significatives dans la recherche d'information dans le monde des réseaux peuvent être obtenues en utilisant des techniques qui apparient des substituts sémantiques aux exigences spécifiques des nouveaux procédés de recherche de ressources d'information. Cela peut être accompli plus facilement par des architectures, qui permettent l'association de substituts multiples avec des objets ou substituts dynamiques qui répondent aux besoins actuels de la recherche de ressources d'information.

Avec ou sans l'ordinateur et l'Internet, le processus de recherche de l'information peut être complexe, comme il peut être très simple. Au départ, quel est le but de la recherche d'information? On peut facilement rechercher la réponse à un certain besoin d'informations ou de réponses à des questions. Dans certains cas, ce qu'une personne cherche est la meilleure réponse possible à une question (nous ignorons les caprices de ce qui est "meilleur"). Dans d'autres cas, à cause des contraintes tels que le temps, le coût, la patience, le chercheur d'informations peut être satisfait avec peu de

réponses spécifiques. Dans certains cas, le processus de recherche d'information peut être lancé sans même un but d'information clairement défini, et la réponse satisfaisante pourrait être une information de valeur par suite de recherche du processus lui-même.

Un chercheur américain, John Kunze [107] pense qu'il est utile d'étudier le processus de recherche de ressources d'information comme une série de mouvements entre deux phases, ou états. La première phase est la phase de localisation, pendant laquelle une personne formule un ensemble de critères de sélection (une question) et à partir de ces critères, nous avons un ensemble de ressources candidates. La deuxième phase est la phase d'examen, qui entraîne l'examen de ces ressources candidates. Il est habituellement préférable de balayer un ensemble de références pour ces ressources plutôt que d'examiner les ressources intégrales. L'utilisateur peut arrêter le processus de recherche d'information, en considérant qu'une réponse adéquate au besoin d'informations a été trouvée, ou peut retourner vers la phase de localisation avec d'autres critères de sélection.

Au cours du processus de recherche d'information, les utilisateurs se déplacent sur un spectre de granularité d'information qui a un impact sur leurs critères d'examen et de localisation. Généralement, au début du processus, les chercheurs d'informations ont des critères de granularité relativement "gros". Ils peuvent composer ou soumettre des demandes relativement imprécises qui, de par leur nature entraînent de grands résultats (« bruit » très important). Dans plusieurs cas, l'examen du résultat peut permettre d'affiner les requêtes par des détails successifs, pour une granularité plus précise, puisque le chercheur essaie de diminuer le nombre de résultats global obtenu. Cependant, le processus n'est pas linéaire et à long terme le chercheur d'information fait un aller retour sur le spectre de granularité.

Une autre dimension relative quoique différente du processus se rapporte à la spécificité du domaine. De ce point de vue, le chercheur d'informations commence souvent le processus de recherche en sélectionnant des critères qui ne sont spécifiques à aucun domaine d'analyse ou de discipline pédagogique. Il peut adopter une perspective plus spécifique d'un domaine ;

et dans ce cas ,il peut utiliser les termes et la sémantique plus spécifiques à certains domaines (par exemple bibliothéconomie, informatique, ou mécanique).

Une description utile de la dimension du processus est la métaphore du “Touriste virtuel” de Ricky Erway (OCLC) [121].

Cette métaphore lance le chercheur de ressources d’information comme " un voyageur virtuel" qui, comme un vrai voyageur, peut naviguer sur les étapes du processus de recherche d’information. Cependant, le voyageur réel se doit d’être au courant de la culture et du langage de la contrée à visiter ; et par conséquent, le chercheur d'informations doit adopter une syntaxe spécifique au domaine et une sémantique pour aborder le processus de recherche de ressources….

II- Les outils de recherche d'information :

Si d'un côté, la quantité de ressources d'information sur Internet se développe à des taux élevés, d'un autre côté il n'y a pas d'avancée proportionnelle dans les outils de recherche d'information. La plupart des outils de recherche d'information en réseaux utilisent un modèle qui s'est développé à partir de l'outil de recherche Archie pour la recherche des ressources FTP. Ce modèle, utilisé surtout par "les Web- crawlers ou les générateurs d'index (Web- indexers)" et caractéristique des services tels que ceux d'Alta Vista , compte sur des explorations globales périodiques des ressources du Web, en utilisant des hyperliens comme guide. Le crawler utilise la requête HTTP GET pour décharger chaque ressource et utilise une variété technique de recherche d'information pour indexer les contenus de cette ressource. Les utilisateurs soumettent des requêtes en texte intégral, au service centralisé (qui peut être distribué parmi plusieurs serveurs et le reproduire sur plusieurs sites).

La génération d'index à l'aide de cette technologie présente quatre principaux problèmes :

--Les problèmes d'échelle:

Les Web- crawlers se basent sur un déchargement complet des ressources à partir de leurs emplacements vers leur service d'indexation. Il est évident que cette procédure ne peut pas être évaluée ou estimée à l'échelle puisqu'il n y a pas de statistiques fiables sur le pourcentage de trafic réseaux parcouru par l'indexation. Les générateurs d'index du Web seront tentés d'accroître la fréquence de leur analyse tout en augmentant le chargement sur les réseaux et les serveurs Web qu'ils visitent sous la pression de la concurrence. Les ressources visitées et indexées seront de plus en plus importantes, sans pour autant que nous soyons assurés d'avoir une information ou des résultats de recherche plus pertinents.

--le problème d'accessibilité :

L'accès au contenu sur le Web n'est généralement pas restreint. La plupart des contenus ne sont pas placés derrière des serveurs spéciaux, à accès restreint. Ainsi, à fortiori, à court terme, les générateurs d'index seront

capables d'accéder et d'indexer tout le contenu diffusé sur le Web. Mais nous pouvons affirmer dès maintenant que nous nous dirigeons vers des accès sélectif restreint et qu'à terme, il deviendra de plus en plus difficile d'avoir librement accès au contenu des ressources intéressantes, parce que la plupart seront derrière des barrières de contrôle d'accès (serveurs payants d'information en ligne).

--*Les problèmes des technologies de recherche de l'information*:

La technologie de recherche de l'information, sur laquelle se basent les outils de recherche d'information du Web, est le résultat de plus de 30 années de recherches. Cette technologie a été perfectionnée de façon constante et a abouti à des résultats satisfaisants. Cependant, la nature du Web en tant que grand réservoir d'information , présente des problèmes difficiles et entraînent souvent des résultats contestables. La taille véritable de ce réservoir est un problème important; l'échec de la précision et du rappel augmente en nombre à cause des incertitudes des approximations mathématiques des algorithmes utilisés. En plus, le Web est un corps ou réservoir non - spécifique, et rend donc les discordances de synonymes inévitables. Certaines discordances peuvent être résolues par des techniques linguistiques qui sont imparfaites et leur utilité constitue actuellement un riche débat dans la communauté qui s'intéresse à la recherche de l'information.

--*Les problèmes du comportement de l'utilisateur* :

L'usage a montré que la plupart des utilisateurs ne gardaient que les premières réponses des pages retrouvées par les systèmes de recherche d'information du Web ; malgré que rien ne peut affirmer que celles-ci soient les plus pertinentes.

Les outils de recherche d'information opèrent actuellement, pour la plupart et en général, sans l'aide d'éléments de description ou de metadonnées. Ils permettent l'indexation automatique et la localisation de n'importe quelles ressources. Cependant ils ne sont pas considérés comme la solution la plus fiable pour la localisation de ressources sur les réseaux ; mais un complément aux méthodes plus structurées qui se basent sur les éléments de description de ressources.

La reconnaissance des limites des outils de recherche d'information courants sur Internet, a amené beaucoup de chercheurs en sciences de l'information et la communauté des bibliothèques numériques, a étudier d'autres solutions et méthodes de recherches d'information basées sur une indexation de documents par des metadonnées. Le travail le plus connu dans ce domaine est le standard de metadonnées du Dublin Core. Ce dernier, défini dans la partie précédente, est un ensemble d'éléments de metadonnées, faciles à créer et à maintenir et qui contient un nombre minimum d'éléments requis pour décrire et ainsi faciliter la localisation de ressources dans un environnement réseaux. Cependant, nous pensons que le plus grand potentiel pour améliorer la recherche d'information en réseaux se base sur l'utilisation des substituts dynamiques ou dérivés. Lynch et Michelson [146], se réfèrent à cette capacité avec le commentaire : "Il est important de reconnaître que l'environnement de l'information en réseaux offre de nouvelles occasions pour dériver (par extraction ou calcul) un ensemble plus riche et plus diversifié de substituts, à partir d'objets réseaux que des substituts ou éléments de description de l'information sur support classique (papier)."

Dans un processus de recherche d'information, nous distinguons les éléments de description de ressources classiques que nous pouvons considérer comme des substituts statiques et ceux associés à des profils d'utilisateurs que nous considérons comme des substituts dynamiques. Evidemment, ils sont entièrement indépendants les uns des autres. Notre intention est de préserver la capacité de faire un ordre de ces substituts en développant un ensemble de prototypes logiques qui simulent les besoins de l'utilisateur dans les différentes étapes de la recherche d'information. La recherche dans ce domaine peut procéder à des études détaillées comportementales de l'utilisateur dans le monde des bibliothèques traditionnelles et dans le monde de l'information en réseaux. Elle est entreprise en partie dans des projets tels NSF/ARPA/NASA DLI . Par conséquent, d'après ces études, nous pouvons énumérer des étapes clés dans le processus de la recherche d'information et développer des éléments de profils standards. Ces profils peuvent être ensuite affinés et leur nombre peut être augmenté pour permettre une granularité plus fine dans l'appariement des besoins par rapport aux ressources décrites.

Conclusion :

Actuellement tous les spécialistes de la recherche d'information admettent que les améliorations substantielles dans la recherche d'informations en réseaux sont possibles, si nous raisonnons en terme de substituts dynamiques plutôt que statiques. Le succès des substituts statiques dans le monde des bibliothèques traditionnelles a été possible avec plusieurs médiations humaines pour confondre les besoins spécifiques de l'utilisateur avec l'information exprimée dans les enregistrements des catalogues. Puisque nous ne pouvons probablement jamais égaler le savoir-faire du bibliothécaire professionnel par les automates de recherche, nous pouvons améliorer la recherche d'information en réseaux en développant des mécanismes qui tiennent compte de la sémantique dans la description de ressources; et cela selon les besoins de l'utilisateur. La réalisation de ce but exigera une recherche substantielle en technologie des interfaces de l'utilisateur et du système. En attendant, nous pouvons améliorer la recherche d'information en réseaux en standardisant les formats des substituts de description, en développant des méthodes pour associer les substituts multiples aux objets du contenu, et en fournissant des méthodes à l'utilisateur pour ajuster son profil aux substituts pris en compte par l'outil de recherche d'information. Nous revenons ainsi, encore une fois à cette notion de profil dynamique.

III - Le filtrage d'information:

Définition :

Généralement, le but des systèmes de filtrage d'information est de trier parmi des volumes d'informations générés dynamiquement (en réseaux) et de présenter à l'utilisateur l'information recherchée et répondant à ses besoins. Dans la suite de cette définition, il faut remarquer qu'il y a une distinction nuancée entre collecte d'information et filtrage d'information. Dans certains domaines (exemple USENET News) , l'effort de collecte est minime; car l'information vous est transmise souvent , selon un profil que l'on vous aura préalablement attribué ou aidé à définir , à travers la messagerie électronique ; et donc le flux d'information dépend entièrement de votre volonté et doit correspondre à vos besoins.

Dans d'autres domaines (exemple Worldwide web), l'effort de collecte peut être considérable car il n'existe pas spécialement de mécanisme très fiables de systèmes de filtrage, permettant de trier de nouvelles informations au rythme d'un processus dynamique de lancement de requêtes.

Le problème de filtrage d'information commence seulement au moment de l'accès à de nouvelles informations.

Le filtrage d'information s'applique à plusieurs domaines et sous une grande variété de techniques d'approches. Les méthodes traditionnelles s'appuyaient sur un service d'alerte non automatisé et qui fournissaient de l'information aux besoins spécifiques des utilisateurs. Cela constituait tout simplement la Diffusion Sélective de l'Information. Actuellement, quelques systèmes adoptent les principes de description de ces méthodes dans leur processus de filtrage ; surtout dans les projets de bibliothèques virtuelles. Avec le développement d'Internet et d'autres réseaux d'informations, ces dernières années, un grand effort de recherche a été entrepris dans les systèmes de filtrage automatique de l'information en réseaux.

A cause de leur faible coût et de la facilité d'identification des nouvelles informations qu'ils véhiculent, les groupes de discussion gérés par USENET et la messagerie électronique sont les domaines où la recherche de systèmes de filtrage a été la plus abondante.

Le récent développement impressionnant du Worldwide web a permit l'émergence d'intéressants domaines où des besoins de recherche de systèmes de filtrage se sont fait sentir ; pour attirer l'utilisateur par une offre d'une information toujours plus pertinente , et toujours plus en rapport avec son profil et ses besoins. L'Annuelle Conférence de la Recherche Textuelle (TREC / Université de Maryland / USA) pendant laquelle des collections de texte standard sont utilisées et une méthodologie d'évaluation imposée; a permis aussi un intéressement considérable à la recherche dans les systèmes de filtrage de l'information.

D'ailleurs le TREC[3] a récemment adopté une procédure de filtrage spéciale qui repose sur différentes méthodes d'évaluation. Des systèmes filtrant des documents divers et des sources d'informations spécialisées dans le commerce sont en voie de développement. Ces nouvelles techniques de filtrage prendront en compte, outre l'information textuelle, des images, du son et des vidéos, dans le futur.

Souvent, la distinction entre le filtrage d'information et des domaines de la recherche d'information bien déterminés induisent bien des confusions. La recherche d'information entraîne toujours une sélection de l'information et beaucoup d'acquis récents des systèmes de recherche d'information (mesure de similarités , sélection booléenne , approximations et distances) sont maintenant intégrés dans le filtrage de l'information.

Si, en général la recherche d'information prend le sens de "sélection de l'information" ; le filtrage d'information est tout simplement un cas particulier dans lequel "l'espace information" est très dynamique. D'autre part, si nous considérons que la recherche d'information induit une sélection d'une information relativement statique, en réponse à des requêtes relativement dynamiques, le filtrage d'information s'impose alors comme une nécessité dans le processus de la recherche d'information. En regard de ces déductions et points de vue, les chercheurs en filtrage d'information se doivent de baser leurs travaux d'abord par une exploration de toutes les particularités et tous les travaux qui ont été menés dans le cadre de la recherche d'information….

[3] Text REtrieval Conferences

Travaux et projets sur le filtrage de l'information dans le monde:

Pour souligner l'importance de cet axe de recherche dans le concept global de la recherche d'information dans le monde, surtout chez les Anglo-saxons ; il nous semble opportun d'énumérer quelques principaux projets, travaux et conférences consacrés au filtrage de l'information:

- ***Dartmouth*** :

Service d'alerte actualisé à interrogation à relances qu'utilisent les moteurs de recherche d'information Alta Vista et Lycos pour informer les utilisateurs des nouvelles pages Web et conformes à leurs profiles définis manuellement.

- ***COBALT*** :

Projet de la Communauté Européenne (I*M Language Engineering Program) sur le filtrage des news (USENET).

- ***Université de Stockholm*** :

Lancement du projet IntFilter et du système GHOSTS pour le filtrage de la messagerie électronique et du service de news USENET.

- ***TREC*** :

Proceedings des troisième et quatrième conférences TREC (Text REtrieval Conferences) lesquelles ont abordé le cheminement de l'évaluation de plusieurs approches de filtrage de l'information et leur comparaison.

- ***University of Aberdeen***:

Plusieurs projets portant sur un agent de filtrage de messagerie électronique, un outil de filtrage de news USENET et un outil d'apprentissage avec une fonction de filtrage appliqué au Worldwide web.

- ***Infoscope*** :

Projet rentrant dans le cadre d'une thèse de PHD à l'université de Boulder , dans l'Etat du Colorado , et portant sur un système de filtrage de news USENET utilisant des agents booléens.

- **Système de filtrage PICS[4]:**

Le système PICS a été développé dans le but de filtrer un certain type d'information véhiculée par le réseau Internet portant sur la violence ou la pornographie et l'éviter aux enfants. PICS fonctionne sur le principe de conventions établies et de labels ou niveaux d'accès. Il est important de noter que le système PICS en lui-même est une méthodologie ou une infrastructure et non un service d'indices ou un système actif pour la censure ou la sélection. Il est neutre en valeurs: ce sont les applications qui dirigent son exécution. Les labels de PICS décrivent le contenu d'une ou de plusieurs dimensions utilisant un vocabulaire fait dans ce but et permettent à un logiciel de sélection de déterminer l'accès. Dans son rôle le plus médiatisé, PICS peut être utilisé pour contrôler l'accès d'Internet à des enfants. Le logiciel de sélection local peut être établi pour interdire l'accès si la violence dépasse un niveau déterminé auparavant par le parent ou le professeur. Selon les accords de part et d'autre, Le propriétaire de ressources ou l'étiqueteur indique le niveau de violence (ou nudité, blasphème, etc) sur leur site et la sélection des barres du logiciel dont les labels égalisent ou dépassent ce niveau. Les sites sans étiquetage peuvent aussi être exclus. D'autres utilisations pour un tel système peuvent être facilement imaginées.

Les premiers labels étaient conçus pour permettre une circumnavigation des sites indécents en réponse à une agence gouvernementale américaine (US Télécommunications Decency Act).

Les labels peuvent être utilisés pour communiquer les caractéristiques qui nécessitent un jugement humain - si une page Web est drôle ou choquante - ainsi qu'une information qui n'est pas claire à partir des mots ou des graphismes, telle que les méthodes de sites Web sur l'utilisation ou la diffusion ou cession de données personnelles.

Des auteurs américains, Resnick et Miller [121] ont déjà constaté que "les nouvelles infrastructures sont souvent utilisées dans des voies non prévues, pour rencontrer des besoins latents", et suggèrent que les articles de revues électroniques peuvent être codés (article spécialisé, article de revue, annonce courte, etc.); des vocabulaires de propriété intellectuelle peuvent être

[4] Plateform for Internet Content Selection

développé ; et des vocabulaires de réputation peuvent "associer des labels aux sites commerciaux qui ont des pratiques de commerce spécialement bonnes ou surtout mauvaises."

Durant les deux dernières années, un travail considérable de développement a eu lieu dans le domaine des metadonnées. Le développement du PICS est soutenu par le Worldwide Web Consortium (W3C), qui est l'organisation responsable du développement des standards Web. Derrière les concepts généraux facilement compris, le travail complexe sur la syntaxe et le vocabulaire se poursuit en même temps que les applications supplémentaires de PICS qui ont été proposées. Les labels PICS peuvent être utilisés pour porter des signatures numériques des ressources ou pour protéger les ordinateurs des virus. Le mécanisme qui doit limiter l'accès et obtenir l'accès représente deux côtés de la même pièce: les requêtes pour LIMITER l'accès à n'importe quel site traitant un sujet donné ou ayant un indice de circulation moins récente, sont similaires aux requêtes pour TROUVER tout site selon ce sujet ou avec un code de date non courant.

L'utilisation de PICS aux USA fut largement rapportée pendant la réunion de PICS Working Group Meeting tenue à Londres du 13 au 14 Janvier 1997. Il semblerait que PICS soit approuvé par les gouvernements européens. En prenant en considération le travail de développement de PICS, il est clair que la communauté d'Internet bénéficierait de ce travail pour stocker des informations sur la qualité des ressources.

- **Outils et systèmes de filtrage de l'information**:

  Plusieurs outils et systèmes de filtrage de l'information sont disponibles gratuitement sur certains sites Internet. Dans la plupart des cas , des documents décrivant ces outils et systèmes et les détails de leurs implémentations sont aussi disponibles sur le Web. Nous citons, ci-après les plus connus :

  - *SIFT :*

    Outil de filtrage de l'information développé par Tak Yan , chercheur à l'université de Stanford ; et comportant deux services de diffusion sélective , l'un pour les rapports techniques d'informatique et l'autre pour

les nouvelles (news) USENET. Le code source et des documents décrivant le développement de SIFT sont disponibles sur Internet. Un service spécialisé de filtrage de l'information à l'Institut Européen de Bio-informatique , fonctionne avec SIFT.

- *Firefly :*

Service de filtrage collaboratif pour la musique et les films. Des documents descriptifs du développement de cet outil ainsi qu'un guide d'utilisation , écrits par Alexander Chislenko sont disponibles sur Internet.

- *InfoScan :*

Logiciel développé par la compagnie Machina Sapiens, Inc. pour filtrer la messagerie électronique et les nouvelles USENET. Des versions de ce logiciel existent pour diverses plate formes (Macintosh et Windows)

- *InfoTicker :*

  Robot (outil intelligent) de surveillance du Web avec une fonction de filtrage, développé par Erik Mueller.

- *NewsSieve :*

  Système de filtrage de nouvelles USENET développé par Elmar Haneke de l'Université de Bonne en Allemagne.

- *iAgent :*

  Agent de filtrage adaptatif pour le web, développé par Kok Lai, à l'Institut des Technologies de l'Information de Singapour.

- *RAMA :*

  Système de filtrage de nouvelles USENET, développé sous Unix par Jim Binkley du Département des Sciences Informatiques de l'Université de l'Etat de Portland.

- *Sift-Mail :*

Système de filtrage de messagerie électronique développé par Laurence Lundblade de la compagnie Virginia Tech.

- *SMART :*

Logiciel sous Unix destiné pour l'évaluation des performances des techniques d'espace vectoriel de la recherche d'information de Gérard Salton et Chris Buckley de l'université de Cornell.

- *MAXIMS :*

MAXIMS est un outil de filtrage collaboratif de messagerie électronique développé par Max Metral de MIT Media (Massachusetts / USA).

- *Procmail :*

Logiciel sous Unix destiné pour filtrer automatiquement la messagerie électronique

- *WebWatcher :*

WebWatcher est un système de filtrage de l'information pour le Worldwide Web , développé par David Zabowski dans le cadre du projet *PLEIADES* (Laboratoire d'apprentissage de l'université de Carnegie Mellon).

- *WebFilter :*

C'est un système conçu pour filtrer le contenu des pages Web en temps réel , par Axel Boldt de l'Université de l'Etat de Californie , à Santa Barbara.

- *Web Filter :*

Second système du même nom que le précédent, mais basant sa fonction de filtrage des pages Web sur le principe du WAIS ; et développé par Steve Gant de l'Ecole de Bibliothéconomie et des Sciences de l'Information de l'Université de la Caroline du Nord.

- *FilterGus :*

C'est un programme simple écrit en langage Java, par Mauro Marinilli , permettant de filtrer des documents textuels.

- *WebWasher :*

Logiciel sous Windows développé par Siemens, pour filtrer des images sur le Web. Il est gratuit pour un usage personnel, non commercial et pour les institutions éducationnelles.

- *ScienceIndex :*

Index de citation valable pour le Web, dont l'objectif est de constituer une bibliothèque virtuelle pour la littérature scientifique sous forme électronique Ses caractéristiques comportent aussi une indexation de citations autonome, une localisation de documents autonome, une extraction des contextes de citations, une indexation de texte intégral, une identification de documents similaires, et une analyse de graphes de citations.

- *Select :*

Projet de l'Union Européenne dont l'objectif est de développer un système de filtrage collaboratif qui puisse aider les scientifiques, les techniciens et autres utilisateurs professionnels d'Internet pour retrouver l'information dont ils ont besoin.

- *ChaffAway :*

Bibliothèque virtuelle où chacun peut chercher ou laisser des documents. Ce système, développé par Ian Ford du *Groupe In-Key Information* , fonctionne sur le principe du filtrage collaboratif par la collecte des votes des utilisateurs.

- *NoCeM :*

Système de filtrage collaboratif de l'information pour les groupes de discussion USENET. Les instructions de filtrage sont transmises dans un formulaire signé *PGP* en accord avec les critères personnels de chaque utilisateur. Il est utilisé aussi pour le filtrage des spams de la messagerie électronique

- **Modélisation de l'utilisateur et filtrage de l'information**:

Les systèmes de filtrage de l'information sont destinés pour examiner le flot de documents générés dynamiquement et de ne permettre la visualisation que de ceux considérés pertinents pour l'utilisateur. Par contre

les systèmes de recherche de l'information sont destinés pour répondre à chaque requête en se basant sur le contenu d'ensembles de documents relativement statiques.

Bien que les deux systèmes de recherche de l'information et de filtrage de l'information aient en commun un ensemble de techniques de représentation et de sélection de texte ; les différences de leurs objectifs ont des implications sur la modélisation de l'utilisateur. La plus significative distinction entre le filtrage et la simple recherche d'information consiste dans la durée sur laquelle le besoin d'information de l'utilisateur doit être modelé. Généralement les utilisateurs d'un système d'information peuvent se prévaloir de plusieurs intérêts ou besoins d'information ; et dans ce cas-là nous pouvons supposer que la durée de vie d'un intérêt particulier ou d'un besoin est sujette à de larges variations , selon les circonstances et les profils des individus.

Par exemple, un utilisateur peut avoir un intérêt constant pour les cotations du marché financier, tandis que son intérêt à une histoire d'actualité peut être considérablement plus transitoire. Même quand des besoins ou des intérêts paraissent constants ; il n'empêche qu'ils peuvent subir des changements significatifs durant leur cycle de vie. De plus, ces changements peuvent survenir graduellement ou d'une façon plus inattendue. Il faut signaler aussi que l'intensité d'un intérêt peut varier sur le temps dans une voie qui est souvent difficile à prévoir. Un utilisateur peut avoir un persistant intérêt dans tous les documents ou information qui peuvent concerner un domaine en vogue, par exemple le réseau X25 (Transpac , Minitel) dans les années 1980 ; puis subitement cet intérêt peut s'effondrer conjoncturellement dès confirmation de l'avènement d'une autre technologie. Malgré les apparences , en réalité cela peut se dérouler dans un laps de temps relativement court. Les nouvelles technologies arrivant vers le grand public très lentement ; par un effet de traîne, un utilisateur peut commencer à s'intéresser à une technologie donnée, qui est en voie d'être supplantée par une autre ; et ne s'aperçoit de son retard qu'après une bonne avance dans ses recherches documentaires. Ses besoins en information, constants pour un temps changeront alors brusquement.

Parce que les systèmes de filtrage de l'information doivent opérer sur des

mesures de temps relativement longues, l'aptitude à observer, modeler et adapter à cette persistance, variation et interaction d'intérêts serait utile. Un certain progrès a déjà été fait dans cette direction. La modification et relance d'interrogation, permises par beaucoup de systèmes de recherche d'information, identifient explicitement des intérêts d'information continus et les changements que ces intérêts ont a subir. Jennings et Higuchi [44] ont établi un modèle de connections d'intérêts à long-termes pour un système d'apprentissage des nouvelles USENET . Leur modèle emploie un apprentissage surveillé pour gérer un grand espace de recherches, sans une ingénierie extensive de connaissances. Fischer et Stevens [47] emploient des techniques basées sur des règles pour suggérer des agents de recherche booléenne pour la même application. Le choix de paramètres pouvant être vraisemblablement compris aisément par les utilisateurs, les performances du filtrage tirent avantage de la participation d'utilisateurs dans le processus de mise en place de filtres. Sheth et Maes [38] ont utilisé un algorithme génétique pour faire évoluer les paramètres booléens d'agent de recherche, combinant les avantages des deux précédentes techniques de filtrage.

Ces trois approches incluent une certaine disposition pour des changements de modélisation de l'intérêt de l'utilisateur. Cependant, dans chaque cas, l'utilisateur "modèle" lui-même est chargé de discerner les intérêts à long terme, de l'usager. Le développement de modèles d'utilisateur pour lesquels la persistance, la variation et l'interaction des intérêts humains sont pris aussi vraisemblables que possible ; permettra d'améliorer la performance et l'utilisation de systèmes de filtrage de l'information. Bien que la structure de l'espace de recherche résultant soit plus complexe, les nouvelles techniques d'apprentissage par ordinateur devraient bénéficier des informations tirées des contraintes observées dans la formulation des intérêts humains réels.

Un riche vocabulaire des concepts sur ces intérêts peut aussi être employé pour améliorer la qualité de participation d'utilisateurs dans le processus de mise en place de filtres. En résumé, tandis que le filtrage de l'information continuera à bénéficier des recherches sur les systèmes de recherche d'information, des modèles qui reflèteront exactement la nature des intérêts humains seraient probablement développés, un jour.

IV - LE CROSS-LANGUAGE :

Le "cross - language" est l'aptitude d'un système de recherche d'information à permettre le lancement d'une requête dans une langue et récupérer les résultats dans une autre langue. Il signifie donc le recouvrement de documents , basé sur des requêtes formulées par un utilisateur dans une langue différente de celle des documents recherchés. Cette aptitude à émettre une requête dans une langue et recevoir le document y afférent dans une autre langue distingue la recherche d'information avec le cross-language (ou tout simplement multilingue) , de la recherche d'information monolingue. Bien que la recherche d'information monolingue soit en dehors de la portée de cette définition, il n'en demeure que les fonctionnalités du cross-language et de la recherche monolingue soient assurées par le même système.

Le nom qui désigne ce nouveau champ de la recherche d'information ne fait pas encore l'unanimité des spécialistes des sciences de l'information. Pourtant la conférence sur les systèmes de recherche d'information *SIGIR 96* a recommandé le terme "Cross-Language" que j'ai préféré garder tel quel sans lui chercher une traduction en français. Il faut noter aussi que le terme "cross-lingual" apparaît assez souvent ; et qu'il est légèrement plus près de la correction grammaticale anglaise. Mais "cross-language" est aujourd'hui adopté par tous les spécialistes du domaine et il est clairement le plus utilisé.
Le congrès AAAI[5] du printemps 1997 sur le cross-language (Cross-Language Text and Speech Retrieval) a clarifié quelque peu ces définitions par l'identification explicite des modalités impliquées.

Le DARPA (organisme de la Défense Américaine) a adopté l'expression "Gestion translingue de l'information " pour décrire un ensemble de fonctions qui incluent en même temps le cross-language , la visualisation de collections de documents multilingues ; et d'autres sujets qui concernent la gestion de l'information multilingue

[5] voir le site : http://www.clis.umd.edu/dlrg/filter/sss/papers/

V- CONCLUSION:

En plus de chercher de l'information sur Internet selon la méthode des outils de recherche tels que Alta Vista ou des services tel que *Excite Web Reviews* et *le Magellan Internet Guide*, les utilisateurs peuvent accéder aux articles recherchés grâce à une variété d'accès spécifiques aux sujets, tels que *ADAM, EEVL, OMNI et SOSIG*, qui les dirigera vers les ressources évaluées et choisies dans leur zone d'intérêt. Tous ceux-ci ont des inconvénients allant des résultats considérables d'Alta Vista vers les accès de sujets qui décrivent simplement soit la ressource, soit offrent un mécanisme d'enregistrement arbitraire. Aucun vocabulaire de qualité standard n'a été développé et les utilisateurs sont incapables d'évaluer les forces et les faiblesses de ces sites.

Un projet étudié par le *IEEE Computer Society Standards Activities Board* se propose d'utiliser les spécifications du système PICS pour améliorer la pertinence des articles mis en ligne. Selon les promoteurs de ce projet , le même mécanisme peut être utilisé pour un étiquetage plus important des ressources et peut offrir aux utilisateurs un mécanisme d'assurance de qualité couvrant un éventail de critères. Ceci évidemment dans le but d'allier un outil ou système de recherche d'information avec une fonction de filtrage.

Les avantages du système basé sur PICS sont relatifs au vocabulaire standardisé et aux limites qui peuvent être imposés selon des jugements de qualité.

Un travail considérable a déjà été entrepris selon l'évaluation et les critères de qualité des ressources. Ceci doit être bien sûr pris en considération en développant tout un ensemble de valeurs (labels PICS). Les deux études parmi les plus importantes sont DESIRE et le projet développé par Alison Cooke de l'Université de Wales Aberystwyth qui se sont terminés par des ensembles de critères d'évaluation pour différents types de ressources d'Internet, alors que les développements dans la communauté médicale, tel que le développement des codes de conduite pour les sites médicaux Web par *the Health on the Net Foundation* et le *British Medical Internet Association* , sont également utiles et devraient être étudiés.

Les labels de base de données du CIQM existants contiennent un mélange d'informations qualitatives et quantitatives: par exemple, non seulement ils détaillent le nombre d'enregistrement sur la base de données et les pourcentages d'enregistrement à partir de régions géographiques différentes, mais ils contiennent plus de vingt rapports d'assurance de qualité tels que, "Tous les domaines du texte sont vérifiés et étudiés", "Tous les auteurs pour chaque article sont indexés" ou "Il n'y a pas d'information reproduite dans cette base de données". Les labels de PICS peuvent contenir un mélange similaire de pratique (par exemple: l'auteur / propriété, type de source corporative, longueur, couverture des sujets, couverture et utilité géographique) et qualitative (indicateur de contrôle, exactitude, indication de l'examen , opportunité, etc).

Même avec ces quelques exemples, il existe des données suffisantes pour permettre une évaluation de la ressource: le nom de l'auteur lié à son standing professionnel et savoir qu'il écrit à partir d'un institut de recherche ou d'une université (par opposition à la maison); la longueur de l'article et son actualité ; le fait qu'il ait été utilisé dans sa production (faits confirmés, citations validées, etc.) et savoir qu'il est récent , revu et mis à jour régulièrement pour attester de la valeur du site.

Selon une partie du projet DESIR mentionné déjà, le projet SOSIG a produit une liste détaillée de critères de sélection de qualité pour l'accès au sujet. Cet outil de catalogage est conçu pour être utilisé par des accès au sujet afin " de définir ou d'affiner leurs critères de sélection de qualité ". Il y a cinq sections: méthode du domaine, critères du contenu, critères de forme, critères du procédé et méthode du traitement de collection. Le critère du contenu, par exemple, contient des sections sur la validité, l'autorité, la réputation , la précision , la compréhension, le caractère unique, la composition et l'organisation , la précision et la circulation du traitement. Sous chaque titre il y a des séries de critères qui sont formulées comme des questions avec des séries de suggestions et d'analyses qui peuvent être utilisées pour voir si une ressource rencontre un critère particulier. Selon la validité, le critère en question se présente ainsi:

Comment est la validité du contenu de l'information?
Est-ce que l'information semble bien recherchée?
Quelles sources de données ont été utilisées ?
Est-ce que les ressources répondent au but établi ?
Est-ce que le format a été dérivé à partir d'un autre format?
Est-ce que l'information semble impartiale (quand en fait, est-elle partiale ?)
Est –ce que l'information paraît ce qu'elle est?
Pourquoi cette information est là ?/ Quelle est la motivation du fournisseur de l'information quand il rend l'information disponible ?/ a-t-il un motif précis?
Est-ce que la ressource indique les autres sources qui peuvent être contactées pour la confirmation?
Est-ce que le contenu de la ressource est vérifiable - pouvez vous contrôler l'information?

Peu de références/ - -> /Beaucoup de référence
Les sources de données sont pauvres/ - -> /très bonnes
Bibliographie/ pas de bibliographie
Rapport de plan / pas de rapport de plan.
Rapport de plan soutenu/ - -> / non soutenu par le contenu
Copie des données disponibles ailleurs ; balance: sur document /sur CD ROM/électroniquement/ pas de copie
L'information a une option géographique /politique /autre.
Le site est fourni par une institution de recherche académique/ de publicité / de commerce /autre
L'information est incomplète / adéquate /complète
Maison de publicité/ - -> / article de référence.
Le contact e-mail de l'auteur / contact postal. /

Ainsi, par exemple, les utilisateurs pourront sélectionner les sites selon les conditions du sujet dans la voie normale, et filtrer ce qui ne contiennent pas un nombre adéquat de références; qui sont disponibles sur un autre milieu de publication; qui sont fournis par une institution qui est plus basse que l'éditeur dans " l'échelle" ; qui ne se rapportent pas; et qui ont des ressources de données moins adéquates. Une recherche sur Alta Vista, par exemple, se

terminerait toujours dans le même nombre de retours qui ont été approvisionnés avant l'étiquetage mais les utilisateurs du label de qualité PICS pensent que la plupart d'entre eux sont bloqués à partir de leur station de travail parce qu'ils n'ont pas trouvé le standard qu'ils ont eux-mêmes établi.

Pour que l'indexation de qualité de PICS fonctionne efficacement, le système général de PICS doit être adopté par les navigateurs afin que les utilisateurs puissent récupérer les données selon le critère de qualité tels que l'utilité géographique, l'autorité, le caractère unique en plus des conditions de recherche utilisées pour découvrir la ressource d'abord en établissant leur logiciel local convenablement.. Un filtre de qualité limite efficacement ce que les utilisateurs croient comme étant les meilleures ressources. Le filtrage étant établi à l'avance par des spécialistes de l'information ou varié selon une base de recherche par la recherche par des utilisateurs capables d'évaluer et d'utiliser les critères de qualité de l'information pour eux-mêmes. L'adoption de la recherche conforme au système PICS par les navigateurs courants d'Internet semble prendre place, probablement parce que celui-ci est déjà soutenu par le W3C.

Si tous les outils de recherche devaient incorporer le mécanisme de filtrage du système PICS, les utilisateurs termineraient avec un outil d'accès extrêmement puissant. Dans ce cas la limitation de qualité peut être entreprise par l'outil de recherche, répondant au besoin d'un filtrage/ traitement local à deux étapes.. L'interface de recherche peut incorporer quelques boutons couvrant un sous-ensemble de critères de qualité PICS et les utilisateurs fixeraient leurs labels en même temps qu'ils font entrer les termes de recherche. S'il reste seul, le filtre de qualité peut être établi selon la fonction à des niveaux moyens ou peut rester inopérant.…

Donc , probablement nous nous dirigeons vers la mutation des actuels outils de recherche de l'information vers de véritables systèmes de recherches d'informations avec des fonctions bien développées de filtrage , de cross-language et d'indexation. Cette dernière fonction , très importante dans la chaîne de traitement de l'information , sera intégrée aussi pour éventuellement permettre aux auteurs de choisir leur propre système de réfèrencement.

Partie C :

LES NOUVELLES PERSPECTIVES DE LA RECHERCHE D'INFORMATIONS

I- Introduction

II- Projets actuels :

5) **le projet NewsAgent**
6) **Formats d'affichage et préférences utilisateurs**
7) **Le projet SmartPush**
8) **Le projet EUrogatherer**

III-Conclusion

I- INTRODUCTION:

Depuis toujours, les bibliothécaires savent qu'une recherche d'information ne vise jamais à obtenir des renseignements pour eux-mêmes ; elle est faite en vue de leur exploitation extérieure dans des conditions précises. Il faut donc que ces conditions soient connues. En particulier, il convient de savoir :

- Qui est le demandeur.
- Quelle utilisation compte-t-il faire des informations.
- De quel délai il dispose.
- Quels documents connaît-il déjà sur la question et, d'une manière générale, ce qu'il sait déjà sur le sujet.
- Quelles langues peut-il lire !
- Sous quelle forme, préfère-t-il obtenir les informations.
- Quelle période et quelle aire géographique la question couvre-t-elle exactement !

La formulation des questions par les utilisateurs risque d'être imprécise ou ambiguë à plus d'un titre :

- ✓ d'abord au niveau de la description du sujet, qui peut être trop large ou trop restreint ;
- ✓ ensuite au niveau de l'utilisation envisagée des informations recueillies. Le même sujet peut être traité différemment par des documents de différents types, dont chacun peut être mieux adapté à une utilisation qu'à une autre. Par exemple, un article résumant les principales orientations d'un plan de développement économique peut en donner une vue d'ensemble, mais ne pourra pas permettre d'entamer un travail spécifique d'analyse économique. Il y faudrait dans ce cas le document du plan même.
- ✓ Enfin au niveau des conditions dans lesquelles les informations devront être employées. Faire établir une bibliographie d'une centaine de références et rechercher les documents correspondants n'a guère

d'utilité quand le demandeur doit produire une note de synthèse sur le sujet en un temps très court (vingt-quatre heures par exemple).

Autrement dit, la question la plus fréquente : « Quelles informations avez-vous sur tel sujet ? » devrait être formulée par une phrase du type : « Avez-vous sur tel sujet tel type d'information me permettant de faire tel travail, dans telles conditions ? ».

Cela nous amène à la notion de profil utilisateur classique tel qu'il est défini en bibliothéconomie ; c'est à dire comme une équation de recherche (ensemble structuré de descripteurs), exprimant les informations que l'utilisateur désire recevoir régulièrement d'un service de diffusion sélective de l'information (DSI). L'utilisateur peut construire lui-même son propre profil, en s'aidant d'un manuel que lui fournit le service de DSI ; mais le plus souvent, le profil est construit par un spécialiste de l'information, qui procède d'abord à un entretien détaillé avec l'utilisateur et à des essais qu'il soumet à son appréciation. Chaque envoi de résultats est accompagné d'une demande d'évaluation qui permet de corriger rapidement les insuffisances éventuelles du profil établi. L'utilisateur est invité à demander des modifications de son profil, au fur et à mesure que ses centres d'intérêt évoluent.

Les points d'accès ou de recherche sont les diverses caractéristiques d'une information ou d'un document à partir desquels peuvent être opérées à la fois, la recherche et la sélection. Ils sont exprimés par l'utilisateur dans sa question, par les indications qu'il donne sur le sujet, les dates, l'aire géographique, le type de document recherché, la langue, etc... Ils sont aussi fonction, d'une part du détail plus ou moins important de la description bibliographique et de la description de contenu (c'est à dire de leur présence dans la base d'information) et, d'autre part de la finesse du système de recherche, c'est à dire de la possibilité ou non de faire des tris dans les fichiers ou index, en fonction de ces caractéristiques. Ils portent en général sur les sujets traités, la date des informations ou des documents, l'aire géographique, l'auteur, le type de document (et donc de traitement des

sujets) ; et ils peuvent aussi parfois porter sur la langue, le volume, l'accessibilité, le numéro de rapport ou brevet, le lieu de publication, etc...

Partant de ce préambule, nous voudrions dans cette partie C présenter quatre projets lancés par des équipes de recherches éloignées géographiquement, les unes des autres, puisque appartenant à différents pays (le projet NewsAgent en Angleterre ; l'étude des préférences d'utilisateurs et des formats d'affichage au Canada ; le projet Smartpush en Finlande et le projet Eurogatherer en Italie avec la participation d'autres pays européens). Pourtant ces équipes sont reliées par le même objectif : améliorer la pertinence des documents retrouvés lors de processus de recherches d'information par une prise en compte du profil ou des préférences de l'utilisateur.

Le choix de ces exemples parmi tant d'autres n'est pas fortuit. Même si le but poursuivi est identique ; il n'en demeure pas moins que ces quatre projets se caractérisent par des options différentes :

- Le projet NewsAgent fonctionnant sur le principe de la diffusion sélective de l'information (DSI) prend en compte le profil de l'utilisateur ;
- L'étude des formats d'affichage s'intéresse aux préférences de l'utilisateur pour une information bibliographique plus en rapport avec les désirs supposés (définis statistiquement) de ce dernier ;
- Le projet Smartpush utilise une théorie des approximations pour une prise en charge des intérêts (théoriques) de l'utilisateur et s'éloigne complètement de la notion de profil classique de l'utilisateur ;
- Le projet Eurogatherer reprend le principe du projet NewsAgent avec des fonctions et des services plus évolués.

II- PROJETS ACTUELS :

1) Le Projet NewsAgent :

L'objectif de ce projet est de créer un service de préférence utilisateur basée sur un système de news (groupe de discussions usenet) et un service actualisé de sensibilisation pour le personnel de l'information et des bibliothèques avec un mélange de contenus courants, incluant des metadonnées (descriptions de documents). Le contenu sera constitué essentiellement d'articles et rapports tirés de revues, telles *Programm*, *Vine*, *Library Technology*, *Ariadne* et *le Journal of Librarianship and Information* Science. Les news et les fiches analytiques transmis aux utilisateurs sont fournis par The Library Association, The Institute of Information Scientists, UKOLN[6], The British Library, et LITC[7].

Le projet a démarré en avril 1996 et la Phase I a pris fin en mars 1998. Les partenaires continuent maintenant avec davantage de travail pour réaliser un service de préférence d'utilisateur, à pleine échelle. Les partenaires principaux du consortium impliqués dans ce projet, sont:

- LITC, South Bank University, Londres (Coordinateur de projet)
- CERLIM, Manchester Metropolitan University (Evaluation)
- Dept of Information and Library Studies, UWA, Aberystwyth
- Fretwell-Downing Informatics Ltd (développement technique)
- UK Office for Library and Information Networking (UKOLN)

Finalement le principe qui permet au projet NewsAgent de prendre en charge le profil utilisateur, est le même que celui de la DSI (Diffusion Sélective de l'Information). Les changements par rapport à la méthode classique, connue chez les bibliothécaires sont :

- L'intégration des éléments de metadata du Dublin Core dans la description du contenu ;

[6] UK Office for Library and information Networking
[7] LITC est une division de LISA/South Bank University, Londres.

- Les éléments qui définissent le profil utilisateur sont décrits par l'utilisateur lui-même ; sur un formulaire mis en réseau ;
- Les résultats sont reçus dans la boite de messagerie électronique de l'utilisateur.

Il faut noter que le projet vient d'entamer sa deuxième phase qui prévoit la généralisation de son emploi à plus grande échelle.

Nous avons tenu à citer ce projet pour le principal fait que c'est le premier, à notre connaissance à avoir associé la notion de profil d'utilisateur à une élaboration de requêtes dans un système d'information, basé sur une description de contenu par des éléments de metadonnées du Dublin Core.

2-) Profil utilisateur et format d'affichage :

Beaucoup d'auteurs ont relevé l'inutilité de certains champs qui se retrouvent dans les formats d'affichage, auprès de la plupart des utilisateurs. Crawford (1992) [40] se réfèrent à « l' arcana » des données montrées dans les enregistrements bibliographiques. Montrer des éléments inutiles et non voulus peut entraîner de la confusion et rebuter les utilisateurs pour poursuivre la suite de l'affichage (plusieurs auteurs cités par Wallace [51], 1984). Le coût de ces enregistrements inutiles a été également soulevé (Walace 1984 [40] ; IFLA, 1992).

Les études de Palmer (1972) [65] ont montré que 5000 utilisateurs à l'université du Michigan ont utilisé, en moyenne, 4,5 éléments des 20 éléments d'un affichage : il a conclu que plusieurs des éléments pouvaient être supprimés.

Des études antérieures (citées par Palmer, 1967) ont soulevé aussi l'utilisation de champs, dans des formats catalographiques, qui n'étaient consultés qu'une fois sur dix...

Ces remarques que nous venons de reprendre ont été citées par deux professeurs de la faculté des sciences de l'information de l'université de Toronto, Lynne C. Howarth et Joseph P.Cox, dans un article [17] qui fait la synthèse d'une étude de comportement d'utilisateurs devant des formats d'affichage bibliographiques. Des enregistrements de dix articles ont été

rendus accessibles par Internet, dans des OPACs de dix bibliothèques canadiennes (enregistrements conformes à l'AACR2[8]). En même temps que la collecte des résultats, des interviews d'utilisateurs ont été organisés. Ces derniers ont été invités à classer les champs suivant leurs préférences. Ainsi a été mis en place un format d'affichage construit sur des préférences d'utilisateurs. L'article écrit, compare ce format avec les formats habituels et en tire des observations.

Cela constitue une autre méthode d'introduction de la notion de préference d'utilisateur dans un processus de recherche d'information, à travers cette fois-ci, un élément cadre qu'est le format bibliographique.

3-) Le projet SmartPush :

Introduction :

Les développements récents dans le domaine des metadonnées suggèrent que le filtrage de l'information pourrait être fait selon la description du contenu réel des documents, par ces dernières. C'est en se basant sur ce principe qu'une équipe de recherche du Centre de Recherche *TAI*, de l'Université d'HELSINKI (FINLANDE) se sont lancés dans un projet dénommé SmartPush [140] et qui repose sur la mise en place de metadonnées basées sur une association de documents et de profils d'utilisateurs.

Dans le filtrage de l'information, les documents sont appariés avec des profils d'utilisateur.

Cela est basé dans une certaine mesure, sur la similarité ou distance entre une représentation de profils d'utilisateur et des documents. Les deux représentations et la mesure de distance devraient permettre des comparaisons significatives.

Dans le projet *SmartPush*, les auteurs suggèrent une représentation hiérarchique pour décrire des documents et des profils d'utilisateur en essayant de modeler les concepts connexes. Leur modèle inclut une

[8] Anglo-American Cataloguing Rules 2

mesure asymétrique de distance qui peut permettre de détecter des documents en tenant compte d'un profil d'intérêt de l'utilisateur.

Une représentation plus compacte de contenu de l'information peut être réalisée en reliant des metadonnées à chaque document. Ces Metadonnées doivent contenir une description du contenu d'un document multimédia ainsi que l'information bibliographique associée. Le filtrage automatique de l'information nécessite aussi une connaissance de l'utilisateur.

Cette connaissance peut être représentée dans un profil d'intérêts. Les metadonnées et le profil de l'utilisateur devraient avoir des représentations compatibles afin de permettre des comparaisons significatives possibles (Fig. 1).

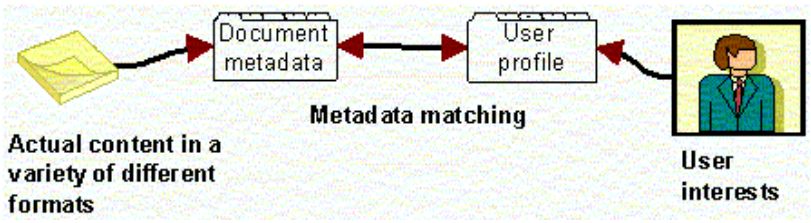


Fig.1 : Metadonnées du document et leur relation avec les intérêts de l'utilisateur

Pour leur projet, les auteurs ont développé une représentation hiérarchique du contenu du document et l'information du profil de l'utilisateur. A partir de cette représentation, les metadonnées et le profil de l'utilisateur sont associés pour pouvoir fournir à chaque utilisateur son besoin d'information. Le processus d'association est basé sur une mesure de distance qui permet de classer un ensemble de documents par rapport à des résultats et un profil donné dans une approximation des intérêts de l'utilisateur pour chaque document.

Evidemment les metadonnées sont pris dans leur sens originel, à savoir la description de divers aspects du contenu d'un document (sujet, mots-clés, type, source, format et les champs bibliographiques traditionnels). Ainsi les

metadonnées donnent une représentation compacte du document et peuvent être employé pour guider le filtrage sans l'accès au document original. Une fois que les metadonnées sont extraites du contenu réel, il devrait être possible de les transférer et de les traiter indépendamment et isolément du contenu originel. Alors cela permet d'opérer seulement sur les metadonnées au lieu du contenu entier. Mais les structures des metadonnées doivent être uniformes et compatibles pour traiter le contenu de sources différentes. Elles doivent aussi permettre de :

- Décrire des données structurées avec différentes valeurs ;
- Décrire différents formats de médias ;
- Supporter plusieurs langues ;
- Reposer sur un vocabulaire contrôlé ;
- Supporter plusieurs types de données.

Un document peut être attribué à seulement une parmi plusieurs catégories (une valeur d'un ensemble discret de possibilités, exemple : les types de médias Fig. 2) ou il peut être attribué à plusieurs catégories en même temps. Une voie plus distincte permettrait une diffusion discrète (un vecteur) avec des valeurs d'une gamme continue (attribution de poids pour les différents mots-clés). Certaines propriétés peuvent être représentées par une valeur d'une gamme continue ou dans le cas général, même avec une diffusion continue. Les données à structure plus compliquées, telle que des hiérarchies ou les réseaux, peuvent être représentés dans un format numérique en utilisant un code convenable.

Le terme *dimension* est employé pour décrire différents aspects du contenu. Les dimensions indépendantes pourraient être par exemple, des concepts d'actualité, des types de médias du document et la période de temps pertinente pour de nouveaux articles (ou documents).
La présence de plusieurs dimensions permet de filtrer les documents selon des aspects différents.

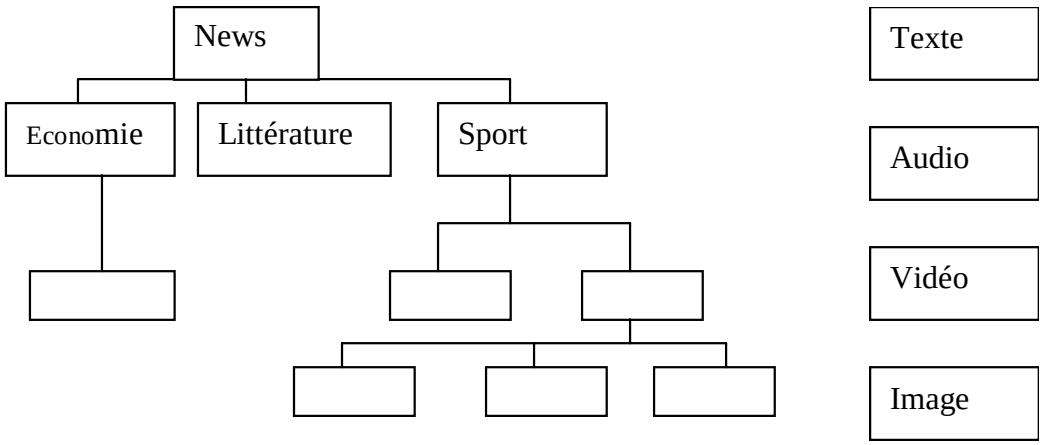


Figure 2 : Exemples de différentes dimensions de metadonnées et leurs structures dans un document d'actualité.

Le contenu du sujet (Subject matter) des *News* est décrit avec une structure hiérarchique. Les types de médias (Media type) sont présentés avec une valeur discrète d'un ensemble de 4 possibilités.

Pour les promoteurs de ce projet, vouloir filtrer l'information et la donner à un utilisateur dans une forme personnalisée requière de représenter les intérêts de celui-ci dans un format numérique ; et cela constitue un profil. Employé dans ce contexte, le profil de l'utilisateur ne contient pas d'information détaillée sur les préférences de l'utilisateur pour un logiciel donné, mais une information générale sur les intérêts de l'utilisateur relatif au domaine du filtrage. Dans le projet SmartPush, le but principal du profil de l'utilisateur est de permettre le filtrage des documents et leur organisation ou classement dans un résultat de requête, qui reflète les intérêts supposés de l'utilisateur. Evidemment il demeure qu'un profil d'utilisateur ne peut pas représenter les vrais intérêts à l'infime détail près ; tels que peut les percevoir l'esprit humain. Le profil de l'utilisateur représente une cartographie du vrai profil d'intérêt dans un modèle

d’espace plus compact (Fig. 3). L'espace est nécessairement abstrait, ce qui signifie que le profil de l'utilisateur soit basé sur une approximation des vrais intérêts de l'utilisateur dans une représentation spatiale du modèle.

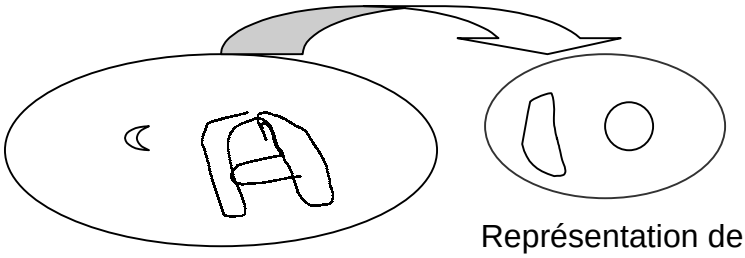


Figure 3 : Profil ou représentation des intérêts de l'utilisateur dans l'espace du modèle.

Les profils d’intérêt des utilisateurs évoluent et donc leurs besoins en information fluctuent aussi. Donc, la représentation de profil d'utilisateur doit être capable de s'adapter graduellement aux changements dans les intérêts réels de l'utilisateur.
Les intérêts d'un utilisateur peuvent être définis (par exemple liés à un certain nom de marque) ou vagues (exemple : les nouveaux développements dans l'intelligence artificielle).

La partie cruciale du système de filtrage du projet SmartPush est l’association de documents et de profils de l'utilisateur.
Pour comparer des documents et des profils, il faut d’abord avoir les moyens de mesurer ou d’estimer la similarité qui les lient ou ce qui les différencient. La mesure de distance est une mesure entre un profil et un document qui produit un nombre réel, non – négatif, reflétant la somme par laquelle ils diffèrent chacun de l’autre. Pour les besoins de la modélisation, les auteurs optent pour une méthode classique chez les mathématiciens : la méthode de calcul vectoriel. Dans ce modèle d’espace vectoriel qu’ils ont construit, les profils d’utilisateurs et les documents sont représentés par des vecteurs, avec des composantes pour chaque terme.

Ces composantes correspondent au poids ou fréquence de chaque terme dans le document et au terme donné représentant l'intérêt dans le profil. Le poids pour un terme donné dépend de la fréquence de ce terme dans le document spécifique en comparaison de la fréquence de ce terme dans la collection entière de documents….

Il est évident que cette méthode de recherche et filtrage de l'information s'applique seulement aux documents textuels. Il faut signaler aussi que pour chaque langue, il faudrait une liste séparée de termes et connaître au préalable la collection de documents sur laquelle doit s'effectuer la recherche. Un autre problème peut provenir de l'emploi de certains mots qui nécessitent des outils d'analyse de langue. Il n'en demeure pas moins que pour le moment, ce projet n'en est qu'à sa phase théorique ; et est encore loin de toute possibilité de donner lieu à une application pratique.

EXEMPLE :

si un profil d'intérêt d'une personne consiste en :

-Horaire de train……0,3 (poids)

-Recherches intelligence artificielle….0,2

-Annonces de vente d'appartement…..0,5

Essayons de faire correspondre des documents doc1, doc2 et doc3 aux parties du profil d'utilisateur. Le rôle du profil est distinct du rôle de chaque document et les mesures de distances sont symétriques.

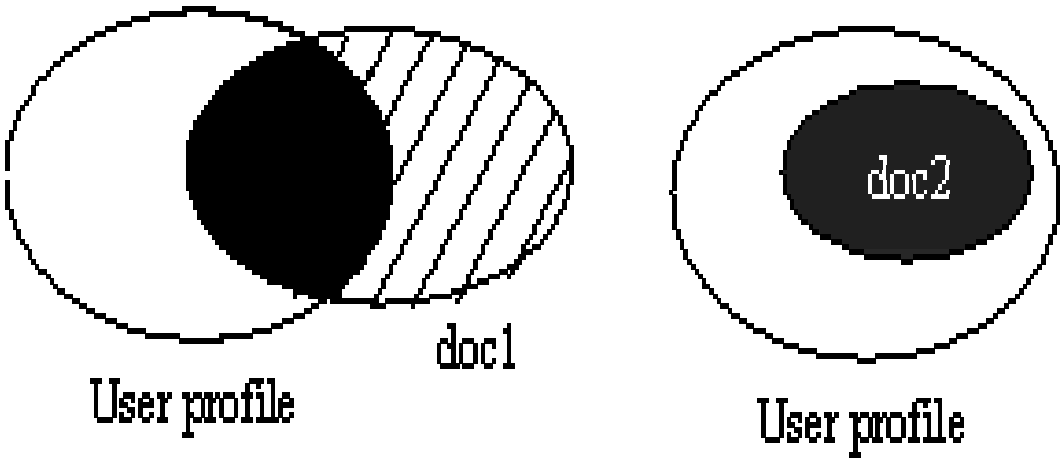
Examinons les documents doc1 et doc2 dans le tableau ci-dessous, montrés aussi dans la figure4. Les deux documents croisent avec le profil seulement dans le sujet d'intelligence artificielle. Mais il y a une différence entre ces deux cas: doc2 est lié seulement aux sujets du profil tandis que doc1 implique l'économie, qui n'est pas présenté dans le profil.

La figure 4 et le tableau qui suivent schématisent l'application de cette méthode à cet exemple. (Figure4 : Interaction de profils utilisateurs avec les documents D0C1 et DOC2).

Concept	Profil utilisateur	Doc1	Doc2	Doc3
Intelligence Artificielle	0,2	0,4	1,0	
Horaire train	0,3			
Annonces vente appartements	0,5			0,4
Economie		0,6		.0,6

TABLEAU 1

Figure 4



4)- Projet EUROgatherer[3]:

Ce projet a été lancé en janvier 1998, par le Groupe des Programmes Télématiques Européens, en partenariat avec les institutions et organismes suivants:

- I.E.I. - C.N.R., Pisa - ITALIE, Coordinateur du projet;
- Italia On-line SpA, Milan – ITALIE;
- Rank Xerox Research Center, Grenoble – FRANCE;
- Eurospider Information Technology AG, Zurich – SWISSE;
- Xarxa CINET SL, Barcelone – ESPAGNE;
- Université de Dortmund, Dortmund – ALLEMAGNE;
- Université de la ville de Dublin, Dublin – IRLANDE.

4.1- Introduction:

Une énorme quantité d'informations est créée et diffusée chaque jour grâce au réseau Internet. Il est quasiment devenu impossible pour les individus, de contrôler et de gérer ces masses d'informations. L'ironie du sort est que justement, il est de plus en plus difficile pour les utilisateurs de repérer et de retrouver une information pertinente. Les traditionnels systèmes de recherche de l'information sont pour la plupart orientés vers la recherche de textes non structurés de documents statiques. Les systèmes de filtrage de l'information prennent en charge surtout les documents courants tels les nouvelles des groupes de discussion, les messages électroniques et les divers documents circulant sur Internet. La collecte d'information est un nouveau domaine qui combine les caractéristiques de la recherche d'information, du filtrage de l'information, du langage naturel et de la représentation des connaissances ; et l'applique au nouveau domaine de documents structurés sous diverses formes (hypertexte, MIME, etc.) et différents formats (texte, PostScript, GIF, MPEG, etc.).

Ce domaine a récemment eu une évolution significative et a suscité beaucoup intérêts, avec l'apparition de plusieurs outils de recherche plus performants. Ces derniers scannent régulièrement le Web et produisent des index qui permettent de répondre aux requêtes des utilisateurs. Ils fournissent ainsi un service d'indexation de toute l'information électronique disponible sur Internet. Ces index textuels, structurés en pages HTML,

permettent une recherche d'information plus ou moins fructueuse pour les gens de la communauté du Web, mais ne fournissent aucun support personnalisé à un utilisateur en tant qu'individu. Certes, ils sont destinés pour un utilisateur indéterminé, et donc ils sont orientés pour répondre à des requêtes dynamiques, plutôt qu'à prendre en compte les exigences particulières de sélection à long terme pour un utilisateur spécifique. La technologie de collecte ou rassemblement d'information peut être appliqué à un nombre énorme de services en ligne, assistant les utilisateurs, par exemple dans la sélection des réserves d'archives ou autres documents de bibliothèques, articles d'actualité ou autres.

Partant de ces observations, les promoteurs du projet EUROgatherer se sont donnés pour objectif la conception et l'implémentation d'un système qui pourra offrir un service de collecte personnalisée d'information, basé sur la technologie des agents intelligents.

En particulier, les buts de ce projet sont :

- De filtrer et de contrôler le potentiel non limité de flux d'information depuis les sources vers les utilisateurs finaux;
- Fournir une information valable aux utilisateurs, sous une forme appropriée et en temps opportun;

Le système EUROgatherer devrait être en mesure de proposer les fonctionnalités suivantes :

- Obtenir et retenir un profil d'intérêt de l'utilisateur ;
- Agir, d'une façon autonome, poursuivant les buts fixés sans tenir compte de la présence de l'utilisateur en ligne (si celui-ci est connecté ou non au système où l'agent intelligent est installé);
- Accéder à diverses sources d'information;
- Créer une significative abstraction des documents retrouvés et les classer conformément aux fondements de leur structure et en accord avec leur contenu, dans un schéma de classification interne basé sur le profil de l'utilisateur ;

- Posséder un mécanisme de retour de pertinence, qui puisse permettre à l'utilisateur de renseigner le système par retour, sur le nombre de documents pertinents retrouvés.

En parallèle :

- Contrôler les changements fréquents des sources d'informations. Le système doit contrôler, à intervalles réguliers les adresses URL qui sont mises à jour. Si les changements sont significatifs, alors l'utilisateur devra être informé.

- Le système se lancera dans les recherches de documents sur le Web en utilisant les outils d'indexation et les meta-moteurs existants afin de retrouver ceux qui sont dans l'intérêt de l'utilisateur. Ensuite le système analysera les documents retrouvés afin que soient sélectionnés les plus proches des préférences de celui-ci.

4.2- Environnement de l'interface utilisateur:

Le système accédera en ligne (en réseau) à des bases de données afin de retrouver des données ou des documents qui intéressent l'utilisateur.

Du point de vue de son architecture, le projet EUROgatherer a pour objectif le développement d'agents intelligents basés sur un système à architecture multi-plans.

L'architecture du système est composée de trois plans :

- L'environnement de l'interface utilisateur;
- L'environnement du filtrage de l'information;
- L'environnement de la recherche d'information; (voir Figure 1).

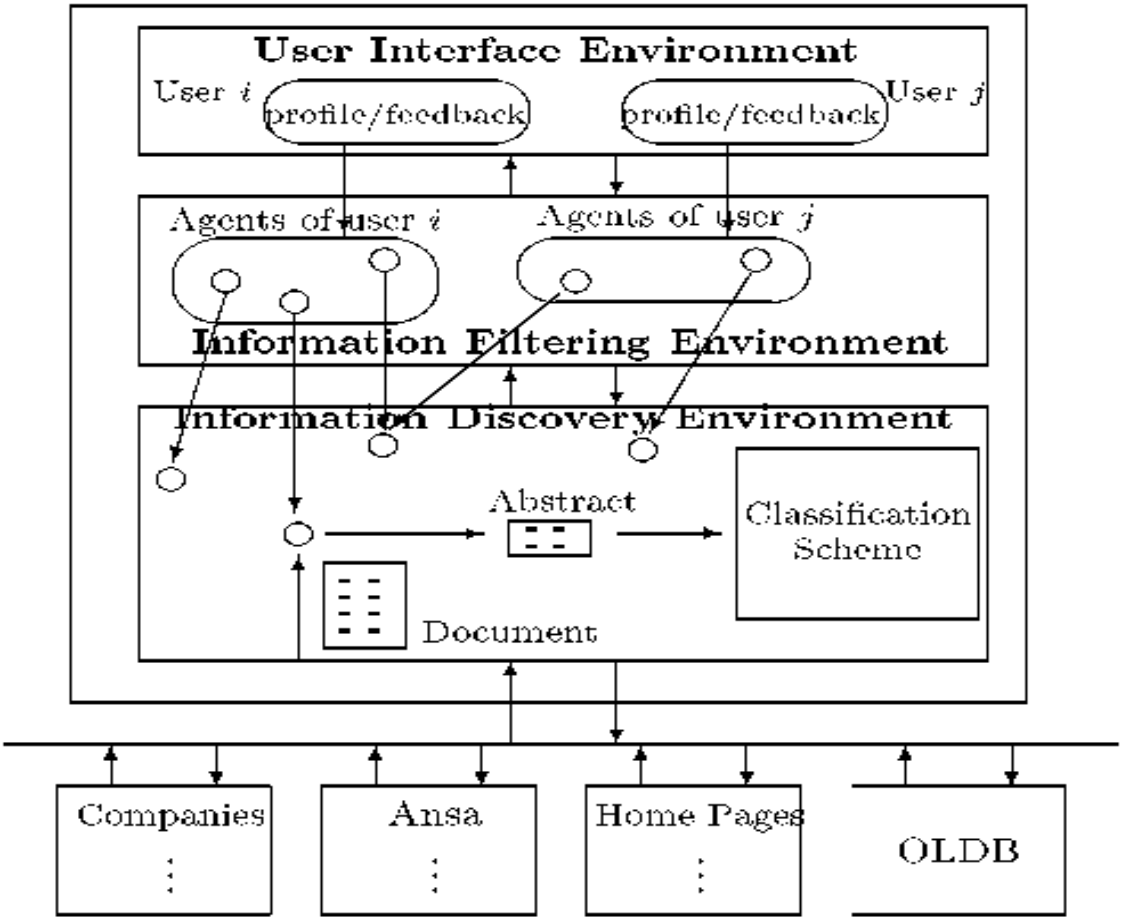


Figure 5: Architecture du système EUROgatherer

Deux différentes catégories d'agents logiciels devront être développées:

- Des agents de filtrage d'information et des agents de recherche d'information.

Les agents de filtrage d'information seront chargés de la personnalisation du système et son adaptation aux intérêts de l'utilisateur. Les agents de recherche d'information seront chargés de retrouver, de rapporter, de résumer et de classer l'information qui intéresse les utilisateurs.

Un important aspect de l'architecture du système est la séparation qu'il y a entre l'environnement du filtrage de l'information et celui de la recherche d'information.

Dans l'architecture du système proposé, la personnalisation de l'information doit être décentralisée du niveau de l'utilisateur au moment où la recherche

de celle-ci s'exécute sur le serveur (en réseau). Ce modèle choisi présente un certain nombre d'avantages:

- Dans un environnement multi-utilisateurs, chaque utilisateur aura son lot d'agents de filtrage, mais ils devront partager ensemble, les agents de recherche de l'information.

- L'introduction de plusieurs processus à niveaux entre l'information réelle et les attentes de l'utilisateur d'une information plus pertinente, en utilisant de nouvelles formes de filtrage et de recherche d'information.

L'environnement de l'interface utilisateur supporte les fonctionnalités suivantes :

1. L'acquisition du profil de l'utilisateur par le système;
2. Une présentation interactive des documents retrouvés par le système à l'utilisateur;
3. Le retour d'information de l'utilisateur vers le système, quant au nombre de documents pertinents retrouvés.

D'un point de vue global, nous distinguons dans ce système deux "acteurs" ou centres intérêts qui sont *les besoins d'informations de l'utilisateur* et *les sources d'informations,* reliés par trois fonctions principales qui sont : *la récupération de l'information d'Internet*, *la sélection de l'information pertinente* et *sa diffusion ou envoi à l'utilisateur*. Pour mettre en relation ces cinq variables, les promoteurs du projet EUROgatherer se proposent de mettre en place neuf services :

- **Gatherer Service (GS):** Ce service est chargé de collecter les pages Web (HTML) grâce à un robot (spider). Les pages ramenées sont transmises aux services Pull et Push pour indexation.
- **News Service (NS) :** Service chargé de gérer les nouvelles transmises par des groupes de discussion ou des agences d'actualités et de les transmettre au service Push pour indexation.
- **Wrapper Service (WS):** Service chargé de retrouver des pages Web par des interrogations de bases de données sur Internet ou en lançant des requêtes avec des moteurs de recherches. Les pages récupérées sont

transmises aux services Pull et Push pour indexation. Ce service se base dans la recherche de l'information sur les thèmes ou besoins définis dans les profils utilisateurs.

- **Pull Service (PullS):** Service spécialisé dans la gestion des documents récupérés du réseau Internet.
- **Push Service (PushS):** Service permettant un filtrage des documents rapportés d'Internet, conformément au profil de l'utilisateur. En fait ce service, en comparant les documents avec le profil de l'utilisateur, peut alerter l'utilisateur en l'informant de la présence de documents pertinents dans les résultats de la recherche.
- **Profile Service (PS):** Ce service est chargé de la gestion (stockage, maintenance et recherche) des profils des utilisateurs.
- **Delivery Service (DelS):** Ce service est chargé de la notification et de l'envoi des résultats de la recherche d'information, vers les utilisateurs ; en accord bien entendu avec leurs préférences Actuellement, la messagerie électronique constitue le principal mode de notification.
- **EUROgatherer User Service (EGUS):** C'est en quelque sorte le service d'accueil du projet EUROgatherer. A travers un navigateur Web classique, les utilisateurs peuvent y accéder, s'inscrire aux services de ce dernier et remplir le formulaire de renseignement de son profil.
- **Dispatcher Service (DispS):** Ce dernier service est chargé de répartir les requêtes vers les services de recherche d'information, principalement vers les services Pull et Push.

4.3- Définition du profil de l'utilisateur dans EUROgatherer:

Le profil personnalisé comportera d'abord le nom, l'adresse (y compris l'émail) et un numéro d'identification de l'utilisateur. D'autres données spécifiques y seront incluses, telles:

(i) Préférences d'envoi de l'information(adresse et horaire de réception);

(ii) Définition des besoins en informations;

(iii) Eventuellement les sources à interroger;

(iv) Signalement des relances d'interrogations ou des retours de pertinences.

Les sujets d'intérêt sont représentés dans le profil, sous forme d'ensembles de données. Chaque profil d'utilisateur doit contenir un lot de thèmes ou de sujets. Chaque sujet ou thème possède un nom et un identificateur unique. Par exemple les données de réception de l'information renseignent sur les modalités d'envoi des documents retrouvés par la recherche, quand ils sont jugés pertinents. Elles contiennent particulièrement des indications quant à l'adresse de réception, le moyen et l'horaire d'envoi. Les données relatives au contenu renseignent sur le contenu des documents, objet d'intérêt de l'utilisateur.

III-CONCLUSION :

Les différentes manières de tenir compte du profil ou des préférences de l'utilisateur dans le processus de recherche de l'information que nous venons de passer rapidement en revue nous font rappeler que les spécialistes de l'information ont pris conscience très tôt de cette nécessité. Certes, il est tout à fait trivial de reconnaître que le meilleur moyen pour récolter des documents pertinents, est de tenir compte du profil du demandeur, dans le processus de recherche de l'information ; mais il n'en demeure pas moins que les voies suivies jusqu'à maintenant n'ont apporté que des solutions partielles. Il est utile de rappeler que :

- ✓ La diffusion sélective de l'information tient compte du profil de l'utilisateur dans l'élaboration des requêtes ;
- ✓ Le projet NewsAgent, dans la mesure où il fonctionne sur le principe de la DSI, n'utilise lui aussi le profil de l'utilisateur que pour affiner et automatiser son système de requêtes documentaires ;
- ✓ Associer des préférences d'utilisateurs à des formats d'affichage constitue une nouvelle étape dans cette approche, dans la mesure où l'étude rapproche des éléments de description de ces préférences.
- ✓ Associer une description de documents avec des profils d'interêts de l'utilisateur constitue une approche originale dans le projet SmartPush ; même si ces derniers ne sont définies que par approximation.
- ✓ Le projet EUROgatherer n'apporte rien de nouveau aux systèmes d'alerte dèjà existants en matière d'information, à part une offre de plus de services qui se complètent. Cette multitude de services le composant, qui se voudrait être un avantage risque de l'alourdir et de le rendre encombrant à l'usage. Il apparaît plus comme un patchwork d'outils préexistants de recherche, de filtrage et de diffusion de l'information, plutôt qu'un système intégré d'alerte en matière de recueil de l'information, basé sur la notion de profil de l'utilisateur….

Il faut signaler par ailleurs que maintenant plusieurs moteurs de recherche d'information sur Internet offrent des possibilités de filtrage de l'information,

basées sur des préférences d'utilisateurs. Ces possibilités se basent souvent sur les mêmes caractéristiques :

- L'utilisateur doit définir les différents sujets ou thèmes qui l'intéressent ;
- Chaque sujet ou thème est défini dans les termes d'une catégorie sélectionnée parmi une liste proposée par l'outil de recherche d'information ;
- Pour chaque sujet, la requête se lance en cliquant sur la catégorie correspondante ; si une relance d'interrogation s'avère nécessaire pour un retour de pertinence, le moteur de recherche considérera cela comme une autre recherche sans lien avec la première. Le processus essentiel d'une interrogation interactive suivie, qui aurait pu garantir l'affinement de la recherche d'information est rompu à ce stade. Tous les moteurs de recherche actuels opèrent selon un processus « borgne » devant de telles situations. A chaque nouvelle interrogation, le moteur de recherche affichera de nouvelles listes de résultats, comportant un plus ou moins grand nombre de documents pertinents, souvent avec plein de redondance ; mais l'utilisateur clôturera sa recherche avec un taux de « bruit » impressionnant. Beaucoup d'utilisateurs se découragent et abandonnent la recherche quand ils se retrouvent devant de telles situations (surtout les utilisateurs habitués aux recherches guidées et fiables effectuées dans les services de références de bibliothèques) ; d'autres se contentent des premiers résultats ….

Partant de toutes ces expériences et tentatives faites dans le but d'améliorer la recherche de l'information, souvent dans le cadre général, parfois pour des communautés spécifiques ; Il nous est apparu tout à fait logique que l'étape qui devait suivre, associerait complètement un profil de l'utilisateur à des éléments de description de ressources: c'est l'objectif même du modèle DREPU que nous allons développer dans la partie suivante.

Partie D:

DESCRIPTION DE RESSOURCES ET PROFIL D'UTILISATEUR

I- Introduction

II- Construction du modèle DREPU

III- Définition des éléments de description de ressources

IV- Définition des éléments de profil d'utilisateurs

V- Définition des éléments de metadonnées de DREPU

VI- Fonctionnement du système DREPU

VII- Evaluation

ANNEXE 1 : Plate-forme logicielle du système DREPU

ANNEXE 2 : Les algorithmes

CONCLUSIONS

I-INTRODUCTION:

Une production d'information en évolution permanente et une incroyable prolifération des ressources augmentent considérablement le volume d'information à consulter pour obtenir une réponse pertinente à une requête donnée.

En fait, dans les systèmes de recherches documentaires en texte intégral, il est toujours possible de retrouver des documents contenant un des termes de la question, sans pour autant qu'ils soient pertinents. Le fait de retrouver de pareils documents est appelé « *bruit* ». Celui-ci peut constituer une gène certaine dès que le volume des réponses dépasse un seuil tolérable, surtout si on récupère l'intégralité des documents, alors que l'information effectivement utile dans ces documents, peut être très courte. Evidemment, ces systèmes tout en proposant presque toujours une réponse à la demande de l'utilisateur, n'arrivent en réalité à répondre que partiellement à ses besoins. En effet, la particularité de la recherche d'information provient du fait que l'utilisateur collecte des données pour un besoin précis, ou pour la résolution d'un problème bien défini. Dans ce cas-là, le système ne possède aucune information sur le contexte dans lequel il effectue sa recherche, ou sur les buts qu'il poursuit.

Rappelons que pour réaliser une classique diffusion sélective de l'information (DSI), le bibliothécaire établit une équation de recherche (souvent constituée d'un ensemble structuré de descripteurs) en procédant à un entretien détaillé avec l'utilisateur. Les points d'accès ou de recherche recueillis sont les diverses caractéristiques d'une information ou d'un document à partir desquels peut s'opérer la recherche ou la sélection. Ils sont exprimés par l'utilisateur dans sa question par les indications qu'il donne sur le sujet, les dates, l'aire géographique, le type de document recherché, la langue du document, l'auteur, l'éditeur, etc. De la même manière, pour qu'un système documentaire automatisé fournisse des réponses satisfaisantes, il faut qu'il ait une certaine connaissance du problème que l'utilisateur se pose. Si l'on cherche l'information sur un réseau, un autre fait vient amplifier le bruit.

Sur un réseau comme Internet où circulent des informations issues de multiples serveurs répartis un peu partout dans le monde, des documents hétérogènes, sont présentés au même niveau, sans distinction particulière de domaine (la physique, la chimie, l'économie...), de nature (on retrouve pêle-mêle des images, du texte, du son), de contenu (pages personnelles, catalogues publicitaires, publications scientifiques,...), ou de format (HTML, Postscript, texte,...).

Pour pallier les insuffisances de l'indexation de texte intégral et avoir une meilleure connaissance du fonds, les systèmes documentaires ont rajouté dans la description des documents, des critères externes à leurs contenus. Ainsi en bibliothéconomie classique, la dimension d'un ouvrage, son nombre de pages, etc..., sont autant de critères supplémentaires permettant d'améliorer la gestion du fonds, mais il est rare qu'un utilisateur se serve de ces critères pour sélectionner des documents.

Grâce aux systèmes de gestion de fichiers ou aux systèmes de gestion de bases de données, la recherche d'une notice par l'ensemble des champs (zones) la décrivant est devenue possible ; des champs définissant des caractéristiques externes au contenu ont ainsi pu être rajoutés : le pays et le champ disciplinaire de l'auteur, le nom du laboratoire, etc.

Ces systèmes documentaires permettent aujourd'hui de guider l'usager, par un processus de recherche d'information affiné (souvent en plusieurs étapes) de remonter au document pertinent, qu'il pourrait ne pas retrouver par une recherche simple.

Ce processus est très bien illustré par le schéma de CARL LAGOZE [109] (Digital Library Scientist / Cornell Universty / USA), que nous proposons ci-dessous :

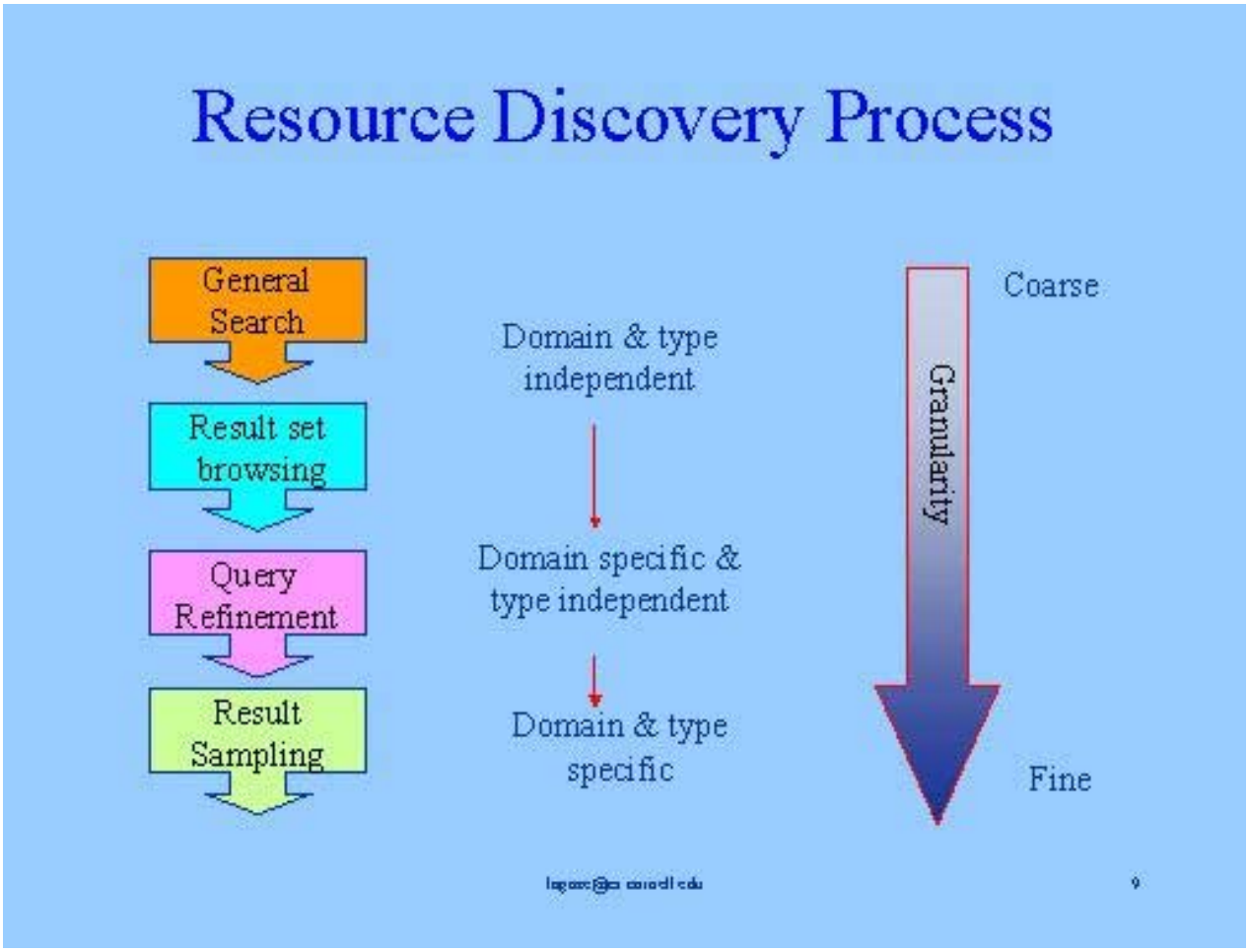


Image No 1

En réalité, le plus souvent, dans une opération de recherche documentaire classique, l'utilisateur se contente de formuler une requête donnée, puis le système apparie les mots de celle-ci avec ceux du dictionnaire qu'il possède et génère ainsi une réponse. Dans le cas des systèmes documentaires en texte intégral, il est toujours possible de trouver des documents contenant un des termes de la question, mais cela ne veut pas dire qu'ils seront vraiment pertinents pour l'utilisateur….

De même que pour les documents sur support "papier" , les documentalistes ont crées des éléments de description de ressources constituées par les notices catalographiques pour suppléer au manque de pertinence des méthodes d'indexation du texte intégral, pour les documents numériques, les spécialistes des sciences de l'information ont proposé la

description des documents par les metadonnées. Les metadonnées sont considérées comme " des données supplémentaires pour décrire les données du Web". Il y a plusieurs approches pour construire ces données supplémentaires. L'objectif est de fournir une description efficace de l'information du Web pour un traitement et une analyse efficaces et rationnels. L'étude de plusieurs standards de metadonnées nous amena à penser qu'il est indispensable de développer un modèle simple , à partir du Dublin Core et de l'utiliser pour indexer l'information numérique. Nous pensons aussi nous baser sur les développements du W3C (World Wide Web Consortium) , notamment nous inspirer des recommandations de Tim Barners L. [139] (créateur du Web) qui prévoit qu'un jour les metadonnées seront complètement intégrées dans les systèmes de recherches de l'information pour améliorer ceux-ci et donc la pertinence des documents retrouvés. Nous avons dans notre esprit aussi les propriétés du PICS (Plateform for Internet Content Selection), développé par le W3C, à l'origine dans un but de filtrage d'un certain type d'information (violence, pornographie) ; qui associe une fonction de description à une fonction de filtrage. Des auteurs (Stu Weibel et Eric Miller d'OCLC [24]) ont déjà fait remarquer la nuance qu'il y a dans le type de description du système PICS, tenant compte de la sémantique du contenu ; et dans le type de description de contenu, réalisé avec des éléments de metadonnées.

Nous avons étudié aussi la majorité des systèmes de recherche d'information, en projet ou dèjà réalisés, intégrant une fonction de filtrage basée sur les préférences ou le profil de l'utilisateur. Nous nous sommes aperçus que selon le système ; il était question parfois de :

- L'intérêt de l'utilisateur ;
- Préférences de l'utilisateur;
- Modélisation de l'utilisateur;
- Profil de l'utilisateur.

Si au premier abord, par extrapolation par rapport à l'objectif final visé par l'intégration d'une de ces notions dans le processus de recherche de l'information, qui s'avère n'être qu'une quête de pertinence, nous pouvons tolérer un sens commun approché à ces quatre expressions ; il n'en demeure

pas moins que dans le fond, de subtiles nuances existent.

L'intérêt de l'utilisateur repose sur une approximation de ce qui pourrait être utile à un utilisateur. Il est défini par le médiateur en information ou le système de recherche d'information automatique, sans participation active de l'utilisateur.

Cette notion cadre bien avec des projets tels que celui que nous avons cité dans le chapitre précédent : SmartPush[140]. Le processus de filtrage de l'information dans ce projet repose d'ailleurs sur une théorie des approximations, basée sur le calcul vectoriel de distances.

Les préférences de l'utilisateur correspondent à un choix de ce que voudrait voir afficher ou avoir comme information, le commun des utilisateurs. C'est une notion que nous retrouvons, par exemple, dans l'étude des formats d'affichage de Lynne Howarth (Université de Toronto/ Canada).

La modélisation de l'utilisateur repose sur une analyse permanente du comportement de celui-ci, pris en pleine phase de recherche d'information. Pendant ce processus de recherche, un système de modélisation collecte les préalables et les buts de l'utilisateur et construit un modèle, basé sur une structure de données bien définies. Ce système de modélisation est bien sûr adaptatif : ses connaissances de l'utilisateur évoluent en permanence et lui permettent de s'adapter aux futures demandes de celui-ci [9]. Finalement ce système d'apprentissage du comportement de l'utilisateur fonctionne selon un processus de captation de données, qui résultent toujours d'une interprétation causale ou cognitive. Mais ces données, malgré qu'elles soient objectives d'apparence, ne sont souvent qu'une transposition des connaissances des constructeurs du système….

Le profil de l'utilisateur est utilisé depuis toujours dans les bibliothèques traditionnelles, par les spécialistes de la Diffusion Sélective de l'Information (DSI), qui y ont recours pour satisfaire des requêtes personnalisées de leurs usagers. Tous les systèmes qui actuellement utilisent cette notion pour traiter les requêtes des utilisateurs, tels les projets NewsAgent et EUROgatherer se basent sur les mêmes principes, à savoir :

- Qui est le demandeur ? quelle est sa spécialité ? son niveau de qualification ?
- Quelle est sa communauté d'appartenance ?

- Quelle est la destination des informations ou documents recherchés?
- De quel délai dispose-t-il pour sa recherche?
- Quels documents connaît-il déjà sur la question et, d'une manière générale, ce qu'il sait déjà sur le sujet ?
- Quelles langues peut-il lire et comprendre ?
- Sous quelle forme ou format, désire-t-il obtenir les informations ?
- Quelle période la question couvre-t-elle exactement ?

Lorsque nous comparons les diverses observations que nous venons d'effectuer, nous constatons la mise en relief de deux approches. Celle de considérer les systèmes de recherche d'information dans une perspective logique et computationnelle : approximation de l'intérêt de l'utilisateur, modélisation de l'utilisateur. Une deuxième approche plus classique s'oriente vers des systèmes ou outils qui sont l'extension de la pensée humaine et de son mode de perception : projets basés sur le principe de la DSI (même la technologie PUSH). Ces deux approches aident certes à concevoir des systèmes de recherche d'information spécifiques ; mais sans toutefois annihiler la complexité du problème. Signalons à décharge pour la première approche qu'il n'est pas aussi aisé qu'il le parait de modéliser les comportements humains et leur mode de perception de l'information. Les systèmes qui en résultent peuvent être très bons sur le plan syntaxique ; mais souvent ignorent complètement la sémantique.

Plutôt que de voir dans ces deux approches des positions antagonistes, nous voudrions nous convaincre d'une possible complémentarité de ces deux perspectives qui nous permettrait de construire notre propre modèle générique que nous appellerons : DREPU….

II-Construction du modèle DREPU :

Bien qu'il y ait plusieurs standards de metadonnées proposés pour décrire les documents et les informations sur le Web, la plupart d'entres eux furent développés pour des buts spécifiques, telles que l'indexation et la sélection des documents, souvent propres à des communautés données (par exemple : standard de metadonnées GEM pour la communauté de l'éducation américaine et le standard de metadonnées EDNA pour la communauté de l'éducation australienne). De par sa simplicité et sa souplesse, le Dublin Core constitue pour nous le meilleur standard de metadonnées pour la description de ressources sur Internet. Ses propriétés d'extensibilité et de répétitivité permettent de compléter ses éléments par d'autres plus spécifiques et ainsi obtenir un standard de metadonnées dérivé, avec des objectifs différents ou tout simplement orientés. Rappelons que les éléments du Dublin Core ont une fonction de description de ressources électroniques, donc d'indexation en vue d'une prise en charge par un système de recherche d'information. Si nous leur rajoutons des éléments de profil d'utilisateur, cela reviendrait à compléter le système de recherche d'information, par une fonction de filtrage.

Ce point de vue a été dèja mis en exergue par deux chercheurs d'OCLC (On-line Computer Library Center), Stuart Weibel et Erik Miller [24] ; qui sont parmi les plus actifs du Groupe qui suit les développements du Dublin Core. Ils ont commencés par faire un parallèle entre le système PICS (Plateform for Internet Content Selection) et les metadonnées, d'un côté, et AACR2 (Anglo-américain Cataloguing Rules 2) et USMARC, d'un autre côté. AACR2 s'intéresse à la sémantique et au sens du document; tandis que USMARC s'intéresse à la structure et à la syntaxe du document; de même, si le système PICS s'intéresse à la sémantique et au sens du document (pour un objectif évident de filtrage d'information), les metadonnées prennent en compte la structure et la syntaxe (pour un objectif de simple description de l'information sans tenir compte de son sens). Ils ont ensuite conclu que l'association de ces deux concepts, serait la meilleure solution pour une meilleure information sur le réseau.

Construire un standard de metadonnées avec une finalité bien précise, à partir des 15 éléments ou d'une partie des éléments du Dublin Core rejoint

un principe qui a commencé à se dégager des développements de ce dernier, dès le deuxième workshop qui s'est tenu à l'université de Warwick, en avril 1996, en Grande Bretagne.

Sandra D. Payette (Digital Library Research Group / Cornell University) [132] décrit cela, dans une communication donnée lors d'un séminaire en novembre 1997, par ces quelques mots :

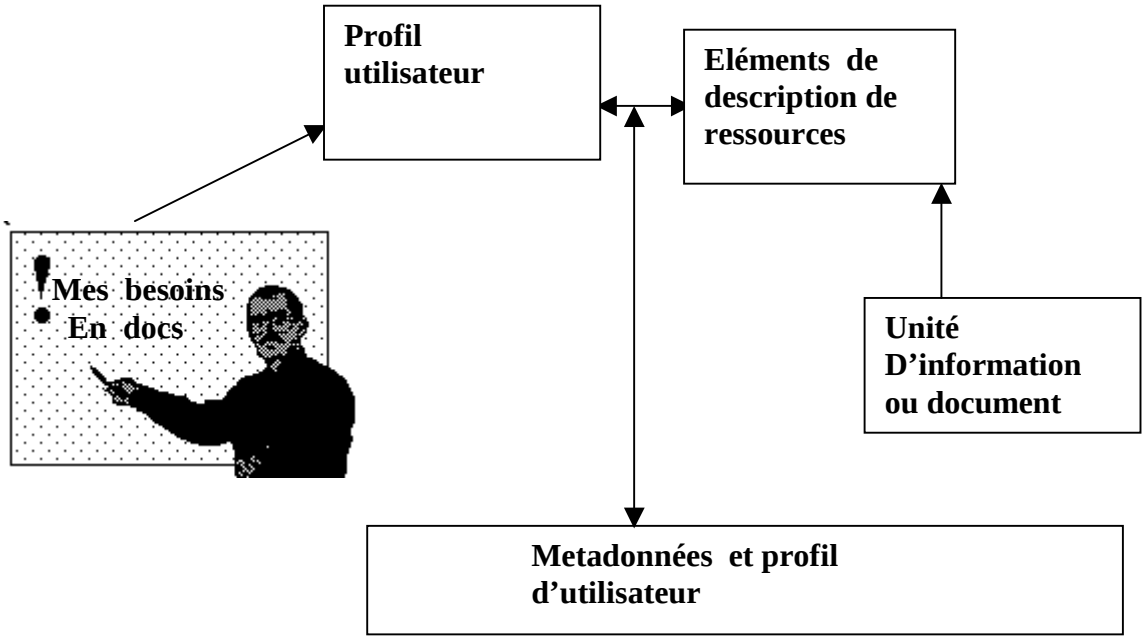
- "*Les enregistrements du Dublin Core sont l'ancre de tout système de description.*
- *Le Dublin Core est le bon choix pour les lignes de base des paquets de description qui peuvent être complétés par d'autres paquets spécialisés, au sein d'un même container.*
- *Les éléments du Dublin Core sont un dénominateur commun pour la recherche d'information dans des ressources disparates.*

Nous rajouterons pour notre part que les propriétés de répétitivité et d'extensibilité des éléments de description du Dublin Core permettent à toute communauté de les utiliser tous ou en partie pour une description de ressource; et d'adjoindre d'autres éléments à portée plus spécifiques.

Le modèle DREPU :

Partant de ces postulats, nous avons conçu notre modèle *DREPU* (Description de Ressources Et Profile Utilisateur) qui s'appuie sur le développement d'un standard de metadonnées basé sur l'association d'éléments de description de ressources pris parmis les éléments du Dublin Core et d'éléments de profil d'utilisateur définis par l'usage.

Evidemment associer dans une description de ressources, une caractérisation de ces dernières avec une caractérisation du profil de l'utilisateur entraînera nécessairement un système d'interrogation à relance.



-Figure 6-

L'interrogation à relance :

Dans un système de recherche d'information classique, l'interrogation est basée sur une élaboration d'une équation de recherche, qui permet d'avoir un premier résultat, lequel pourrait éventuellement être affiné par une deuxième requête ; mais cette dernière opération constitue une option non obligatoire. Dans ce que nous appelons un système d'interrogation à relance, nous supposons que la première interrogation ne constitue en elle-même qu'une ouverture de session interactive avec le système, avec qui s'ensuivra un dialogue. Nous imaginons que les relances permettraient de rentrer dans le système, la caractérisation de l'utilisateur. Le système de recherche d'information comparera les informations rentrées avec les informations contenues dans les éléments de metadonnées, et fournira à l'utilisateur les références du ou des documents répondant à l'interrogation.

En Australie, un organisme de recherche, le RDN-CRC rattaché au DSTC (Distributed System Technology Center) , a mis en place un outil de recherche d'information *HotMeta* [137] , basé sur le Dublin Core , qui utilise les interrogations à relance pour tenir compte des préférences de l'utilisateur, dans un processus de recherche d'information. Nous pensons que si les éléments de metadonnées rentrant dans la description de

ressources comportent des éléments pertinents de profil d'utilisateur, nous pourrons limiter le processus de recherche d'information à une seule relance. C'est ce que nous avons tenté de réaliser avec le système DREPU.

Le lancement de la requête se fait en orientant la recherche sur un ou plusieurs éléments de description de ressources. Nous obtenons ainsi un premier résultat brut qui recense tous les documents pouvant répondre à la question, et contenu dans la base interrogée. La relance de l'interrogation permettra une mise en correspondance du ou des éléments de profil d'utilisateur choisis par l'usager , avec les éléments de profil contenus dans les metadonnées des documents repérés par la première interrogation. Cette opération constitue la phase de filtrage du système qui éliminera tous les documents ne correspondant pas au profil choisi par l'utilisateur qui a lancé la requête.

Genèse du modèle DREPU :

Ce modèle est en fait le fruit du mûrissement de plusieurs années d'observations des systèmes de traitement de l'information en général et des systèmes de recherche de l'information en particulier. Depuis que j'ai commencé à m'intéresser aux sciences de l'information, j'ai toujours prêté une attention particulière à la manière avec laquelle certains documentalistes ou bibliothécaires prenaient en charge une recherche d'information guidée en présence d'un utilisateur ou établissaient un profil d'utilisateur dans le cadre de la préparation d'une DSI (Diffusion Sélective de l'Information). Il m'arrivait au début de faire moi-même des recherches d'informations infructueuses et de reprendre ces recherches avec ces personnes et le résultat me surprenait à chaque fois. La façon avec laquelle ces personnes associaient des éléments de profils utilisateurs classiques dans le processus de recherche d'information était pour moi un rituel porteur et efficace dans ces sanctuaires de l'information que sont les services de références des bibliothèques. J'ai vainement cherché à retrouver cette fiabilité dans tous les systèmes automatisés de recherche d'information que j'ai eu la chance d'utiliser par la suite tels que ceux des bases de données de QUESTEL , du CEDOCAR accessibles les premiers temps par vidéo

texte (Minitel) et en asci , des bases de données médicales (Medline de la National Library of Medecine ; bases de données de Silver-Platter et d'Embase ; base textuelle Adonis etc…) et finalement de plusieurs autres systèmes de recherche d'information en ligne, depuis l'avènement d'Internet. Et j'ai toujours pensé que cela ne pourrait advenir qu'avec un système qui simulerait le même cheminement ou processus dans l'interrogation qu'un spécialiste de la DSI ; c'est-à-dire que la requête tienne compte des éléments du profil utilisateur. Pour cela il fallait que ces derniers soient pris en compte lors de l'indexation ; facilité qui ne pouvait se concevoir avant l'apparition des metadonnées et surtout du Dublin Core. Il faut signaler que les canons du catalogage expugnait complètement l'éventualité de l'existence de champs contenant de pareille information dans un format catalographique classique. Il est vrai aussi que le catalogage a été créé au départ pour confectionner des fichiers de notices et des catalogues pour les besoins d'une consultation manuelle. Même avec l'apparition des catalogues en ligne (OPACs), les intitulés des champs des formats catalographiques sont restés pratiquement les mêmes. Les quelques études[9] qui ont été faites sur ce sujet se sont concentrées plutôt sur l'utilité des champs existants que l'éventualité d'en rajouter d'autres. Pour l'information numérique, le Dublin Core est venu bousculer complètement ces normes ; et quand j'ai pris connaissance de ce standard de metadonnées, j'ai commencé alors à envisager la faisabilité d'une intégration d'éléments de profil utilisateur dans un format de description de ressources, sans toutefois entrevoir la manière.

Le véritable déclic qui m'a mis sur ma voie s'est produit lorsque j'ai commencé à travailler mon mémoire de DEA sur la structuration des éléments de PROFILDOC (projet du laboratoire RECODOC de l'université Lyon1) en standard de metadonnées, dérivé du Dublin Core.

Mon modèle est né à cette époque-là ; mais il m'a fallu le mûrir, le comparer avec tous les projets existants dans le même axe ; trouver les moyens de le formuler ; réfléchir à des outils logiciels de mise en oeuvre ; introduire la notion d'interrogation à relance pour imager le documentaliste qui prépare

[9] Citées par Lynne Howarth (voir page 90)

sa DSI ; argumenter le choix des éléments de description de ressources et des éléments de profil utilisateur. Pour ce faire j'ai commencé d'abord par étudier tous les systèmes réalisés ou en projet qui partage les mêmes objectifs, à savoir tenir compte du profil de l'utilisateur dans un processus de recherche d'information. Cela m'a permis d'enrichir mon modèle, tout en me confortant dans mes choix. J'ai fais appel aux avis éclairés des professionnels de la DSI, aux niveaux de bibliothèques et de structures documentaires que j'ai considéré être les plus en vue en ALGERIE quant au choix des éléments de description de ressources et des éléments de profil utilisateur. J'en ai profité lors de ces consultations, pour exposer aussi ma théorie à mes interlocuteurs. Certes il m'a fallu expliquer cela longuement, avec beaucoup de pédagogie et en me referant à tous les travaux qui ont été faits dans le même domaine. Mais en fin de compte les échos positifs que j'ai eus à chaque fois de professionnels généralement conservateurs et très critiques vis à vis de tout ce qui bouleverse les règles habituelles, constitue pour moi une véritable validation de mon modèle….

III- DEFINITION DES ELEMENTS DE DESCRIPTION DE RESSOURCES:

Dans le système DREPU, les éléments de description de ressources ont pour objectif de décrire le document ou l'objet-document (document-like objet pour les Anglo-saxons), afin de permettre son indexation. L'objet-document correspond en fait à une unité d'information telle qu'une page web, une image, un texte, une séquence vidéo ou audio, etc.

Ces éléments sont analogues aux étiquettes des champs d'une notice catalographique. Or une communauté de professionnels de l'information s'est concertée depuis 1995 pour mettre en place un standard, le Dublin Core, comportant 15 étiquettes de champs pour prendre en charge le même type d'information. Donc il m'est apparu inutile de redéfinir d'autres éléments de description de ressources pour mon système ; par contre la souplesse du Dublin Core qui fait que chaque élément est optionnel m'a permis de ne pas prendre tous les éléments de ce dernier.

J'ai considéré dans une première approche les éléments suivants :

- Titre ;
- Auteur ou créateur ;
- Sujet et mots clés ;
- Description ;
- Source ;
- Type de ressource ;
- Format ;
- Date ;
- Langage ;
- Editeur ;
- Identifiant de la ressource.

Bien entendu, parmis tous les professionnels avec qui j'ai eu des consultations, certains m'ont reproché le fait de n'avoir pas pris tous les éléments du Dublin Core. En fait, j'ai pris les éléments qui se retrouvent généralement dans les formats bibliographiques ; sans exclure la possibilité d'intégrer d'autres si des impératifs d'usage les recommandent, dans un deuxième temps. Evidemment j'ai eu aussi droit aux avis de ceux qui

pensent que j'ai pris trop d'éléments et qu'il fallait me limiter aux sept éléments communs à tous les formats bibliographiques. Toutefois ce débat m'a poussé à laisser le système DREPU ouvert quant au nombre d'éléments de description à prendre en charge….

Définition des éléments de description de ressources de DREPU:

Les éléments de description de ressources du système DREPU étant pris parmis les éléments de metadonnées du Dublin Core, nous pouvons donc garder les mêmes définitions et les proposer dans le même ordre et la même forme. Nous tenons à rappeler encore une fois que le standard de metadonnées du Dublin Core doit être considéré comme un standard générique permettant à toute communauté de l'adapter dans des projets spécifiques[10]. C'est en nous basant sur ce postulat que nous nous sommes permis donc de rapporter à DREPU les définitions suivantes :

1.Titre :

Etiquette: titre

Le nom donné à la ressource par le créateur ou l'auteur.

2.Auteur ou Créateur :

Etiquette: créateur

La personne ou l'organisation principalement responsable de la création du contenu intellectuel de la ressource.

3. Sujet et mots-clef :

Etiquette: sujet

Le sujet de la ressource , qui sera décrit par un ensemble de mots-clefs ou de phrases qui précisent le sujet ou le contenu de la ressource.

4. Description :

Etiquette: description

[10] Voir page 138 la référence à Sandra D. Payette [132]

Une description textuelle du contenu de la ressource, y compris un résumé, dans le cas d'objets tels que des documents,ou une description du contenu dans le cas de ressources visuelles.

5. Source :

Etiquette: source

Une chaîne de caractère ou un nombre, utilisé pour identifier de façon unique le travail d'où la ressource est dérivée, si applicable. Par exemple une version PDF d'un roman peut avoir un élément source contenant un numéro ISBN correspondant à la version physique du livre à partir de laquelle la version PDF a été réalisée.

6. Type de ressource :

Etiquette: type

La catégorie de la ressource, telle qu'une page personnelle, un roman, un poème, un document de travail, un rapport technique, une dissertation ou un dictionaire.

7.Format :

Etiquette: format

Le format de la ressource, utilisé pour identifier le logiciel et, éventuellement, le matériel qui peuvent être nécessaires pour afficher ou traiter la ressource.

8.Date :

Etiquette: date

La date à laquelle la ressource a été publiée dans sa forme actuelle. L'usage recommandé est sous la forme d'un nombre de 8 chiffres tel que YYYY-MM-DD, comme défini par la norme ISO8601. Dans ce schéma, l'élément date 1994-11-05 correspond au 5 Novembre 1994. Beaucoup d'autres schémas sont possibles, mais si un autre schéma est utilisé, il devrait être précisé de façon non ambigüe.

9. Langage :

Etiquette: langage

Langage(s) du contenu intellectuel de la ressource. Si approprié, le contenu de ce champ devrait correspondre à la norme RFC 1766.

10. Editeur :

Etiquette: editeur

L'entité responsable de la diffusion de la ressource dans sa forme actuelle, telle qu'une maison d'édition, un département universitaire, une entreprise.

11. Identifiant de la ressource :

Etiquette: identifiant

Chaîne de caractère ou nombre utilisé pour identifier de façon unique la ressource. Exemples pour des ressources réseau incluent URLs et URNs (si implementé). D'autres identificateurs globaux et uniques, tels que ISBN (International Standard Book Numbers), ou d'autres noms formellement définis, sont des candidats potentiels pour cet élément, dans le cas de ressources privées.

Remarque : Une nouvelle version 1.1 de la description de réference des éléments du Dublin Core a été publiée dernièrement et même traduite en français par Anne-Marie Vercouste (INRIA). Je suis membre du Groupe Datamodele du Dublin Core et je suis assidûment et autant que je peux , tous les développements qui sont proposés et ceux qui sont adoptés. Mais pour le système DREPU , j'ai préferé garder l'ancienne présentation des éléments du Dublin Core , beaucoup plus simple , que j'ai dèja d'ailleurs proposée dans mon mémoire de DEA en juillet 1998. De toute façon , les définitions des éléments n'ont pas changé.

IV- DEFINITION DES ELEMENTS DU PROFIL D'UTILISATEUR :

Pour définir les éléments du profil d'utilisateur à intégrer dans le système DREPU, je me suis inspiré des éléments qui rentrent dans la construction d'un profil d'utilisateur classique pour le compte d'une DSI (Diffusion Sélective de l'Information). J'ai tenté de résumer les traditionnelles questions que le spécialiste du service de références d'une bibliothèque, pose en général aux utilisateurs, pour établir leurs profils, par les éléments suivants :

- Discipline ou spécialité couverte ;
- Profession ou qualification ;
- Niveau éducationnel ;
- Affiliation ou communauté d'appartenance ;
- Type de publication.

Evidemment, étant donné que ces éléments résultent de mon propre choix même si je n'ai fait que condenser en quelques expressions une procédure classique de confection de profil d'utilisateur ; j'ai entrepris de mener des enquêtes directes auprès des documentalistes des services de références de dix bibliothèques universitaires, six bibliothèques d'entreprises et quatre bibliothèques publiques, réparties dans les villes d'Alger, Blida, Oran, Constantine et Tizi-Ouzou. J'ai interviewé en tout, cinquante-deux bibliothécaires chargés des opérations de diffusion sélective d'information (DSI) au niveau de ces bibliothèques. Mon enquête s'est déroulé sur le principe d'un entretien guidé très succinct, comportant les questions suivantes :

1- Pensez-vous que des éléments de profil de l'auteur d'un document puissent servir à catégoriser ce dernier ?
2- Dans l'affirmatif, pensez-vous qu'ils ont une importance dans une opération d'indexation de ce document ?
3- Maintenant nous voulons intégrer ces éléments dans l'indexation de

documents donnés pour qu'au cours d'une relance d'interrogation lors d'un processus de recherche d'information, ils soient mis en correspondance avec des éléments de profils de l'utilisateur qui a lancé la requête. Ces éléments que l'indexeur prend chez l'auteur pour catégoriser son document et qui permettent en même temps au chercheur d'information pour établir son profil devront être désignés selon vous, par l'expression :

- Eléments de profil de l'auteur ?
- Eléments de profil de l'utilisateur ?
- Eléments de profil d'utilisateur ?

4- Pensez-vous qu'une mise en correspondance d'éléments de profil de l'utilisateur avec des éléments de profil de l'auteur, dans un processus de recherche de l'information puisse améliorer la pertinence des documents retrouvés ?

5- Pensez-vous qu'un système basé sur un tel principe sera :

5.a) Inintéressant et inutile ?

5.b) Une solution supplémentaire dans l'évolution de la recherche d'information ?

5.c) Une solution spécifique de recherche et de filtrage de l'information ?

5.d) La meilleure solution de recherche d'information ?

6- Classez les éléments de profil suivants par ordre d'importance, en leur attribuant un poids compris entre 0 et 100 :

- Profession ou qualification ;
- Niveau éducationnel ;
- Affiliation ou communauté d'appartenance ;
- Type de publication.

Remarque :

Je dois signaler ici que j'ai préféré l'enquête par interrogation directe au lieu de l'enquête par questionnaire à distance pour les raisons suivantes:

_ Possibilité de mieux expliquer à la personne interrogée l'objet et le contenu du questionnaire ;
_ Les réponses seront plus étudiées et plus sérieuses ;
_ Au-delà des simples réponses, le débat avec l'interlocuteur peut amener des informations pertinentes ;
_ Réponses assurées par rapport à un questionnaire à distance qui ne génère généralement que très peu de réponses.

Résultats de l'enquête :

1) *Première question* : 52 réponses affirmatives. Toutes les personnes interrogées admettent qu'effectivement les éléments de profil de l'auteur d'un document peuvent aider à catégoriser celui-ci.

2) *Deuxième question* : Toutes les personnes interrogées ont répondu par l'affirmative, tout en notant que ces éléments ne sont toutefois pas nécessaires à l'indexation d'un document ;

3) *Troisième question* :
   -20 bibliothécaires m'ont répondu directement qu'il faut adopter la première expression (éléments de profil de l'auteur) ;
   -15 autres m'ont répondu après mûre réflexion, qu'il faut juste les appeler éléments de profil, puisqu'ils peuvent aussi bien servir à catégoriser le document lors du processus d'indexation, en étant des éléments de profils de l'auteur ; et servir à définir l'utilisateur en étant des éléments de profil de celui-ci lors du processus de filtrage de l'information (à la relance de l'interrogation) ;
   - 17 bibliothécaires m'ont répondu qu'il faut plutôt les appeler éléments de profil d'utilisateur; expression indéfinie, nuancée par rapport à l'expression éléments de profil de l'utilisateur .

J'ai décidé d'adopter cette dernière expression dans le système DREPU parce que je considère que les éléments de profil introduits dans les

metadonnées du document lors de l'indexation de celui-ci, même s'ils sont pris à l'auteur, correspondent en fait aux éléments de profil d'un utilisateur modèle, dont l'objectif de recherche est de retrouver précisément ce document. Par extension, à chaque fois que l'expression éléments de profil d'utilisateur sera utilisée, elle concernera les éléments de profil proposés par l'interface de recherche et de filtrage d'information de DREPU, pour aider l'utilisateur à sélectionner les éléments de profil qui seront mis en correspondance avec ceux contenus dans les metadonnées des documents ciblés. Je dois signaler ici que je ne suis pas le premier à proposer la superposition de ces deux profils. D'autres projets (par exemple SmartPush[11]) utilisent cette approximation selon d'autres approches.

4) *Quatrième question* : 52 réponses affirmatives.
   Tous les professionnels des bibliothèques interrogés ont répondu unanimement qu'effectivement une mise en correspondance des éléments de profil de l'utilisateur avec des éléments de profil de l'auteur d'un document peut permettre de retrouver plus facilement ce dernier dans un processus de recherche d'information.

5) 5.a) 52 réponses négatives.
   Toutes les personnes interviewées ont réfuté ces qualificatifs pour le système DREPU.

   5.b) 30 réponses affirmatives et 22 réponses négatives.
   Les avis sont partagés ; mais la majorité des spécialistes des bibliothèques interrogés pensent que DREPU constitue une nouvelle approche de systèmes de recherche de l'information, qui peut s'inscrire dans l'évolution logique de ces derniers.

   5.c) 50 réponses affirmatives et 2 réponses négatives.
   A cette question, la grande majorité des spécialistes des bibliothèques consultés ont répondu par l'affirmative, étant

[11] voir le projet SmartPush dans la partie C

convaincus que le système DREPU est réellement une solution spécifique de recherche et de filtrage d'information.

5.d) 52 réponses négatives.

Pour cette question les personnes interrogées ont répondu à l'unanimité qu' il n'y a pas de système qui peut être considéré comme la meilleure solution de recherche de l'information. Chaque nouveau système peut apporter éventuellement des améliorations et de nouvelles performances selon certains points de vue. Le système DREPU peut s'inscrire évidemment dans cette lignée.

6) Les réponses obtenues m'ont permis de dresser le tableau hiérarchique suivant ; où j'ai porté dans la première colonne le classement des éléments de profil d'utilisateur, dans la deuxième colonne la désignation de chaque élément selon son ordre d'importance et dans la troisième colonne le poids moyen pondéré pour chaque élément.

N°	***Désignation de l'élément***	***Poids moyen calculé***
1	Discipline ou spécialité couverte	100
2	Profession ou qualification	89
3	Niveau éducationnel	74
4	Affiliation ou communauté d'appartenance	60
5	Type de publication	20

Tableau No 2

Interprétation des résultats :

Les résultats obtenus dans le tableau ci-dessus sont très explicites. Dans un milieu professionnel tel que celui des bibliothèques ou la pratique et l'expérience forge de véritables savoir-faire, la discipline ou la spécialité

couverte constitue l'élément fondamental d'un profil et donc ce dernier a remporté le consensus de toutes les personnes interrogées. L'ensemble de ces personnes pense que le deuxième élément relatif à la profession ou qualification de l'auteur est un indice non avéré de la valeur ou qualité du contenu du document recherché ou à indexer. Il en est de même du niveau éducationnel ; bien que certains pensent qu'il y a redondance par rapport à l'élément précédent. Toutefois, la majorité des personnes consultées m'ont conseillées de garder cet élément optionnellement. L'affiliation ou la communauté d'appartenance de l'auteur est très importante pour certains, mais les avis demeurent partagés pour cet élément quant au type de document considéré. Cet élément sera important pour tous les documents scientifiques et techniques dont les auteurs sont des chercheurs ; mais sera de moindre importance pour des documents de vulgarisation ou d'informations générales. Pour la majorité des personnes interrogées, le type de publication ne constitue nullement un élément à intégrer dans un profil d'utilisateur, selon l'entendement classique. Pris seul, comme élément intrinsèque, il est vrai qu'il peut renseigner l'utilisateur sur la valeur du document recherché ; selon que le support de publication soit un magazine, une lettre d'information ou une revue scientifique à comité de lecture. Pris avec l'ensemble des éléments du profil, son importance s'amoindrit par le fait que les autres éléments renferment toutes les informations nécessaires à l'appréciation de la valeur ou qualité du document recherché.

La notion de profil (d'utilisateur) dynamique :

En commençant mon enquête auprès des professionnels des bibliothèques, j'appréhendais avec inquiétude leurs réactions par rapport à mon modèle et surtout par rapport à la notion de profil dynamique que j'introduisais. Certes j'ai dû commencer à expliquer que le système DREPU est basé sur une association d'éléments de description de ressources, empruntés au standard de metadonnées du Dublin Core, et d'éléments de profil d'utilisateur. J'ai pris la précaution de leur parler avec beaucoup de détail des projets qui associent des éléments de profil d'utilisateur dans des processus de recherche d'informations. Grande fut ma surprise de trouver chez la majorité d'entre eux un intéressement inattendu à mes explications et à ma théorie.

L'avènement d'Internet et l'évolution vertigineuse des technologies de l'information ont dèjà imprimé chez ces professionnels un esprit vivace plein de curiosités envers tous ces nouveaux systèmes qu 'on leur propose…. Dans le système DREPU, le document est décrit par les éléments de description de ressources ; et catégorisé par les éléments de profil de l'auteur. Durant le processus de relance de l'interrogation, la phase de filtrage de l'information se fondera sur une mise en correspondance des éléments du profil de l'usager qui a lancé la requête, avec les éléments de profil contenus dans les metadonnées des documents indexés avec G-MET[12]. Dans la classique diffusion sélective de l'information (DSI) et tous les systèmes automatisés qui en dérivent tels NewsAgent, Eurogatherer (y compris la technologie PUSH), nous trouvons des profils d'utilisateurs caractérisant des individus donnés et presque invariables dans le temps, si nous occultons quelques mises à jour souvent irrégulières. Dans le cas du système DREPU, le chercheur d'information se comporte comme « le voyageur virtuel » de Ricky Erway (OCLC) [121]. Celui-ci tel un vrai voyageur, peut naviguer sur les étapes du processus de recherche d'information. Comme le voyageur réel se doit d'être au courant de la culture et du langage de la contrée à visiter ; le chercheur d'informations doit adopter une syntaxe spécifique au domaine et une sémantique pour aborder le processus de recherche de ressources. Il assume ainsi un rôle propre à cette étape.

D'après Erving Goffman[13], en tant qu'individus, nous nous retrouvons souvent à assumer plusieurs rôles. Dans les interactions quotidiennes, nous sommes constamment en mouvement, en déplacement dans des rôles et en les intercalant, dans plusieurs permutations. Ce comportement dynamique se retrouve aussi dans les processus de recherche d'information. Quand on tente d'assumer son propre rôle à travers une activité de recherche d'informations spécifiques, les changements de rôle peuvent survenir même dans le contexte. Par exemple, je suis un chercheur en sciences de l'information, père d'une fille de douze ans, et résident à Alger (Algérie). Je

[12] G-MET est l'outil d'aide à l'indexation du système DREPU (voir page 160)

[13] "La mise en scène de la vie quotidienne" ; publié en 1958 aux Etats-Unis et en 1973 en France, aux Editions de Minuit. ISBN 2-7073-0063-2.

peux lancer une requête en tant que chercheur en sciences de l'information ; chercher l'information sur des logiciels que je peux utiliser dans mon travail. Pendant le processus, je peux aussi chercher de l'information sur un logiciel approprié à ma fille. Ceci peut provoquer un transfert de rôles, à partir duquel je travaille. Enfin, je peux ensuite revenir à ma première perspective ou même à un autre rôle. Pour chaque rôle, je dois adopter un vocabulaire différent, une sémantique de recherche, et une sémantique de filtrage en choisissant des éléments de profil appropriés pour faire aboutir le processus de recherche. J'adopte ainsi un profil (d'utilisateur) dynamique.

V- Définition des éléments de metadonnées de DREPU:

Dans le paragraphe précédent, nous avons défini les éléments de description de ressources et les éléments de profil d'utilisateur de DREPU. Afin de faire de celui-ci un véritable standard de metadonnées dérivé du Dublin Core, nous allons réécrire tous ses éléments selon la structure de ce dernier. Cette opération nous permet d'avoir le tableau de concordance suivant entre l'écriture simple des éléments de DREPU et leur écriture structurée selon le Dublin Core :

Eléments simples de DREPU	**structurés selon le Dublin Core**
-Titre	-Titre Etiquette:Titre • DC.Titre
-Auteur ou créateur -Co-auteur(s) - Discipline ou spécialité couverte - Profession ou qualification - Niveau éducationnel - Affiliation ou communauté d'appartenance	-Auteur ou créateur Etiquette:Créateur • DC.Créateur -Sous-éléments(type): • DC.Créateur • DC.Créateur. Discipline • DC.Créateur.Profession • DC.Créateur. Niveau • DC.Créateur.Affiliation

Eléments simples de DREPU	structurés selon le Dublin Core
-Sujet et mots clés	-Sujet et mots clés Etiquette : Sujet • DC.Sujet
-Source	-Source Etiquette : Source • DC.Source
-Type de ressource	-Type de ressource Etiquette : Type • DC.Type
-Description	-Description Etiquette : Description • DC.Description
-Editeur -Type de publication	-Editeur Etiquette:Editeur -Sous-élément (type): • DC.Editeur.Publication
-Date	-Date Etiquette:Date • DC.Date
-Format	Etiquette : Format • DC.Format
-Langage	-Langage Etiquette:Langage • DC.Langage
-Identifiant de la ressource	-Identifiant de la ressource Etiquette:Identifiant • DC.Identifiant

Tableau No3

Ainsi structuré, DREPU devient un standard de métadonnées dérivé du Dublin Core ; et donc héritant de toutes ses propriétés[14] notamment :

- ***Extensibilité*** :

le mécanisme d'extension permettra l'inclusion de données intrinsèques pour des objets qui ne peuvent pas être décrits suffisamment par un petit ensemble d'éléments. L'extensibilité est importante parce qu'elle permet de rajouter des descriptions supplémentaires à certains champs, si necessaire.

Indépendance de Syntaxe :

Cette indépendance permettra à DREPU d'être transcrit en HTML, RDF ou XML et de pouvoir être associé éventuellement à une large gamme de programmes d'application.

Optionalité :

Tous les éléments sont optionnels, pour une meilleure simplicité d'emploi.

Répétabilité :

Tous les éléments de DREPU sont répétables. Par exemple, plusieurs éléments 'auteur' seraient employés quand une ressource a plusieurs auteurs.

Modifiabilité :

Chaque élément dans DREPU a une définition qui est sensée être évidente. Cependant, il est aussi nécessaire que les définitions des éléments satisfassent aux besoins de communautés différentes. Ce but est accompli en permettant à chaque élément d'être modifié par un qualificateur optionnel. Si aucun qualificateur n'est proposé, l'élément prend son sens commun.

Définition de qualificateurs propres à DREPU:

1-Définitions:

Nous reprenons pour DREPU un ensemble de qualificateurs définis dèja pour le Dublin Core[15]. Nous distinguons :

[14] voir page 54

- ***Sous-élément (ou type):***

Il affine et clarifie la définition de l'élément auquel il est rattaché. Par cette fonction, il exprime aussi la propriété d'extensibilité du système.

- ***Schéma***:

Il permet à la valeur d'un élément d'être identifiée en tant que donnée extraite d'un système de classification, d'un code, d'un glossaire ou d'un thésaurus existant.

- ***Langage:***

Ce dernier qualificateur, introduit au quatrième workshop sur le Dublin Core, qui s'est tenu à Camberra en Australie, en Mars 1997, indique le langage du contenu **d'un élément** de métadonnées. Il faut bien comprendre ici que chaque élément de métadonnées peut être écrit dans une langue différente ou non, du contenu décrit; mais surtout qu'il peut être aussi écrit dans une langue différente de celle d'autres éléments qui entrent dans la même description..

Exemple: Le titre d'un document en anglais peut être reporté dans une métadonnée en français; avec une répétition de l'élément " *titre"* contenant le titre originel; et dans la même description de ce document (description par des éléments de métadonnées en français), nous pouvons retrouver un résumé en français et des descripteurs en anglais.

2-Qualificateurs proposés pour les éléments de DREPU:

- ***Titre:***:

Etiquette : Titre

-*Sous-élément*:

  - DC.Titre
  - DC.Titre.Alternatif (pour tout autre titre ou sous-titre)

[15] voir page 54

-*Shéma* :texte libre par défaut

-Langage : indiqué généralement par deux lettres (RFC 1766)

REMARQUE:

Le qualificateur "langage" sera identifié par le même contenu (RFC 1766) pour tous les éléments; donc nous nous abstiendrons de le répéter à chaque fois.

❖ ***Auteur ou créateur:***

Etiquette : Créateur

-*Sous-élément*:

- DC.Créateur
- DC.Créateur.Discipline
- DC.Créateur.Profession
- DC.Créateur.Niveau
- DC.Créateur.Affiliation

-*Schéma* : texte libre

❖ ***Sujet et mots-clés* :**

Etiquette : Sujet

-Sous-élément ; aucun ;

-Shéma : Texte libre.

❖ ***Description :***

Etiquette : Description

-Sous-élément : Aucun ;

-Shéma : Texte libre.

❖ ***Editeur***

Etiquette : Editeur

-Sous-élément :

- DC.Editeur
- DC.Editeur.Publication

-Schéma : texte libre

❖ ***Date***:

Etiquette : Date

-Sous-élément : Aucun

- Schéma : ISO 8601
  - ANSI X3.30
  - IETF RFC 822

❖ ***Type de la ressource*** :

Etiquette : Type

-Sous-élément : Non défini ;

-Shéma : texte libre;

❖ ***Format*** :

Etiquette : Format

-Sous-élément : non défini ;

-Shéma : - txt ; doc ; rtf ; pdf ; post-script ; MIME ; etc….

❖ ***Identifiant de la ressource***:

Etiquette : Identifiant

-Sous-élément :

- DC.Identifiant

- Schéma : - URI ;
  7- URL ;
  8- DOI ;
  9- ISBN ;
  10- ISSN ;
  11- AUTRES à préciser

❖ ***Source***:

Etiquette : Source

- Sous-élément : non défini
- Schéma : texte libre

❖***Langage*** :

Etiquette : Langage

-Sous-élément : non défini

- Schéma : - RFC1766 ;
  - ISO639 ;
  - ISO3166.

Syntaxe de transcription en HTML:

Définitions:

Pour transcrire les éléments de DREPU en un document électronique, il est évident qu'il est nécessaire de disposer d'une syntaxe spécifique.

A partir du moment que nous avons structuré notre système DREPU selon le standard de metadonnées du Dublin Core, nous adopterons donc pour ses éléments la même transcription en HTML[16].

Le choix d'un codage en HTML permet d'avoir un format ou une forme de représentation interprétable aussi bien par l'homme que par la machine. Les metadonnées sont extraites automatiquement et ne doivent pas apparaître à la visualisation du document dans le navigateur ou à sa sortie sur imprimante.

En HTML, chaque définition d'élément d'enregistrement commence avec "<META" et fini par ">".

Pour transcrire des éléments de metadonnées en HTML, la balise <META> est placée entre le <HEAD> et le </HEAD> de fin de paragraphe en utilisant la syntaxe suivante:

<META NAME = "DC.NomElément" CONTENT = "Valeur">

Dans l'écriture ci-dessus, 'NomElément' et 'Valeur' contiennent une désignation de l'élément et sa 'valeur'.

Exemple :

<META NAME = "DC.Créateur" CONTENT = "Youcef AMEROUALI">

♦ ***Principes de base des éléments de description*** :

✓ La notation que nous décrivons est basée sur les balises META. Dans le cas d'utilisation de caractères non-ASCII, il faudrait utiliser la même disposition des mots que dans le corps du document.

✓ Chaque élément est optionnel et répétable. Les éléments de Metadonnées peuvent apparaître dans n'importe quel ordre. La disposition de multiples occurrences du même élément peut avoir une signification voulue par le fournisseur de l'information ; mais cette disposition n'est pas assurée d'être préservée dans chaque environnement utilisateur.

✓ La Convention proposée pour la transcription des metadonnées dans HTML prévoit leur identification et leur groupement. Cette convention compte sur l'utilisation de préfixe pour indiquer que l'élément utilisé appartient au Dublin Core ou à un autre standard. Pour le Dublin Core, le préfixe "DC" doit être écrit avant l'étiquette de l'élément et en lettres capitales.

exemple:

META NAME="DC.Titre"

META NAME="DC.Créateur"

⇒ Pour le cas de DREPU, nous proposons de ne pas utiliser de préfixe pour le moment, comme c'est le cas pour la plupart des standards de metadonnées dérivant du Dublin Core :

Exemple :

META NAME=" Titre"

META NAME=" Créateur"

[16] voir page 64

♦ **Transcription des éléments de DREPU en HTML** :

1. Titre :

Etiquette : Titre

Description :

Nom donné à la ressource par l'auteur ou le créateur

Transcription :

<META NAME=" Titre" CONTENT="Nom de la ressource">

Exemples:

<META NAME=" Titre" CONTENT="Les fleurs du mal">

<META NAME=" Titre" CONTENT="L'albatros">

2.Auteur ou créateur:

Etiquette: Créateur

Description :

Personne ou organisme principalement responsable de la création du contenu intellectuel de la ressource. Les auteurs doivent être repris séparément, dans le même ordre qu'ils apparaissent dans la publication. Le champ *auteur* peut être répété dans le cas de l'existence de co-auteurs.

Transcription :

<META NAME=" Créateur" CONTENT="Nom de l'auteur">

Exemples:

<META NAME=" Créateur" CONTENT=" Baudelaire, Charles">

<META NAME=" Créateur" CONTENT=" Marx, Karl">

<META NAME=" Créateur" CONTENT="Engels, Friedrich">

*Nous pouvons regrouper ces deux derniers co-auteurs dans un seul champ :

<META NAME=" Créateur" CONTENT=" Marx, Karl ; Engels, Friedrich">

*Nous pouvons aussi introduire les sous-éléments rattachés à l'élément *auteur :*

<META NAME="Créateur.Discipline" CONTENT=" donnée descriptive">
<META NAME=" Créateur. Profession" CONTENT=" donnée descriptive">
<META NAME=" Créateur. Niveau" CONTENT=" donnée descriptive">
<META NAME=" Créateur. Affiliation" CONTENT=" donnée descriptive">

Exemples :

<META NAME=" Créateur. Discipline" CONTENT=" informatique">
<META NAME=" Créateur. Profession" CONTENT=" enseignant">
<META NAME=" Créateur. Niveau" CONTENT=" doctorat">
<META NAME="Créateur.Affiliation" CONTENT=
" ENSSIB. Systèmes d'Information et Interfaces">

3. Sujet et mots-clés :

Etiquette : Sujet

Description : Cet élément décrit le sujet du contenu de la ressource ou comporte des mots-clés.

Transcription :

<META NAME="Sujet" CONTENT="donnée ou attribut">

Exemple:

<META NAME="Sujet" CONTENT="Analyse socio-économique….">

4. Source :

Etiquette : Source

Description :

Cet élément indique une référence d'une ressource à partir de laquelle a été tirée la ressource actuelle.

Transcription :

<META NAME="Source" CONTENT="donnée ou attribut">

Exemple:

<META NAME="Source" CONTENT="les fleurs du mal de Baudelaire">

5. Type de ressource :

Etiquette : Type

Description :

Ce champ décrit la catégorie de la ressource.

Transcription :

<META NAME=" Type" CONTENT="attribut">

Exemple :

Si la ressource est un fichier PDF :

<META NAME=" Type" CONTENT="Fichier PDF">

6. Description :

Etiquette : Description

Description :

Cet élément donne une description ou un résumé du contenu de la ressource.

Transcription :

<META NAME="Description" CONTENT="donnée ou attribut">

Exemple:

<META NAME="Description" CONTENT="Dans cet article, nous abordons la notion de profil utilisateur ….">

7. Editeur:

Etiquette : Editeur

Description :

Entité responsable de la publication de la ressource dans sa forme actuelle, telle une maison d'édition, un département universitaire, ou une société savante. Le but, en spécifiant ce champ est de permettre l'identification de l'entité fournissant l'accès à la ressource décrite. Nous rattachons à cet élément, le sous-élément Type de publication qui défini le type de publication ou le type de production éditoriale.

Transcription :

<META NAME="Editeur" CONTENT=" donnée ou attribut">

<META NAME="Editeur.Type" CONTENT=" donnée ou attribut">

Exemples:

<META NAME="Editeur" CONTENT="ENSSIB">

<META NAME="Editeur.Type" CONTENT=" thèse">

8. Date :

Etiquette : Date

Description :

Cette date est associée à la création ou à la validité de la ressource décrite. Il est recommandé de toujours l'écrire dans le format ISO8601.

La note technique du W3C :

[http://www.w3.org/TR/NOTE-datetime]

inclue le format suivant : YYYY et YYYY-MM-DD.

Ainsi, par exemple, la date 1998-06-15 correspond au 15 juin 1998.

Si des éléments de la date ne sont pas connus, écrire seulement le mois et l'année, ou uniquement l'année ; comme par exemple :

<META NAME=" Date" CONTENT="1998-06-15">

<META NAME=" Date" CONTENT="1998-06">

<META NAME=" Date" CONTENT="1998">

9.Format :

Etiquette : Format

Description :

Cet élément est caractéristique de la matérialisation physique de la ressource.

Transcription :

<META NAME="Format" CONTENT="donnée ou attribut">

Exemple :

<META NAME="Format" CONTENT="MIME">

10. Langage:

Etiquette : Langage

Description :

Cet élément indique le langage du contenu intellectuel de la ressource.

Transcription :

<META NAME="Langage" CONTENT=" donnée ou attribut">

Exemples:

<META NAME="Langage" CONTENT="Fr">

11.Identifiant de la ressource:

Etiquette : Identifiant

Description :

Cet élément est défini par une chaîne de caractères ou un nombre, utilisé pour uniquement identifier le document en tant que ressource. Il peut être une référence locale, attribuée dans le cadre de DREPU ; ou un autre type d'identifiant tel l'ISSN, l'ISBN, l'URI, l'URL, l'URN, le DOI, etc...

Transcription :

<META NAME="Identifiant" CONTENT="attribut >

Exemple :

<META NAME="Référence" CONTENT=
"http://www.enssib.fr/dort.html">
<META NAME="Identifiant" CONTENT="0385424728" [ISBN] >

VI-FONCTIONNEMENT DU SYSTEME DREPU :

INTRODUCTION ;

Nous rappelons que nous avons conçu DREPU comme un standard de metadonnées dérivé du Dublin Core, à l'instar de tous les projets[17] qui se sont construits sur les principes de ce dernier. Toutes les équipes qui ont eu à travailler sur ces projets ont été amenées à créer des outils logiciels spécifiques pour la génération des éléments de metadonnées dans l'entête du document html pour éviter de le faire manuellement avec un éditeur html. La plupart de ces équipes ont eu aussi à développer des moteurs de recherche spécifiques à leurs projets. Pour notre système, il nous fallait d'abord un outil d'aide à l'indexation selon le principe de DREPU ; c'est à dire qui peut générer les éléments de metadonnées de DREPU directement dans l'entête du document html. Cet outil que nous avons appelé G-MET permettra la constitution d'une base de données propriétaire qui contiendra tous les éléments de metadonnées pris dans l'opération d'indexation pour chaque document indexé, avec le chemin logique vers le document lui-même. Nous pouvons avoir un document sur un site distant (en réseau) indexé avec l'outil logiciel G-MET ; nous aurons dans la base propriétaire (meta.idx) en plus des éléments d'indexation, l'hyperlien vers ce document qui permettra d'afficher celui-ci au besoin.

Outil d'aide à l'indexation G-MET :

L'outil d'aide à l'indexation G-MET permet d'indexer des documents en générant les éléments de metadonnées de DREPU directement dans le document source du document html à indexer ; à l'aide d'un formulaire qui peut être appelé de la page d'accueil du système DREPU. IL peut indexer tout document en ligne sur le réseau (Internet ou Intranet) ou sur un poste local si un serveur Web[18] y est installé. Les éléments d'indexation seront stockés dans la base de données propriétaire (meta.idx), créée sur le poste local ou sur un serveur distant.

---

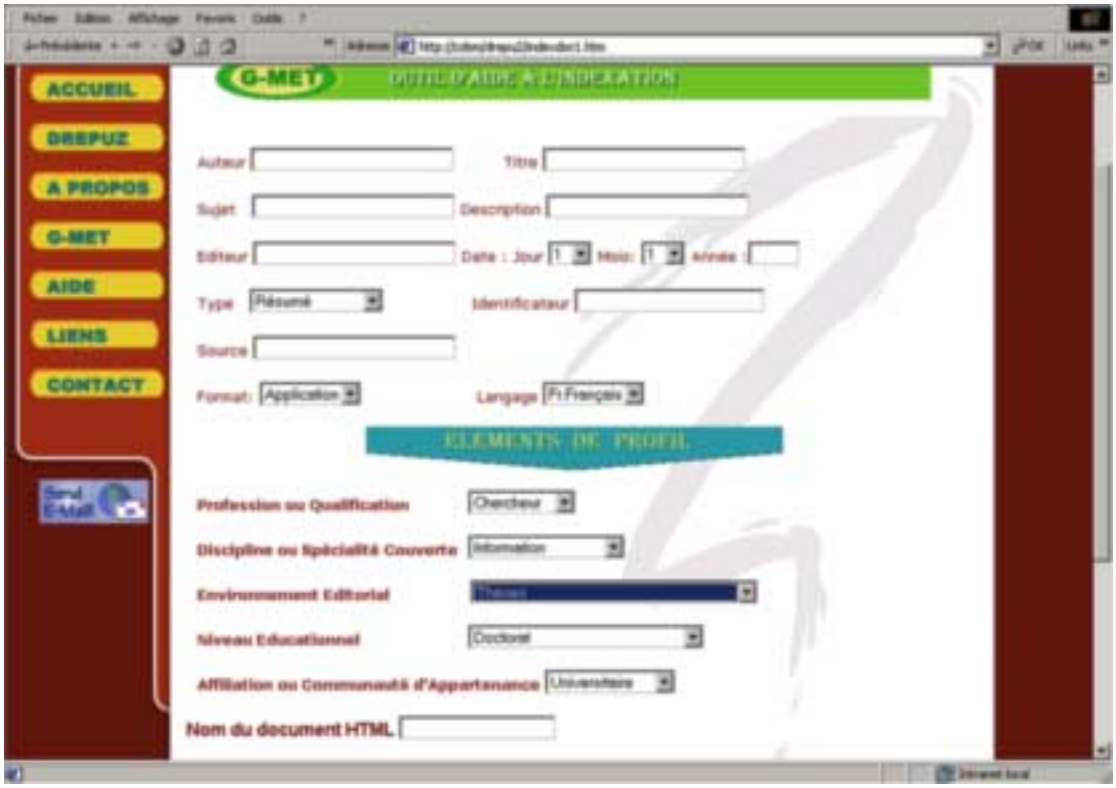
[17] Voir page 56

[18] Websever de Microsoft pour les micro-ordinateurs sous Windows

Eléments de description de ressources	**Eléments de profil d'utilisateur**
Titre	Discipline ou spécialité couverte
Auteur ou créateur	Profession ou qualification
Sujet et mots-clés	Niveau éducationnel
Source	Affiliation ou communauté d'appartenance
Type de ressources	Type de publication
Description	
Editeur	
Date	
Format	
Langage	
Identifiant de la ressource	

-**Tableau 5**

Image 2



Outil de recherche d'information DREPUZ :

Le système DREPU portant sur une indexation préalable de documents à l'aide de metadonnées ; il était évident pour nous qu'il fallait, pour donner corps à ce modèle:

- Un outil ou une méthode d'indexation spécifique portant sur les metadonnées de DREPU (constituées de l'association des onze éléments de description de ressources et des cinq éléments de profil d'utilisateur : tableau 5) ;
- Un outil de recherche et de filtrage d'information spécifique qui puisse cibler directement les éléments de metadonnées de documents.

Après avoir développé l'outil d'aide à l'indexation G-Met, nous permettant de constituer une base (catalographique) d'éléments d'indexation selon le système DREPU, nous avons entrepris de développer un outil de recherche et de filtrage d'information appelé *DREPUZ*[19], avec le langage Perl, qui offre dans son interface de navigation des possibilités de lancer des requêtes en ciblant, dans un premier temps un ou plusieurs des onze éléments de

[19] La lettre Z est rajoutée pour différencier l'outil de recherche du système entier

description de ressources. Le processus de recherche (Figure 6)[20] est analogue à tout processus de recherche d'information au lancement de la procédure. L'utilisateur peut lancer sa recherche en écrivant le mot ou l'expression, porteur du thème de sa recherche dans la fenêtre prévue à cet effet. Une autre fenêtre lui permet de sélectionner les éléments de description de ressources sur lesquels il peut axer sa recherche. Un premier résultat est affiché, consistant en un nombre de documents répondant à l'interrogation, sans affichage des références de ces documents. L'outil de recherche DREPUZ propose dans une deuxième fenêtre d'affiner ce résultat en proposant de tenir compte d'un ou de plusieurs des cinq éléments de profil utilisateur, précédemment définis. Alors le système ira chercher parmi les documents déjà trouvés, ceux répondant au profil rentré ; et affichera tous les détails permettant de les visualiser (chemins de parcours et hyperliens). Dans le cas où l'utilisateur voudrait se passer de cette possibilité de filtrage ; alors le système lui affichera, par défaut tous les documents trouvés préalablement.

Image 3

[20] page 139

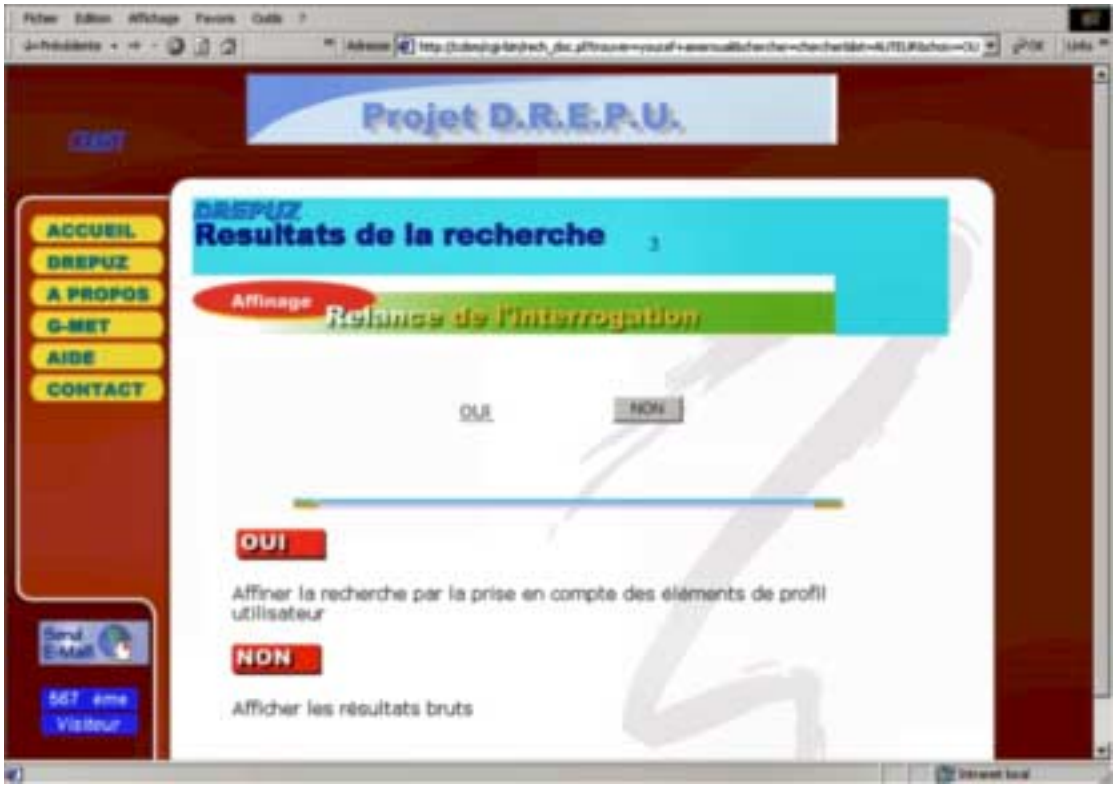


Image 4

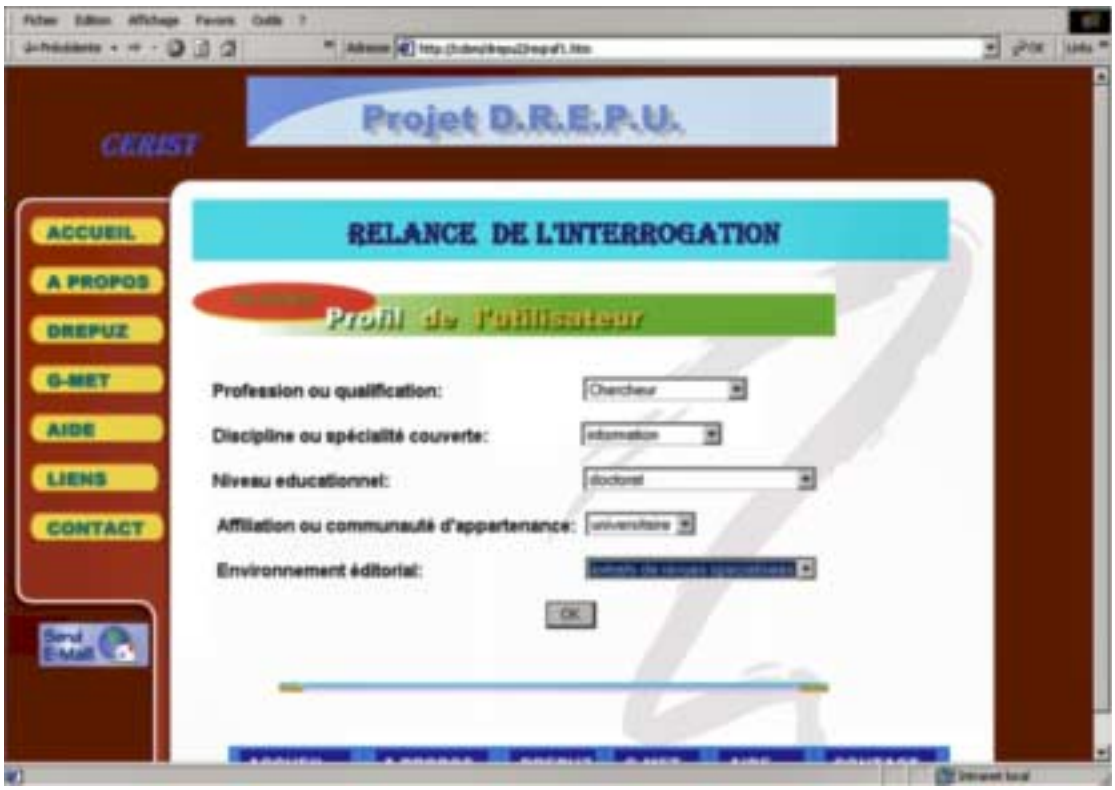


Image 5

Image 6

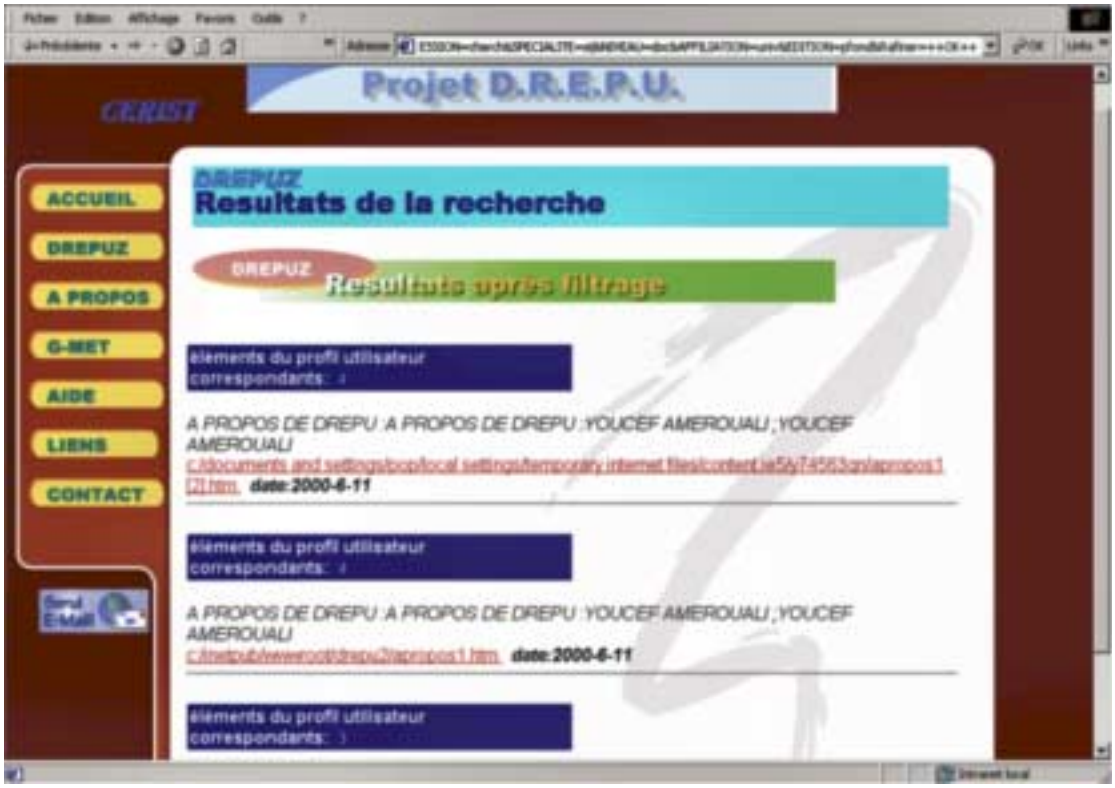


Image 7

VII- EVALUATION :

L'évaluation des systèmes de recherche de l'information consiste souvent à mesurer divers paramètres qui expriment la plus ou moins grande capacité d'un système de retrouver les documents répondant aux questions posées. L'efficacité du système est la première donnée à connaître : la réponse du système à une requête est-elle aussi complète, exhaustive et pertinente que possible ? ou que voudrait l'utilisateur ? Actuellement la pertinence demeure encore une notion controversée dans la mesure où elle est exprimée par un jugement individuel de l'utilisateur ; même si c'est ce jugement qui détermine en fin de compte la satisfaction de ce dernier. Nous avons tenu compte de ce point de vue dans le système DREPU, en permettant à l'utilisateur de filtrer sa recherche d'information pour une meilleure pertinence et donc un taux de bruit réduit. C'était notre objectif au départ de la conception de DREPU et c'est ce que nous voudrions que cette évaluation confirme.

7.1*) Evaluation des outils logiciels* :

Dans ce domaine, la littérature est prolifique ; et plusieurs méthodes sont souvent proposées. Nous nous contenterons d'évaluer nos outils selon les plus simples. Généralement l'objectif du test logiciel est de détecter les inadéquations ou fautes d'un logiciel. Il existe deux grandes classes de méthodes de test :

8.1a – Les méthodes de test statique :

Elles traitent le texte du logiciel sans exécuter ce dernier sur des données réelles. Elles peuvent être appliquées à des textes de spécifications ou de programmes. Les principales méthodes de test statiques sont :

Les lectures croisées et l'inspection :

C'est à dire la vérification « collégiale » d'un document (programme ou spécification du logiciel). L'inspection est une relecture en groupe des textes de programmes. Sur la base de scénarios et d'une liste de points à vérifier, l'inspection doit trouver des fautes éventuelles. Si des fautes ont été détectées, des tests dynamiques visant à vérifier leur rectification sont à envisager.

- ***L'analyse des anomalies :***

Des anomalies telles qu'une incohérence des interfaces de modules, des portions de code isolées, un mauvais usage de variables ou des variables oubliées, des utilisations impropres de pointeurs etc…. peuvent être facilement détectées dans les programmes grâce à des analyses faites à la compilation.

- ***L'évaluation symbolique*** :

Elle simule l'exécution du programme sur des données symboliques: on obtient ainsi des expressions symboliques correspondant au texte des programmes.

8.1b- Les méthodes de test dynamique :

❖ Le test dynamique repose sur:

- La sélection qui consiste à choisir judicieusement un sous ensemble des entrées possibles du logiciel appelé « jeu de test » ;

- La soumission du jeu de tests afin d'exécuter les tests retenus lors de l'étape précédente ;

- Le dépouillement des résultats qui consiste à décider du succès ou de l'échec du jeu de tests ;

8.1C- Efficacité des tests :

L'efficacité d'un jeu de tests est son aptitude à trouver des fautes. En général l'arrêt du test est un arrêt sur objectif et n'est pas basé sur une véritable mesure d'efficacité. Pour chaque méthode de test utilisée, on se définit un critère adapté :

- L'arrêt des lectures croisées ou inspections est effectué lorsque la liste de points à vérifier est remplie ;
- L'arrêt de l'analyse d'anomalies est décidé lorsqu'il n'y a plus d'anomalies non expliquées ;

Dans le cas de notre travail, nous avons suivi l'ensemble de points d'évaluation cité ci-dessus sur notre logiciel. En ce qui concerne les jeux d'essai, nous avons constitué une base de données à titre d'essai en indexant avec l'outil logiciel G-MET, 500 documents contenus dans le disque dur d'un PC (Pentium 100 Mhz, 16 Mega-octets de RAM) émulé en Serveur Web Personnel (installation de Webserver de Microsoft).

Un jeu de requêtes triviales a été effectué : le temps de réponse était d'une moyenne de 3 secondes.
Selon la plupart des spécialistes que nous avons consultés, un temps de réponse de 3 secondes pour un tel système est assez bon.

8.2)Evaluation du système de recherche d'information:

8.2.1)Introduction

Un grand effort dans la recherche a été fait pour résoudre le problème d'évaluation des systèmes de recherche d'information. De nombreux articles ont été publiés sur ce sujet. Néanmoins, de nouvelles méthodes sur l'évaluation sont constamment éditées (Cooper [61] ; Jardine et Rijsbergen [69]; Heine [62]).
Devant un tel problème en perspective ; il y a trois questions qui s'offrent au premier abord selon Rijsbergen[69] :

1. Pourquoi évaluer?
2. Que doit-on évaluer?
3. Comment évaluer? .

La réponse à la première question est d'ordre social et économique. La partie sociale est sincèrement intangible, mais se rapporte principalement aux avantages (ou inconvénients) relatifs aux systèmes de recherche de l'information. Le mot «avantage» ici à un sens beaucoup plus large. Par exemple, quel avantage les utilisateurs obtiennent-ils (ou quels torts subiront-ils) en remplaçant les sources traditionnelles d'information par un système de recherche complètement automatique et interactif ? Pour évaluer tout cela, des études ont été effectuées, mais les résultats sont difficiles à

interpréter. Pour certains systèmes de recherche, l'avantage pourrait être plus facilement évalué que pour les autres. La réponse économique représente ce qui va vous coûter d'utiliser l'un de ces systèmes, suivie d'une question qui est: est - ce que cela vaut la peine? Même un relevé du coût est difficile à faire. Les prix de l'ordinateur peuvent être faciles à évaluer, mais les prix en termes d'effort du personnel sont plus difficiles à déterminer.

Dans l'évaluation d'un système de recherche d'information, nous traitons surtout l'obtention des données afin que les utilisateurs puissent prendre une décision à savoir :

1. S'ils préfèrent un tel système (question sociale) ;
2. Si cela vaut la peine.

Par ailleurs, ces méthodes d'évaluation sont utilisées de façon comparative pour évaluer si certains changements entraîneraient une amélioration du rendement. En d'autres mots, quand une demande est faite pour une stratégie de recherche particulière, la valeur d'évaluation peut être appliquée pour déterminer si la demande est valable.

La deuxième question (que doit-on évaluer?) concerne l'évaluation qui se rapporte à la capacité du système qui répondra aux besoins de l'utilisateur. Au début de l'année 1966, Cleverdon [68] a répondu à cette question. Il désigna six principales entités quantifiables:

1/ La collection de documents retrouvée par le système de recherche d'information qui comprend ceux qui sont pertinents ;

2/La durée d'attente, qui est l'intervalle moyen entre l'instant du lancement de la requête et l'instant d'obtention de la réponse;

3/ La forme de présentation des résultats en sortie;

4/ L'effort de recherche demandé à l'utilisateur pour l'obtention de réponses à ses requêtes;

5/ Le rappel du système, qui est la proportion de documents pertinents retrouvés réellement en réponse à la demande de recherche d'information ;

6/La précision du système, c'est à dire la proportion de documents retrouvés qui sont réellement pertinents.

La question finale (comment évaluer ?) a une réponse technique plus détaillée. Il est intéressant de noter que la technique de mesure de l'efficacité de recherche est très dépendante de la stratégie de recherche particulière adoptée et de la forme de sa sortie.

8.2.2) Exemples de modèles d'évaluation :

a)Le modèle de Swets:

Au début de l'année 1963, Swets [95] a exprimé son mécontentement sur les méthodes d'évaluation existantes. Son expérience sur la détection du signal lui a permis de concevoir un modèle d'évaluation basé sur une théorie statistique de décision. En 1967, il évalua une cinquantaine de méthodes selon son modèle. Les résultats de son évaluation étaient encourageants, mais non concluants. Plus tard, Brookes [82] a proposé des modifications à la mesure d'efficacité de Swets [88] ; Robertson [76] a démontré que les modifications proposées étaient en fait, relatifs à une mesure alternative déjà proposée par Swets.

Brookstein [54] a récemment réexaminé ce modèle expliquant comment Swets se basait sur une supposition de « contradiction égale ».

Bien que le modèle de Swets soit théoriquement attrayant et qu'il lie les mesures de recherche d'information à la théorie statistique bien développé et prête à l'emploi, il n'a pas trouvé de preneurs parmi les spécialistes. Au début de son rapport de 1967, Swets [88] affirmait :

Une mesure de la performance de recherche attendue devrait avoir les propriétés suivantes : Premièrement, elle doit uniquement exprimer la capacité du système de recherche à reconnaître les articles voulus et les articles non voulus, ce qui veut dire qu'elle doit être une mesure d'efficacité seulement, laissant les facteurs distincts de considération relatifs au coût ou « capacité ». Deuxièmement, la mesure attendue ne doit pas être confondue par l'empressement du système à ignorer des articles – elle doit exprimer une

force de discrimination indépendante de tout critère d'approbation employé, si le critère est caractéristique du système ou ajusté par l'utilisateur. Troisièmement, la mesure doit être un chiffre unique - de préférence, à une paire de chiffres qui peut co- varier de façon imprécise, ou à une courbe représentant une table de plusieurs chiffres doubles, afin qu'elle soit simplement transmise et immédiatement comprise. Quatrièmement et finalement, la mesure doit permettre un ordre complet de différentes performances et évaluer la performance d'un seul système en termes absolus, c'est à dire que le système métrique doit représenter une échelle avec une unité, un zéro réel et une valeur maximum. Avec une mesure ayant ces propriétés, nous pouvons être sûrs d'avoir un index valide où un système de recherche (ou méthode) réalise la fonction qu'il devait accomplir, et nous pourrons poser les questions telles que : payerons - nous tant pour tant d'unités d'efficacité?

b)Le modèle de Cooper : Longueur de recherche attendue

En 1968, Cooper [81] a déclaré : « la fonction principale d'un système de recherche est d'éviter aux utilisateurs l'étude et le rejet des documents non pertinents pendant la recherche des documents pertinents. C'est cette 'sauvegarde' qui est mesurée et qui est le seul index de mérite pour les systèmes de recherche. En général, l'index est appliqué aux systèmes de recherche avec des sorties ordonnées ou classées. Il mesure approximativement l'effort d'une recherche qu'on espère sauvegarder, en utilisant le système de recherche au lieu de chercher la collection au hasard. Un essai est effectué pour étudier les différents problèmes posés au moment de trouver les documents pertinents pour différentes requêtes. L'index est calculé pour une demande de type spécifié. On suppose que les utilisateurs peuvent quantifier leur besoin en informations selon l'un des types suivants :

a) seul un document pertinent est recherché ;
b) un chiffre arbitraire n est demandé ;
c) tous les documents pertinents sont recherchés ;
d) une proportion donnée des documents pertinents est recherchée.

- Le résultat d'une stratégie de recherche est supposé être un ordonnancement faible des documents. Hélas, le classement produit par la fonction d'appariement est rarement un ordre simple, mais plus communément un ordre faible. Ceci signifie qu'à n'importe quel niveau de classement, il y a au moins un document (probablement plusieurs) qui rend la longueur de la recherche inappropriée, puisque l'ordre des documents est aléatoire. Si les informations que nous voulons sont ordonnées à un certain niveau et dépendent de la disposition des documents pertinents, nous obtiendrons des longueurs de recherche différentes. Cependant, nous pouvons utiliser une quantité analogue qui est la longueur de la recherche attendue. Pour cela, nous avons besoin de calculer la probabilité de chaque longueur de recherche en jonglant (mentalement) avec les documents pertinents et non pertinents selon le besoin de l'utilisateur

8.2.3)Evaluation du système DREPU:

Nous venons de passer en revue quelques méthodes d'évaluation de système de recherche d'information ; avec une spécificité pour chacune d'entre elles. Certes, elles sont assez anciennes et surtout traitent des anciens systèmes de recherche automatisée d'information. Néanmoins la philosophie et les principes restent sensiblement les mêmes.

Pour l'évaluation du système DREPU, nous voudrions nous en tenir aux trois questions fondamentales de Van Rijsbergen [69], même si elles ont soulevé quelques controverses que l'auteur reconnaît lui-même dans une nouvelle édition de son ouvrage[21] .

Pour rappel, ces questions sont :

1- Pourquoi évaluer?

2- Que doit-on évaluer?

3- Comment évaluer?

1-La première question nous amène à situer notre système par rapport aux autres systèmes existants déjà. Si nous voulons l'évaluer, c'est pour voir s'il

[21] " Information retrieval" de C.J. van RIJSBERGEN (Londres, Butterworths, 1979)

est plus performant et plus efficace que d'autres systèmes ; ou du moins permet-il d'avoir des résultats pertinents et permet-il une nette dimunition du « bruit » ? Nous avons conçu et développé le système DREPU notamment pour ces deux objectifs ; nous voudrions donc l'évaluer pour valider notre modèle et ainsi vérifier notre théorie.

2-Nous tenterons de répondre à la deuxième question en mettant en relief les points essentiels qui distinguent le système DREPU.

Nous pouvons d'abord tenter de retrouver quelques unes des entités quantifiables définies par Cleverdon :

-La durée d'attente, qui est l'intervalle moyen entre l'instant du lancement de la requête et l'instant d'obtention de la réponse :Nous avons calculé un intervalle moyen de trois secondes pour le système DREPU.

- L'effort de recherche demandé à l'utilisateur pour l'obtention de réponses à ses requêtes : Nous avons minimisé cet effort en permettant à l'utilisateur d'être entièrement guidé dans sa recherche, avec des menus déroulant lui évitant même des saisies d'éléments de description ou de profil d'utilisateur.

- La proportion des documents pertinents retrouvés par rapport à l'ensemble des documents pertinents présents dans la collection (taux de rappel) est très importante pour tous les essais que nous avons faits. Ces derniers ayant été fait dans une petite base propriétaire constituée de documents que nous avons nous même indexé, nous ne pouvons donner d'évaluation numérique pour le taux de rappel. Nous pouvons toutefois conclure qu'avec le système DREPU le "silence" est considérablement réduit. Evidemment seuls des essais à grande échelle avec une installation du système DREPU sur un serveur en réseau (Internet ou intranet) permettant à

plusieurs utilisateurs d'indexer des documents suivant ce modèle ; donc de participer à la constitution d'un grand réservoir d'éléments d'indexation ; pourraient nous confirmer cela.

- La proportion des documents pertinents par rapport à l'ensemble des documents fournis par la recherche (taux de précision) pour les mêmes tests s'est avéré aussi très importante. De même que pour le taux de rappel, nous ne pouvons pas donner de valeur numérique pour ce taux, pour les mêmes raisons que précédemment. Le taux de « bruit » est assez réduit.

Nous avons testé aussi les mesures secondaires classiques d'efficacité telles:

- Le rejet ou la proportion des documents non pertinents retrouvés par rapport au nombre total de documents non pertinents existants dans la base. Dans notre cas, sur les quelques recherches faites, nous avons estimé le rejet considérable; c'est à dire que la majorité les documents non pertinents ont été éliminés. Nous signalons que nous avons effectué les essais avec plusieurs bibliothécaires rompus aux recherches d'information pour un besoin d'évaluation objective des performances du système DREPU. Les estimations que nous donnons ici correspondent à une moyenne d'avis de plusieurs utilisateurs.

- La sélectivité ou proportion de documents non pertinents qui n'ont pas été retrouvés. Cette variable est l'inverse de la précédente. Les tests ne nous ont fait que confirmer les résultats obtenus précédemment, avec les mêmes conclusions donc.

Nous avons par ailleurs conduits notre évaluation sur un échantillon de recherches, avec d'autres utilisateurs habitués aux systèmes de recherches d'informations, à qui nous avons expliqué le fonctionnement de DREPU. Après quelques tests sur une base de données de 500 documents, nous avons recueilli leur satisfaction quant aux temps de réponse et à la notion d'effort de l'utilisateur dans le processus de recherche. Nous avons aussi

croisé les recherches avec les utilisateurs ; c'est à dire chaque utilisateur reprenant la recherche d'un autre et nous avons comparé les résultats qui furent pratiquement identiques pour évidemment un jeu de test limité….

Nous rappelons encore une fois que ces résultats très encourageants sont obtenus sur une petite base propriétaire de 500 documents. Il n'y a aucune loi qui peut nous permettre d'extrapoler pour les grandes bases ou réservoirs de ressources d'informations du Web, pour affirmer que notre système sera autant performant ou un peu moins, si ce n'est de tenter l'expérience en grandeur réelle. C'est à cet effet que nous nous proposons de le mettre en ligne sur Internet pour un retour d'écho des éventuels utilisateurs qui l'auront testés….

CONCLUSIONS :

Toutes les stratégies de recherche d'information sont basées sur la comparaison entre la demande et les documents stockés dans la base de données ou le réservoir de ressources d'information. Parfois cette comparaison est seulement réalisée indirectement quand la demande est comparée aux groupes (ou plus précisément aux profils représentant les groupes).

Les distinctions faites entre les différents types de stratégies de recherche d'information peuvent parfois être comprises selon le langage de la demande, c'est à dire le langage dans lequel l'information est transmise. La nature du langage de la demande dicte souvent la nature de la stratégie de recherche. Par exemple, le langage de la demande qui permet aux énoncés de la recherche d'être exprimés selon les combinaisons logiques des mots clés, dicte normalement la recherche Booléenne. C'est une recherche qui donne des résultats par des comparaisons logiques de la demande avec les documents.

D'autres stratégies de recherche d'information plus sophistiquées sont mises en application par les moyens de la fonction d'appariement (ou similarité). Le plus connu des projets, basé sur cette stratégie est le projet SMART, appelé aussi corrélation cosinus, qui suppose que le document et la demande sont représentés comme des vecteurs numériques. Cette stratégie est appliquée dans le projet smartpush, que nous avons cité dans la Partie C.

Il serait utile de mentionner aussi que le mot retour, introduit par certaines stratégies de recherche d'information, est normalement utilisé pour décrire le mécanisme par lequel un système peut améliorer sa performance en étudiant sa dernière performance. En d'autres termes, un système entrée – sortie simple retransmet l'information à partir de la sortie de sorte que cela puisse être utilisé pour améliorer la performance sur la prochaine entrée.

Le but de chaque stratégie de recherche est de trouver les documents pertinents et d'éliminer les documents non pertinents. Malheureusement, la pertinence est définie selon l'interprétation sémantique de la demande de l'utilisateur.

Les expériences ont démontré que le retour de pertinence peut être très efficace. Malheureusement, l'importance de l'efficacité est difficile à évaluer puisqu'il est plutôt difficile de séparer la contribution à accroître l'efficacité de recherche produite quand les documents individuels sont transférés par classement à partir de la contribution engendrée dés que les nouveaux documents sont retrouvés.

Finalement, il apparaît que la mise en œuvre du retour de pertinence sur une base opérationnelle peut être plus problématique. Nous ne savons pas comment les utilisateurs évaluent la pertinence ou la non-pertinence d'un document à partir de références. Dans un système opérationnel, il est plus facile de classer les résumés qui doivent sortir, mais l'utilisateur a besoin de balayer les documents retrouvés afin de déterminer leur pertinence et de reformuler sa demande selon une meilleure position.

Mais en règle générale tous les systèmes de recherche d'informations sont basés sur le mot-clé. Certains sont devenus plus sophistiqués dans leur utilisation des mots-clés, par exemple, ils peuvent inclure une forme de normalisation et une sorte de pondération. Certains utilisent l'information de distribution pour mesurer la force des rapports entre les mots-clés et les descriptions des mots-clés des documents; quand des rapports sémantiques entre les mots sont définis et exploités, nous avons alors atteint les limites de l'ingéniosité avec les mots-clés.

La plupart des preuves expérimentales de la dernière décennie ont démontré la supériorité de cette méthode sur les alternatives possibles. Néanmoins, il y a un espace pour des améliorations plus performantes. Il semble qu'à la base de l'utilité de recherche, il existe la capacité ou (l'incapacité) de la représentation informatique des documents….

Nous avons commencé par faire une étude rétrospective des principaux thèmes et systèmes rentrant dans le cadre de la recherche d'information. Pour les systèmes de recherches d'informations, nous avons tenté de voir comment chacun d'eux prend en charge les requêtes des utilisateurs ; tout en scrutant bien, les différentes possibilités avec lesquelles, ils tiennent comptent du profil de l'utilisateur ; afin de retrouver d'éventuelles similitudes avec notre théorie que nous avons établie, il faut peut-être le préciser, en juin

1998, au moment où je terminais mon DEA. C'est à ce moment là qu'est née dans mon esprit l'idée d'associer des éléments de description de ressources du Dublin Core avec des éléments de profil d'utilisateur, dans un standard de metadonnées spécifique. Evidemment l'idée n'est pas venue spontanément, sans aucun préalable. Sa genèse repose en fait sur plusieurs observations et hypothèses :

- Mes observations dans mon milieu professionnel de la fiabilité des recherches de documents quand le profil de l'utilisateur est associé ou pris en compte dans le processus, soit dans le cadre d'une recherche manuel, dans un service de références ; soit sur un poste de recherche automatique ;
- Les hypothèses de tenir compte du profil de l'utilisateur dans un processus de recherche d'information du projet PROFILDOC du laboratoire RECODOC de l'université LYON1 ;
- Les hypothèses sur lesquelles se sont construits plusieurs projets, que j'ai cité d'ailleurs dans la partie « C » de cette thèse ;
- Mes observations des développements et de l'évolution rapide du Dublin Core que je suivais au sein d'un Groupe de travail (Datamodele) dont je suis membre ; et qui me permettaient d'entrevoir toutes ses possibilités d'utilisations ;
- Mes observations sur l'utilisation des préférences utilisateur par certains moteurs de recherche et leur impact sur la recherche d'information;
- Le succès du système PICS (Plateform for Internet Content Selection) qui est basé sur une association d'éléments de description et des notions d'intérêts d'utilisateur (même restreint à certains domaines uniquement) ;
- Et enfin une ferme conviction que l'avenir d'une recherche d'information fiable et efficace sur Internet ne pourrait se faire sans l'association d'éléments de description de ressources et d'éléments de profil d'utilisateur.

Mais je n'avais dans mon esprit, en ce temps-là que l'embryon de mon modèle. Il m'a fallu le mûrir pendant deux ans, m'informer de la majorité

des systèmes d'information existants ; étudier les caractéristiques et les particularités de chacun. Je me suis intéressé à tous les projets de développement de systèmes prenant en compte l'intérêt ou le profil ou les préférences de l'utilisateur. J'ai découvert que depuis deux ou trois ans, beaucoup d'équipes de recherche s'intéressent à ce domaine et ont mis en place des projets de recherche (voir partie C). J'ai tenté de me situer parmi tous ces projets ; et de prendre ma propre voie. Evidemment je ne pouvais à mon sens proposer une théorie sans prévoir d'outil de test ou de validation ; et c'est là où j'ai lancé le développement de l'outil de recherche et de filtrage de l'information DREPUZ et de l'outil d'aide à l'indexation (ou générateur d'éléments META) G-MET. Je me suis rapproché un peu plus des procédures des projets que j'ai cités, tout en me maintenant sur mon propre axe. Je pense ainsi avoir contribué avec le système DREPU, autant que les autres projets, dont certains n'ont pourtant pas connu de réalisation pratique, à faire entrevoir d'autres possibilités pour la recherche et le filtrage d'information. Certes le système DREPU est encore perfectible sur plusieurs plans ; et je suis entièrement convaincu, au vu des résultats obtenus et de la comparaison avec certains systèmes de recherches d'information, qu'il peut être une alternative pour les directories telles que OpenDirectory ou Yahoo. Il pourrait grandement faciliter le travail manuel d'indexation des ressources du Web, des nombreux réferenceurs qui travaillent continuellement pour ces systèmes. La recherche serait aussi améliorée et deviendrait plus efficace….

Pour le moment nous espérons tout simplement que ce travail sera repris sous forme de projet par un laboratoire universitaire pour le valoriser et nous permettre de lui apporter de nécessaires améliorations.

ANNEXES

Annexe 1 : Plate-forme logicielle du système DREPU :

1)- Outil de recherche d'information DREPUZ :

Introduction:

DREPUZ est un outil de recherche et de filtrage d'information basé du système DREPU. Développé avec le langage Perl, il est constitué d'une suite de scripts qui s'exécutent grâce à une plate-forme Perl installé sur le serveur. Nous avons opté pour cette solution au lieu de compiler directement des exécutables sous Perl pour gagner en rapidité d'exécution et pour répondre aux soucis de la plupart des webmestres qui préfèrent des scripts au lieu d'exécutables plus lourds et plus difficiles à maintenir.

DREPUZ permet de faire des recherches dans une base de données propriétaire préalablement créée, et fonctionne sur le principe de l'interrogation à relance. C'est à dire que la première requête nous donne le nombre total de documents y répondant et que la relance permet de les affiner en tenant compte du profil de l'utilisateur.

A-Presentation de DREPUZ :

Pages HTML Statiques

Page d'accueil (index1.htm) :

Cette page contient : Un logo, une bannière « DREPUZ site de démonstration », un cadre à gauche pour orienter vers d'autres pages html du site, un cadre au milieu pour saisir la requête et sélectionner les éléments Meta. Dans cette page, l'utilisateur peut saisir sa requête dans le champ '**Trouve**r' et faire une sélection unique ou multiple (touche de CTRL enfoncée+ clic sur-le-champ désiré) dans la liste déroulante, des éléments méta sur lesquels il désire orienter sa recherche. L'utilisateur peut faire appel aux opérateurs booléens « OU » (par défaut) et « ET » dans le cadre de sa recherche.

Image 9

Ensuite, il pourra activer sa recherche en appuyant sur le bouton '**chercher**'. L'utilisateur peut également lancer une requête sur plusieurs thèmes en les séparant par un '# ' (Diez blanc).

Remarques :

Si la recherche est basée sur le facteur booléen « OU » aucune restriction n'est faite quant au choix des éléments META et la requête de l'utilisateur. Si le facteur booléen « ET » est sélectionné, l'utilisateur est tenu de suivre les règles suivantes :

- *Si un seul terme ou mot clé est entré, la recherche peut se faire sur un ou plusieurs éléments META sélectionnés ;*

- *Si plusieurs termes ou mots clés sont choisis, il faut que l'ordre d'entrée des valeurs cherchées et leur nombre coïncident avec l'ordre et le nombre des éléments META affichés dans la liste déroulante*

Exemple *:*
Chercher les termes et expression « ***amerouali# metadata et profil utilisateur# 2000****» en tenant compte des éléments META « AUTEUR, TITRE, DATE »*
La recherche aura la forme : AUTEUR= « amerouali» ***ET*** *TITRE= « metadata et profil utilisateur»* ***ET*** *DATE= « 2000». Uniquement les documents répondant à cette requête seront pris.*

- *Si le nombre des termes est supérieur au nombre d'éléments META sélectionnés ; ceux qui n'auront pas de correspondance directe dans la liste des éléments META sélectionnés ne seront pas pris en considération.*

Exemple *:*
Chercher les termes et expression « ***amerouali# metadata et profil utilisateur# 2000# français****» en tenant compte des éléments META « AUTEUR, TITRE, DATE »*
La recherche aura la forme : AUTEUR= « amerouali» ***ET*** *TITRE= « metadata et profil utilisateur»* ***ET*** *DATE= « 2000». Le mot clé français sera ignoré.*

DREPUZ, contrairement à beaucoup de moteurs de recherche, ignore la casse : l'utilisateur n'a pas à se soucier de l'emploi des majuscules ou des minuscules durant la saisie de sa requête car le résultat sera le même. Les blancs répétitifs seront ignorés.

Par exemple **: You&cef Am@e*rOuali# mOte>ur de rec<>ùH%erche# 1998.**

DREPUZ effectuera automatiquement une correction et prendra comme thème de recherche : **youcef amerouali# moteur de recherche# 1998.**

La page d'accueil peut également nous renvoyer vers les pages indexdoc1.html, apropos1.htm, aide1.htm, en sélectionnant respectivement « **indexer un document »**, **« A propos de… »** et **« Aide ».**

Les champs : **titre**, **auteur**, **identificateur**, **date**, **nom** du document HTML à indexer et tous les éléments caractérisant le profil d'utilisateur sont obligatoires. Pour les champs **auteur**, **sujet**, **source**, **identificateur**, l'utilisateur peut entrer plusieurs attributs d'un même élément descriptif du document en les séparant par une virgule.

1. **Page "G-MET»**(indexdoc1.htm) :
   Cette page (image10) contient :Un logo, une bannière « outil d'aide à l'indexation» , un cadre à gauche pour orienter vers la page d'accueil, un formulaire de saisie comprenant l'ensemble des champs à renseigner pour le document à indexer.
   Les données saisies dans ce formulaire sont traitées par le script CGI **genmeta.pl**

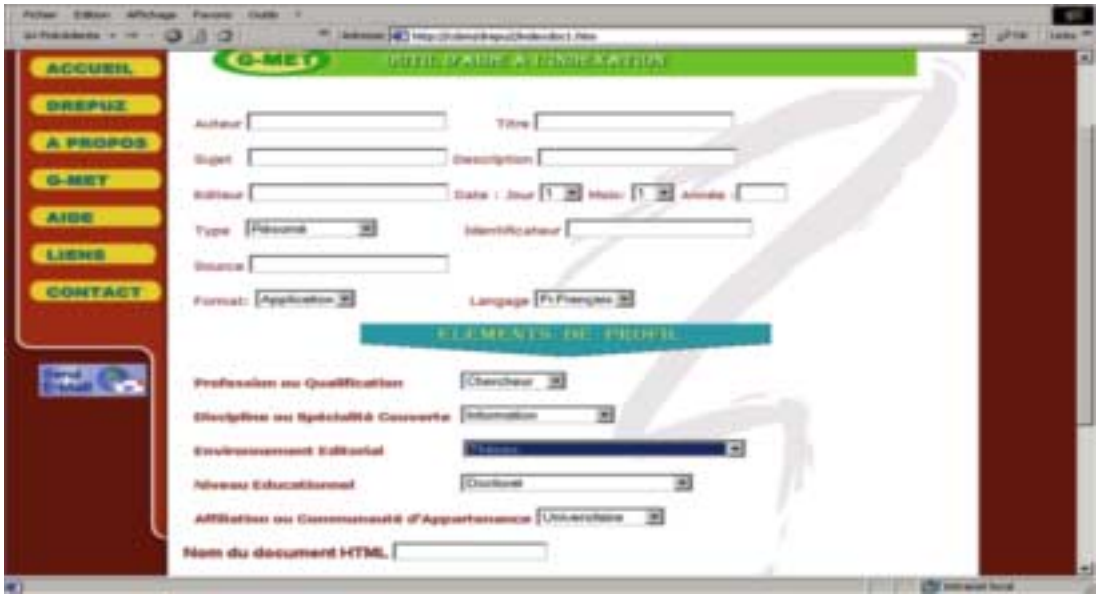


Image10

2. **Page « a Propos de… »** (Apropos1.htm) :

Cette page contient : Un logo, une bannière « A Propos de… ", un cadre à gauche pour orienter vers les autres pages du site.

Dans cette page l'utilisateur peut prendre connaissance du système DREPU.

Image11

3. **Page aide** (Aide1.htm) :

Cette page contient :Un logo, une bannière « Aide » , un cadre à gauche pour orienter vers les autres pages du site.

Dans cette page, l'utilisateur est informé sur la manière d'utiliser l'outil de recherche et de filtrage d'information DREPUZ.

Image12

B- Pages HTML générées :

1- Page de relance de l'interrogation :

Cette page contient :Un logo, une bannière « Relance de l'Interrogation» et une barre dans laquelle est affiché le nombre de documents trouvés.

Dans cette page, l'utilisateur est amené à choisir entre la possibilité de filtrer les résultats obtenus en appuyant sur le bouton « oui » (relance de l'interrogation) ou d'afficher la totalité des documents retrouvés.

Cette page est générée par le script CGI **rech_doc.pl**

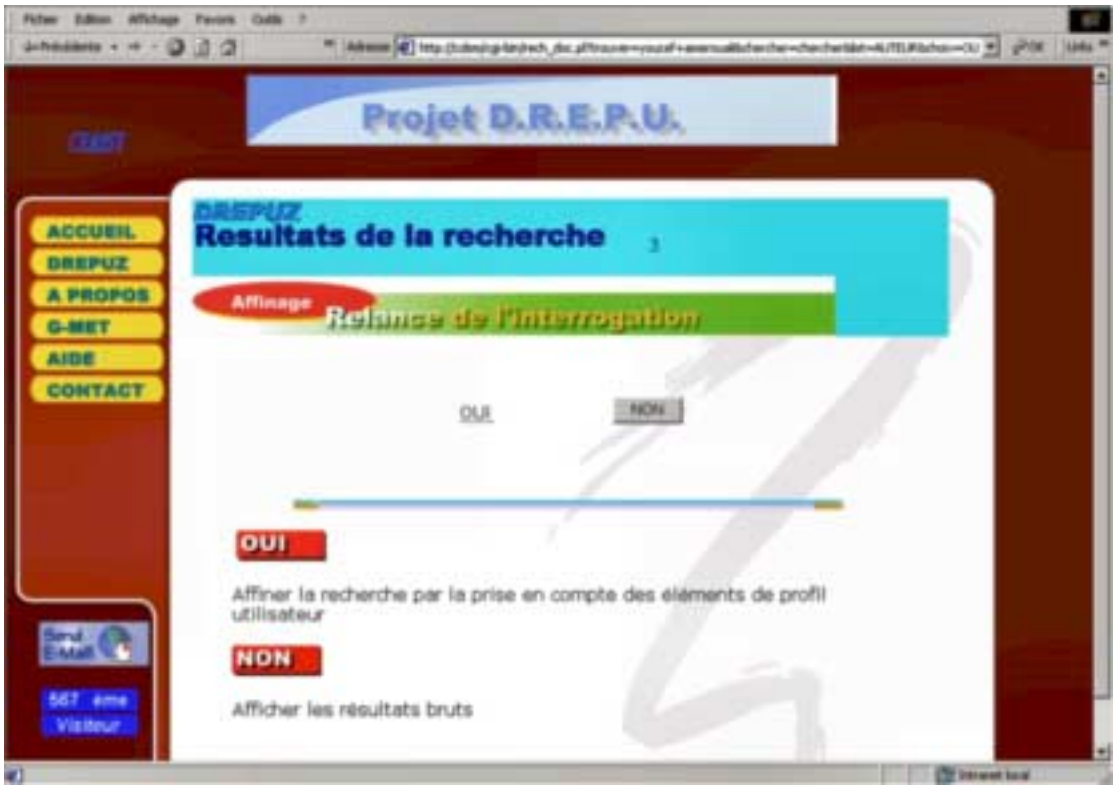


Image13

2- Page (s) Résultats sans filtrage (resultat_brut.html) :

Cette page (ou plusieurs) contient :Un logo, une bannière « Résultats sans filtrage» et une barre dans laquelle est affiché le nombre de documents trouvés.

Chaque page contient dix (10) documents référencés par le titre, l'auteur et son URL. Au bas de la page, l'utilisateur peut actionner un cadran pour voir la suite des documents restants s'il y a lieu.

Cette page est générée par le script CGI **aff_pg.pl.**

Image14

3- Filtrage suivant le profil utilisateur :

Cette page contient :Un logo, une bannière « relance de l'interrogation » et « filtrage suivant profil utilisateur».

Dans cette page, l'utilisateur choisit dans la liste déroulante pour chaque élément du profil utilisateur, l'attribut lui convenant et valide son choix en appuyant sur le bouton « Affiner ».

Cette page est générée par le script CGI **aff_pg.pl.**

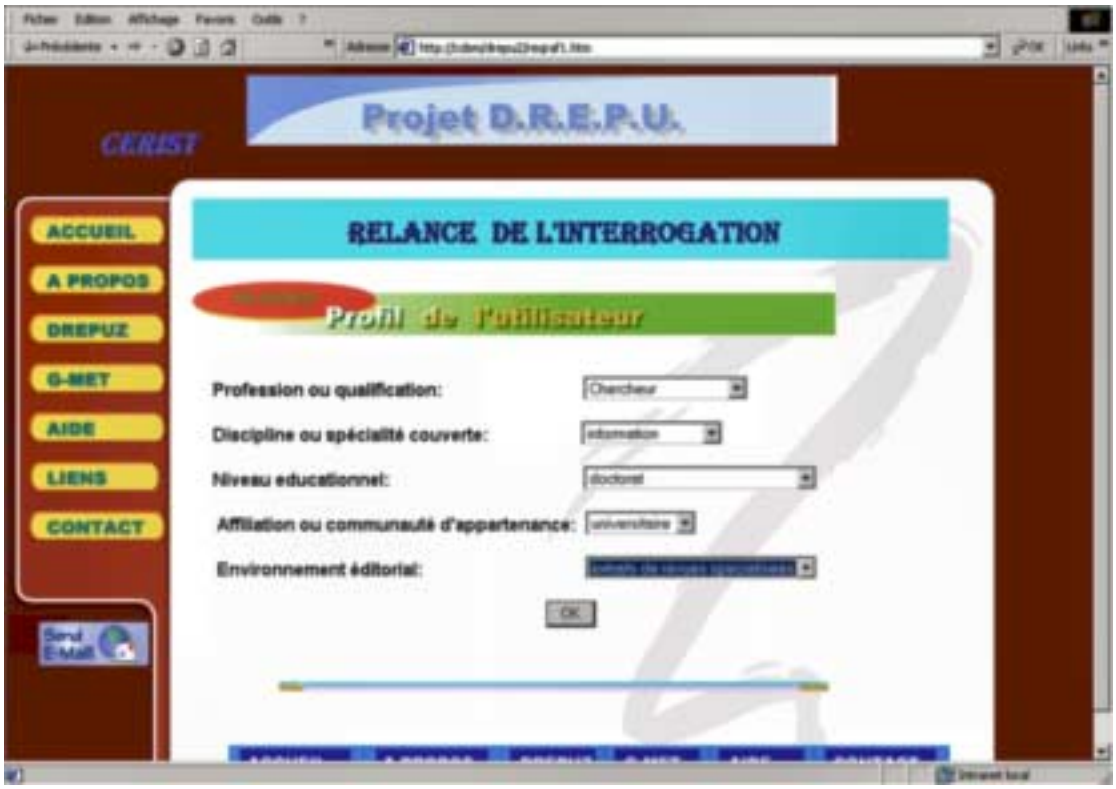


Image 15

4.Page des résultats après filtrage:

Cette page (ou pages en fonction du nombre de documents trouvés) contient :Un logo, une bannière « Résultats Après filtrage » et une barre dans laquelle est affiché le nombre de documents trouvés.

Dans cette page (ou ces pages), l'utilisateur peut trouver la liste affinée des documents répondants à sa requête.

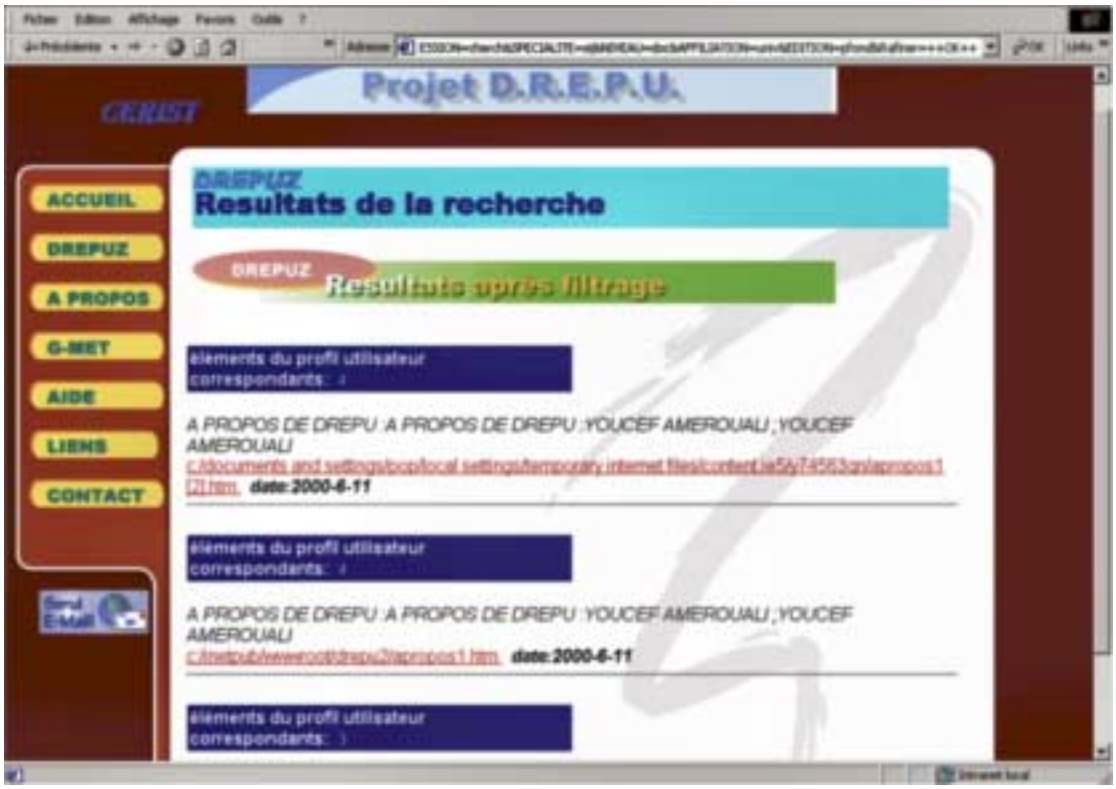


Image16

Cette liste est triée suivant le nombre décroissant d'éléments de profil d'utilisateur pour chaque document. Si deux ou plusieurs documents sont à égalité, ils seront triés suivant leur date de création. Chaque page contient dix (10) documents référencés par le nombre d'éléments de profil d'utilisateur, le titre, l'auteur, son URL et la date de création. Au bas de la page, l'utilisateur peut actionner un cadran pour voir la suite des documents restants s'il y a lieu.

Cette page est générée par le script CGI **raffiner.pl.**

C- Structure du fichier meta.idx :

META.IDX est un fichier catalographique contenant les éléments méta de tous les documents indexés par G-MET. En cas de perte, il peut être regénéré par le robot ertila.

Il sert de base de recherche pour DREPUZ.

Sa mise à jour est également faite grâce au script CGI genmeta.pl.

Sa structure est la suivante :

Numéro d'ordre, élément méta, contenu de l'élément méta.

Exemple :

2-AUTEUR=Youcef AMEROUALI

2-TITRE=drepuz

2-SUJET=moteur de recherche

2-DESCRIPTION=moteur de recherche site de démonstration

2-EDITEUR=*CDSM*

2-DATE=2000-5-3

2-IDENTIFICATEUR=isbn 2-86595-467-6

2-LANGAGE=fr

2-PROFFESSION=Etudiant

2-URL= http://www.drepuz.chezmoi.html

Annexe 2 : Les algorithmes :

La plate-forme logicielle de DREPU comporte six scripts écrits sous le langage de programmation Perl, dont nous reproduisons ci-après les algorithmes:

- **Algorithme Ertila** :

Ertila est un robot parcourant un site à la recherche des documents HTML répondant au modèle DREPU :

Début

Ouvrir (meta.idx) /* *fichier index* */

Ouvrir (ertila.idx) /* *fichier contenant l'ensemble des chemins à parcourir par le robot* */

i←1 /* *le nombre de chemins à prendre en compte par le robot* */

num_ord_doc←1 /* *numéro d'ordre du document dans meta.idx* */

Tant que ¬eof(ertila.idx)

Faire

Seachdir(i) ←chemin(i)

i←i+1

Finfaire

j←1

Tant que j<=i

Faire

Tant que ¬ fin searchdir(j) /* *parcourir le répertoire choisi jusqu'à la fin* des fichiers ayant l'extension HTML ou HTM */

Faire

Si extension(seachdir(j), fichier)='.html'

ou extension (seachdir(j)), fichier='.htm' **Alors**

Empiler (allfiles, fichier) ;

Finsi

Finfaire

j←j+1

Finfaire

```
j←1
Tant que ¬ vide(allfiles)
  Faire
  /* Appel de la procédure index_fichier pour la recherche des éléments
   meta dans fichier j */
   file ← dépiler (allfiles, fichier)
   Index_fichier  (file)
  Finfaire

  Procédure index_fichier (fichier);
   Début
      Lire (attribut_fichier)
      Si attribut_fichier = '-r'  Alors
      Ecrire ('fichier protégé en lecture')
      Sinon
      Ouvrir(fichier)
      Lire (fichier, chaîne)
     Tant que    chaîne <> '\head'
        Faire
          Metaname←nom_balise_meta
          Metacontent← contenu_balise_meta
          Empiler (metalist, metaname+'='+metacontent) ;
        Finfaire
  Test_profil←faux       /* pour s'assurer que le document est décrit suivant
                            le profil utilisateur */
   Test_robot←faux       /* tester si le document accepte d'être indexé par
                            un   robot */
     Tant que ¬      vide (metalist)
        Faire
          W← dépiler (matalist) /*  exemple :w←"AUTEUR=amerouali" */
          Verif_ident_gauche←partie_gauche (w) /* partie_gauche p.r.p.
          au signe '='*/
          Verif_ident_droite←partie_droite (w) /* partie_droite p.r.p
          au signe '=' */
```

Si verif_ident_gauche=’PROFILUSER’ **Alors**

Test_profil←vrai

Finsi

Si (verif_ident_gauche= ’ROBOT’) **ET** (verif_ident_droite) = ’NOINDEX’ **Alors**

Test_robot← vrai

Finsi

Si (test_profil=vrai) ET (test_robot=faux) **Alors**

Ecrire(meta.idx, num_ord_doc+’-‘+w)/*

*exemple :10-AUTEUR=amerouali */*

Finsi

Finfaire

Si (test_profil=vrai) ET (test_robot=faux) **Alors**

num_ord_doc← num_ord_doc +1

Finsi

Finsi

Fin procédure

Fermer (meta.idx)

Fin Algorithme

- **Algorithme rech_doc :**

C’est le script qui permet de faire la recherche de documents html répondant à la requête de l’utilisateur.

Début

Si méthode =POST **Alors**

/ data contient une chaîne de caractères lue*/*

leng = variable d’environnement (‘CONTENT_LENGTH’)

Tant que leng > 0

Faire

data ← char */* char représente un caractère de la requête entrée */*

leng←leng-1

Finfaire

Sinon

*/ * il s’agit de la méthode GET */*

```
        data = variable d'environnement ('QUERY_STRING')
Fsi
  Ouvrir (meta.idx) /* fichier index */
  Ouvrir ( fichier temporaire) /* fichier temporaire */
   /*Décodage des données saisies sur la forme nom1
    =valeur1&nom2=valeur2&…*/
   i←1 ;
   Tant que ¬ fin (data)
      Faire
      query_string←élément &(data)
     /*query_string est une variable chaîne de caractères telle que : */
/* query_string ←'nom1=valeur1' ; query_string←'nom2=valeur2' ; etc.*/
express_requet(i)←requete_décode(query_string) /* requete_décode est
 une procédure */
 /* exemple : express_requet[1]← Auteur=Amerouali Youcef# Afif Mohamed */
 /*              express_requet[2] ← Auteur=Benhammadi Farid  */
      i←i+1
      Finfaire
 j←1
 l←1
Tant que j<=i
     Faire
     exp_gauche← partie_gauche (expression_requete (j))  /*partie_gauche
    p.r.p.  au signe'=' */
     exp_droite ← partie_droite (expression_requete (j))    /* partie_droite p.r.p
     au signe '='*/
     nombre_diez←Chercher ('#  ') dans Exp_droite /* chercher si exp_droite
     comprend des '#  '    et sauvegarder le nombre
     Exemple : valeur1#  valeur2 …*/
   Si nombre_diez =0 Alors
        requete_finale (l)←express_requet (j)
        l←l+1
   Sinon
        Tant que nombre_diez > 0
```

```
            Faire
            Requet_finale (l) ←exp_gauche+'='partie gauche # '
            l←l+1
          nombre_diez← nombre_diez-1
          Finfaire
   Finsi
   j←j+1
Finfaire
 i←1
 k←1
 var=0
  Tant que i<= l
     Faire
          Test ← chercher (meta.idx, requete_finale(i))   /* la recherche se fait
          ligne par ligne */
          Si  test = vrai Alors
              var←numero_ordre (ligne, meta.idx)
             Pour m←1 jusqu'à k faire  /* i : longueur du tableau comprenant
             les numéros des documents  */
             Chercher variable dans tableau_ordre
               Si trouver Alors
                  Tableau_ordre(k)←var
                   K← k+1
               Finsi
      Finfaire
    Finsi
    i←i+1
  Finfaire
 j=1
 Tant que j<=k
    Faire
          Lire(meta.idx, ligne)
          Si tableau_ordre[j]=ligne (numero_ordre) Alors
            Ecrire (fichier temporaire, ligne)
```

Finsi

j←j+1

Finfaire

Fin Algorithme

- **Algorithme ajoutesite** :

C'est le script qui permet d'ajouter les sites au fichier ertila.idx contenant l'ensemble des chemins à parcourir par ertila.pl:

Début

Si méthode =POST **Alors**

/ data contient une chaîne de caractères lue*/*

leng = variable d'environnement ('CONTENT_LENGTH')

Tant que leng > 0

Faire

data ← char */* char représente un caractère de la requête entrée */*

leng←leng-1

Finfaire

Sinon

*/ * il s'agit de la méthode GET */*

data = variable d'environnement ('QUERY_STRING')

Fsi

Ouvrir (ertila.idx) */* fichier ertila.idx contenant l'ensemble des chemins à parcourir */*

Query_string← élément &(data) */ * qs est un tableau de chaîne de caractères tel que :*

Query_string (1) ='nom1=valeur1';

Query_string (2) ='nom2=valeur2' ; etc./*

Query_string ←requete_decode(Query_string) */ * enlever les caractères '+' , '%' de la requête*/*

qs_droite ←partie droite (Query_string) ; */* qs_droite p.r.p. au signe '=' */*

Si qs_droite <> ' ' **Alors**

Ecrire (ertila.idx, qs_droite)

Finsi

Fermer (ertila.idx)

Fin Algorithme

- **Algorithme genmeta** :

C'est l'outil d'aide à l'indexation G-Met qui permet la génération d'éléments META et la mise à jour du fichier meta.idx :

Début

Si méthode =POST **Alors**

/ data contient une chaîne de caractères lue*/*

leng = variable d'environnement ('CONTENT_LENGTH')

Tant que leng > 0

Faire

data ← char */* char représente un caractère de la requête entrée */*

leng←leng-1

Finfaire

Sinon

*/ * il s'agit de la méthode GET */*

data = variable d'environnement ('QUERY_STRING')

Fsi

Ouvrir (meta.idx) */* fichier index */*

Ouvrir (fichier_temporaire) */* fichier temporaire */*

/ Décodage des données saisies sur la forme nom1 =valeur1& nom2=valeur2&…*/*

/ et mise à jour du fichier index */*

i←1 ;

Tant que ¬ fin (data**)**

Faire

query_string←élément &(data)

*/*query_string est une variable chaîne de caractères telle que : */*

/ query_string ←'nom1=valeur1' ; query_string←'nom2=valeur2' ; etc.*/*

qs_gauche←partie gauche (query_string) ;

qs_droite←partie droite (query_string) ;

Si qs_gauche = 'ficname' **Alors**

Fichier_utilisateur ← requete_decode(qs_droite)

Sinon

Si gs_gauche <> 'soumettre' **Alors**

Valeur_ajouter ← requete_decode(qs_droite)

Finsi

numero_ordre← dernière_position+1

Chaîne(i) ← '<meta name='+qs_gauche+'content='+qs_droite+'>'

Ecrire (meta.idx, numero_ordre +'-'+chaine)

i←i+1

Finsi

Ouvrir (Fichier_utilisateur)

j=1

Tant que j<=i

Faire

écrire (Fichier_utilisateur, chaine(j))

j← j+1

Finfaire

Finfaire

Fermer (meta.idx)

Fermer (Fichier_utilisateur)

Procedure requete_decode;

Début

/ transformer le '+' en blanc et % XX à sa valeur correspondante */*

Si caractère = '+' **Alors**

Remplacer + par ' '

Sinon

Remplacer valeur hexadécimale par équivalence

Finsi

Fin procédure

Fin Algorithme

- **Algorithme aff_pg :**

C'est le script qui traite et met en forme les résultats à afficher

Début

Calcul ← 1 */*pour vrai*/*

Doc_affiches ← 0

Nbr_doc_affiche← 0

Last_aff_doc ← 0

Dat ← variable d'environnement */* variable qui reçoit les paramètres de la procédure récursive */*

Si calcul = 1 **Alors**

data ← variable d'environnement

tab_requet(i) ← les éléments de data séparés par rapport au signe &

/ tab_requet est un tableau dont les éléments sont de la forme :*

tab_requet(1)= « nom1=valeur1 » ; tab_requet(i)= « nom i=valeur i »/*

Pour k←1 **jusqu'à** i

Faire

tab_requet_Gauche ← partie gauche de tab_requet(k)

tab_requet_Droite ← partie droite de tab_requet(k)

/ gauche et droite par rapport au signe'= ' */*

Si tab_requet_Gauche= « Raffiner » **Alors**

Affiner_ou_afficher ← requet_decode(tab_requet_Droite)

k← i

Finsi

Finpour

Sinon

tab_requet(i) ←les éléments de dat séparés par rapport au signe &

Pour k=1 **jusqu'à** i

Faire

tab_requet_Gauche ← partie_gauche de tab_requet(k)

tab_requet_Droite ← partie_droite(tab_requet(k))

/ gauche et droite par rapport au signe = */*

Si tab_requet_Gauche= ' DERNDOC ' **Alors**

Last_aff_doc← tab_requet_Droite

Sinon

Si tab_requet_Gauche= ' NBDOCRES ' **Alors**

```
                    Nbr_doc_rest← tab_requet_Droite
                    Sinon
                        Si tab_requet_Gauche= ‘ calcul ‘ Alors
                            Calcul←tab_requet_Droite
                        Sinon
                            Si tab_requet_Gauche= ‘ AFFOURAFF ‘ Alors
                                Affiner_ou_afftout← tab_requet_Droite
                            Sinon
                                Si tab_requet_Gauche= « CPT » Alors
                                    Compteur← tab_requet_Droite
                                Finsi
                            Finsi
                        Finsi
                    Finsi
                Finsi
            Finfaire
        Finsi
Si affiner_ou_afftout= ‘ OUI ‘ Alors
    affafin       /* appel de la procédure affafin */
Sinon
    afftout                 /* appel de la procédure afftout */
Finsi

Procédure afftout
    Début
        Ouvrir (fichier temporaire)
                affiche_list_doc

Si nbr_doc_rest > 10 Alors
    Nbr_doc_rest← nbr_doc_rest-10
    aff_pg (variable d’environnement)
   /* appel récursif de aff_pg les paramètres sont envoyés dans la variable
    d’environnement   par la méthode GET */
 Fsi
```

```
    Ecrire (« /BODY »)
  Fin procédure

Procédure affiche_list_doc
  Début
        Nbr_doc_aff ← 0
        Tant que ¬ Eof ( fichier_temporaire)  ET nbr_doc_aff < 10
                    Faire
                       Lire(fichier temporaire, ligne)
                       Si last_doc_aff < ligne(numero_ordre) Alors
                            W  ←ligne
                            W_Gauche ← partie gauche de W
                            W_Droite  ← partie droite de W
                            /* gauche, droite p.r.p au signe = */
                            Si W_Gauche= ‘ IDENTIFICATEUR ‘Alors
                               Resultat_3  ← W_Droite
                            Sinon
                                  Si W_Gauche= ‘ TITRE ‘Alors
                                     Resultat_1 ←  W_Droite
                                  Sinon
                                     Si W_Gauche= ‘ AUTEUR ‘ Alors
                                        Resultat_2  ← W_Droite
                                  Finsi
                            Finsi
                    Finsi
                 Afficher(résultat_1)
                 Afficher(résultat_2)
                 Afficher(résultat_3)
                Finsi
               Nbr_doc_aff ← nbr_doc_aff+1
        Finfaire
        Last_aff_doc ← ligne(numero_ordre)
  Fin Procédure
 Procédure affafin
```

Début

Affichage du formulaire de sélection des éléments de profil utilisateur (reqraf.html)

Raffiner /* *Appel du programme rafiner.pl* */

Fin

- **Algorithme raffiner :**

C'est le script qui affiche les résultats après filtrage:

Début

calcul ←1 /* 1 :pour vrai

doc_affichés ←0

last_aff_doc ← 0

i ←0

dat ← variable d'environnement

requet(i)←éléments de dat séparés p.r.p au signe &

/* *requet(i) : tableau de la forme nom1=valeur1 ; nom i= valeur i*

Pour j= **1 jusqu'à** i

Faire

Requet_Gauche←partie gauche de requet(j) p.r.p au signe =

Requet_Droite ←partie droite de requet(j) p.r.p au signe =

Si Requet_Gauche= ' NBDOCRES ' **Alors**

Nbr_doc_rest ←Requet_Droite

Sinon

Si Requet_Gauche= ' PROFESSION' **Alors**

Data1← Requet_Droite

Sinon

Si Requet_Gauche= ' SPECIALITE ' **Alors**

Data2 ←Requet_Droite

Sinon

Si Requet_Gauche= ' AFFILIATION ' **Alors**

Data3 ←Requet_Droite

Sinon

Si Requet_Gauche= 'NIVEAU' alors

```
                    Data4 ←Requet_Droite
                Sinon
                  Si Requet_Gauche= ‘ ENVIRONNEMENT ‘ Alors
                    Data5 ←Requet_Droite
                  Sinon
                    Si Requet_Gauche= ‘DERNDOC’ Alors
                      Last_aff_doc  ←Requet_Droite
                    Sinon
                      Si Requet_Gauche= ’calcul’ Alors
                        Calcul←Requet_Droite
                      Finsi
                  Finsi
              Finsi
            Finsi
          Finsi
        Finsi
      Finsi
    Finsi
  Finsi
Finfaire
 Data  ←  « data1&data2&data3&data4data5 »
 Ouvrir(fichier temporaire)
Expr_requet(i) ←les éléments de data séparés p.r.p au signe &
/* Expr_requet(1)= « data1 »,……., expr_requet(5)= « data5 »*/
 I ←0
Pour j=0 jusqu’à i
  Faire
    Tant que ¬  eof(fichier temporaire)
      Faire
          Lire(fichier temporaire, ligne)
          Si expr_requet(j)=ligne Alors
            Tab_result(I) ←ligne(numero_ordre)
            I ← I+1
          Finsi
      Finfaire
```

Finfaire

/ une fois les documents répondants à la requête trouvés et leurs numéros d'ordre ajoutés au tableau tabrésult */*

trier(tab_result) */* par ordre croissant */*

/ compter le nombre d'éléments méta « profil utilisateur » par document */*

W← 0

Pour y=0 **jusqu'à** l

Faire

compt ← 0

i ← y

Tant que i < l

Faire

Si tab_result(i)=tab_result(y) **Alors**

compt ←compt+1

Finsi

i ← i+1

Finfaire

tab_final(W)←' tab_result(y) '+ ' – '+ compt

W ← W+1

y ←y+compt */* j'ai trouvé compt fois le nombre du document, donc, compt fois les éléments du profil utilisateur pour ce document */*

/ exemple : tab_final(w)=10-4 documents n°10 et 4 éléments profil utilisateur*/*

Finfaire

/ ajouter les date pour pouvoir les utilisées dans le classement des documents */*

Pour m=0 **jusqu'à** W

Faire

tab_final_Gauche ← partie_gauche (tab_final(m))

tab_final_Droite ←partie_droite (tab_final(m)) /* prp au signe-*/

val ← ' tab_final_Gauche '+ '-'+'DATE =' /* 10-DATE=*/

trouver ←faux

Tant que¬ eof (fichier temporaire) et trouver = faux

Faire

Lire(fichier temporaire, ligne)

Si val= ligne **Alors**

Spresult_Droite ←partie_droite de ligne p.r.p au signe = /*10-12-1999 */

Trouver ← vrai

Finsi

Finfaire

Hassh(tab_final_Gauche) ← ‘ tab_final_Droite ‘+ ‘-‘+ spresult_ Droite

/* hassh(10)=4-10-12-1999 */

Finfaire

/* donc les indices de hassh sont les numéros des documents et les valeurs sont le nombre d’éléments meta « profil utilisateur » avec les dates des documents */ trier hassh(i) /*par ordre croissant*/

affiche_list_doc /* *lancer la procédure d’affichage des résultats*/

Fin Algorithme

BIBLIOGRAPHIE :

1. [2000] Youcef Amerouali et Richard Bouché. Metadata based on elements of ressources'description and user profiles. Dans les Proceedings des Conférences SCI'2000/ ISAS'2000, Orlando, USA, 23-26 juillet 2000.

2. [2000] Youcef Amerouali. Metadata et profil utilisateur. Dans les proceedings de la Conférence CAIS/ACSI 2000, Edmonton/ Alberta, Canada, 28-30 mai 2000.

3. [2000] Amato, Giuseppe and Thanos, Costantino and Straccia, Umberto. EUROgatherer: a Personalised Gathering and Delivery Service on the Web. In Proc. of the 4th World Multiconference on Systemics, Cybernetics and Informatics (SCI-2000), Orlando, USA, 2000.

4. [1999] Hubert Fondin. La recherche d'information dans les mémoires électroniques. Documentaliste-Sciences de l'information, vol.36, numéro 4-5, pages 242-248, 1999.

5. [1999] Amato, G. and Straccia, U. User Profile Modeling and its Application to Digital Libraries. Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, LNCS, 1696, pp. 184-197, 1999.

6. [1999] Youcef Amerouali. Metadonnées basées sur des éléments de description de ressources et des éléments de profil d'utilisateur. Dans les proceedings du 2° colloque du chapitre français de l'ISKO, ENSSIB / Lyon, 21-22 octobre 1999.

7. [1999] Straccia, Umberto. Foundations of a Logic based approach to Multimedia Document Retrieval. PhD thesis, Department of Computer Science, University of Dortmund, June 1999.

8. [1999] Amato, Giuseppe and Straccia, Umberto. User Profile and Applications to Digital Libraries. In Proceedings of the Third European

Conference on Research and Advanced Technology for Digital Libraries (ECDL-99), LNCS 1696, pages 184-197, Paris, 1999.

9. [1998] Amos A. David. Modélisation de l'utilisateur et recherche coopérative dans les systèmes de recherche d'informations. Proceedings de la 5ème Conférence Internationale de l'ISKO ; Université de Lille ; 25 au 29 août 1998.

10. [1998] Youcef Amerouali, Prise en compte du système descriptif de documents Profildoc dans le cadre des metadata (Dublin Core). 126 pages. Mémoire de DEA en Sciences de l'Information et de la Communication, ENSSIB, juillet 1998.

11. [1998] Faouzi Tchenar, *Modélisation de l'utilisateur fondée sur ses croyances et ses buts en vue d'améliorer l'efficacité des systèmes de recherche d'information*. Proceedings de la 5e Conférence Internationale de l'ISKO ; Université de Lille ; 25 au 29 août 1998.

12. [1997] Boris Chidlovski, Uwe M. Borghoff and Pierre-Yves Chevalier. Towards Sophisticated Wrapping of Web-based Information Repositories. In Proc. 5 Th Int'l RIAO Conf., Montreal, Canada, June 25-27, 1997, pp. 123-135.

13. [1997] Roman S.Panchyshyn et France Bouthillier *Cataloguer le cyberspace :le defi des ressources electroniques* Documentation et bibliotheques, juillet/septembre 1997; pp.137-146.

14. [1997] Norbert Fuhr. A decision-theoretic approach to database selection in networked IR. ACM Transactions on Information Systems, 17, 1997.

15. [1997] Braun, I. Konig, A. Wichmann, T.:"PICS-SE: a proposed standard for the annotation of internet documents using a string extension to PICS», Draft, Février 1997. Http://www.kulturbox.de/aid.pics-se.dc.html

14. [1997] Younger, J.A.
Resource description in the digital age
dans: "Library trends", 45:3, pp.462-487 .

15.[1997] Desai, B.C. ; Supporting discovery in virtual libraries
dans: "Journal of the American Society for Information Science",
48:3, 1997 ; pp.190-204.

16.[1996] Luk, Annie T. . Evaluating bibliographic displays from the users' point of view:a focus group study. Master of information studies research project report. Faculty of information studies , Universté de Toronto; 1996.

17.[1996] Lynne C. Howarth , Joseph P. Cox. Facilitating access to electronic bibliographic record content with client preferences. In proceedings of 24th Annual Conference of the Canadian Association for Information Science. Toronto , juin 1996.

18.[1996] Jan Smits
Digital Metadata,Standards for Communication and Preservation
European Research Libraries Cooperation :The Liber Quaterly
6(1996); pp.383-406

19. [1996] Dempsey, L. et Weibel, S.L.
The Warwick Metadata Workshop: A framework for the
deployment of resource description
dans: "D-lib Magazine", July/August 1996.

20. [1996] Dempsey, L.
ROADS to Desire: Some UK and Other European Metadata
and Resource Discovery Projects dans: "D-lib Magazine", July/August 1996.

21.[1996] Smith, Terence R.
The Meta-information Environment of Digital libraries

dans: "D-lib Magazine", July/August 1996.

22. [1996] Bloedorn, E., Mani, I., MacMillan R. (1996). Representational Issues in Machine Learning of User Profiles. In proc. of AAAI Spring Symposium on Machine Learning in Information Access.

23.[1996] Heery, Rachel
Review of Metadata Formats
dans: "Program" 30:4, pp.345-373. October 1996

24.[1996] Weibel, S. et al
Advances in metadata. (1996),
dans: "D-lib Magazine, July/August 1996"

25.[1995] Dempsey L. RAD*AR reflexions :internet ressource,access,discovery and retrieval systems and libraries.* Library networking in Europe. Conference europeenne/Bruxelles/12-10-1994, Londres :TFPL Publishing 1995.

26. [1995] G.S. Jung and V.N. Gudivada, Autonomous tools for information discovery in the worldwide web. Technical Report CS-95-01, School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, 1995.

27. [1995] K. Decker, V. Lesser, et al., Macron: An Architecture for multi-agent cooperative information gathering. In CIKM Conference, Workshop on Intelligent Information Agents, 1995.

28. [1995] A. ORiordan and C. Buckley, An intelligent agent for high-precision information filtering. In CIKM Conference, Workshop on Intelligent Information Agents, 1995.

29. [1995] R. Armstrong et al., Webwatcher: A learning apprentice for the world-wide web. In Proc. of the Symposium on Information Gathering from

Heteroge-neous, Distributed Environments. AAAI Press, 1995.

30. [1995]. H. Lieberman Letizia, an agent that assists web browsing. In Proceedings of the IJCAI-95. AAAI Press, 1995.

31. [1995] M. Balabanovic and Y. Shoham, Learning information retrieval agents: Experiments with automated web browsing. In AAAI Technical Report SS-95-08, Proc. of the 1995 AAAI Spring Symposium Series, 1995.

32.[1995] Mangan Elizabeth U. *The making of a standard* Information technology and libraries, 2(14), 1995.

33.[1994] Y. Labrou and T. Finn, A semantics approach for kqml - a general purpose communication language for software agents. In Proc. of Conference on Information and Knowledge Management 1994. MIT press, 1994.

34.[1994] S. Laine-Cruzel. Vers de nouveaux systèmes d'information prenant en compte le profil des utilisateurs. Documentaliste-sciences de l'information, 1994, vol. 31, n° 3, pp. 143-147.

35.[1994] Riekert Wf. Management of data and services for environmental applications. Environmental knowledge organisation and information management : Bratislava, 14-16 Septembre 1994 Knowledge Organisation in Subject Areas, DEU 1994.

36. [1994] Sheth, Beerud (1994). A Learning Approach to Personalised Information Filtering, M.I.T.; http://agents.www.media.mit.edu/groups/agents/papers/

37.[1994] Yan, Tak W.,Garcia-Molina, Hector (1994). Index Structures for Information Filtering under the Vector Space Model. IEEE Conference on Data Engineering.

38. [1993] B. Sheth and P. Maes. "Evolving agents for personalized information filtering." In Proceedings, The Ninth Conference on Artificial Intelligence for Applications, pages 345-352. IEEE Computer Society, 1993.

39. [1993] B. Sheth and P. Maes, Evolving agents for personalized information filtering. In Proc. of the ninth Conference on Artificial Intelligence for Applications, IEEE Computer Society Press, 1993.

40. [1992] Crawford, Walt. Starting over: current issues in online catalog user interface design. Information Technology and Libraries. Mars 1992 ; pp.62-76.

41.[1992] Foltz, P.W., Dumais, S. (1992). Personalised Information Delivery : An Analysis of Information Filtering Methods. Communications of the ACM 35(12)) http://www--psych.nmsu.edu/~pfoltz/cacm/cacm.html

42. [1992] N. Belkin and B. Croft, Information filtering and information retrieval. Communications of the ACM, 35, No. 12, 1992.

43. [1992]. Nicholas J. Belkin and W. Bruce Croft. Information Filtering and Information Retrieval: Two Sides of the Same Coin? In Communication of the ACM, December 1992, Vol 35, No. 12, pp. 29-38.

44. [1992]. Peter W. Foltz and Susan T. Dumais. Personalized Information Delivery: An Analysis of Information Filtering Methods. In Communication of the ACM, December 1992, Vol 35, No. 12, pp. 51-60.

45. [1992] A. Jennings and H. Higuchi. "A personal news service based on a user model neural network." IEICE Transactions on Information Systems, 75(2):198-209, 1992.

46.[1991] Hubert Fondin. Ergonomie des systèmes d'information documentaire. Thèse de Doctorat d'Etat en Lettres et Sciences Humaines. 698 pages. Décembre 1991. Université Michel de Montaigne, BordeauxIII.

47. [1991] G. Fischer and C. Stevens. "Information access in complex, poorly structured information spaces." In Human Factors in Computing Systems, CHI '91 Conference Proceedings. ACM, 1991.

48. [1989] Gerard Salton and J. Michael McGill. Introduction to Modern Information Retrieval. Addison Wesley Publ.Co., Reading, Massachusetts, 1989.

49.[1988] Richard Bouché. Sciences de l'information : sciences de la mise en forme. Infomédiatique, avril 1988 ; pages 11-18.

50.[1987] Hubert Fondin. L'évolution des systèmes et des métiers du traitement de l'information: La crise du monde documentaire. Documentaliste, Janvier-Février 1987 ; vol.24, numéro 1, pages 3-10.

51. [1984] Wallace, Danny P. The user friendlinessof the library catalogs. Université de l'Illinois; Graduate School of Library and Information Science. Occasional papers N°163.

52. [1983] Salton, G., McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill Eds.

53. [1977] PAICE, C.D., Information Retrieval and the Computer, Eds Macdonald and Jane's, London (1977).

54. [1977] BOOKSTEIN, A., 'When the most "pertinent" document should not be retrieved - an analysis of the Swets Model', *Information Processing and Management,* **13**, pp.377-383 (1977).

55. [1975] CAWKELL, A.E., 'A measure of "Efficiency Factor" - communication theory applied to document selection systems', *Information Processing and Management,* **11**, pp.243-248 (1975).

56. [1975] DOYLE, L.B., Information Retrieval and Processing, Eds Melville Publishing Co., Los Angeles, California (1975).

57. [1975] SALTON, G., Dynamic Information and Library Processing, Eds Prentice-Hall, Englewoods Cliffs, N.J. (1975).

58. [1974] ROBERTSON, S.E. and TEATHER, D., 'A statistical analysis of retrieval tests: a Bayesian approach', *Journal of Documentation,* **30**, pp. 273-282 (1974).

59. [1974] KOCHEN, M., Principles of Information Retrieval, Ed Melville Publishing Co., Los Angeles, California (1974).

60. [1973] BARBER, A.S., BARRACLOUGH, E.D. and GRAY, W.A. 'On-line information retrieval as a scientist's tool', Information Storage and Retrieval, 9, pp. 429-44- (1973).

61. [1973] COOPER, W.S., 'On selecting a measure of retrieval effectiveness', Part 1: 'The "subjective" philosophy of evaluation', Part 2: 'Implementation of the philosophy', *Journal of the American Society for Information Science,* **24**, pp.87-100 and pp.413-424 (1973).

62. [1973] HEINE, M.H., 'Distance between sets as an objective measure of retrieval effectiveness', *Information Storage and Retrieval,* **9**, pp.181-198 (1973).

63. [1973] HEINE, M.H., 'The inverse relationship of precision and recall in terms of the Swets' model', *Journal of Documentation,* **29**, pp.81-84 (1973).

64. [1972] SALTON, G., Paper given at the 1972 NATO Advanced Study

Institute for on-line mechanised information retrieval systems (1972).

65. [1972] Palmer Richard P. Computerizing the card catalogue in the university library: a survey of user requirements. Littleton, Colo.: Libraries Unlimited.

66. [1972] WINOGRAD, T., Understanding Natural Language, Edinburgh University Press, Edinburgh (1972).

67. [1972] COOPER, M.D., 'A cost model for evaluating information retrieval systems', *Journal of the American Society for Information Science,* **23**, pp.306-312 (1972).

68. [1972] CLEVERDON, C.W., 'On the inverse relationship of recall and precision', Journal of Documentation, 28, pp.195-201 (1972).

69. [1971] JARDINE, N. and van RIJSBERGEN, C.J., 'The use of hierarchic clustering in information retrieval', *Information Storage and Retrieval,* **7**, pp.217-240 (1971).

70. [1971] SPARCK JONES, K., Automatic Keyword Classification for Information Retrieval, Eds Butterworths, London (1971).

71. [1970] CLEVERDON, C.W., 'Progress in documentation. Evaluation of information retrieval systems', Journal of Documentation, 26, pp.55-67, (1970).

72. [1970] SALTON, G., 'Automatic text analysis', Science, 168, pp.335-343 (1970).

73. [1970] VICKERY, B.C., Techniques of Information Retrieval, Eds Butterworths, London (1970).

74. [1970] SARACEVIC, T., Introduction to Information Science, P.R. Bowker, New York and London (1970).

75. [1969] SENKO, M.E., 'Information storage and retrieval systems'.
In Advances in Information Systems Science, (Edited by J. Tou)
Plenum Press, New York (1969).

76. [1969] ROBERTSON, S.E., 'The parameter description of retrieval tests',
Part 1; the basic parameters, Journal of Documentation, 25,
pp.11-27 (1969).

77. [1969] ROBERTSON, S.E., 'The parameter description of retrieval tests',
Part 2; Overall measures, Journal of Documentation, 25, pp.93-107 (1969).

78. [1969] LESK, M.E. and SALTON, G., 'Relevance assessments and retrieval
system evaluation', Information Storage and Retrieval, 4, pp.343-359 (1969).

79. [1968] LANCASTER, F.W., Information Retrieval Systems: Characteristics,
Testing and Evaluation, Wiley, New York (1968).

80. [1968] MINSKY, M., Semantic Information Processing, MIT Press,
Cambridge, Massachusetts (1968).

81. [1968] COOPER, W.S., 'Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems', *Journal of the American Society for Information Science,* **19**, pp.30-41 (1968).

82. [1968] BROOKES, B.C., 'The measure of information retrieval effectiveness proposed by Swets', *Journal of Documentation,* **24**, pp.41-54 (1968).

83. [1968] SALTON, G., *Automatic Information Organization and Retrieval,* McGraw-Hill, New York (1968).

84. [1967] KOCHEN, M., The Growth of Knowledge - Readings on Organization and Retrieval of Information, Wiley, New York (1967).

85. [1967] BORKO, H., Automated Language Processing, Wiley, New York 1967.

86. [1967] SCHECTER, G. Information Retrieval: A Critical View, Academic Press, London (1967).

87. [1967] CUADRA, A.C. and KATTER, R.V., 'Opening the black box of "relevance"', Journal of Documentation, 23, pp.291-303 (1967).

88. [1967] SWETS, J.A., *Effectiveness of Information Retrieval Methods,* Bolt, Beranek and Newman Eds, Cambridge, Massachusetts (1967).

89. [1966] ROCCHIO, J.J., 'Document retrieval systems - optimization and evaluation', Ph.D. Thesis, Harvard University. Report ISR-10 to National Science Foundation, Harvard Computation Laboratory (1966).

90. [1966] CLEVERDON, C.W., MILLS, J. and KEEN, M., Factors Determining The Performance of Indexing Systems, Vol. 1, Design, Vol.II, Test Results, ASLIB Cranfield Project, Cranfield (1966).

91. [1966] GOFFMAN, W. and NEWILL, V.A., 'A methodology for test and evaluation of information retrieval systems', *Information Storage and Retrieval,* **3**, pp.19-25 (1966).

92. [1964] BAR-HILLEL, Y., Language and Information. 'Selected Essays on their Theory and Application, Addison-Wesley, Reading, Massachusetts (1964).

93.[1964] SHANNON, C.E. and WEAVER, W., The Mathematical Theory of Communication, University of Illinois Press, Urbana (1964).

94.[1964] STEVENS, M.E., GIULIANO, V.E. and HEILPRIN, L.B., Statistical

Association Methods for Mechanized Documentation, National Bureau of Standards, Washington (1964).

95. [1963] SWETS, J.A., 'Information retrieval systems', *Science,* **141**, pp.245-250 (1963).

96. [1963] GARVIN, P.L., Natural Language and the Computer, McGraw-Hill, New York (1963).

97. [1961] FAIRTHORNE, R.A., 'The mathematics of classification', Towards Information Retrieval, Butterworths, London, 1-10 (1961).

98. [1961] STILES, H.F., 'The association factor in information retrieval', Journal of the ACM, 8, pp.271-279 (1961).

99. [1960] MARON, M.E. and KUHNS, J.L., 'On relevance, probabilistic indexing and information retrieval', Journal of the ACM, 7, pp.216-244 (1960).

100. [1958] GOOD, I.J., 'Speculations concerning information retrieval', Research Report PC-78, IBM Research Center, Yorktown Heights, New York (1958).

101. [1957] LUHN, H.P., 'A statistical approach to mechanized encoding and searching of library information', IBM Journal of Research and Development, 1, pp.309-317 (1957).

➢ **Documents retrouvés sur Internet :**

1- Documents généraux:

[102] Dempsey, Lorcan.
ROADS to Desire: Some UK and Other European Metadata and Resource Discovery Projects.
D-Lib Magazine, Juillet/Aout 1996.
http://www.dlib.org/

[103] IEEE.
The Metadata and Data Management Information Page.
http://www.llnl.gov/liv_comp/metadata/metadata.html

[104] UKOLN.
Metadata.
http://www.ukoln.ac.uk/metadata/

[105] UKOLN.
Metadata: Mapping between metadata formats.
http://www.ukoln.ac.uk/metadata/interoperability/

[106] EPA(Environmental Protection Agency)
Scientific Metadata Standards Project.
http://www.lbl.gov/~olken/epa.html

[107] Weibel, S., Kunze, J. and Lagoze, C.
Dublin Core Metadata for Simple Resource Description.
ftp://ds.internic.net/internet-drafts/draft-kunze-dc-01.txt

[108]Gill, Tony, Grout, Catherine and Smith, Louise.
Visual Arts, Museums and Cultural Heritage Information Standards: a domain specific review of relevant standards for networked information discovery.
http://vads.ahds.ac.uk/standards.html

[109] Lagoze, Carl.
The Warwick Framework: A Container Architecture for Diverse Sets of Metadata. Juillet/Aout 1996.
http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

[110] Library of Congress.
Metadata, Dublin Core and USMARC:

A Review of Current Efforts.
gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc

[111] Miller, Paul.
Metadata for the masses.
Ariadne, Issue 5, Septembre 1996.
http://www.ukoln.ac.uk/ariadne/issue5/metadata-masses/

[112] Smith, Terence R.
The Meta-Information Environment of Digital Libraries.
D-Lib Magazine, Juillet/Aout 1996.
http://www.dlib.org/dlib/july96/new/07smith.html

[113] Warwick, Cathro. (Aout 1997)
Metadata: An Overview.
http://www.nla.gov.au/nla/staffpaper/cathro3

[114] American Society for Testing and Materials.
ASTM Section D18.01.05 Draft Specification Content Specification for Digital Geospatial Metadata.
http://info.er.usgs.gov/research/gis/standard/index.htm

[115] *Federal Geographic Data Committee (FGDC)*.
http://fgdc.er.usgs.gov/fgdc2.html

[116] Federal Geographic Data Committee.
Metadata Standards Development.
http://www.fgdc.gov/Metadata/metahome.html

[117] U.S. Geological Survey.
Government Information Locator Service.
http://www.usgs.gov/public/gils/

[118] American Society for Testing and Materials.

ASTM Section D18.01.05 Draft Specification Content Specification for Digital Geospatial Metadata.
http://info.er.usgs.gov/research/gis/standard/index.htm

[119] Federal Geographic Data Committee.
Metadata Standards Development.
http://www.fgdc.gov/Metadata/metahome.html

[120] U.S. Defense Information Technical Center.
Technology Transfer Center GILS Toolbox.
http://skydive.ncsa.uiuc.edu/toolbox/

[121] Ricky Erway (CNI/OCLC), Workshop on Metadata for Networked Images - Executive Summary,
http://www.oclc.org:5046/research/dublin_core/summary.html

[122] (AACR2) American Library Association, Anglo-American Cataloging Rules, 2nd edition.

[123] (MARC) Library of Congress, MARC Standards,
http://lcweb.loc.gov/marc/marc.html

[124] David Levy, Cataloging in the Digital Order, Digital Libraries '95,
http://csdl.tamu.edu/DL95/contents.html

[125] (ARCHIE) Alan Emtage and Peter Deutsch, Archie – an Electronic Directory Service for theInternet, USENIX Winter 1992 Technical Conference Proceedings,
http://www.bunyip.com/research/papers/1992/archie-usenix.ps

[126] Stuart L. Weibel and Carl Lagoze, An Element Set to Support Resource Discovery: The State of the Dublin Core, to appear in Journal of Digital Libraries, Draft Copy available at :
http://www2.cs.cornell.edu/lagoze/papers/jodl.html

[127] Jim Miller, Paul Resnick and David Singer, Rating Services and Rating Systems (and their Machine Readable Descriptions), Platform for Internet Content Selection Version 1.1, May 1996,
http://www.w3.org/pub/WWW/PICS/services.html

[128] Bob Schloss and Eric Miller, PICS 1.x Changes to support digital libraries, a talk at PICS WG meeting, January 1997,
http://www.w3.org/pub/WWW/PICS/970113/DigiLib/pics970113.htm

[129] Carl Lagoze, Clifford A. Lynch, and Ron Daniel Jr., The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata, Cornell University Technical Report TR 96-1593,
http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR96-1593

[130] (FGDC) Federal Geographic Data Committee, Content Standards for Digital Geospatial Metadata,
http://geochange.er.usgs.gov/pub/tools/metadata/standard/metadata.html

[131] Ann Peterson Bishop, Working Towards an Understanding of Digital Library Use, D-Lib Magazine, October 1995,
http://www.dlib.org/dlib/october95/10bishop.html

[132] Sandra D. Payette and Oya Y. Rieger, Supporting Scholarly Inquiry: Incorporating Users in the Design of the Digital Library, to appear in Journal of Academic Libraries

[133] Kristen Summers and Daniela Rus, Using Non-Textual Cues for Electronic Document Browsing, in Digital Libraries: Current Issues, Lecture Notes in Computer Science, Springer-Verlag 1995,
http://www.cs.cornell.edu/Info/People/summers/segment.html

[134] Gerard Salton and Amit Singhal, Automatic Text Theme Generation and the Analysis of Text Structure, Cornell Computer Science Technical Report TR94-1438,
http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR94-1438

[135] Michael P. Wellman, Edmund H. Durfee and William P. Birmingham, The Digital Library As Community of Information Agents, to appear in IEEE Expert, June 1996,
http://ai.eecs.umich.edu/people/wellman/pubs/expert96.html

[136] Khoa Doan, Catherine Plaisant, and Ben Scheiderman, Query Previews in Networked Information Systems, Technical Report CAR-TR-788, University of Maryland, September 1995,
ftp://ftp.cs.umd.edu/pub/papers/papers/3524/3524.ps.Z

[137] HotMeta
Moteur de recherche développé par le RDN-CRC (Australie)
Http://flare.dstc.edu.au :8017/search.html

[138] IFLA (1998) Metadata Resources,
http://www.nlc-bnc.ca/ifla/II/metadata.htm

[139] Berners-Lee, T. (1997). Metadata Architecture,
http://www.w3.org/designIssues/Metadata

[140] Savia, Eerika (1998). Metadata Based Matching of Documents and Users Profiles. http://smartpush.cs.hut.fi/pubdocs/

[141] Distributed Systems Technology Centre.
Resource Discovery Unit (RDU).
http://www.dstc.edu.au/RDU/

[142] Web Developers Virtual Library.
META Tagging for Search Engines.

http://WWW.Stars.com/Search/Meta/Tag.html

[143] Ahronheim, Judy.
Judy and Magda's List of Metadata Initiatives.
http://www-personal.umich.edu/~jaheim/alcts/bibacces.htm

[144] Cameron, Robert D.
Towards Universal Serial Item Names.
Rapport technique 97-16, Ecole d'informatique, Simon Fraser University,
3 Decembre, 1997.
http://elib.cs.sfu.ca/USIN/USIN.html

[145] Dempsey, Lorcan.
ROADS to Desire: Some UK and Other European Metadata
and Resource Discovery Projects.
D-Lib Magazine, Juillet/Aout 1996.
http://www.dlib.org/

[146] Lynch, Clifford.
Searching the Internet.
Scientific American, Mars 1997
http://www.sciam.com/0397issue/0397lynch.html

[147] W3 Consortium.
WWW Names and Addresses, URIs, URLs, URNs, URCs.
http://www.w3.org/hypertext/WWW/Addressing/Addressing.html

[148]EPA(Environmental Protection Agency)
Scientific Metadata Standards Project.
http://www.lbl.gov/~olken/epa.html

2- DUBLIN CORE:

[149] *DC-5 Dublin Core Metadata Workshop*
http://linnea.helsinki.fi/meta/DC5.html

[150]Miller, Paul and Tony Gill.
DC5: The Search for Santa.
Ariadne, Issue 12, Novembre 1997.
http://www.ariadne.ac.uk/issue12/metadata/

[151]Weibel, Stu, et. al.
The 4th Dublin Core Metadata Workshop Report.
D-Lib Magazine, Juin 1997
http://www.dlib.org/dlib/june97/metadata/06weibel.html

[152]DC-4: *NLA/DSTC/OCLC Dublin Core Down Under / The 4th Dublin Core Metadata Workshop*.
http://www.dstc.edu.au/DC4/

[153]Heery, R., et. al.
The 4th Dublin Core Workshop: Notes from UK participants.
http://www.ukoln.ac.uk/metadata/resources/dc4-notes.html

[154]Miller, P. and Gill, T.
Down Under with the Dublin Core.
Ariadne. Avril, 1997.
http://www.ukoln.ac.uk/ariadne/issue8/canberra-metadata/

[155]*CNI/OCLC Metadata Workshop: Workshop on Metadata for Networked Images*.(24-25 Septembre, 1996)
http://purl.oclc.org/metadata/image

[156]Weibel, Stuart and Eric Miller.
"Image Description on the Internet: A Summary of the CNI/OCLC Image Metadata Workshop."
D-Lib Magazine. Janvier 1997.
http://www.dlib.org/dlib/january97/oclc/01weibel.html

[157]Dempsey, Lorcan and Weibel, Stuart L.

The Warwick Metadata Workshop: A Framework for the Deployment of Resource Description.
D-Lib Magazine, Juillet/Aout 1996.
http://www.dlib.org/dlib/july96/07weibel.html

[158]Burnard, L., et. al.
A Syntax for Dublin Core Metadata: Recommendations from the Second Metadata Workshop. (Avril 1996)
http://www.uic.edu/~cmsmcg/tech/metadata.syntax.html

[159]Hakala, Juha H., et. al.(1996)
Warwick framework and Dublin core set provide a comprehensive infrastructure for network resource description.
http://www.ub2.lu.se/tk/warwick.html

[160]*OCLC/NCSA Metadata Workshop Report*.
URL: http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html

[161]*OCLC/NCSA Metadata Workshop: The Essential Elements of Network Object Description.* (1995)
http://www.oclc.org:5046/conferences/metadata/metadata.html

[162]*Meta2 Mailing List Archive*.
http://weeble.lut.ac.uk/lists/meta2/

[163]Weibel, S., Kunze, J. and Lagoze, C.
Dublin Core Metadata for Simple Resource Description.
ftp://ds.internic.net/internet-drafts/raft-kunze-dc-01.txt

[164]*Dublin Core Metadata*.
http://purl.org/metadata/dublin_core

[165]Baker, Thomas.
Metadata Semantics Shared Across Languages:

Dublin Cores in languages other than English
http://www.cs.ait.ac.th/~tbaker/Cores.html

[166]Beckett, Dave.
Proposed Encodings for Dublin Core Metadata.
http://www.hensa.ac.uk/pub/metadata/dc-encoding.html

[167]Gill, Tony, Grout, Catherine and Smith, Louise.
Visual Arts, Museums and Cultural Heritage Information Standards: a domain specific review of relevant standards for networked information discovery.
http://vads.ahds.ac.uk/standards.html

[168]Grout, Catherine and Tony Gill.
Visual Arts, Museums & Cultural Heritage Metadata Workshop Report.
http://vads.ahds.ac.uk/Metadata1.html

[169]Guenther, Rebecca.
Dublin Core Qualifiers/Substructure : a proposal.
http://www.loc.gov/marc/dcqualif.html

[170]Hakala, Juha.
Dublin Core Metadata Element Set and it's applications.
http://linnea.helsinki.fi/meta/present.html

[171]Knight, Jon and Martin Hamilton.
Dublin Core Standard Resource Types.
http://www.roads.lut.ac.uk/Metadata/DC-ObjectTypes.html

[172]Knight, Jon and Martin Hamilton.
Dublin Core Qualifiers.
http://www.roads.lut.ac.uk/Metadata/DC-SubElements.html

[173]Kunze, John.
Metadata User Guide group report.
http://weeble.lut.ac.uk/lists/meta2/0220.html

[174]Lagoze, Carl.
The Warwick Framework: A Container Architecture for Diverse Sets of Metadata. Juillet/Aout 1996.
http://www.dlib.org/dlib/july96/lagoze/07lagoze.html

[175]Lagoze, Carl, Lynch, Clifford A., and Daniel,RonJr.
The Warwick Framework:A Container Architecture for Aggregating Sets of Metadata.TR96-1593,21 Juin, 1996.
http://www.ifla.org/documents/libraries/cataloging/metadata/tr961593.pdf

[176]Library of Congress.
Metadata, Dublin Core and USMARC: A Review of Current Efforts.
gopher://marvel.loc.gov/00/.listarch/usmarc/dp99.doc

[177] Miller, Paul and Daniel Greenstein.
*Discovering Online Resources Across the Humanities: A Practical Application of the Dublin Core;*Oct.1997
http://ahds.ac.uk/public/metadata/discovery.html

[178]Powell, Andy.
Dublin Core Management; Ariadne. Juillet 1977.
http://www.ariadne.ac.uk/issue10/dublin/

[179]Sperberg-McQueen, C.M. (Avril 1996)
On Information Factoring in Dublin Metadata Records.
http://www.uic.edu/~cmsmcq/tech/metadata.factoring.html

[180]Warwick, Cathro. (Aout 1997)
Metadata: An Overview.

http://www.nla.gov.au/nla/staffpaper/cathro3.html

[181]Weibel, Stuart.
A Proposed Convention for Embedding Metadata in HTML.
http://www.oclc.org:5046/~weibel/html- meta.html

3- Projets sur le Dublin Core :

Australie

[182] *DSTC/Resource Discovery Unit*.
http://www.dstc.edu.au/RDU/

[183] *Australian Geodynamics Cooperative Research Centre (AGCRC*).
http://www.agcrc.csiro.au/

Canada :

[184]*searchBC: Vancouver Webpages*.
http://vancouver-webpages.com/VWbot/searchBC.html

Allemagne :

[185]*Metadaten-Projekt = Metadata Project*.
http://www2.sub.uni-goettingen.de

[186]*SSG-Fachinformation (SSG-FI) Mathematick = Subject Area Information for Mathematics*.
http://www.sub.uni-goettingen.de/ssgfi/

[187]*The German Educational Resources Server / Deutscher Bildungs-Server*.
http://dbs.schule.de/indexe.html

[188]*Math-Net*.
http://elib.zib.de/math-net/

[189]*Electronic Information Management and Metadata in Physics*.
http://www.physik.uni-oldenburg.de/EPS/EurophysNet/PhysDep/dep-links.html

Les Pays-Bas :

[190]*Koninklijke Bibliotheek/ The National Library of the Netherlands*.
http://www.konbib.nl:8000/

Scandinavie :

[191]*The Nordic Metadata Project*.
http://linnea.helsinki.fi/meta/

Suede :

[192]*Swedish EnviroNet*.
http://smn.environ.se/smnproj/proj/summary.htm

Danemark :

[193] *Netpublikationer*.
http://www.fsk.dk/fsk/publ/online-pub/

Royaume-Uni :

[194]*Art, Design, Architecture & Media Information Gateway and the Visual Arts Data Service*.
http://adam.ac.uk/
http://vads.ahds.ac.uk/

[195]*AHDS Arts & Humanities Data Service*.
http://ahds.ac.uk/

[196]*Project BIBLINK*.
http://www.ukoln.ac.uk/metadata/BIBLINK/

[197]*Project DESIRE*.
http://www.nic.surfnet.nl/surfnet/projects/desire/desire.html

[198]*SCRAN (Scottish Cultural Resources Access Network).*
http://www.scran.ac.uk

[199]*NewsAgent for Libraries*.
http://www.sbu.ac.uk/litc/newsagent/

[200]*Electronic Library Image Service for Europe (ELISEII)*
http://severn.dmu.ac.uk/elise/

[201] *Resource for Urban Design Information*.
http://rudi.herts.ac.uk/catsrch/catsrch.html

Etats-Unis :

[202]*Monticello Electronic Library*.
http://www.solinet.net/monticello/monticel.htm

[203] *Medical Metadata Project*.
http://medir.ohsu.edu/~maletg/MedMetadata.HTM

[204] *Florida International University Digital Library*.
http://www.fiu.edu/~diglib/

[205]*Internet Scout Project's Signpost*.
http://www.signpost.org/signpost/index.html

[206] *University of Washington Digital Library*.
http://content.engr.washington.edu/

[207]*Everglades Information Network & Digital Library.*
http://everglades.fiu.edu/

[208]*University of Michigan Digital Library Registry*

Database.
http://dns.hti.umich.edu/registry/

[209] *Digital Library Catalog.*
http://sunsite.berkeley.edu/Catalog

4-ENCODED ARCHIVAL DESCRIPTION (EAD):

[210]*Encoded Archival Description (EAD) DTD*.
http://lcweb.loc.gov/loc/standards/ead/

[211]*Development of the Encoded Archival Description Document. Type Definition*.
http://lcweb.loc.gov/loc/standards/ead/eadback.html

[212]*Finding Aids for Archival Collections*.
http://sunsite.berkeley.edu/FindingAids/

[213]*Berkeley Finding Aids Conference*. 4-6 Avril, 1995.
http://sunsite.berkeley.edu/FindingAids/EAD/bfac.html

[214]*Society of American Archivists. Committee on Archival Information Exchange. Encoded Archival Description Working Group*.
http://sunsite.berkeley.edu/FindingAids/EAD/eadwg.html

5-STANDARDS DE METADATA GEOSPATIALS :

[215]American Society for Testing and Materials. *ASTM Section D18.01.05 Draft Specification Content Specification for Digital Geospatial Metadata*.
http://info.er.usgs.gov/research/gis/standard/index.htm

[216]*ERIN Directories and Metadata*.
http://kaos.erin.gov.au/technical/retrieval/directory/directory.html

[217]*ERIN Australian Glossary of Geographic Information Systems and Metadata Terms.*
http://kaos.erin.gov.au/gis/gis_gloss.html

[218]*Federal Geographic Data Committee (FGDC*).
http://fgdc.er.usgs.gov/fgdc2.html

[219]Federal Geographic Data Committee.
Metadata Standards Development.
http://www.fgdc.gov/Metadata/metahome.html

[220]Federal Geographic Data Committee.
Frequently Asked Questions concerning the FGDC's Content Standard for Geospatial Metadata.
http://www.its.nbs.gov/nbs/meta/faqa.htm

[221]*Directory Interchange Format (DIF) Writer's Guide.*
http://gcmd.gsfc.nasa.gov/difguide/difman.htmll

[222]*Global Change Master Directory (GCMD).*
http://gcmd.gsfc.nasa.gov/difguide/difman.html

[223]*MetaData and WWW Mapping Home Page*.
http://www.blm.gov/gis/nsdi.html

6-GOVERNMENT INFORMATION LOCATOR SERVICE (GILS) :

[224]U.S. Geological Survey.
Government Information Locator Service.
http://www.usgs.gov/public/gils/

[225] GILS Forum Archives. (Depuis Aout 1994)
gopher://gopher.cni.org:70/11/cniftp/pub/forums/gils

[226]National Archives and Records Administration(NARA)

Guidelines for the Preparation of GILS Core Entries.
http://www.ifla.org/documents/libraries/cataloging/metadata/naragils.doc
http://www.dtic.dla.mil/gils/documents/naradoc/

[227]Moen, William and McClure, Charles.
An Evaluation of the Federal Government's
Implementation of the Government Information Locator Service.
Préparé par les Services Généraux de l'administration U.S. June 30, 1997.
http://www.unt.edu/slis/research/gilseval/gilseval.htm

[228]Turner, Fay.
The U.S. Government Information Locator Service:
Description and Status.
http://gils.gc.ca/gils/backg_e.html

[229]U.S. Defense Information Technical Center.
Technology Transfer Center GILS Toolbox.
http://skydive.ncsa.uiuc.edu/toolbox/

7- Implémentations internationale de GILS :

Australia:

[230]*Information Management Steering Committee (IMSC).*
Architecture For Access To Government Information.
http://www.adfa.oz.au/DOD/imsc/imsctg/imsctg1a.htm

Canada :

[231]*Government Information Locator Service.*
http://gils.gc.ca

[232]*Guidelines for the Preparation of GILS Records*.
http://gils.gc.ca/gils/guide_e.html
[233] *Creating GILS Records in an SGML Environment.*
http://gils.gc.ca/gils/creatingg_e.html

8- IAFA/WHOIS++ TEMPLATES :

[234]Beckett, David.
IAFA Templates in use as Internet Metadata.
http://www.w3.org/pub/Conferences/WWW4/Papers/52/
http://www.hensa.ac.uk/tools/www/iafatools/slides/

[235]*Field Descriptions for Document, Software, Image, Sound, Video, Mailarchive, USENET and FAQ IAFA Template Types*.
http://www.man.ac.uk/MVC//SIMA/MMFFDB/IAFA-help/document.html

[236]ROADS. *Resource Organisation And Discovery in Subject-based services*.
http://ukoln.bath.ac.uk/roads/

[237]Heery, R. ROADS:
Resource Organisation and Discovery in Subject-based Services.
Ariadne, No. 3, 1996.
http://www.ukoln.ac.uk/ariadne/issue3/roads/

[238]Heery, R. (September 1996). *ROADS templates: how they are used.*
http://www.ukoln.ac.uk/metadata/templates.html

[239]Hiom, D., Dempsey, L. and Norman, F. *Road to resource discovery.*
http://ukoln.bath.ac.uk/roads/lar_arti.html

9- MARC (MACHINE-READABLE CATALOGUE) :

[240]Library of Congress.
Library of Congress MARC Office.
http://lcweb.loc.gov/marc/marc.html

[241]Library of Congress.
Machine-readable cataloging (MARC).
http://lcweb.loc.gov/marc/

[242]Library of Congress.
USMARC formats and documentation.
gopher://marvel.loc.gov:70/11/services/usmarc

[243]OCLC. *Bibliographic formats and standards*.
http://www.oclc.org/oclc/bib/about.htm

10-META CONTENT FORMAT (MCF) :

[244]Guha, R.V.
Meta Content Framework.An overview of MCF.
http://mcf.research.apple.com/hs/mcf.html

[245]*MCF Tutorial*.
http://www.textuality.com/mcf/MCF-tutorial.html

[246] *Meta Content Framework Using XML*.
http://www.textuality.com/mcf/NOTE-MCF-XML.html

11-PLATFORM FOR INTERNET CONTENT SELECTION PICS :

[247] Armstrong, Chris. (1997). *Metadata, PICS and Quality*.
http://www.ariadne.ac.uk/issue9/pics/

[248]W3 Consortium. *PICS.*
http://www.w3.org/pub/WWW/PICS/

[249]Resnick, Paul. *Filtering Information on the Internet.*
http://www.sciam.com/0397issue/0397resnick.html

[250]Resnick, Paul and James Miller. (Octobre 1996)
PICS: Internet Access Controls Without Censorship.
http://www.w3.org/pub/WWW/PICS/iacwcv2.htm

[251]PICS-SE -

A Proposed Standard for the Annotation of Internet Documents using a String Extension to PICS.
http://www.kulturbox.de/aid/pics-se/

12-HARVEST SUMMARY OBJECT INTERCHANGE FORMAT (SOIF) :

[252]*The Harvest Home Page*.
http://harvest.transarc.com/

[253]*The Harvest Summary Object Interchange Format*
http://harvest.transarc.com/Harvest/brokers/soifhelp.html

[254]*Harvest User's Manual.*
http://harvest.transarc.com/afs/transarc.com/public/trg/Harvest/doc.html

[255]University of Colorado.
The Harvest Information Discovery and Access System.
http://harvest.transarc.com/

13-TEXT ENCODING INITIATIVE (TEI) :

[256] *The TEI Header.*
http://etext.virginia.edu/bin/tei-tocs?div=DIV1&id=HD

[257]Burnard, Lou. (Juillet 1995)
Text encoding for information interchange: an introduction to the Text Encoding Initiative.
ftp://info.ox.ac.uk:80/~archive/teij31/WHAT.html

[258]Burnard, Lou and C. M. Sperberg-McQueen.
TEI Lite:An Introduction to Text Encoding for interchange.
Document No: TEI U 5, Juin 1995.
http://sable.ox.ac.uk/ota/teilite/

[258]Sperberg-McQueen, C.M. and Lou Burnard.
Guidelines for Electronic Text Encoding and Interchange.

Chicago / Oxford, 1994.
http://etext.lib.virginia.edu/TEI.html

14-AUTRES DOCUMENTS :

[259]Bearman, David and Sochats, Ken.
Metadata Requirements for Evidence.
http://www.lis.pitt.edu/~nhprc/BACartic.html

[260]Berners-Lee, Tim. *Document Naming*.
http://www.w3.org/pub/WWW/DesignIssues/Naming.html

[261]Burnard, Lou and Richard Light.
Three SGML metadata formats: TEI, EAD, and CIMI .
BIBLINK Work Package 1.1. December 1996.
http://www.ukoln.ac.uk/metadata/BIBLINK/wp1/

[262]Daniel, Jr., Ron.
A Global Distributed Directory Service for the World Wide Web.
http://www.acl.lanl.gov/URI/URC_proposal/index.html

[263]Day, Michael. (1997). *Extending metadata for digital preservation*.
http://www.ariadne.ac.uk/issue9/metadata/

[264]Dempsey, L.(1996). *Meta Detectors*.
http://www.ukoln.ac.uk/ariadne/issue3/metadata/

[265]Desai, Bipin C.
The Semantic Header and Indexing and Searching on the Internet.
http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html

[266]Green, Ann Gerken, and Dionne JoAnn. (20 Dec 1996)
Preserving the Whole: A two-track Approach to
Rescuing Data and Metadata.
http://www-cpa.stanford.edu/cpa/misc/preswhol.html

[267]Heery, R. (Juillet 1996)
Resource description: initial recommendations for metadata formats.
http://www.ukoln.ac.uk/metadata/DESIRE/recommendations/

[268]Iannella, Renato and Waugh, Andrew.
Metadata: Enabling the Internet.
http://www.dstc.edu.au/RDU/reports/CAUSE97/

[269]Koch, Traugott and Day, Michael.
The role of classification schemes in Internet resource description and discovery. Projet DESIRE, February 1997.
http://www.ukoln.ac.uk/metadata/DESIRE/classification/

[270]Lasher, Rebecca and Cohen, Danny.
A Format for Bibliographic Records. RFC 1807. June 1995.
http://www.dsg.cs.tcd.ie:1995/rfc1807.html

[271]Lynch, Clifford, et. al.
CNI White Paper on Networked Information Discovery and Retrieval.
http://www.cni.org/projects/nidr/www/toc.html

[272]Miller, Eric J.
Issues of Document Description in HTML.
http://www.oclc.org:5046/~emiller/tmp/issues.html