

# **Un Système à base de métadonnées pour la création d'un cache communautaire**

## **Cas de la communauté pédagogique**

**Habib SMEI, Abdelmajid BEN HAMADOU**

Laboratoire de recherche LARIM, Institut Supérieur d'Informatique et de Multimédias de  
Sfax - Tunisie

habib.smei@isetsf.rnu.tn , abdelmajid.benhamadou@isimsf.rnu.tn

### **Résumé :**

Le Web est une base de données gigantesque où il est difficile de trouver, de façon précise et rapide, l'information souhaitée. En effet les outils de recherche et de filtrage existants ne sont pas assez performants pour répondre aux exigences des utilisateurs demandeurs de documents.

Ceci est plus vrai pour une communauté d'utilisateurs où les demandes en documents sont très pointues et où les membres d'une communauté donnée souhaitent disposer des outils leur permettant de disposer des possibilités de recherche ciblées.

Dans cet article nous présentons un Système à base de Métadonnées pour la création d'un cache communautaire appliqué au contexte pédagogique. L'objectif de ce système est de créer des liens entre les membres de la communauté dans un esprit de collaboration, d'échanger des ressources et des idées, faciliter l'accès à ces ressources par un moteur de recherche qui s'articule autour d'une approche de filtrage par appariement entre Métadonnées et profil utilisateur pour améliorer la pertinence lors de la recherche des documents pédagogiques hébergés dans le système de cache que nous proposons. Ainsi nous présentons l'architecture globale de ce système et son implémentation.

**Mots-clés :** profil utilisateur, métadonnées, cache communautaire, document pédagogique, Système de Recherche d'Information.

### **Abstract:**

The Web is a gigantic data base where it is difficult to find, in a precise and fast way, desired information. Indeed the existing tools of research and filtering are not enough powerful to fulfil the users requirements of documents. This is truer for a community of users where the demands for documents are very pointed and where the members of a given community wish to have the tools their allowing having research targeted possibilities. In this article we present a System containing metadata for the creation of a Community server applied to the teaching context. The objective of this system is to create bonds between the members of the community in a collaborative spirit, to exchange resources and ideas, to facilitate the access to these resources by a search engine which uses an approach of filtering by coupling metadata and user profile to improve the relevance of the teaching documents. Thus we present the total architecture of this system and its implementation.

**Key words:** user profile, metadata, Community server, teaching document, Information Search System.

## 1. Introduction

Actuellement, on assiste à une énorme quantité de documents qui est créée, publiée et diffusée chaque jour grâce au réseau Internet. En effet, on compte plus de 3 milliards de pages Web [BrPl 03] [MEATM] indexés par Google [GOGL2] sans prendre en compte les 550 milliards de pages estimé comme Web profond "Deep Web" par BrightPlanet [BrPl 03].

Ceci engendre une augmentation continue de la taille des bases de données des outils de recherche, estimée par le site metamend [METAM] à 10 millions de pages par jour, ce qui rend par conséquent difficile la localisation de la "bonne" information recherchée par les utilisateurs. En effet, la moindre des requêtes sur un outil de recherche retourne fréquemment un nombre important de réponses, et il est très difficile de localiser l'information pertinente dans cette masse énorme de documents. L'utilisateur est confronté donc aux problèmes de bruit (beaucoup de documents retournés) générés par les outils de recherche et passe beaucoup de temps pour trouver l'information pertinente.

D'autres part, les documents qui existent dans le Web manifestent une pauvreté considérable dans leurs descriptions, on ne trouve pas dans un document Web d'indicateurs explicites décrivant par exemple sa position par rapport à la problématique qu'il traite, ni la nature du public cible auquel est destiné le document, ni encore son degré d'interactivité,... Cette insuffisance de description sémantique complique le processus de recherche et le rend incapable parfois à analyser et comprendre le contenu des documents et donc ne satisfait pas les demandes des utilisateurs en terme de recherche pointue dans un domaine particulier.

D'un autre côté, les systèmes de recherche traditionnels utilisent le vocabulaire de la langue comme lien de correspondance entre l'utilisateur et le document (correspondance requête/contenu). Néanmoins, cette stratégie basée sur la sémantique de la langue, se révèle dans la pratique insuffisante comme le montre l'augmentation du bruit lors d'une recherche. Or, l'utilisateur est une entité porteuse de connaissance et a des caractéristiques (profil) qui peuvent être prises en considération dans le processus de recherche d'informations. En effet

une connaissance des caractéristiques de l'utilisateur (son niveau intellectuel, son domaine d'action, ses centres d'intérêts, etc.) peut (en complément d'une recherche classique avec les mots clés) améliorer la qualité des systèmes de recherche d'informations.

Notons que ce travail s'inscrit dans le cadre du projet SYFAX [SFX1] [SFX2] qui s'intéresse au contexte pédagogique. SYFAX assure la fouille du Web, l'extraction des métadonnées concernant les documents qui pourront être hébergé dans le cache communautaire pour un éventuel accès rapide, il assure aussi la prise en charge des données de l'utilisateur ainsi que ses centres d'intérêt (profil utilisateur). Ce système propose une approche de filtrage par appariement métadonnées – profil utilisateur utilisé dans le processus de recherche pour éliminer les documents non intéressants. SYFAX offre aux utilisateurs des mécanismes de coopération à fin de leur permettre de partager aussi bien leurs expériences que les informations ou documents à leur disposition. La méthodologie utilisée n'est pas restreinte au contexte pédagogique, elle est tout à fait applicable à toute autre contexte d'utilisation.

Dans ce qui suit, nous présentons les solutions existantes pour améliorer la recherche, puis nous abordons l'utilisation des métadonnées et leurs apports dans le processus de recherche, et leurs adaptations au contexte pédagogique. Par la suite nous proposons une approche de filtrage par appariement métadonnées – profil utilisateur utilisée dans un système de cache communautaire pédagogique, et en fin l'architecture globale du système de cache proposé et son implémentation.

## **2. Les solutions existantes pour améliorer la recherche :**

Pour rechercher une information sur le Web, on utilise les outils de recherche automatiques (moteurs de recherche, annuaires, méta-moteurs, etc.). Ce sont des logiciels puissants permettant de parcourir tout le Web à la recherche de nouveaux sites pour les indexer et les intégrer dans leurs bases de données. Par conséquent, lorsque l'internaute formule sa requête

via l'interface d'interrogation d'un outil de recherche, ce dernier procède, par la suite, à la recherche dans les sites référencés dans sa base, pour fournir en sortie les documents en rapport avec la question posée.

Ces systèmes de recherche dits "traditionnels" utilisent le vocabulaire de la langue comme lien de correspondance entre l'utilisateur et le document (correspondance requête/contenu).

Cette stratégie appelée « indexation pleine texte en aveugle », au sens de Michard [Michard 99], se révèle dans la pratique insuffisante comme le montre l'augmentation du bruit lors d'une opération de recherche.

Récemment, on a pu voir développé d'autres idées pour améliorer la recherche des documents comme :

- la recherche multi critères en ajoutant la possibilité de raffiner la recherche par d'autres éléments que les mots clés (comme la langue, le format du document...),
- l'intégration des données de type sémantique pour décrire les documents (les métadonnées),
- la prise en charge des centres d'intérêts de l'utilisateur ainsi que son profil au processus de recherche pour mieux répondre à ses besoins,
- la restriction des documents à une communauté à travers des portails communautaire.

Chacune de ces idées pris à part présentes des limites, ainsi, l'intégration des métadonnées toute seule n'améliore que partiellement la recherche surtout pour un utilisateur non expérimenté. En effet, la multitude de choix des métadonnées lors de la recherche constitue pour lui une solution complexe et un obstacle supplémentaire pour trouver rapidement les documents. Le profil utilisateur ne constitue pas en lui seul une solution adéquate pour améliorer la recherche, l'ajout des centres d'intérêt des systèmes de recherche classiques améliore peu les résultats et n'est utilisé que pour affiner et automatiser son système de requêtes documentaires. Enfin, les portails communautaires diminuent le nombre de documents disponibles pour la communauté, sans améliorer le processus de recherche basé sur

le contenu textuel de ces documents. En plus, il y a un problème lié à l'alimentation de la base de données du portail, qui se fait de manière manuelle et où l'utilisateur est simple spectateur ou visiteur. Tout cela nous conduit à dire que le problème de pertinence lors d'une recherche persiste lorsqu'une des solutions citées plus haut est utilisée toute seule.

Notre système de cache [CACH] communautaire se base sur la plupart des idées citées en dessus : L'intégration des métadonnées et du profil utilisateur en les utilisant dans le processus de recherche par un système d'appariement métadonnées – profil utilisateur ainsi qu'un système de recommandation automatique basé aussi sur le système d'appariement déjà cité en haut.

Nous préconisons d'associer l'ajout des métadonnées et le profil utilisateur en l'appliquant à une communauté, cela permet d'améliorer considérablement la recherche en adoptant des mécanismes qui alimentent automatiquement la base de données des documents disponibles à partir du web.

### **3. Les Métadonnées :**

Le terme de métadonnées est utilisé pour définir l'ensemble des informations techniques et descriptives ajoutées aux documents pour mieux les qualifier. Pour que ces données soient utilisables par d'autres, elles doivent s'inscrire dans des modèles largement reconnus par les acteurs du Web. Plusieurs organismes de standardisation ont donc proposé et publié des schémas de métadonnées susceptibles d'être utilisés par le plus grand nombre.

Le schéma de métadonnées le plus utilisé est proposé par l'organisation Dublin *Core Metadata Initiative (DCMI)* [DUB 03a]; on l'appelle le plus souvent le *Dublin Core*.

Le *Dublin Core* vise depuis sa création à résoudre le problème de la description unifiée des ressources d'information électroniques et de leur localisation. Il est devenu une norme ISO 15836 depuis février 2003.

Selon la norme ISO 15836, le Dublin Core propose une quinzaine d'éléments descriptifs. Ces éléments sont les suivants : *Title, Creator, Subject, Description, Publisher, Contributor, Date,*

*Type, Format, Identifieur, Langage, Relation, Couverture, Droits, Source.*

Le réel intérêt de ces métadonnées est l'ajout de contenu de nature sémantique aux documents publiés, ce qui augmente fortement la qualité de l'information disponible pour les outils de recherche. Cela devrait déboucher, à terme, sur une amélioration de la pertinence des résultats des outils de recherche, en particulier pour les documents comportant ce type d'information.

Il est à noter que ces éléments peuvent apparaître dans n'importe quel ordre et que chacun d'eux est optionnel et répétitif. Nous pouvons en plus constater que ces éléments ont été prévues pour un usage purement sémantique et non par conséquent aucune hypothèse sur les langages formels ou les outils logiciels qui peuvent être employés pour créer des descriptions et des associations entre les ressources. Michard [Michard 99].

Weibel et Lagoze [WeiL 97] affirment que *“l’association de métadonnées descriptives standardisées avec des objets en réseau offre un potentiel d’amélioration substantiel des possibilités de découverte de ressources: en permettant des recherches basés sur des champs (ex.: auteur, titre, description), en permettant l’indexation d’objets non-textuels...”*

Les métadonnées garantissent l'interopérabilité en assurant le partage et l'échange d'information rendant son contenu lisible et compréhensible par les machines.

Des modèles de recherche sémantique à base de métadonnées ont également été développés. On ne s'intéresse plus aux mots saisis en eux-mêmes mais au sens qu'ils véhiculent (via une ontologie constitué d'une liste de termes proches dans le sens pour chaque terme utilisé des métadonnées, et ceci pour les utiliser dans le processus de recherche). Les documents sont donc automatiquement indexés suivant les concepts qu'ils renferment les requêtes sont analysées et leur sens est comparé à ceux des documents, ce qui permet d'établir un score de similitude entre la requête et les documents sélectionnés, et de construire une liste de réponses qui soit la plus pertinente possible.

#### **4. Adaptation de la structure des métadonnées au contexte pédagogique**

Avec l'avènement des nouvelles technologies de l'information, le développement des systèmes d'informations pédagogiques s'avère une nécessité à fin de favoriser l'implémentation de nouveaux outils et services pédagogiques, et de créer un environnement rassemblant les acteurs du contexte pédagogique (enseignants, étudiants, chercheurs,...). Un tel système peut être réalisé en exploitant les ressources pédagogiques présents déjà dans le Web, mais aussi en exploitant d'autres ressources issues des acteurs spécialistes dans le domaine d'enseignement. Ces documents ont besoin d'être bien décrits avec les métadonnées tout en tenant compte des spécificités des ressources pédagogiques. En effet les ressources pédagogiques utilisent plusieurs formats (hétérogènes), surtout des documents multimédias. Les métadonnées sont particulièrement importantes pour les documents multimédias qui, sans elles, peuvent demeurer pratiquement inexploitable et impossibles à retrouver.

Pour l'éducation, d'autres schémas de métadonnées, que le *Dublin Core*, ont été conçus, par exemple pour prendre en compte des particularités de certains métiers ou domaines d'application. Loin de s'opposer au standard *Dublin Core* (qui a le grand mérite d'être stable depuis 1999), ils le complètent en réutilisant les éléments de base et en ajoutant des éléments spécifiques. Dans le domaine de l'éducation et du *e-learning*, des groupes de travail ont tenté de définir et de spécifier des éléments de description adaptés. On peut citer principalement :

- des travaux spécifiques au sein du *Dublin Core* lui-même, avec le groupe de travail *Dublin Core Education* [DCEd] qui propose des extensions au noyau de base ;
- le modèle LOM (*Learning Object Model*) [LOM] qui s'est constitué progressivement au niveau international par un groupe de travail du *Learning Technology Standards Comitee* (LTSC) et qui est proposé à la standardisation de l'IEEE.
- un modèle récemment proposé par le réseau *European Schoolnet* [ESclNet].

- le projet européen ARIADNE Educational Metadata Recommendation [ARDN],
- le groupe de Travail Métadonnées Education (GTME) [GTME 01a]...

Nous avons retenu le schéma du GTME 1.1 [GTME 01b] au quel on a modifié quelques éléments, le schéma de métadonnées présenté ici prend place parmi ces outils. Dans sa forme actuelle il a été produit par un petit groupe de travail du CRDP qui s'est réuni 8 fois au cours de la période décembre 2000 – avril 2001. Ce groupe a été constitué dans le cadre d'une mission d'opérateur pour les ressources en ligne de l'académie de Montpellier, suite au constat de l'absence de schéma normalisé de métadonnées adapté au système éducatif français.

En effet, si de toutes part se multiplient aujourd'hui les initiatives visant à définir des métadonnées, les systèmes proposés sont rarement conçus pour les ressources éducatives.

*a) éléments du Dublin Core [DUB 03b] :*

Constatant la notabilité croissante de l'ensemble de métadonnées du Dublin Core (Dublin Core Metadata Element Set, DCMES), le groupe a sélectionné 12 des 15 éléments proposés et les a localisés, c'est-à-dire adaptés aux nécessités de la diffusion en ligne de ressources éducatives en France.

Une des premières qualités d'un tel ensemble de métadonnées étant son interopérabilité, c'est-à-dire sa capacité à rendre universelles la visibilité et l'accessibilité de toutes les ressources, il importait d'utiliser autant que possible des métadonnées dont la sémantique puisse être universellement comprise et acceptée, ce qui est précisément l'objectif du DCMES. La contrepartie de ce choix (malheureusement incontournable) est l'usage de la langue anglaise pour les identifiants des éléments et des qualifiants...

Ces éléments comportent des qualifiants qui soit précisent le contenu de la métadonnée, soit proposent des schémas d'encodage c'est-à-dire des listes d'autorité ou des listes de vocabulaire recommandé ou des notations formelles. Dans la mesure du possible, et pour la même raison d'interopérabilité, les qualifiants retenus sont ceux du DC. Les éléments du DC

non retenus par le groupe, principalement dans un but de simplification, peuvent évidemment être utilisés sous réserve du respect des consignes d'utilisation du groupe DC.

#### *b) autres éléments*

A ces 12 métadonnées, le groupe en a ajouté 7 autres dont 5 pour décrire les ressources d'un point de vue plus particulièrement éducatif. A titre provisoire, cet ensemble de 7 métadonnées est baptisé GTME. Les identifiants et les schémas d'encodage proposés sont en français.

Comme cela a déjà été souligné, ce schéma est évolutif et en aucun cas définitif. Il est destiné à être modifié, puis progressivement enrichi par l'ajout de nouvelles métadonnées, de nouveaux qualifiants, de nouveaux schémas d'encodage. Dans cet esprit, le groupe n'a donc retenu que des éléments de métadonnées (y compris des qualifiants) dont l'utilité lui a paru indiscutable, même si leur présence n'est pas toujours obligatoire pour décrire telle ou telle ressource. De même, pour cette version initiale, le groupe a choisi de retenir un nombre très limité d'éléments de "métadonnées éducatives" à savoir 5, à comparer avec les 11 éléments du LOM (Learning Object Model) [LOM] auquel se réfèrent les projets ARIADNE [ARDN] et IMS, ou les 8 du GEM (Ministère de l'éducation, USA).

Dans le cadre de notre système de cache on a adopté le GTME1.1 [GTME 01b] dont on a retenu 13 éléments, modifié les propriétés de plusieurs éléments et ajouté 2 éléments pour obtenir 15 éléments de métadonnées pour décrire les documents pédagogiques. Ces éléments sont : Titre, Auteur, Mots-clés, Description (résumé), Date (de la dernière modification), Type de document, Format, Langue, Statut, Niveau, Discipline, Sous discipline, Nature, Media, Public.

### **5. Profil utilisateur**

Selon le lexique du projet MAIA [LexMAIA], un profil utilisateur est une collection d'informations sur l'utilisateur. Cette collection peut être vue comme un ensemble de caractéristiques avec des valeurs associées contenant par exemple ce que l'utilisateur préfère, ou ce qu'il est capable de faire sa profession (son niveau d'instruction...). On peut également

prendre en compte l'historique des actions de l'utilisateur, voir leur évolution dans le temps.

Il y a plusieurs façons d'obtenir ces caractéristiques en fonction du degré d'autonomie du système, de ses capacités d'observation et d'adaptation :

- Profils statiques : cette méthode nécessite que l'on dresse la liste des profils et que l'on décrive chacun d'entre-eux. Cette approche est statique, donc une fois que le système à démarré, il est difficile de mettre à jour les profils utilisateurs.
- Profils réflexifs : L'utilisateur doit remplir des formulaires pour configurer son propre profil. Cette approche permet plus de précision et d'adaptation.
- Profils corrigés dynamiquement : Un sous-système de modélisation observe l'utilisateur de derrière l'interface et apprend le profil de l'utilisateur à partir de ses actions.

Amerouali [AMRLI 01a] indique que la représentation de profil d'utilisateur doit être capable de s'adapter graduellement aux changements dans les intérêts réels de l'utilisateur.

Dans notre système nous utilisons les profils réflexifs et nous tendons à adopter les profils corrigés dynamiquement pour s'adapter aux nouvelles exigences des utilisateurs.

En faite, un Moteur de recherche se limitant à la requête formulé par l'utilisateur donne beaucoup de bruit (documents non pertinents). En ajoutant le profil utilisateur au processus de recherche, c'est-à-dire qu'on va prendre en compte les données, les centres d'intérêts et les préférences de l'utilisateur, le nombre de documents répondants aux besoins de l'utilisateur devient plus important. Donc l'intérêt d'associer le profil utilisateur au processus de recherche de l'information est d'améliorer nettement la pertinence des documents retrouvés.

Amerouali [AMRLI 01b] affirme aussi 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.

Ainsi, 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 et pour chaque sujet, la requête se lance en cliquant sur la catégorie correspondante.

Notre système de cache essaie d'associer complètement le profil de l'utilisateur à des éléments de description de ressources (les métadonnées). C'est l'objectif même de l'approche de filtrage par appariement entre métadonnées et profil utilisateur que nous expliquons dans le prochain paragraphe.

## **6. Proposition d'une approche de filtrage par appariement : métadonnées – profil utilisateur**

Avec la naissance de la Recherche d'Information (RI) et des Systèmes de Recherche d'Information (SRI), Salton [Salton 71], [Salton 83] et van Rijsbergen [vR 79] développent des modèles de RI sur lesquels sont basés les moteurs de recherche actuels du Web, autour du triplet : < document, besoin, correspondance >. Ce qui correspond à donner une importance à l'utilisateur (c'est lui qui formule ses besoins d'information) qui est l'acteur essentiel du système: il est la source, le déclencheur d'une recherche d'information et il valide le résultat de cette recherche.

Belkin [BEL 82] constate que l'utilisateur déclenche une recherche documentaire lorsqu'il est confronté à un manque dans sa connaissance sur un sujet. Wilson [WIL 81] considère que le besoin d'information est en fait un sous besoin de trois besoins fondamentaux: les besoins physiologiques (besoin de manger, boire, respirer, etc.), les besoins affectifs (parfois appelés besoins psychologiques ou émotionnels) et les besoins cognitifs (besoin d'apprendre, de structurer etc.). La satisfaction d'un besoin d'information correspond ainsi à la satisfaction des besoins fondamentaux.

L'identification du besoin d'information nous amène à nous pencher sur une modélisation de l'utilisateur lors d'une recherche d'information. Daniels [DAN 86] considère que la

modélisation de l'utilisateur, dans un contexte plus général, est constituée de cinq sous-fonctions:

USER: le statut de l'utilisateur,

UGOAL: détermine les buts de l'utilisateur,

KNOW: les états de connaissances de l'utilisateur dans un domaine,

IRS: la familiarité de l'utilisateur avec le système documentaire,

BACK: l'expérience de l'utilisateur.

Ces cinq sous fonctions appartiennent toutes aux systèmes cognitifs de l'utilisateur. Dans le cadre d'une recherche d'information, la prise en compte d'un profil contextuel lié à une recherche d'information ponctuelle sera un ensemble de valeurs de ces cinq sous fonctions.

Dans le projet SYFAX [SFX1], [SFX2], nous souhaitons modéliser l'utilisateur pour une prise en compte de ses besoins dans le domaine de la documentation pédagogique.

Ainsi on peut représenter un Système de Recherche d'Information (SRI) par ces trois éléments : l'utilisateur qui recherche de l'information, les ressources documentaires et un intermédiaire. Cette intermédiaire a pour fonction d'interpréter la requête de l'utilisateur en terme de documents pertinents. Actuellement, nous pouvons distinguer deux types de SRI: le SRI « traditionnel » et le SRI « évolué ». Le premier système est le système de référence fréquemment utilisé sur le réseau Internet (moteur de recherche altavista [ALT], google [GOGL] yahoo [YAH] ...), par les bibliothèques et la recherche sur base de données. Le deuxième type de systèmes se situe dans le cadre de notre système de cache en tant que prototype d'expérimentation. Ces systèmes proposent des améliorations notoires vis-à-vis du premier type de systèmes.

De nombreuses études démontrent l'importance de l'expertise de l'utilisateur dans un SRI. Nous considérons que l'expert du SRI souhaite avoir une interface élaborée lui permettant de construire lui-même une requête complexe notamment grâce à l'accès aux métadonnées des documents. Par contre, l'utilisateur non expérimenté sera démuni devant l'étendue des possibilités de constitution d'une requête. Nous pensons que les deux types d'utilisateurs

doivent être satisfaits. Le mécanisme de fonctionnement du SRI et de la structure du corpus documentaire : représentation des documents, types de métadonnées, stratégie d'indexation, mécanismes d'appariement entre requêtes et documents sont des concepts que l'utilisateur expert maîtrise parfaitement. Dans ce contexte, il est parfaitement légitime de penser que l'utilisateur a la capacité de diriger lui-même la recherche d'information. Pour cela une interface « ouverte », permettant l'accès à toutes les méta-informations est recommandée. Mais pour un utilisateur non expérimenté, cette recherche doit être assistée par un système qui fait la correspondance automatique entre ses besoins et les métadonnées des documents pour ne lui ramener que les documents qui lui intéressent, ce qui se traduit par l'intégration du profil utilisateur dans le processus de recherche. Dans cet article nous nous pencherons sur l'utilisateur non expérimenté et donc sur la conception d'une interface permettant la saisie des métadonnées du document et la prise en compte du profil de l'utilisateur (ce que nous appelons système d'appariement métadonnées - profil utilisateur). Le rôle de l'interface est de capter ce profil et de sélectionner en fonction de ce profil les documents pertinents pour l'utilisateur. Pour cela elle dispose de l'accès aux méta-informations du document : les métadonnées.

Pour atteindre notre objectif, nous adoptons la méthodologie suivante :

- 1- Analyser les besoins d'information de l'utilisateur afin de déterminer les caractéristiques de l'utilisateur.
- 2- En déduire les métadonnées utiles à associer aux documents.
- 3- Etablir une correspondance entre caractéristiques de l'utilisateur et métadonnées des documents.

Notre système de cache [CACH] est basé sur un corpus de documents pédagogiques. Nous ciblerons donc notre étude du profil de l'utilisateur dans ce contexte.

Sylvie Lainé-Cruzel [LAI 99] propose trois questions pour définir l'utilisateur :

**Qui est-il ?** Quels sont les caractéristiques cognitives, situationnelles, l'état de connaissances de l'utilisateur au moment de la recherche d'information ? Ce qui nous ramène à l'activité de l'utilisateur.

Dans le contexte pédagogique les professions sont : *étudiants, enseignants, chercheurs...*

**Que veut-il ?** Quelle information l'utilisateur souhaite obtenir ? Cette activité s'inscrit dans le cadre de la profession, ce qui correspond au type de documents souhaités : *exposé, cours, travaux pratiques, travaux dirigés...*

**Pour faire quoi (quelle information) ?** Quel est le but final de l'utilisation des documents sélectionnés ? Ce qui nous amène aux disciplines (exemple : *informatique*) des utilisateurs et éventuellement les sous-disciplines ou modules qui est constitué par un ensemble de matières. Nous nous servons de ces trois questions comme trois étapes successives pour déterminer les caractéristiques de l'utilisateur.

Nous avons dégagé de cette étude plusieurs caractéristiques de l'utilisateur. Nous pouvons en dénombrer trois :

1. Profession de l'utilisateur
2. Expertise de l'utilisateur dans le domaine (son niveau d'instruction)
3. But de la recherche de l'utilisateur (types des documents souhaités)

Nous pouvons remarquer que les trois caractéristique correspondent à trois sous fonctions de la modélisation de Daniels [DAN 86] : respectivement «USER» que l'on peut associé à la profession de l'utilisateur, « KNOW » pour l'expertise de l'utilisateur dans le domaine et «UGOAL» pour le but de la recherche.

On peut ainsi caractériser la correspondance entre les métadonnées et le profil utilisateur, nous utiliserons cette correspondance pour préciser la recherche documentaire sur métadonnées. Comment interpréter et utiliser ces informations dans le cadre de notre prototype ? Nous avons établi précédemment une liste récapitulative des caractéristiques des utilisateurs. Ces données (dégagés ci-dessus) nous permettrons de compléter les métadonnées des documents afin d'obtenir une liste étendue de métadonnées par document.

Ainsi, nous allons présenter une correspondance entre les éléments du profil utilisateur et les métadonnées :

- la profession de l'utilisateur correspond au public cible du document,

- le niveau d'instruction de l'utilisateur correspond au niveau du document,
- le but de la recherche (type de documents souhaités) : cette caractéristique est sans doute la plus riche en apport de d'information sur l'utilisateur, elle correspond à la discipline et éventuellement la sous-discipline (modules et matières) et au type de document (cours, exposé, travaux dirigés, travaux pratiques...).

Nous possédons à présent deux ensembles d'informations majeures : les métadonnées du document et la liste des caractéristiques potentielles décrivant les besoins de l'utilisateur. Nous pouvons concevoir deux méthodes basées sur deux types d'associations :

- Une association fixe, prédéfinie entre valeurs et métadonnées, représentée sous la forme d'un tableau binaire d'association. Par exemple la valeur type des documents souhaités coté profil utilisateur correspond à la sous discipline du document coté métadonnées. Nous parlons alors de modèle d'utilisateur statique (définie par Sparck Jones [DAN 86]).
- Une association variable, modulable entre profil d'utilisateur et métadonnées. Nous parlons alors de modèle d'utilisateur dynamique. Chaque valeur possible d'une caractéristique du profil utilisateur est une association fixée de métadonnées. Néanmoins, la précision de la valeur par une saisie progressive lors de la formulation des requêtes permet d'associer un nombre croissant de métadonnées. On peut aussi associer une ontologie pour chaque métadonnée pour une association plus complète aux éléments (caractéristiques) du profil utilisateur.

## **7. Architecture du système de cache proposé**

Le système de cache repose sur un modèle Client/Serveur dans lequel chaque utilisateur est un client du cache qui collabore avec d'autres clients via le Serveur du cache.

Un client est assimilable à un assistant personnel ; il tourne chez chaque utilisateur du système de cache. Le client du cache propose aux utilisateurs des interfaces de recherche, de gestion de documents ainsi que des outils de filtrage collaboratif. Le serveur du cache gère les documents et méta-données qui sont associées. Ils coopèrent pour offrir aux utilisateurs du

système l'abstraction d'un cache Web sémantique distribué performant aussi bien en terme de latence d'accès que de pertinence des résultats de recherche. Ils offrent aux utilisateurs des outils adaptés de recherche et de filtrage. Le filtrage se base essentiellement sur les avis des utilisateurs. En effet, chaque utilisateur peut donner son avis (aspect annotation) sur chaque document accédé. Les annotations des utilisateurs pour un document donné peuvent être consultées par les utilisateurs et faire même l'objet de nouvelles annotations.

Pour alimenter la base de données de documents du cache communautaire à partir du web (voir *figure 1*), le serveur du cache fait appel à un robot qui va parcourir le Web et ramener des documents, puis par une étape de préfiltrage on discrimine les documents non intéressants et qui ne correspondent pas au profil de la communauté (défini sous forme de caractéristiques à travers une méta-description de la communauté), ces documents présélectionnés vont passer par une étape d'analyse et d'extraction de métadonnées (automatique ou semi-automatique). Ainsi on obtient une base de données contenant des documents et leurs métadonnées respectives, cette base va être triée manuellement pour ne retenir qui intéresse réellement la communauté.



*Figure 1 : architecture du système filtrage des documents et du système l'extraction des métadonnées.*

Le système de cache communautaire s'articule au tour de cinq services :

- La prise en charge du profil utilisateur (saisie) lors de la phase d'inscription (voir *figure 3*).
- L'hébergement des document et l'extraction de métadonnées (voir *figure 4*).
- La recherche de documents par un système d'appariement profil utilisateur – métadonnées (voir *figure 5*).
- Recommandation à base du système d'appariement profil utilisateur – métadonnées.
- Outils communautaires (forums de discussion...)

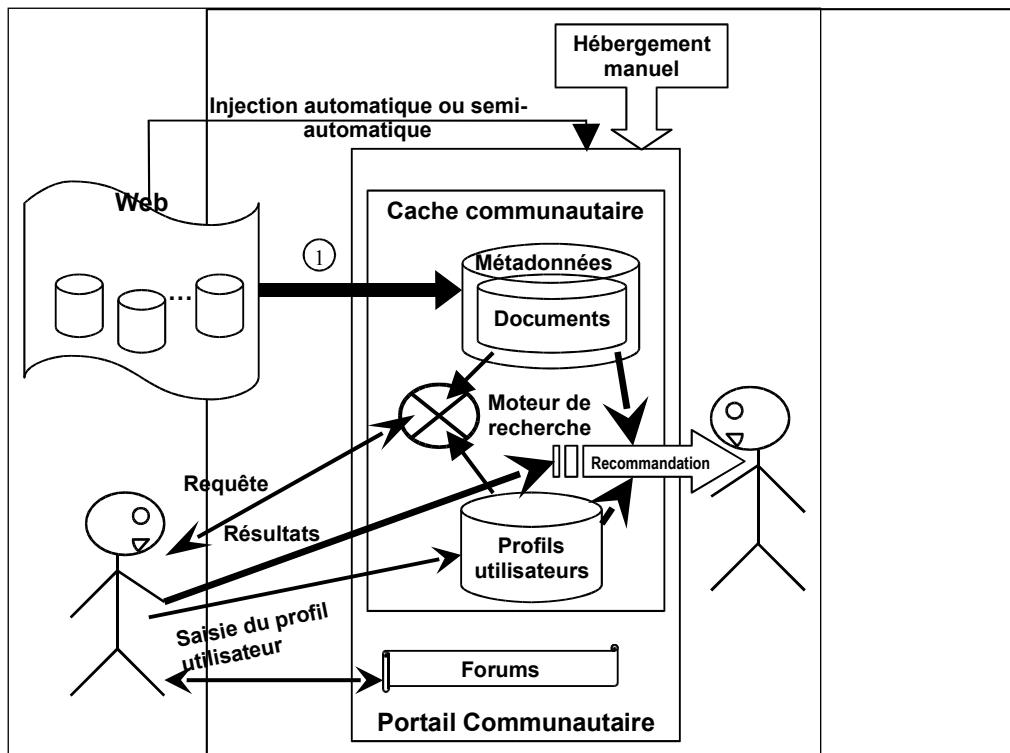


Figure 2 : architecture Globale du système de cache communautaire

Lors de la phase d'inscription au système, l'utilisateur doit saisir son profil utilisateur ainsi que ses centres d'intérêt, ensuite lorsque cet utilisateur se connecte au système de cache il a la possibilité d'héberger ses propres documents pour alimenter la base de données tout en ajoutant les métadonnées pour chaque document (l'auteur est la personne la plus qualifiée pour décrire son document). Cette étape de description peut s'effectuer de trois manières : manuelle (par l'auteur du document), automatique (à travers un système d'extraction automatique de métadonnées) ou semi-automatique (si après l'extraction automatique il manque des éléments, on va les compléter manuellement). La base de données de documents peut être alimentée automatiquement ou d'une manière semi-automatique à partir du Web (voir ① dans *figure 2*), cette étape est détaillée en dessus.

On dispose maintenant d'une base de ressources pédagogique que l'utilisateur peut consulter pour rechercher les documents qui lui intéressent à travers un moteur de recherche basé sur les métadonnées et éventuellement sur le profil utilisateur ainsi que ses centres d'intérêt. La recherche est multi-critères et s'effectue par la saisie des mots clés qui peuvent

être raffiné par la sélection d'autres critères basés sur les métadonnées, le profil utilisateur peut être activé (pour une recherche centrée sur le domaine et les centres d'intérêts de l'utilisateur) ou désactivé pour élargir le champs de la recherche. L'utilisateur peut aussi recevoir automatiquement les nouveaux documents hébergés dans le système de cache communautaire et recommandés par d'autres utilisateurs experts et qui correspondent à son profil utilisateur et ses centres d'intérêts à travers le système d'appariement profil utilisateur – métadonnées.

La communauté dispose aussi d'outils de communication et de collaborations comme les forums afin de permettre aux membres de la communauté d'échanger des documents et des idées sur des thèmes spécifiques pour consolider l'esprit de la communauté.

## **8. Implémentation**

Le prototype a été réalisé par EasyPHP v1.5 (Serveur Apache, langage PHP et Base de données MySQL) ce qui permet de l'héberger sur plusieurs plateformes. Le prototype a été implémenté sous forme de portail communautaire, la partie extraction des métadonnées est développée pour une saisie semi-automatique des métadonnées, le nombre de champs extraits automatiquement varie suivant le format du document (ex : l'extraction des métadonnées est plus facile pour les documents bien structurés comme les documents HTML ou XML et qui comporte des informations textuelles).

Nous sommes entrain de concevoir et réaliser un prototype de framework pour l'extraction automatique des métadonnées pour différents formats et types de documents utilisés dans le contexte pédagogique qu'on va intégrer dans notre système de cache communautaire SYFAX. Le prototype du cache subit actuellement des tests et validation afin de valider les résultats obtenus et éventuellement les améliorer.

Voici quelques exemples de services offerts par notre prototype (figures 3, 4, 5)



Figure 3 : inscription d'un nouveau membre



Figure 4 : extraction des métadonnées



Figure 5 : Moteur de Recherche des documents basé sur l'approche d'appariement entre métadonnées et le profil utilisateur



## 9. Conclusion et perspectives

Dans cet article nous avons présenté un système d'informations dédié à des communautés

d'utilisateurs données. Ceci est un des défis majeurs à relever pour faciliter et rendre plus efficace la recherche d'informations sur le Web. Cela favorise le partage et la collaboration entre les utilisateurs appartenant à une même communauté

Le système de cache propose à la fois une méthodologie et des outils pour construire et exploiter des systèmes d'informations dédiés à des communautés d'utilisateurs.

La méthodologie préconisée consiste à attacher des métadonnées aux informations partagées par la communauté. Le système de filtrage intégré dans le système de cache repose sur une approche de filtrage par un appariement entre métadonnées et profil utilisateur.

Le prototype ainsi développé nous a permis de remédier à certains problèmes des moteurs de recherche actuels, et peut constituer un noyau pour l'hébergement des documents pédagogique pour le service de la communauté pédagogique afin d'échanger les expériences et servir de support pour l'enseignement. Ce travail peut être étendu et amélioré afin de renforcer le concept de collaboration et de communication entre les membres d'une communauté. Ceci peut être réalisé par l'utilisation des protocoles de coopération comme le WebDAV [WEBDV], un des protocoles qui permet la création d'entreprises virtuelles, et qui est capable de servir une large gamme d'applications collaboratives.

#### **Références :**

[ALT] Altavista <http://www.av.com> .

[AMRLI 01] [Youcef Amerouali](#), Thèse de doctorat, Métadonnées basées sur l'association d'éléments de description de ressources et d'éléments de profil d'utilisateur, 2001, (a) p117, (b) p127-128.

[ARDN] ARIADNE Educational Metadata Recommendation, [http://www.ariadne-eu.org/3\\_MD/main.html](http://www.ariadne-eu.org/3_MD/main.html)

[BEL 82] BELDIN N.J., ODDY R.N., BROOKS H.M., Ask for Information retrieval : Part I Background and Theory. Journal of Documentation. Vol 38 n°2, 1982, p 61-71/

- [BrPl 03] BrightPlanet  
[http://www.brightplanet.com/deepcontent/deep\\_web\\_faq.asp#DeepWebSize](http://www.brightplanet.com/deepcontent/deep_web_faq.asp#DeepWebSize)  
 (consulté le 12 Décembre 2003).
- [CACH] H. SMEI, A. BEN HAMADOU, M. MAKPANGOU, Les caches Web sur Internet, GEI 2001 – Mars 2001 – Sousse – Tunisie
- [DAN 86] DANIELS J.P., « Cognitive Models in Information Retrieval – An Evaluation Review », *Journal of Documentation*, vol 42, n° 4, December 1986, p 272-304.
- [DCEd] Dublin Core Education <http://dublincore.org/groups/education/>
- [DUB 03a] Dublin Core Online Computer Library Center. Dublin Core Metadata Initiative. <http://purl.oclc.org/dc/>. (page consulté le 20 Novembre 2003)
- [DUB 03b] DCMI, <http://dublincore.org/resources/faq/>
- [ESclNet] European Schoolnet, <http://www.en.eun.org/>
- [GTME 01a] Groupe de Travail Métadonnées Education (GTME), CRDP de Montpellier Avril 2001, <http://www.ac-montpellier.fr/ressources/GTME1-0.doc>
- [GTME 01b] Groupe de Travail Métadonnées Education (GTME), CRDP de Montpellier Juin 2001, <http://www.ac-montpellier.fr/ressources/GTME1-1.doc>
- [GOGL] Google France <http://www.google.fr>
- [LAI 99] LAINE-CRUZEL Sylvie, « ProfilDoc, filtrer une information exploitable », *BBF*, T 44, n°5, 1999, p 60-64.
- [LexMAIA] *R.Charton*, *Lexique du projet MAIA : MACHine Intelligente Autonome, 2001*, [http://www.loria.fr/equipes/maia/lexique/profil\\_utilisateur.html](http://www.loria.fr/equipes/maia/lexique/profil_utilisateur.html)
- [LOM] Learning Technology Standards Comitee, <http://ltsc.ieee.org/>
- [METAM] How Big Is The Internet? How Fast Is The Internet Growing?  
<http://www.metamend.com/internet-growth.html>
- [Michard 99] A. Michard.- *XML langage et applications*.- Paris : Eyrolles, 1999.- 361p.

- [Salton 71] Salton (Gerald). – *The SMART retrieval system: experiments in automatic document processing*. – Prentice Hall, 1971.
- [Salton 83] Salton (Gerald) et McGill (Michael J.). – *Introduction to Modern Information Retrieval*. – McGraw-Hill, Janvier 1983.
- [SFX1] H. SMEI, M. MAKPANGOU, A. BEN HAMADOU, SYFAX : Un système de cache Web Sémantique pour des communautés Distribuées, Conférence Internationale – MediaNet 2002 – Juin 2002 – Sousse – Tunisie
- [SFX2] H. SMEI, M. MAKPANGOU, A. BEN HAMADOU, Vers un SYstème de Fouille Automatique et d'eXtraction d'information : SYFAX, Cas de l'enseignement. GEI 2002 – Mars 2002 – Hammamet – Tunisie
- [vR 79] van Rijsbergen (Cornelis Joost). – *Information Retrieval*. – Butterworths, London, Janvier 1979.
- [WIL 81] WILSON T.D., « On user studies and information needs », 1981, *Journal of Documentation*, vol 37, n 1, pp 3-15.
- [WeiL 97] Weibel et Lagoze, 1997 <http://www.sosig.ac.uk/>
- [YAH] Yahoo France <http://www.yahoo.fr>
- [WEBDV] WEBDAV [<http://www.webdav.org>]