CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

le cnam

Ecole Doctorale ABBE GREGOIRE

DICEN-IDF

THÈSE DE DOCTORAT

présentée par : **Manel ANKOUD**

soutenue le : **19 décembre 2014**

pour obtenir le grade de : **Docteur du Conservatoire National des Arts et Métiers**

Discipline / Spécialité : **Sciences de l'information et de la communication**

Gestion de l'hétérogénéité d'un SI de classification documentaire multifacette et positionnement dans l'environnement des ECM

THÈSE dirigée par

M. ZACKLAD Manuel — *Professeur, Cnam-Paris*

RAPPORTEURS

M. BEN YAHIA Sadok — *Professeur, Faculté des sciences de Tunis*

Mme. SALZANO Gabriella — *Maître de conférences HDR, Université Paris-EST Marne-La-Vallée*

EXAMINATEURS

M. BARKAOUI Kamel — *Professeur, Cnam-Paris*

A mes chers parents Dalila et Mehrez, pour leur amour, leur soutien et leur patience ...

A mes sœurs, mon frère et mes beaux-frères : Mejda, Marwa, Maha, Mohamed, Rached et Hamdi, pour leur encouragement ...

A mon cher Walid, pour son soutien et sa patience ...

A mon cher oncle Anouar et sa femme Fatma, pour leur soutien et leur encouragement ...

A mes petits anges Ayoub, Yassin, Zayneb et Maram, pour tout le bonheur qu'ils m'ont apporté dans les moments difficiles ...

A mes amis : Manel, Maya, Nouha et Houssem, pour leur encouragement ...

Remerciements

Au terme de ce travail, je tiens à présenter mes remerciements à Monsieur le président du jury Kamel Barkaoui, Professeur au CNAM, qui me fait l'honneur de présider le jury de ma thèse, qu'il trouve ici l'expression de mes vifs remerciements.

Je remercie également les membres du jury, Sadok Ben Yahiya, Professeur à la faculté des sciences de Tunis, et Madame Gabriella Salzano, Maître de conférences HDR à l'université Paris-EST Marne-La-Valée, qui ont bien voulu accepter d'évaluer le rapport de ma thèse, qu'ils trouvent ici l'expression de ma profonde gratitude.

Je voudrais aussi exprimer ma gratitude à mon directeur de thèse, Manuel Zacklad, Professeur au CNAM et directeur du laboratoire DICEN, pour ses conseils précieux et pour la disponibilité qu'il m'a toujours montrée.

Je remercie enfin tous les membres du DICEN-IDF pour leur soutien ainsi que tous ceux qui m'ont aidé dans l'accomplissement de ce travail.

A vous tous, merci.

Résumé

L'organisation des connaissances est une discipline investie par des bibliothécaires, documentalistes, archivistes spécialistes de l'information, informaticiens et tous professionnels du document. Elle englobe toutes activités, études et recherches qui élaborent et traitent les processus d'organisation et de présentation des ressources documentaires utiles dans une organisation. Dans ce contexte, le projet ANR Miipa-Doc a pour objectifs d'explorer de nouvelles méthodes d'indexation ascendantes, en utilisant des termes descripteurs formulés par les individus plutôt que choisis parmi une liste préétablie, pour l'organisation des contenus documentaires complexes au sein des entreprises de large taille, et concevoir l'architecture logicielle correspondante.

Dans ce projet, notre contribution consiste à gérer l'hétérogénéité d'un système d'information d'organisation des contenus documentaires, basé sur une approche orientée métier et un SOC (système d'organisation des connaissances) folksonomique à facettes. Nous proposons dans cette gestion une approche incrémentale dirigée par les modèles, issue de l'IDM (ingénierie dirigée par les modèles), basée sur des méta-modèles pour garantir l'aspect d'évolutivité. Après l'implémentation du prototype HyperTaging qui met en place ces deux approches, nous proposons un processus d'évaluation permet de positionner ce prototype et tous SI de classification documentaire dans l'environnement des ECM, en se basant sur des critères d'évaluation fins et particuliers.

Mots-clés : organisation des connaissances, indexation, classification documentaire, IDM, SOC, folksonomie à facettes, système d'information, processus d'évaluation, critères d'évaluation, ECM

Abstract

The knowledge organization is invested by librarians, archivists, information specialists, IT professionals and all discipline of document. It includes all activities, studies and research which develop and treat organization process and presentation of relevant information resources in an organization. In this context the Miipa-Doc project aims to explore new ascendants indexing methods, using descriptors made by individuals rather than selected given list for complex contained in the organization document, in large size companies, and design the corresponding software architecture.

Our contribution in this project is to manage the heterogeneity of an information system of document organization, based on a business-oriented approach and a KOS (knowledge organization system) of folksonomy facet. We propose an incremental approach this management model driven, outcome of MDE (Model Driven Engineering), based on metamodels to ensure scalability appearance. After implementing the HyperTaging prototype, that implements both approaches, we propose an evaluation process used to position the prototype and all IS of documentary classification in the environment of ECM based on purposes of delicate and particular evaluation criteria.

Keywords : knowledge organization, indexing, document classification, MDE, KOS, folksonomy facet, information system, evaluation process, evaluation criteria, ECM

Table des matières

Liste des tableaux

Table des figures

Introduction

Contexte de recherche Le traitement des informations sur des supports numériques constitue aujourd'hui un enjeu essentiel. Il s'agit de repenser les démarches de gestion de connaissances afin de pouvoir d'une part en assurer une meilleure conservation et d'autre part en faciliter l'accès pour en permettre une meilleure exploitation.
Les ressources informationnelles et documentaires sont stockées, classées, organisées, publiées et partagées au sein d'espaces et de systèmes informationnels dispersés et distribués. En effet, après l'organisation documentaire manuelle, le classement traditionnel et l'indexation classique, aujourd'hui la majorité des plateformes et environnements de gestion et d'organisation des documents intègrent un système d'organisation des connaissances (SOC).
Dans des environnements technologiques de plus en plus ouverts et générateurs d'informations, les activités d'organisation des connaissances nécessitent d'accéder à des volumes croissants de données, produites et gérées par des écosystèmes numériques hétérogènes. Sous l'impulsion du web 2.0 au sein même des entreprises, ces écosystèmes englobent de plus en plus les documents produits, échangés, partagés parmi les membres de collectifs de diverses tailles (équipes, départements, directions). Afin de tirer profit de ces nouvelles possibilités et en même temps gérer cette complexité croissante, les systèmes d'organisation des connaissances (SOC) « englobent tous les types de schémas permettant d'organiser des informations et de promouvoir la gestion des connaissances » [HUDON et HEDI 2010].
Les SOC doivent faire face à la croissance et aux évolutions rapides des connaissances, législations et régulations ainsi que des processus de création et de diffusion de ressources informationnelles accessibles par le Web. Au sein des entreprises, les SOC sont confrontés à l'évolution des métiers et à la diversification des usages, dont les notions préliminaires

sont le travail collaboratif et participatif, l'échange informationnel et la communication (le partage) interne et externe. Ils s'efforcent de tirer profit des normes, standards technologiques et des initiatives autour des métadonnées et du Web socio-sémantique. Les SOC d'entreprises opérant en santé environnementale ou dans la prévention de risques naturels et technologiques doivent par exemple supporter de multiples points de vue (veille scientifique, stratégique et territoriale) et plusieurs normes et standards sectoriels (santé, géographie, ...), pour favoriser la génération de nouvelles connaissances.

Le projet Miipa-Doc Dans ce contexte, le projet Miipa-Doc [1] s'inscrit dans le domaine de l'indexation de ressources documentaires. Il a pour objectifs d'explorer des nouvelles méthodes d'indexation ascendantes, en utilisant des termes descripteurs formulés par les individus plutôt que choisis parmi une liste préétablie, pour l'organisation des contenus documentaires complexes au sein des entreprises de large taille, et concevoir l'architecture logicielle correspondante. Ce projet entend étudier les spécificités des procédures et des activités de travail des utilisateurs ainsi que les pratiques individuelles de classement.
Il permet aussi de concevoir une méthode de classification multidimensionnelle stable en proposant un accès unifié à l'ensemble de ressources documentaires et informationnelles de l'entreprise ; dispersées dans différents espaces informationnels, entre les postes de travail individuels, les disques partagés, le Cloud, les GED [2], etc.
Miipa-Doc part de l'idée selon laquelle la combinaison de différentes techniques d'indexation et l'enrichissement progressif des index qui en résulterait, amélioraient la pertinence des réponses proposées par le système de recherche d'information. Plus on ajoute de renseignements, d'entrées dans les index, plus on pourra finement répondre aux requêtes des utilisateurs. Cela permettrait d'enrichir la description des ressources informationnelles et documentaires, à partir des contextes d'activités dans lesquels celles-ci sont produites, reçues, partagées et utilisées, telle est l'approche portée par ce projet.
Cette approche de l'organisation des connaissances s'inscrit dans la continuité des réflexions développées par le Web dit socio-sémantique, qui favorise la complémentarité

1. Méthodes et Services Intégrés Institutionnels et Participatifs pour la Classification à Facettes des Contenus Documentaires Complexes
2. Outils de Gestion Electronique des Documents

entre la sémantique référentielle et la sémantique interprétative dans un environnement participatif.

Contributions Nos contributions sont présentées dans la figure 1 sous la forme de deux approches :

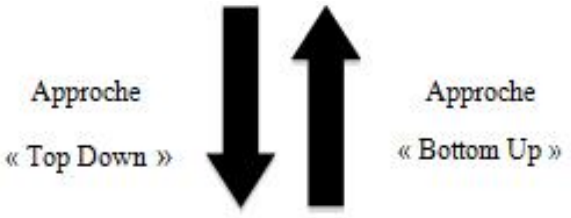

FIGURE 1 – Nos contributions

— Une approche « Top Down » qui cherche à faciliter la gestion des hétérogénéités d'un système de classification documentaire, basé sur une approche orientée métiers et un système d'organisation des connaissances du type « folksonomie à facettes », en allant du processus métier jusqu'à son infrastructure ;
— Une approche « Bottom up » qui cherche à positionner ce système dans l'environnement des ECM par un processus d'évaluation.

Au sein du projet Miipa-Doc, notre travail de recherche s'inscrit dans l'approche disciplinaire de la gestion des connaissances qui représente l'ensemble des initiatives, méthodes, et techniques. Ces derniers permettent de percevoir, identifier, analyser, organiser, mémoriser et partager des connaissances entre les membres d'une organisation en permettant aux acteurs de ne pas se limiter à la consommation des informations brutes mais aussi de veiller [3]leurs usages.
Dans ce cadre, les systèmes d'information constituent un élément central pour tout type de stratégie de gestion des connaissances. En effet, nous nous intéressons à la gestion de l'hétérogénéité entre les différentes préoccupations d'un SI de classification documentaire. Il intègre une démarche d'indexation multidimensionnelle (multifacette) qui s'appuie sur l'organisation des activités de la classification documentaire, et sur une approche parti-

3. Interprétation, structuration, capitalisation et partage

cipative exploitant la dimension subjective et interprétative de l'acteur en « contextes ». Cette approche orientée métiers suppose l'implication des acteurs dans le processus de caractérisation des sources d'information et des documents qu'ils créent, utilisent et sont amenés à classer.

Pour cette gestion, nous utilisons une approche incrémentale et dirigée par les modèles en spécifiant les liens sémantiques. Elle s'inscrit dans l'ingénierie dirigée par les modèles (IDM) qui vise à systématiser le développement logiciel et permet d'intégrer au sein d'un même environnement des techniques modernes et une grande variété de langages de modélisation. Elle offre au génie logiciel un cadre conceptuel homogène pour la formalisation des langages et de leurs relations, ainsi que des techniques de réutilisation pour la conception logicielle. Nous adoptons le point de vue technologique de l'IDM pour penser, d'une manière incrémentale et itérative, les concepts et les préoccupations de l'interaction homme machine, l'évolution du système d'organisation des connaissances, la tendance architecturale et l'interopérabilité avec d'autres systèmes de classification documentaire.

Dans le cadre de cette recherche, le prototype HyperTaging est développé en se basant sur l'approche méthodologique intégrée du SOC, orientée métiers. Il permet de classer, organiser, indexer, rechercher des documents en utilisant l'approche à facettes et le tagging, avec des possibilités de personnalisation et selon divers modes de collaboration (dispositif partagé ou personnel). Ces fonctions s'appuient sur les éléments : vue, facette, tag et document.

La gestion de contenu nécessite une convergence entre la gestion des systèmes d'information et la gestion de l'information, qui est assurée par les ECM [4] en construisant la liaison entre les contenus métier de l'entreprise et les fonctions techniques disponibles. En effet, après avoir franchi la méthodologie conceptuelle et la mise en œuvre de notre SI de classification documentaire, une nécessité d'assurer sa pertinence et sa position stratégique dans les environnements des ECM se pose. Pour répondre à cette problématique, nous établissons une stratégie d'évaluation du point de vue de la classification documentaire. Elle est fondée sur un modèle à base des critères représentant les préoccupations classificatoires documentaires. L'originalité de cette contribution se manifeste dans le processus d'évalua-

4. Enterprise Content Management : outil de gestion des contenus dans les entreprises

tion, notamment la détermination des critères fins sur le SOC et le document traité dans ce type du système. Dans l'environnement ECM, nous nous concentrons sur les fonctionnalités documentaires comme les règles d'organisation, de classement et de stockage, ainsi que le cycle de vie du document.

Organisation du manuscrit Ce manuscrit est structuré en quatre chapitres. Le premier chapitre fournit un état de l'art sur le domaine d'organisation des connaissances dans lequel s'inscrit le projet Miipa-Doc ainsi que nos travaux. Le deuxième chapitre présente les problématiques abordées, nos objectifs et les méthodologies à suivre dans nos contributions.
Dans le troisième chapitre, nous montrons comment notre approche conceptuelle incrémentale permet de gérer les hétérogénéités du SI de classification documentaire. Pour finir, le quatrième chapitre détaille le processus et les critères d'évaluation de ce type du SI dans un cadre particulier : l'environnement des ECM. Nous présentons les perspectives de recherche envisagées en conclusion.

Les résultats présentés dans ce rapport s'appuient en partie sur le travail collectif de développement des partenaires du projet ANR dont nous avons été partie-prenante. C'est le cas en particulier pour le modèle de facettes, les interfaces homme-machine et les exemples d'application. Nos contributions spécifiques au sein du projet ont donné lieu à plusieurs publications citées en référence 4.5.

Première partie

Etat de l'art

Chapitre 1

De la connaissance qui s'organise !

1.1 Introduction

L'organisation des connaissances (OC) est une problématique liée aux domaines de l'entreprise, au Web et aux bibliothèques. Les communautés de l'ingénierie des connaissances, la science d'information et de la documentation, la gestion des connaissances et l'informatique proposent différentes méthodologies, des modélisations et des systèmes d'informations pour organiser les contenus et les documents selon les besoins du domaine d'application. Dans ce chapitre, nous présentons les notions liées à l'OC, nous commençons par le domaine de l'ingénierie des connaissances et sa liaison avec l'OC, la notion de la connaissance et des systèmes d'organisation des connaissances (SOC). Nous abordons aussi la classification en tant que méthode d'organisation, ainsi que la classification documentaire et les structures organisationnelles utilisées dans la bibliothèque et les systèmes de gestion des documents codifiés (comme les environnements de développement informatique et les systèmes de gestion de base des données). A la fin de ce chapitre, nous résumons les différences entre les structures d'organisation et leurs différents utilisateurs.

1.2 Ingénierie des connaissances

Le domaine de l'ingénierie des connaissances est apparu à la fin des années 80 comme une branche du domaine de l'intelligence artificielle pour faciliter à l'être humain la réalisation des tâches qui mettent en œuvre des connaissances. Dans ce domaine, les connais-

sances linguistiques et cognitives sont représentées dans des formes adéquates et adaptés à l'exploitation par des systèmes informatiques. En effet, il regroupe différentes réflexions :

- Linguistique : la formalisation linguistique des connaissances ;
- Terminologique : la conception des connaissances ;
- Psychologique : l'élaboration des méthodes de mise en valeur des connaissances ;
- Logique : l'élaboration des modèles formels ;
- Informatique : l'opérationnalisation des modèles ;
- Sémiotique : l'interprétation de comportements des systèmes [CHALET et al. 2004].

L'ingénierie des connaissances est définie comme la discipline qui correspond à « *l'étude des concepts, méthodes et techniques permettant de modéliser et/ou d'acquérir les connaissances pour des systèmes réalisant ou aidant des humains à réaliser des tâches se formalise a priori peu ou pas* » [BACHIMONT [2000]]. Elle représente l'embranchement de plusieurs domaines comme l'informatique, la linguistique, la logique et la psychologie.
L'objectif de ce domaine est de construire des artéfacts permettant l'organisation des connaissances et l'exercice de la pensée. En effet, les chercheurs et les professionnels de cette discipline proposent des méthodes, des outils, des instruments et des systèmes informatiques pour l'acquisition, la modélisation, la gestion et la capitalisation des connaissances dans différents domaines de recherche liés à l'entité « connaissance » comme l'acquisition des connaissances à partir des corpus de textes, l'organisation des connaissances dans l'entreprise, l'ingénierie des documents, la recherche d'information sur le Web, etc. [BACHIMONT 2000][BACHIMONT 2004][CHALET et al. 2004]
Dans son livre « Ingénierie des connaissances et des contenus », Bachimont considère l'ingénierie des connaissances comme une ingénierie des inscriptions numériques des connaissances ayant comme objectif : permettre l'organisation des connaissances, leur interprétation et leur manipulation.
Il aborde deux tendances distinctes :

- L'ingénierie de représentation : qui vise « *à formaliser le sens des inscriptions pour instrumenter leur exploitation* » ;
- L'ingénierie des contenus : qui vise « *à formaliser la forme d'expression des inscriptions et non leur sens, pour obtenir des outils et méthodes permettant de les*

manipuler et transformer ». [BACHIMONT 2007]

1.3 Connaissance, son organisation et les systèmes d'organisation des connaissances

1.3.1 Connaissance

La connaissance ne représente pas une discipline à part mais l'objet d'étude de plusieurs réflexions comme la modélisation des connaissances, l'organisation des connaissances, l'acquisition des connaissances, etc. Elle représente le savoir faire et la capacité de réalisation d'une action du corps, de l'esprit ou d'un corps complexe comme la société, qui sont transmis à travers des informations inscrites sur des supports numériques et technique, ou bien par la cognition et l'apprentissage (passage oral de l'expérience, mémoire, raisonnement). La représentation de la connaissance dépend du domaine d'exploitation de cette entité, il existe deux modes d'inscription : une représentation formelle avec des langages formels comme les ontologies et une représentation documentaire avec des langages documentaires comme les facettes et les thésaurus. [BACHIMONT 2004]

1.3.2 Organisation des connaissances

D'après Hjorland , il existe deux sens de l'organisation des connaissances. Le premier est le sens large, dans lequel l'organisation des connaissances est liée à l'organisation sociale du travail intellectuel comme l'organisation des universités et d'autres institutions de recherche et d'éducation, l'organisation sociale des médias, la production et la propagation du savoir, la structure des disciplines et des métiers.
Dans le deuxième, sens restreint l'organisation des connaissances évoque des activités comme la description de documents, indexation et la classification [HJORLAND 2008].
En effet, cette discipline est investie par des bibliothécaires, documentalistes, archivistes spécialistes de l'information, informaticiens et tous professionnels du document. Elle englobe toutes activités, études et recherches qui élaborent et traitent les processus d'organisation et de présentation des ressources documentaires utiles dans une organisation. [HJORLAND 2008][HUDON et HEDI 2010]

Ce domaine s'étend vers d'autres réflexions dans la linguistique, la sociologie et l'informatique notamment qui permettent de concrétiser ce concept par la gestion et la diffusion de l'information numérique inscrite sur un support matériel.
D'après Cotte la notion d'organisation des connaissances « *se préoccupe de recenser, qualifier, classifier, organiser, partager le volume global de connaissances que le salariés produisent, brassent, manipulent* » [COTTE 2007]. Il est communément admis que les connaissances dans des différents domaines de compétences représentent un capital pour l'organisation, une ressource à gérer comme les ressources matérielles et personnelles. En revanche, il faut mettre en place des méthodes et démarches spécifiques de gestion, appelées démarches de gestion des connaissances (GC) ou knowlege management (KM), qui mettent en œuvre des systèmes d'organisation des connaissances variés [DUDEZERT 2012].

1.3.3 Les Systèmes d'Organisation des Connaissances (SOC)

La problématique de SOC s'inscrit à la fois dans le domaine de sciences d'information et celui d'ingénierie des connaissances. Le terme Système d'Organisation des Connaissances (SOC) ou Knowledge Organisation System (KOS) en anglais vise à regrouper dans une « *dénomination unique aussi bien les langages documentaires, les schémas de classification que les langages de représentation des connaissances issus de l'Intelligence Artificielle* » [ZACKLAD 2011b]. Il s'intègre dans les plateformes de GED, les plateformes de Web 2.0, les plateformes basées sur le Web Sémantique (comme les moteurs de recherche) et les systèmes de classement et d'indexation documentaire. Les SOC peuvent être différenciés selon plusieurs typologies, nous choisissons celle du degré de formalité proposée par ZACKLAD [2011b]. Il classe les familles des SOC du plus au moins formel :

- Les langages documentaires et les thésaurus ;
- Les ontologies formelles et le web sémantique ;
- Les classifications épistémiques universelles de la bibliothéconomie et les approches à facettes universelles ;
- Les approches multidimensionnelles : les ontologies sémiotiques (web socio-sémantique) et les approches à facettes locales ;
- Les annuaires de ressources internet collaboratifs et les folksonomies ;

— Les indexes automatiques des moteurs des recherches. [ZACKLAD 2011b]

1.3.3.1 Les Langages documentaires et les thésaurus

Le langage documentaire est un système de signes qui permet de représenter le contenu intellectuel d'un document. Il existe deux formes : les langages alphabétiques d'indexation avec un vocabulaire contrôlé et les langages numériques de classification avec les indices. Ils visent à remplacer le langage naturel pour éviter les ambiguïtés et servent à catégoriser, indexer, classer, faire une recherche documentaire et faciliter la sélection et la consultation des documents. [ZACKLAD 2011b]

Les thésaurus, nés dans les années 50, sont définie par la norme internationale ISO 2788 de l'année 1986[5] , comme suit : les thésaurus représentent le « *vocabulaire d'un langage d'indexation contrôlé organisé formellement de façon à expliciter les relations a priori entre les notions (par exemple relation générique spécifique)* » et le langage d'indexation est un « *ensemble contrôlé de termes choisis dans une langue naturelle et utilisés pour représenter sous forme condensée, le contenu des documents* » [SAADANI et BERTRAND-GASTALDY 2000]. Le thésaurus est l'ensemble des termes lexiques et des définitions des structures classificatoires exprimées par des relations sémantiques (relation d'équivalence intra linguistique (synonymie), relation d'équivalence inter linguistique (traduction), relation hiérarchique, relation d'association). [ZACKLAD 2011b]

1.3.3.2 Les ontologies formelles et le web sémantique

L'ontologie, apparue au début des années 90, est une approche de modélisation des connaissances qui vise à décrire des concepts et des représentations mentales plus ou moins universelles. Dans le contexte de l'ingénierie des connaissances, les ontologies s'inscrivent dans plusieurs projets comme la représentation des connaissances, les réseaux sémantiques, les cartes conceptuelles, les graphes conceptuels et le développement du Web Sémantique [GUARINO 1997]. Ce dernier est proposé par Tim Berners Lee en 1994, au sein de lequel les ontologies jouent un rôle essentiel dans la recherche et la mise en relation des infor-

5. ISO, Organisation internationale de normalisation. Documentation - Principes directeurs pour l'établissement et le développement de thésaurus monolingues : norme internationale ISO 2788, 1986.

mations en utilisant le langage formel OWL (Ontology Web Language) inspiré de RDF (Ressource Description Framework), comme langage de présentation, et la syntaxe XML (Extensible Markup Language), comme langage de spécification.
Contrairement à la classification documentaire et aux thésaurus, l'ontologie formelle n'est pas exploitable directement par les utilisateurs, elle est utilisée par des programmes informatiques comme les agents de recherches automatiques sur le Web et l'interrogation se fait à travers un formulaire ou un autre type de langage de requête. Elle manipule des données structurées en répondant à des requêtes sur la base d'une sémantique formelle. D'après Guarino, elle est une théorie de distinction formelle entre les éléments d'un domaine indépendamment de leurs réalités.[GUARINO 1997]
Il existe différents niveaux d'ontologies selon le degré de généricité :

- Les ontologies supérieures représentent des concepts généraux comme l'espace, le temps ou la matière ;
- Les ontologies générales représentent des concepts moins généraux mais ne sont pas propres à un domaine. Leur précision est moyenne ;
- Les ontologies de domaine ont un faisceau plus large, une bonne précision et ne sont pas propres à une tâche particulière ;
- Les ontologies d'application ont un domaine de validité restreint et correspondent à l'exécution d'une tâche.[CHANDRASEKARAN et al. 1999][BURGUN et BODENREIDER 2001]

1.3.3.3 Les classifications épistémiques universelles de la bibliothéconomie et les approches à facettes universelles

Dans le domaine de la bibliothéconomie, les schémas de classification les plus répandus pour le classement documentaire au sein des bibliothèques sont la classification décimale de M. Dewey et la classification décimale universelle de P. Otlet et H. La Fontaine. En se basant sur une approche épistémologique, ils visent à classer le savoir humain selon une division hiérarchique. L'objectif de ces schémas est à la fois de faciliter le rangement des exemplaires physiques et de fournir une organisation systématique des ouvrages permettant « *au chercheur de repérer des documents pertinents qu'il ne connaît pas encore* » [HUDON

2001][ZACKLAD 2011b].

L'approche de S. R. Ranganat, qui sera présentée plus en détails dans d'autres sections de ce chapitre, correspond à une méthodologie à facettes universelles dont le jeu de facettes est stable et universel. Elle est basée sur le principe d'addition de facettes classificatoires, qui sont définies comme un « *ensemble des divisions obtenu par l'application d'une seule caractéristique* » [SALVAN 1962], et la définition d'un vocabulaire pour chaque domaine universellement accepté qui sera utilisé pour le rangement des livres [MANIEZ 1999].

1.3.3.4 Les approches multidimensionnelles : les ontologies sémiotiques (web socio-sémantique) et les approches à facettes locales

Contrairement aux approches à facettes universelles, les approches à facettes locales considèrent que chaque domaine est associé à un jeu des facettes spécifique. Elles se basent comme les approches de web socio-sémantique (les ontologies sémiotiques) sur la notion des points de vue et sur plusieurs dimensions pour faciliter et enrichir le travail de la classification et l'indexation. Le web socio-sémantique s'inscrit dans le projet de structuration des ressources documentaires partagées sur le web pour faciliter les activités coopératives distribuées. Il s'adresse à des communautés d'utilisateurs poursuivant des objectifs similaires en représentant une approche qui combine le Web Social et le Web sémantique et maintient la structuration progressive des réseaux cognitifs et sociaux. Ce Web introduit un format méta-sémiotique pour représenter l'information permettant de construire et partager les ontologies sémiotiques. Ces dernières sont « *des productions sémiotiques cohérentes qui regroupent des expressions stéréotypées extraites des transactions communicationnelles et organisées selon des axes paradigmatique et syntagmatique* » [ZACKLAD 2005a].

1.3.3.5 Les annuaires de ressources internet collaboratifs et les folksonomies

Ils mettent en commun des applications d'indexeurs coopératifs du web en s'approchant au principe du web socio-sémantique. Les annuaires de ressources sont définis comme l'un des schémas de classification bibliothéconomie. Les annuaires collaboratifs sont ceux qui sont pris en charge par des communautés de bénévoles comme l'annuaire libre de dmoz [6] qui adopte le référencement immédiat. Ce dernier est différent de celui de Yahoo qui est

6. http ://dmoz.org/, visité en juin 2013

défini par les managers en appliquant une politique de référencement explicite et un processus d'indexation social contrôlé.
Les folksonomies représentent le modèle de l'usage participatif du Web 2.0, ils sont basés sur des dispositifs informatiques en donnant « *la possibilité à l'usager d'indexer des documents afin qu'il puisse plus aisément les retrouver grâce à un système de mots-clés* » [DEUFF 2007]. Comme sur la plateforme Del.icio.us[7], chaque utilisateur peut ajouter des ressources, des marques de pages ou des photos en les associant à des mots-clefs et les partager avec d'autres utilisateurs malgré la faible cohérence des descripteurs (synonymie, polysémie, non explicitation des facettes prises en compte, absence de relation sémantique....) [ZACKLAD 2005a].

1.3.3.6 Les index automatiques des moteurs des recherches

Les index des moteurs représentent un exemple de SOC automatique qui est généralement invisible pour les utilisateurs mais parmi les plus sollicités. Par exemple pour les documents indexés sur le web, seuls les premiers résultats d'une requête qui sont généralement consultés. Ils sont construits par les moteurs de recherche à travers le traitement spécifique des métadonnées et les opérations successives permettant d'associer à chaque document d'un corpus considéré une liste de termes : lemmatisationL'analyse lexicale du contenu d'un document, filtrage des termes significatifs à l'aide d'anti-index, pondération selon leur pertinence à l'aide d'algorithmes statistiques mettant en relation la fréquence du terme dans un document avec sa fréquence dans le corpus [ZACKLAD 2005a].

1.4 La classification au sens général

« *Le processus d'organisation suppose presque toujours une opération de classification* » [HUDON et HEDI 2010]. Cependant cette dernière représente une méthode d'organisation parmi d'autres, une opération raffinée de « *segmentation de la réalité, de mise en ordre et de production de cadres de référence utiles* », consiste à organiser des entités sous la forme de classe. Chaque classe rassemble les entités ayant des caractéristiques communes. Les classes sont qualifiées des plus générales aux plus spéciales selon le nombre des caractéristiques

7. http ://delicious.com/, visité en juin 2013

essentielles et communes entre les entités de la classe. Plus le nombre des caractéristiques communes est important plus le nombre des entités est limité.
Le processus de classification s'élabore en deux principes fondamentaux :

— Le regroupement : le rapprochement des entités ayant au moins une caractéristique commune représente la caractéristique de division ;
— Le classement : la mise en séquences des groupes ainsi constitués. [HUDON et HEDI 2010]

Dans la littérature, il existe d'autres définitions de la classification. Hjorland la considère comme une opération d'organisation des objets selon leurs renseignements descriptifs dans des classes [HJORLAND 2007] et Vickery la définit par le rangement des entités dans une structure des classes qui est basée initialement sur la relation générique/spécifique et représente les disciplines enseignées [VICKERY 2008].

1.5 La classification documentaire

Comme son nom l'indique, la classification documentaire est la classification des documents. Les documents représentent les entités à regrouper et l'opération de classification consiste à regrouper les documents ayant une caractéristique commune comme le sujet, l'auteur ou le format en une seule classe.
C'est à partir du XVIe siècle que la classification des documents par leurs sujets a commencé à s'implémenter progressivement et le développement des systèmes d'organisation est devenu possible après l'invention de l'imprimerie et la multiplication des ouvrages à organiser par collections et sous-collections.
Une multitude de systèmes de classification ont été développés au début du XIXe siècle suite à l'apport de Melvil Dewey qui simplifie et systématise le rangement des documents. En effet, il a implémenté une classification documentaire appelée «relative », dans laquelle chaque document classé dans une collection a une relation pas seulement avec son successeur et son prédécesseur mais avec tous les documents de la collection. Il a développé par la suite l'idée de William T. Harris qui consiste à classer le document par son contenu, c'est-à-dire à classer non pas l'objet imprimé mais les informations et les connaissances

qu'il transporte. L'indice de la classification est élaboré de telle manière qu'il reflète à la fois la localisation physique exacte du document et l'importance et la place de son sujet dans l'univers des connaissances. [8]

La classification des documents sur un modèle de structure d'organisation des savoirs avait deux objectifs :

- « *l'éducation des masses par l'utilisation d'une carte de navigation des connaissances accumulées au fil des temps* » ;
- « *le repérage et la localisation de documents susceptibles de fournir la réponse à une question précise* ». [HUDON et HEDI 2010]

Le deuxième représentait l'objectif le plus difficile à atteindre. Aujourd'hui, la localisation de documents s'oriente vers les opérations des recherches effectuées sur l'information, cette action représente une constituante fondamentale de la science de l'information.

Au début du XXe siècle, pour répondre aux évolutions scientifiques et sociales et assurer des structures classificatoires plus profondes et satisfaisantes aux besoins plus précis, les classificateurs s'orientaient vers les documents eux-mêmes au lieux des modèles d'organisation des connaissances existants pour identifier des nouveaux sujets et les placer par la suite dans la structure classificatoire en usage. A partir de cette époque, le sujet est considéré comme un élément rattaché à une classe de la structure classificatoire et il représente un attribut du document comme son format ou le titre de la collection à laquelle il appartient. [HUDON et HEDI 2010]

1.5.1 Les schémas de classification documentaire

L'opération de classification des documents s'effectue en se basant sur une structure classificatoire appelée schéma de classification documentaire. Il est défini comme « *un inventaire structuré de sujets organisés logiquement en domaines, sous-domaines, disciplines, sous-disciplines, et ainsi de suite. Cette organisation logique se reflète dans une notation plus ou moins signifiante, laquelle, en plus de représenter un sujet traité de façon générale ou spécifique, indique également dans quel contexte et dans quelle perspective il a*

8. [HUDON et HEDI 2010, p. 14-16] synthèse de l'histoire de la classification documentaire, des systèmes élaborés et des principes qu'ils les sous-tendaient

été traité ». [HUDON et HEDI 2010]

D'après HUDON et HEDI [2010] l'objectif d'un schéma classificatoire documentaire est d'assurer les fonctions suivantes :

— une fonction cognitive : organiser des concepts, clarifier des idées et positionner des sujets ;
— une fonction bibliographique : systématiser les représentations documentaires et harmoniser les opérations relatives au traitement bibliographique ;
— une fonction bibliothéconomique : organiser les objets documentaires et ranger physiquement les documents.

Pour garantir son fonctionnement, un schéma classificatoire documentaire doit être :

— complet, « *c'est-à-dire qu'il intègre tous les sujets qui ont été abordés ou pourraient éventuellement être abordés dans un ou plusieurs documents* » ;
— flexible, « *c'est-à-dire qu'il s'ajuste aux besoins et aux habitudes de recherche des utilisateurs, au volume des collections à classifier, à la rapidité d'évolution du domaine d'activités représenté, aux fonctionnalités informatiques disponibles, etc.* » ;
— hospitalier, « *c'est-à-dire qu'il autorise l'intégration à la structure existante, sans qu'il soit nécessaire de la modifier ou de la repenser, des sujets nouveaux qu'engendrent les développements disciplinaires et l'évolution des savoirs* ». [HUDON et HEDI 2010]

Le schéma de la classification documentaire normalise cette classification, il définie les terminologies et les concepts liés à cette opération, ce qui facilite l'utilisation d'un même schéma classificatoire dans différentes organisations et institutions, par conséquent l'accès aux documents s'en trouve facilité. Un nombre des schémas de la classification documentaire est construit sur la base d'une structure hiérarchique, unidimensionnel et peu flexible d'où la complexité d'intégration et de gestion des nouveaux domaines.

1.5.2 La structure hiérarchique

D'après Maniez, la hiérarchie est une forme de mise en ordre logique des composantes de l'univers des connaissances et des ressources documentaires, favorisée par la majorité

des systèmes d'organisation des connaissances [MANIEZ 2002]. Ce modèle de structure de la classification représente une organisation utile et efficace, dans la mesure où il fournit plusieurs renseignements sur les entités et les classes à travers ses localisations dans la structure et ses relations avec l'entourage (par exemple la hiérarchie suivante fournit des renseignement sur une race des chiens, Monde_Vivant/Vertèbres/Carnivores/Canidés/Chiens /Rottweiller/Rottweiller_moins_de_5_ans, chaque classe représente un type ou une instance de la classe qui lui précède).

Ce modèle de structuration permet de faciliter la découverte, la compréhension et l'extension d'un champ de connaissance puisqu'il contribue à l'enrichissement de l'existant et la détection des vides à combler [KWASNIK 1999]. La rigidité l'empêche de prendre en considération l'appartenance simultanée d'une entité à plusieurs classes, il n'admet qu'une seule façon de classer et décrire. Il ne peut pas convenir à l'organisation des domaines des connaissances en émergence, ayant des bases conceptuelles et terminologiques instables.

La structure arborescente est une forme d'organisation hiérarchique plus souple que la structure hiérarchique pure, elle s'adapte plus facilement aux différentes situations, environnements et aux besoins particuliers. Cependant, elle est moins riche sur le plan sémantique et n'admet qu'un seul modèle de navigation logique entre les entités. La transmission des caractéristiques d'une classe mère aux classes filles n'est pas automatique, elle dépend du type de relation entre les classes : partitive, causale ou basée sur une caractéristique non essentielle de l'entité à classer. Plus l'arborescence se développe plus elle devient rigide et l'intégration des nouveaux sujets ou classes devient difficile si l'on veut éviter la restructuration de l'arbre [HUDON et HEDI 2010]. D'autres formes d'organisation ont été mises en place comme le modèle de la mise en ordre ponctuelle qui se base sur des liaisons temporaires entre les composantes de la structure [9] et le modèle à facettes qui répond aux besoins plus complexes en respectant la logique de navigation entre les domaines des connaissances et en traitant la dépendance conceptuelle et la multiplicité des points de vue.

9. En science de l'information, il est associé à l'organisation des documents selon les centres d'intérêts [HUDON et HEDI 2010, p. 14-16]

1.5.3 Le modèle à base de facettes

1.5.3.1 Le terme facette

Le terme facette a été défini par différents chercheurs. Parmi eux, VICKERY [1963], la considère comme l'ensemble produit par la division d'un sujet à partir d'une seule caractéristique. Et d'après KWASNIK [1999], « *la notion de facette repose sur la croyance qu'il y a plus qu'une seule manière de voir le monde et que même les classifications qui sont vues comme stables sont en fait provisoires et dynamiques* ».

1.5.3.2 La classification à facette

La classification à base de facettes est un modèle d'organisation des connaissances élaboré en 1924 par S.R. Ranganathan, un bibliothécaire et mathématicien indien. Ce modèle facilite l'expression des sujets composés ou complexes, permet de se libérer de la structure hiérarchique, préserve une logique de navigation entre les différentes ressources documentaires et introduit le principe de la multi-dimensionnalité d'un objet.
Après les modifications apportées sur le premier modèle à facettes proposé, cette classification est considérée comme une approche analytico-synthétique qui s'apparente à un dispositif syntaxique et non sémantique. Cette approche procède en deux étapes :

- une étape analytique : consiste à analyser un sujet et le décomposer en différentes facettes ;
- une étape synthétique : permet de générer un indice significatif intégrant les composantes du sujet en respectant un ensemble des règles syntaxiques complexes.

La première application de ce modèle a été proposée en 1933 par Ranganathan aux scientifiques de la communauté bibliothéconomique sous le nom de Classification Colon (CC). Elle est basée sur un modèle à cinq facettes, jugées nécessaires et suffisantes pour l'analyse et la représentation de tout sujet, bien que Ranganathan reconnaisse qu'il n'y a pas de classification idéale ou universelle. Les cinq facettes sont :

- Personnalité : le concept principal du sujet ;
- Matière : la propriété du sujet ;
- Energie : l'action décrite ;

— Espace : la localisation géométrique ;
— Temps : la localisation temporelle [HUDON et HEDI 2010][DORIA 2013].

Malgré la richesse de ce modèle dans les représentations qu'il permet de générer, il a été très peu exploité dans les bibliothèques, dans la mesure où il ne remplit pas l'une des fonctionnalités principales de classification, le rangement des documents. La difficulté de définir des facettes (telles que Personnalité et Energie) qui semblent peu fonctionnelles n'a pas empêché le développement des principes sur lesquels se base ce modèle qui a trouvé un environnement d'application intéressant à l'ère numérique et des systèmes d'information ouverts et interopérables. En effet, les ressources documentaires numériques n'ont plus besoin d'être associées à une seule classe pour être rangées.
Au cours de l'évolution de ce modèle plusieurs définitions ont été publiées comme celle de DENTON [2003]. Celui-ci définit la classification à facettes comme : « *un jeu de catégories mutuellement exclusives et conjointement exhaustives, chacune construite en isolant une perspective sur les items (une facette) qui se combinent pour décrire entièrement tous les objets en question, et que l'utilisateur peut utiliser en cherchant ou en parcourant pour trouver ce dont il a besoin* ». Elle est connue aussi comme une approche qui permet d'explorer, naviguer et rechercher une information grâce à sa structure logique et ce qu'elle offre comme les points d'accès multiples vers les documents, la navigation à l'intérieur du corpus et l'utilisation d'un filtre progressif basé sur des critères multiples pour affiner la recherche. Inspirés par ces principes, plusieurs environnements informationnels ont été récemment élaborés, dans lesquels les facettes établissent des liens fonctionnels et pragmatiques. [HUDON et HEDI 2010][DORIA 2013]

1.5.3.3 Les avantages de la classification à facette

Parmi les avantages de la classification à facette, on note :
— la possibilité de créer autant des sujets qu'on le souhaite et d'ajouter de nouvelles entités, théoriquement il n'existe aucune limite et la structure classificatoire peut s'étendre à l'infini ;

— la flexibilité : interprétée à travers l'indépendance qui existe entre les composantes de le structure jusqu'au moment de la synthèse d'un sujet et la génération de l'indice qui le représente ;
— l'expressivité : chaque facette associe librement le vocabulaire et la structure qui convient le mieux au savoir qu'elle présente ;
— « *leur conception ne nécessite pas un savoir complet sur le domaine, ce qui est utile pour des champs émergents ou qui évoluent* ». [DORIA 2013]

Toutefois, on note quelques limites aux classifications à base de facettes comme :
— la difficulté de visualisation de la structure de la classification ;
— la complexité de la définition des facettes appropriées et leurs fonctions ;
— la diversité des points de vue peut entraîner une perspective dominante ou une vision locale et un degré de précision trop élevé dans la classification qui la rendait non interopérable ;
— ainsi que l'absence des relations entre les facettes. [HUDON et HEDI 2010][DORIA 2013][KWASNIK 1999]

1.5.4 Les évolutions avec le Web : vers les nouvelles approches collaboratives

Devant la progression et la propagation du Web, l'intérêt de l'organisation des ressources mises à disposition a été perçu par certains gestionnaires du Web qui mettent en place des structures classificatoires (plans de classement, listes des facettes, folksonomies, taxinomies ou ontologies) flexibles permettant la navigation et l'accès facile aux ressources dans les portails, les encyclopédies et les bibliothèques virtuelles grâce à des caractéristiques communes. Avec l'avènement du Web 2.0, l'individu n'est plus un utilisateur et récepteur passif, il devient un acteur à part entière appelé individu-internaute. Il contribue aux opérations de classement, d'indexation et d'évaluation de l'information tout en interagissant avec les systèmes collaboratifs d'annotation de contenu, d'indexation sociale et de référencement, un des exemples les plus marquants est l'encyclopédie collaborative Wikipédia.

Devant ces évolutions, de nouvelles approches et de nouveaux outils sont apparus comme les folksinomies et le tagging collaboratif qui sont basés sur une activité d'étiquetage des contenus. En effet, les internautes forment progressivement des collections d'étiquettes, appelées « tags », qui les associent à des ressources ou des situations dans l'objectif de les partager. C'est une approche collaborative d'enrichissement des contenus informationnels à travers des métadonnées librement choisies par les acteurs, elle procure une forme d'auto-organisation permettant l'adaptation des systèmes à la diversité des classements établis par les utilisateurs. Aujourd'hui, la construction collective de formes plurielles de classement, les folksonomies et les activités d'indexation des ressources par les usagers sont le cœur d'étude de plusieurs communautés de recherche dans le Web sémantique et socio-sémantique.

1.6 Les applications de la classification documentaire

Dans le domaine des technologies d'information et de la communication qui regroupent les techniques principales de l'informatique, de l'audiovisuel, des multimédias, de l'Internet et des télécommunications, ZACKLAD [2007] propose un classement de documents selon deux couples de variables. Le premier couplage se base sur deux variables, la première distingue entre deux sous-systèmes d'information, les systèmes de réseaux informatique et de télécommunications (permettent la circulation et le stockage, sur des supports d'enregistrement, des signaux) et les systèmes de documents numériques (organisent le contenu de l'information, son stockage et son accès). La deuxième variable sépare les deux modes de diffusion : diffusion indifférenciée (trouvée dans la distribution de documents ressources [10] comme les livres et les journaux et dans les média audiovisuels) et diffusion directionnelle et interactive (correspondant aux usages des moyens de communication et aux documents articulatoires [11], bases des données, fichiers annotés, messageries ...). Le tableau suivant résume cette typologie :

10. Représentent des œuvres ou des contenus d'information stabilisés sur un support formant des sources ou des références sans subir des modifications (film, livre, rapport, photo, enregistrement musical...) [ZACKLAD 2007]

11. Sont des documents utilisés pour tracer et enregistrer en permanence toutes modifications et tous les événements qui surviennent dans un flux d'informations transactionnel (les bases des données, les mécanismes de coordinations comme le workflow...) [ZACKLAD 2007]

Type de système d'information	Les systèmes de réseaux informatiques et de télécommunication	Les systèmes à documents numériques
Diffusion indifférenciée	les média audiovisuels (radio, télévision)	les documents ressources destinés au large public (livres, journaux ...)
Diffusion directionnelle et interactive	les moyens de communication interpersonnelle (téléphonie, visiophonie)	les documents articulatoires ou les Documents pour l'Action [12] (bases des données, fichiers annotés, messageries ...)

TABLE 1.1 – Typologie de deux sous-systèmes d'information [ZACKLAD 2007]

Le deuxième couplage de variables croise le degré de codification [13] du document et la fréquence de son actualisation. La première variable sépare les documents codifiés qui sont produits dans des situations transactionnelles, le plus souvent répétitives et automatisées, et les documents non codifiés issus des productions sémiotiques. La deuxième établit une séparation entre les documents stabilisés, qui ne subirent aucune modification, et les documents mis à jour en permanence permettant la traçabilité de toute transaction. Cette typologie est présentée dans le tableau suivant :

1.6.1 La classification documentaire bibliothécaire

Le domaine bibliothécaire est le premier domaine qui exploite la classification documentaire et ses schémas classificatoires, dans lequel les documents sont non codifiés. Selon BEGHTOL [2001], généralement, les classifications bibliographiques sont énumératives, organisées sous forme hiérarchique et basées sur la relation générique/spécifique. Elles sont basées fréquemment sur la caution bibliographique, une notion inventée par Hulme en 1911. Cette notion « *consiste à fonder le système d'accès sur le volume de documents sur des thèmes contenus dans la littérature* » [DORIA 2013]. D'après BEGHTOL [2001], une classification basée sur ce principe ne devrait pas chercher à énumérer tous les sujets possibles mais de traiter les sujets des livres existants. HJORLAND [2007] explique, en

13. Dans le sens de « *système de symboles permettant de représenter une information dans un domaine technique* » (http ://atilf.atilf.fr/tlf.htm) [ZACKLAD 2007]

Type de document	Document codifié	Document non codifié
Actualisation rare	Document de codage : — Liste des nomenclatures — Fichiers de programme	Document ressource : — œuvre fixée sur un support : livre, rapport, enregistrement musical, film ...
Actualisation fréquente	Document articulatoire : — Fichiers papiers — Bases de données informatiques — Mécanismes de coordinations (par exemple : workflow)	Document pour l'action : — Fichiers annotés collectivement — Messages électroniques — Forums, blogs, Wiki ...

TABLE 1.2 – Typologie de document [ZACKLAD 2007]

fonction des buts de la classification, que cette notion consiste à sélectionner la meilleure affirmation parmi d'autres présentes dans les textes et ayant le plus haut degré d'autorité, et ne pas identifier un texte pour en tirer une relation sémantique. Il existe d'autres types de caution dans autres domaines, par exemple la caution culturelle est mentionnée dans le domaine socioculturel [14].

RANGANATHAN [1967] dans son ouvrage « Prolegomena to Library Classification », qui a été édité trois fois en 1937, 1957 et 1967 et qui présente les différentes versions de classification à facettes, précise que le cœur de la classification en bibliothèques est la classification par sujet. Il propose plusieurs types de classification en bibliothèque :

— les classifications énumératives ;

— les classifications quasi-énumératives ;

— les classifications quasi-à-facettes ;

— les classifications à facettes rigides ;

14. Dans laquelle la classification est développée et pour lequel elle est opératoire et qui influence le développement du système. Dans cette perspective, une classification peut être vue comme un document qui révèle les préoccupations, croyances, opinions de la culture qui l'a créée. [DORIA 2013][BEGHTOL 2001]

— les classifications à facettes libres.

D'après RANGANATHAN [1967], elles sont capables de traiter des différents types de sujets :

— les isolats : un isolat est une idée ou un ensemble d'idées, il ne forme pas un sujet en lui-même mais un composant du sujet qui peut être commun ou spécifique. Par exemple, l'ensemble « Euro, France, femme, énergie » peut concerner différents sujets comme une zone géographique ou un état du développement social ou économique du pays ;
— les sujets basiques : un sujet basique ne contient pas d'isolat. Ils correspondent aux classes génériques de premier niveau des classifications bibliographiques (littérature, psychologie, sciences politiques, ingénierie, mathématiques ...) ;
— les sujets composés : un sujet composé est formé par la combinaison d'un sujet basique avec un ou plusieurs isolats, par exemple la psychologie de l'enfant ou l'ingénierie financière ;
— les sujets complexes : un sujet complexe est basé sur la relation entre deux sujets, par exemple la psychologie politique en sciences politiques. Ranganathan distingue les sujets en se basant sur deux notions : l'extension et l'intention. Elles correspondent respectivement aux degré de développement du sujet dans les documents (la forme de courte à longue) et degré de la généralité du sujet (du générique au spécifique) ;
— les macro-sujets : sont des sujets avec une grande extension et une petite intention ;
— les micro-sujets : sont des sujets avec une petite extension et une grande intention. DORIA [2013]

Dans ce qui suit, nous présentons les différents types des classifications bibliographiques et nous citons l'exemple le plus connu pour chacun.

1.6.1.1 Les classifications énumératives

1.6.1.1.1 Présentation

Un schéma énumératif est caractérisé par l'énumération de tous les sujets qui décrivent les documents à classe dans une séquence continue, tout comme les classifications quasi-énumératives et quasi-à-facettes. Cette liste des sujets est triée du plus générique au plus spécifique en associant une notation développée pour chaque terme [BEGHTOL 2008] [RANGANATHAN 1967]. Ces classifications sont difficiles à alimenter et dépassées rapidement (faible résistance). D'autres limites de ce type de classification se manifestent à la longue de la liste des sujets à parcourir et la rigidité des indices des classes. La classification de la « *Library of Congres* » en est un exemple.

1.6.1.1.2 *Library of Congres Classification (LCC)*

Ce type de classification est utilisé à la bibliothèque nationale de France (BNF) et dans autres pays comme les États-Unis (utilisé par plusieurs bibliothèques universitaires). Cette classification encyclopédique est conçue à la BNF par Nicolas Clément entre 1675 et 1684. Elle est formée de 23 classes suivant les lettres de l'alphabet latin. A partir de 1997, la BNF utilise une réorganisation en département thématique alphanumérique, par exemple une catégorie qui appartient à une classe peut être précisée successivement par une ou deux lettres et une série des chiffres. [DELAUNEY 2010][WILD et al. 2009]

1.6.1.2 Les classifications quasi-énumératives

1.6.1.2.1 Présentation

Ce type de classification diffère du précédent par l'ajout des catégories communes à la liste principale des sujets. Ces catégories énumèrent des sujets communs, d'où l'importance de la formation des classes composées. Cette classification porte les mêmes limites que la classification énumérative sauf que les indices des classes ne sont plus rigides grâce aux combinaisons possibles entre eux. La Classification Décimale de Dewey (CDD) en est un exemple.

1.6.1.2.2 La Classification Décimale de Dewey (CDD)

Ce schéma de classification est conçu par le bibliothécaire américain Melvil Dewey en 1876. Il est dominant dans près de 130 pays. Cette classification en est à sa 23ème édition (2011). Elle comprend 10 classes principales de 0 à 9, 0 représente la division « généralités », 1 «la philosophie », 2 « la religion », etc., avec une subdivision en 10 sous-classes. Par convention, si un indice dépasse trois chiffres, un point les sépare des suivants. [WILD et al. 2009]
Par exemple :

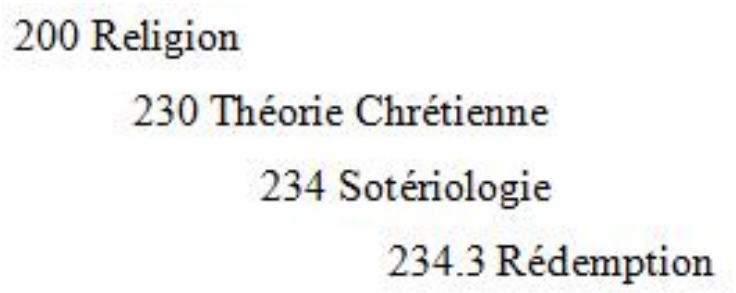

FIGURE 1.1 – Exemple de la Classification Décimale de Dewey

1.6.1.3 Les classifications quasi-à-facettes

1.6.1.3.1 Présentation

Différemment des deux classifications précédentes, la liste principale des sujets énumère des sujets basiques et composés, enrichie par des catégories composées des sujets communs et spécifiques. Grâce aux combinaisons possibles des classes, les indices des classes sont moins rigides et la résistance est plus importante mais la liste des sujets reste longue à parcourir. La Classification Décimale Universelle (CDU) en est un exemple

1.6.1.3.2 La Classification Décimale Universelle (CDU)

Apparue en 1905 par Paul Otlet et Henri Lafontaine, cette classification est un développement de la 5ème édition de la CDD en adoptant la hiérarchie de la structure des classe et la notion décimale mais elles différent sur la formulation d'indice. Contrairement à la CDD, la CDU autorise la combinaison des indices en utilisant une gamme d'opérateurs,

ce qui forme des classes composées et plus complexes permettant l'expression des sujets composés. Les indices se divisent en indices principaux, qui sont associés aux classes principales (de 0 à 9), et indices supplémentaires ou auxiliaires qui sont annexés aux indices principaux sous deux formes :

— des auxiliaires communs : décrivent les types communs des concepts, en conséquence, un document peut être décrit par une combinaison de types de classes ;
— de auxiliaires spéciaux : comme des suffixes à sens précis. Ils décrivent les types spéciaux des concepts, donc un document peut être décrit par une classe spécifique.

La classification à facettes exploite la notion de combinaison de cette approche, dans le fait qu'un sujet peut être composé de différents concepts et qu'une facette peut être construite par une combinaison de classes appropriées. [WILD et al. 2009]

1.6.1.4 Les classifications à facettes rigides

1.6.1.4.1 Présentation

Dans ce type de classification, la liste des sujets ne contient pas des sujets composés mais elle est composée par des sujets basiques avec des catégories composées de sujets communs et spécifiques. Les indices de classes ne sont pas rigides puisqu'ils se basent sur différentes combinaisons de classes, les listes des sujets sont courtes, la résistance est améliorée et l'ajout d'un nouveau sujet est limité. La Colon Classification (CC) en est un exemple.

1.6.1.4.2 La Colon Classification (CC)

Dans cette classification, conçue par Ranghanathan, présentée précédemment dans la section 1.5.3.2 de ce chapitre comme la première application du modèle à facette, les indices de classes sont construits par couplage en utilisant le symbole « : » (colon en anglais) pour connecter les parties de l'indice. Les facettes, l'ordre de leur enchaînement et la formule de facettes pour chaque classe basique sont prédéterminés pour tous les sujets. Dans les trois premières éditions de la CC, la rigidité vient de la prédétermination des classes de composés spécifiques qui sont utilisables avec une classe basique. [RANGANATHAN 1967]

1.6.1.5 Les classifications à facettes libres

1.6.1.5.1 Présentation

Ce type de classification se base sur les principes de la classification à facettes rigides mais il néglige les formules prédéterminées de facettes en proposants un ensemble des théories et des principes qui permettent de construire la séquence de facettes conforme aux sujets.

1.6.1.5.2 La 4éme édition de CC

En 1952, La 4éme édition de CC, appelée aussi la classification analytico-synthétique , a aboutit à l'élaboration de ce type de classification. Dans cette édition, les facettes, les indices et les formules des facettes ne sont plus prédéterminés. Les facettes utilisées dans des sujets composés sont retrouvées grâce à la technique d'analyse par facettes. En conséquence, cette technique est nécessaire pour la recherche d'un sujet d'où la difficulté de ce type de classification car une formation est nécessaire non seulement aux bibliothécaires mais aussi aux utilisateurs de la bibliothèque. Pour construire l'indice, il faut procéder comme suit :

— Décomposer le sujet en ses constituantes fondamentales ;
— Dans les listes appropriées du schéma classificatoire, enlever les indices appropriés ;
— Utiliser une méthode précise pour faire la synthèse des indices appropriés. La généricité de la facette lui permet de décrire les différents composants : le sujet basique et le sujet composé. [DORIA 2013]

1.6.1.5.3 L'analyse par facette

La classification à facettes, qui est définie comme une structure, peut être construite par la technique d'analyse par facette. C'est une méthode itérative, elle permet d'énumérer un ensemble des caractéristiques qui forment la base de la classification. [RANGANATHAN 1951]
La différence entre cette technique et la classification énumérative réside dans les règles de

division logique :

— « *toute catégorie logique distincte doit être isolée, toute nouvelle caractéristique de division doit être clairement formulée, toute relation nouvelle doit être définie* » ;
— les facettes peuvent se combiner et toute relation entre les sujets (ou les termes) est possible ;
— le schéma classificatoire est ouvert aux relations (pas de restriction à la relation hiérarchique genre/espèce). [VICKERY 1963]

D'après BARRE [2010], l'analyse par facette est nécessaire pour conduire une création de la classification à facettes. Il présente les étapes de cette création, dont les deux premières représentent la technique d'analyse, comme suit :

— Définition du champ de la classification (quels sont les intérêts des utilisateurs) ;
— Formulation des facettes en examinant les ressources (des documents en général) qui représentent les intérêts des utilisateurs à fin d'extraire les termes pertinents, les définir dans le champ de la classification et les ranger par la suite dans des groupes (les facettes) ;
— Organisation des termes collectés dans un ordre hiérarchique, dans chaque facette, en éliminant les doublons et créant des fusions si nécessaire ;
— Création des notes, seulement si le terme le nécessite ;
— Organisation des facettes ;
— Création du système de notation ;
— Association du système de notation aux facettes.

HJORLAND [2013] critique cette méthode du point de vue sémantique, il trouve qu'elle n'est pas suffisante pour décomposer les sujets complexes, le même ensemble de facettes peut être utilisé pour des contextes différents, les éléments de la classification ne sont pas définis par des spécialistes de l'information mais par des chercheurs scientifiques.

1.6.1.6 Les classifications bibliothécaires et le numérique

A l'heure du numérique, les besoins classificatoires ont évolués et différents systèmes et outils sont développés en se basant sur des schémas classificatoires traditionnels ou de

nouvelles approches. Traditionnellement, la classification bibliothécaire des livres est basée généralement sur une classification énumérative en appliquant la règle suivante : « *si 20% du livre est à propos d'un sujet, alors il est assigné à une classe correspondant à ce sujet* », par contre un livre peut décrire plusieurs sujets [DORIA 2013][HJORLAND 2012]. En 2005, à propos de ce sujet, Shirky montre dans son article « Ontology is Overrated : Categories, Links, and Tags » que seul l'objet livre se trouve dans un seul endroit dans la bibliothèque, mais pas les idées qu'il contient. Contrairement à la classification énumérative, le numérique ne met pas en œuvre la contrainte physique des classes d'objet, il utilise beaucoup moins de contraintes sur les lieux de stockage [WILD et al. 2009]. SHIRKY [2005] explique aussi que le numérique ne se base pas sur les stratégies de prédéfinition des catégories en couvrant les cas possibles comme les classifications bibliographiques mais il introduit des stratégies itératives et collaboratives permettant le classement des sujets émergeants de manière pertinente comme l'approche du Web Socio-Sémantique. Cette dernière est une évolution d'un modèle d'une bibliothèque qui repose sur « *l'idée de stabilité du savoir [HJORLAND 2012] et de maturité du contenu [WILD et al. 2009], vers l'informatisation des contenus et la mise à disposition d'un savoir dont la pérennité est plus mouvante sur le Web* » [DORIA 2013].

SHIRKY [2005] rajoute qu'une autre contrainte de la classification énumérative est négligée dans le numérique, le fait que la conception du schéma classificatoire peut être influencée par l'optimisation de nombre des livres dans les étagères car les principes de la caution bibliographique (sur les quels se basent la classification énumérative) entraînent une représentation minoritaire des thèmes contenant un petit nombre de documents qui lui sont associés.

Pour WILD et al. [2009], dans les bibliothèques numériques, les systèmes de notations sont simplifiés ou supprimés puisque un document numérique peut être lié à une classe spécifique par un URI (Uniform Resource Identifier) [15].

Contrairement à la classification énumérative qui a trouvé des limites avec le numérique, la classification à facettes est déployée par des outils informatiques et sur le Web, notamment pour la structuration des espaces de tagging comme :

15. Une technologie de Word Wide Web qui permet d'identifier une ressource de manière permanente, même si la ressource est déplacée ou supprimée

— TagMe !, est un système dédié au tagging des images qui sont sur Flickr [16], en exploitant les URI de DBpedia [17], il intègre une fonctionnalité de catégorisation par facette des tags [ABEL et al. 2011] ;

— FaceTag , est développé pour améliorer le système de tagging social de Bookmarks [18] en enrichissant sémantiquement les informations et en offrant une navigation multidimensionnelle à travers les facettes proposées [QUINTARELLI et al. 2007].

1.6.2 La classification des documents codifiés

Dans cette section, nous nous intéressons à des systèmes d'information qui traitent les documents numériques, principalement les documents articulatoires et de codage. Ces systèmes permettent l'organisation et la classification des informations en utilisant des outils informatiques utilisés par les professionnels de l'informatique ou de simples utilisateurs comme les systèmes des fichiers, les bases des données et les navigateurs Web.
En informatique, les langages et les index qui organisent l'information sont devenus des technologies intégrées aux dispositifs de stockage et aux usages de la recherche d'information. Cette dernière s'intéresse aux moyens de trouver une information parmi un grand nombre de ressources (document, fichier, bibliothèque logicielle, base des données ...). Elle est représentée sous deux modalités :

— par navigation : cette forme de recherche consiste à se déplacer d'un endroit à un autre en suivant des liens jusqu'à atteindre l'objectif cible. Les endroits et les liens forment un graphe et dans la plupart des cas un arbre. L'objectif représente les informations recherchées qui peuvent être distribuées sur plusieurs endroits, dans ce cas plusieurs recherches successives sont nécessaires. La navigation se base sur un concept fondamental, qui est le chemin, une liste ordonnées des liens permettant d'accéder à un endroit précis.

— par interrogation : cette forme de recherche consiste à élaborer une requête qui sélectionne suffisamment d'objets contenant la réponse recherchée. Cela nécessite

16. https ://www.flickr.com, visité en mars 2014
17. http ://dbpedia.org, un projet d'exploration et extraction automatique des données dérivé de Wikipédia, http ://fr.wikipedia.org, visité en mars 2014
18. http ://www.bookmarks.fr, visité en mars 2014

une indexation des objets dans un espace de représentation et une fonction de comparaison entre cet espace et le langage de la requête [FERRE 2002].

Dans ce qui suit, nous présentons les structures d'organisation les plus connues en informatique en mettant l'accent sur les critères suivants :

— la génération de la structure qui peut être automatique, semi-automatique ou manuelle (avec une contribution humaine) ;
— le niveau de la *codification* [ZACKLAD 2007] des documents ou des informations traités. L'intitulé des données, leurs contenus ou leurs chemins d'accès qui sont accessibles à une communauté concernée ou à tous les utilisateurs ;
— le niveau de l'ambiguïté de la recherche de l'information et l'accessibilité de l'objectif.

1.6.2.1 Les structures de la navigation

1.6.2.1.1 Les systèmes hiérarchiques

La hiérarchie est la structure de navigation la plus répandue en informatique. Dans cette structure, les documents et les informations sont organisés sous forme d'un arbre, un seul chemin d'accès est associé à chaque document. Les systèmes hiérarchiques représentent des systèmes d'organisation physique dans lesquels un document est placé dans un seul endroit tels qu'ils existent dans les bibliothèques. Contrairement à la classification bibliothécaire, nous détaillons le fonctionnement informatique de cette structure et non pas son aspect sémantique, en considérant comme exemple les systèmes des fichiers comme les systèmes d'exploitation (Unix, Windows ...) et les environnements de développement (Eclipse [19], Microsoft Visual Studio [20], Netbeans [21] ...). Les endroits sont des répertoires, où se trouvent soit des sous-répertoires contenant les fichiers qui forment les nœuds terminaux de la structure de navigation, soit directement ces fichiers. Ils contiennent les informations utiles qui représentent les documents de codage ou les documents ressources.

19. Un projet organisé en ensemble de sous-projets de développement logiciels, englobe un environnement de développement, frameworks, outil de modélisation et conception, outil de test ...
20. Un ensemble complet d'outils de développement formant un environnement de développement intégré
21. Un environnement de développement intégré open source

Dans cette structure arborescente, chaque répertoire a un répertoire parent sauf le répertoire racine, donc les liens de navigation vont des répertoires aux sous-répertoires et aux fichiers [FERRE 2002].

Dans les environnements de développement, les fichiers sont des fichiers de programmation, des bibliothèques informatiques (des standards utilisés dans les langages de programmation : « Library »), des fichiers de configuration et des fichiers ressources (comme les images, les fichiers texte ...) utilisés par les programmes. Les intitulés de ces fichiers sont généralement codifiés et symboliques, ils ne sont accessibles qu'aux développeurs informatiques, ainsi que la structure hiérarchique de ces environnements. Cette structure est standard et uniforme (généralement la même structure), elle permet un accès facile et direct à l'information sans l'ambiguïté sémantique qui existe dans les structures hiérarchiques bibliothécaires, présentées précédemment. Par exemple dans Netbeans, en choisissant le modèle MVC (Modèle Vue Contrôleur) [22], une structure hiérarchique est générée automatiquement, les fichiers de programmation qui définissent les objets classes [23] ou les modèles de données sont placés dans le dossier « Models », les fichiers contenant les fonctions associés aux classes, la gestion des événements, la synchronisation et la logique de contrôle sont dans le dossier « Controllers » et les fichiers permettant la présentation et l'affichage des interfaces utilisateurs sont dans le dossier « Views ». Le schéma suivant présente la structure hiérarchique d'un projet de développement appelé « test » qui applique le modèle MVC dans l'environnement de développement Netbeans.

FERRE [2002] mentionne l'existence de liens symboliques qui représentent des liens transversaux à la hiérarchie, ils sont définis par des chemins menant à la destination qui peuvent être créés par les raccourcis sous Windows ou la commande ln sous Unix (syntaxe de la commande : ln –s 'nom_de_l'élement_visé' 'nom_effectif_du_lien'). Avec ce type des liens plusieurs chemins peuvent mener au même répertoire ou fichier donc la structure de navigation devient un graphe. FERRE [2002] cite le système des fichiers Unix comme un exemple de cette structure. Comme tous les systèmes d'exploitation, la structure de la navigation est générée semi-automatiquement, une partie générée automatiquement repré-

22. Un modèle d'architecture qui cherche à séparer les couches de présentation, métiers et d'accès aux données et minimiser la dépendance entre eux.

23. En programmation orientée objet, une classe représente une catégorie des objets en déclarant des propriétés communes, elle définie un ensemble des attributs et un ensemble des méthodes.

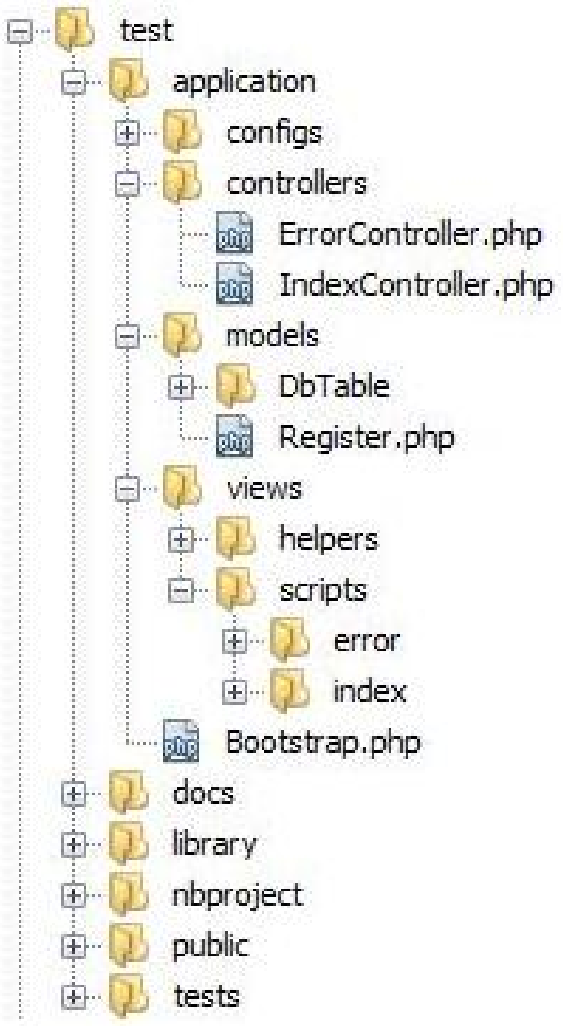

FIGURE 1.2 – Exemple de la structure hiérarchique dans l'environnement de développement Netbeans

sente le noyau du système et une deuxième construite manuellement par l'intervention de l'utilisateur. Les intitulés des répertoires et des fichiers (comme les fichiers de configuration) générés automatiquement nécessitent des connaissances professionnelles du système.

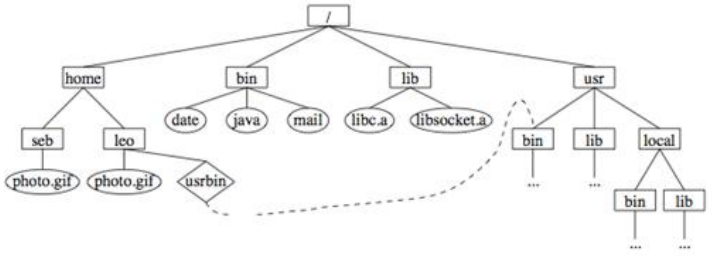

FIGURE 1.3 – Exemple de la hiérarchie dans le système des fichiers Unix [FERRE 2002]

Dans cette figure, les rectangles représentent les répertoires, les ovales représentent les fichiers et les losanges représentent les liens symboliques, ils sont liés par des traits représentant les liens orientés du haut vers le bas. Les chemins commencent par le répertoire racine (noté '/') ou par un sous-répertoire de home. Il est possible de naviguer en saisissant directement le lien (path) comme '/home/seb/photo.gif' ou en utilisant des commandes comme 'cd..' pour remonter au répertoire parent et 'cd sous-répertoire' pour se positionner sous un sous-répertoire.

1.6.2.1.2 Le Web

En s'intéressant de l'aspect structurel du Web, d'après FERRE [2002], il se base sur une structure de navigation sous la forme d'un graphe des pages Web dont le contenu peut contenir des liens vers d'autres pages. Ces liens des liens hypertextes ou des boutons de navigation. Chaque document est référencé par une adresse standard appelé URL (Unifrom Resource Location) par exemple : http ://www.ensiie.fr/ pons/NFA017/COURS/lectureXX-oo-avance-php.shtml. Ce graphe est construit manuellement lien par lien, ses nœuds sont les titres des différentes pages Web. Les noms des liens, selon leur orientation, sont « page principale » pour liens ascendants, « page suivante » pour les liens vers la droite du graphe, « page précédente » pour les liens vers la gauche du graphe et le titre de la page référencée pour les autres liens [FERRE 2002].
Les structures de navigation présentées ci dessus, comme les structures physiques, sont

simples à concevoir et à comprendre, mais il est difficile à les faire évoluer devant les évolutions des besoins de la recherche d'information. Dans bien des cas, l'utilisateur est mené à explorer toute la profondeur de l'arborescence pour consulter un répertoire ou un fichier donc cette structure ne s'adapte pas à ses besoins.

1.6.2.2 Les structures de l'interrogation

Comme présenté précédemment, ce type de structure se base sur l'élaboration des requêtes selon un langage de requêtes précis qui interrogent des objets représentant les informations et les données dans un espace de représentation. Les systèmes d'interrogation les plus connus sont les moteurs de recherche et les systèmes de bases des données. Cette structure est représentée par un modèle d'interrogation comme le modèle vectoriel, le modèle booléen et le modèle logique et un modèle de données comme le modèle des données logiques et le modèle objet [FERRE 2002]. Dans cette structure, nous considérons les bases des données comme un document articulatoire qui peut contenir des documents ressources sous forme de structure des données.

1.6.2.2.1 Les modèles d'interrogation

1.6.2.2.1.1 Modèle booléen

Dans ce modèle, les requêtes sont des combinaisons des descripteurs des objets dans l'espace de représentation avec des opérateurs booléens comme et, ou et non. Le fonctionnement de ce modèle consiste à évaluer la requête en fonction de la description de l'objet, pour chaque terme, une valeur « vrai » lui est attribuée s'il appartient à la description de l'objet et « faux » sinon, de même pour la requête d'après la sémantique des opérateurs booléens [FERRE 2002].

1.6.2.2.1.2 Modèle vectoriel

D'après FERRE [2002], « dans le modèle vectoriel, les requêtes et les descriptions d'objets sont des vecteurs dans un espace de descripteurs. Les coordonnées des vecteurs indiquent l'importance de chaque descripteur, soit pour l'objet décrit, soit pour la requête formulée ».

Le fonctionnement de ce modèle consiste à mesurer la distance entre une requête et une description permettant l'ordonnancement des résultats de plus proche au plus éloigné de cette requête

1.6.2.2.1.3 Modèle logique

Dans ce modèle, « une même logique sert d'espace de représentation et de langage de requêtes. Les descriptions d'objets et les requêtes sont donc des formules logiques, les premières étant extraites des objets au cours de la phase d'indexation, et les deuxièmes étant formulées par les utilisateurs » [FERRE 2002]. Le fonctionnement de ce modèle consiste à mesurer la quantité des informations nécessaires à une description pour satisfaire une requête.

1.6.2.2.2 Les modèles de données

1.6.2.2.2.1 Modèle logique

Dans le monde des formes, ce modèle se concentre sur la formalisation de ces formes avec un langage riche en relations et en opérations permettant de raisonner. Donc ce monde constitue une logique dont les formes sont des formules. CODD [1970] a séparé le niveau logique et physique des bases des données. Le niveau logique représente un modèle des données abstrait permettant une représentation logique des informations (différente de la représentation sous la forme des chaînes de bits). Le premier modèle logique, défini par Codd, est l'algèbre relationnel qui permet d'exprimer ce qu'on veut obtenir (sans donner la façon de l'obtenir) et de composer facilement les requêtes. Il organise les données sous la forme de relations et les gère à l'aide d'opérations relationnelles. Il se base sur des opérations comme le produit cartésien, la sélection, la projection, l'union et la différence ensembliste. Dans les bases de données relationnelles (BDR), les langages déclaratifs sont de nature logique, par exemple, en algèbre relationnel, le présentant logique d'une relation est un prédicat [FERRE 2002]. Géré par un système de gestion de base des données relationnelle (SGBDR), ce type de base des donnée organise les différents types de donnée dans des tables comme les entiers, les chaines des caractères, les documents (avec le type BLOB), etc.

La représentation de la logique dans les BDR est limitée, donc elles ont été étendues vers les bases de données déductives en appliquant le « Datalog » qui est une version du langage de programmation logique des prédicats «Prolog » adapté aux bases des données. La différence entre le Datalog et le modèle relationnel est que ce premier définit les relations de deux manières :

— une relation extensionnelle : une relation, liée à un prédicat, qui est stockée dans la base des données ;
— une relation intensionnelle : une relation, liée à un prédicat, qui est définie par des règles logiques.

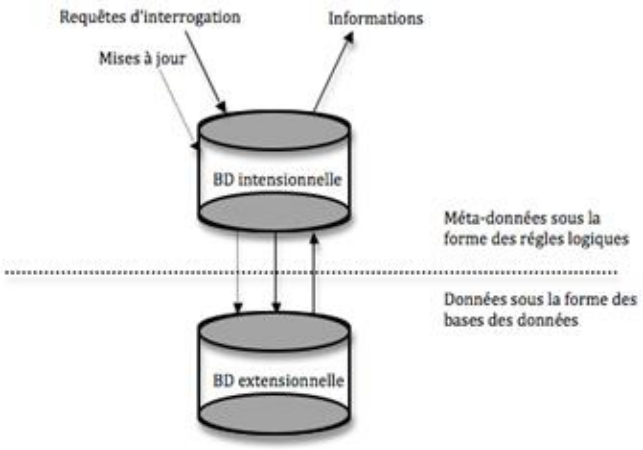

FIGURE 1.4 – Base de données intensionnelle et extensionnelle [GARDARIN 2003]

Cette figure montre la notion de la base des données logique qui est constituée par une base des données intensionnelle (BDI) et une base des données extensionnelle (BDE) contenant respectivement : les règles logiques et les relations extensionnelles. La BDE est dérivée de la BDI par les règles [GARDARIN 2003].
Dans ce modèle logique, une valeur est identifiée par elle-même ce qui peut introduire des anomalies comme la redondance ou la disparition des informations. Ces problèmes sont traités par les formes normales utilisées à la préposition du schéma de la base, ainsi que par l'utilisation d'une valeur spéciale interprétée tel que identifiant. Ce problème n'existe pas dans le modèle objet.

Des nouveaux besoins en terme d'applications informatiques dans différents domaines étaient exposés, comme les besoins de représenter et traiter un volume important de données complexes (dans les domaines de recherche en biologie, médecine et chimie), de gérer des données multi média (dans des applications de gestion), des données cartographiques (dans l'informatique géographique) et assurer la gestion des versions des données (dans le suivi de contenus des entreprises avec le workflow). Ces évolutions des besoins fonctionnels se tournent vers les modèles de données. Le modèle objet est apparu pour répondre à ces besoins et résoudre les limites du modèle logique. [DUCOUNAU et al. 2008]

1.6.2.2.2.2 Modèle objet

Appelé aussi modèle orienté objets, il diffère du modèle logique par la notion d'identité qui distingue les objets même s'ils sont le même contenu. Le modèle de données objets est apparu initialement dans les langages de programmation objets (C++, Java ...) et a été étendu par la suite au domaine des bases des données. Nous présentons la base des données objets comme un exemple de ce modèle. Elle est issue du rapprochement des systèmes de la gestion des bases des données (SGBD) et des langages de programmation orientés objets. Sa notion fondamentale est l'objet qui est caractérisé par :

- Une identité : elle permet de lui faire référence ;
- Un état : une structure de données dont le type peut être un objet désigné par son identité, un n-uple contenant des composants de différents types ou un type de base (chaîne de caractères, entier, BLOB pour les documents ...) ;
- Un comportement : si la structure de données est associée à une identité, elle peut évoluer selon les comportements disponibles qui sont des méthodes utilisées pour obtenir des informations sur l'objet ou le modifier. [FERRE 2002]

Ces caractéristiques permettent à l'objet de représenter de données complexes et facilitent leur gestion à travers un système de gestion de base des données objets (SGBDO). Il assure aussi la gestion et l'accès concurrents à de larges volumes de données.

```
Class Personne
tuple     (nom : string,
           prenoms : list(string),
           adresse : tuple(rue : string, num: integer, ville : string))
Method  Creer (n : string) : Personne,
           Lire_nom : string
End Personne

Class Etudiant inherit Personne
tuple     (numero_carte : integer,
           cours_suivis : set(Cours),
           releve_ notes : blob)
Method  Inscrire (c : Cours),
           Abandonner (c : Cours)
End Etudiant

Class Enseignant inherit Personne
tuple   (grade : string,
          salaire : real,
          cours enseignés : set(Cours))
Method Embaucher()
End Enseignant

Class Ater inherit Enseignant,Etudiant
tuple (domaine recherche : string)
End Ater
```

FIGURE 1.5 – Schéma d'une base des données universitaire (Modèle Objet) [DUCOUNAU et al. 2008]

Ce schéma représente un exemple d'une base des données objets d'un institut universitaire. Au sein d'une université les personnes sont regroupées en deux catégories : des étudiants et des enseignants. Certains étudiants peuvent accéder à un statut d'Ater (attaché temporaire d'enseignement et de recherche) pour enseigner. Chaque objet (personne, étudiant, enseignant et Ater) est défini par un tuple (la structure des données) et des méthodes qui désignent le comportement de chaque classe. [DUCOUNAU et al. 2008]
La structure des données de l'objet Etudiant contient un attribut relevé des notes de type BLOB, un objet binaire de grande taille qui accepte la saisie des fichiers et des images.

La compatibilité entre les bases des données objets et les langages de programmation orientée objet facilite l'écriture d'applications et offre des accès aux données définissables par les programmeurs.

Les SGBDO offrent une interface graphique pour naviguer dans les structures de donnée, ils gèrent la persistance des objets d'où la distinction entre un objet temporaire ayant une durée de vie limitée et un objet persistant stocké dans la base des données et ils permettent aussi aux utilisateurs d'interroger la base avec différentes sortes de requêtes décrites par un langage spécifique, parmi les quelles on cite :

— La requête de base qui donne accès à l'état de l'objet, par exemple la requête de la figure 1.6 qui répond à l'interrogation « quel est le nom des enseignants » ;
— La requête associative qui réalise des sélections d'objets appartenant à des collections en se basant sur des conditions relatives aux valeurs des attributs, par exemple la requête de la figure 1.7 qui répond à l'interrogation « l'ensemble des étudiants de nom François résidant à Paris » ;
— La requête constructive ayant comme résultat une nouvelle structure construite, par exemple la requête de la figure 1.8 qui répond à l'interrogation « l'ensemble des renseignements (nom et ville) de tous les Aters. [DUCOUNAU et al. 2008]

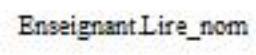

FIGURE 1.6 – Exemple de requête de base

```
Select e
From e in Etudiant
Where e.nom = "François" or e.adresse.ville = "Paris"
```

FIGURE 1.7 – Exemple de requête associative

```
Select tuple (nom : a.nom, ville : a.adresse.ville)
From a in Ater
```

FIGURE 1.8 – Exemple de requête constructive

Actuellement, le modèle objet atteint une large partie des objectifs fixés, il offre une meilleur représentation de la réalité en utilisant des modèles de données complexes, une grande facilité pour l'écriture et la maintenance d'application à travers la modularité et la réutilisabilité de la programmation par objets et une efficacité importante dans des nombreux domaines d'application (médecine, biologie ...).

1.6.2.2.2.3 Les bases de données et les documents

Face à l'explosion du volume des documents et aux évolutions des besoins d'organisation des documents la majorité des systèmes de gestion électronique des documents (GED) utilisent les BD classiques (relationnelles), les BDO ou des BD orientées document pour stocker physiquement les documents, les informations sur les documents (auteur, date création, date de modification ...), les mots clé, les index et les tags. La BDR et la BDO sont présentées dans les sections précédentes, la BD orientée documents est présentée par Stearn dans le cadre du projet MongoBD [24] (l'une des BD orientées document parmi les plus populaire actuellement). Elle est destinée à stocker les documents et la collection des documents sous la forme d'une structure libre, sans schéma prédéfini, identifiée par une clé d'identité. Elle offre les fonctionnalités suivantes :

— L'ajout, la modification, la lecture ou la suppression seulement de quelques champs du document ;

24. http ://www.mongodb.org, visité en mars 2014

— L'indexation des champs de document permettant un accès rapide à ces champs sans avoir recours à la clé qui identifie le document au quel ils appartiennent ;
— L'élaboration des requêtes avec des prédicats sur les champs.

Elle se base sur une structure hiérarchique qui représente les relations de type un-à-un et un-à-plusieurs entre les documents. Chaque document peut avoir des propriétés multiples et des sous-documents.
Finalement, la structure interrogatoire ne nécessite pas obligatoirement une organisation spécifique pour stocker les données sauf exception comme la BD des documents. Les structures des données, tables, objets et documents sont décrites individuellement et manuellement. L'utilisation de ces structures nécessite une connaissance du langage des requêtes (syntaxe et sémantique), la terminologie utilisée (les termes utilisés dans la description) et le contexte pour évaluer la pertinence d'un résultat. Les utilisateurs occasionnels peuvent rencontrer des problèmes pour exprimer une requête. Un autre problème de ce type de structure réside dans le volume des réponses qui peut être parfois vide et parfois trop volumineux d'où le risque de ne pas satisfaire tous les utilisateurs et toutes les utilisations.

1.6.3 Synthèse

Le tableau suivant récapitule les différences entre les structures d'organisation des documents bibliothécaires et les documents codifiés.

Critères	Classification énumérative	Classification à facette	Structure de navigation	Structure d'interrogation
Génération de la structure	Génération manuelle (contribution humaine) de la liste des sujets	Conception et construction des facettes	- Génération automatique et semi-automatique des structures hiérarchiques - Construction manuelle des graphes du Web	Construction manuelle de la structure (création des tables, objets, prédicats ...)

Critères	Classification énumérative	Classification à facette	Structure de navigation	Structure d'interrogation
Dépendance	Héritage : une sous-classe est un membre de toutes les classes supérieures	Les composantes restent indépendantes	- Dépendance d'appartenance de fichiers, sous-répertoires au répertoire parent - Dépendance de référencement entre les pages Web	- Dépendance relationnelle dans les bases des données relationnelles - Indépendance totale dans les bases des données objets - Dépendance hiérarchique dans les bases des données orientées documents
Rigidité	Structure trop rigide : les règles d'inclusion dans les classes sont rigides [DORIA 2013]	Structure lâche et extensible [DORIA 2013]	Structure extensible : étendre facilement les systèmes hiérarchiques et les graphes de Web	Structure extensible : étendre facilement le modèle des données
Exclusivité mutuelle	Une entité appartient à une seule classe [DORIA 2013]	Un sujet est présenté par une seule facette [DORIA 2013]	- Dans les systèmes hiérarchiques, un document appartient à un seul endroit - Une page Web (nœud de graphe) peut être liée à plus qu'une page	Une donnée est placée dans un seul endroit (les différents types des bases des données)
Point de vue d'organisation	Un seul point de vue d'organisation (objet appartient à une seule classe du sujet)	Une variété de points de vue : les différents liens qui peuvent exister entre les sujets	Un seul point de vue d'organisation (un document existe dans un seul endroit)	Une variété de points de vue : les différentes liaisons qui peuvent exister (relationnelle, héritage ...), sauf le modèle des données objet qui admet une organisation libre des objets
Niveau de codification	Fichiers non codifiés : documents ressources	Fichiers non codifiés : documents ressources	- Fichiers codifiés accessibles par la communauté concernée : fichiers de programmation, bibliothèques de développement, fichiers de configuration - Fichiers non codifiés : fichiers de ressources (images, textes, vidéos ...) accessibles par tous les utilisateurs	- Données codifiés - Requêtes d'interrogation sont écrites par les experts du domaine avec une syntaxe spécifique

Critères	Classification énumérative	Classification à facette	Structure de navigation	Structure d'interrogation
Ambiguïté à la recherche	Ambiguïté à l'accès aux documents (un document peut présenter plus qu'un sujet mais il appartient à un seul)	Des points d'accès multiples vers les documents	Il n'y a pas d'ambiguïté	Il n'y a pas d'ambiguïté mais risque d'insuffisance du résultat de l'interrogation
Connaissance du domaine	La nécessité d'avoir des connaissances complètes sur le domaine de la classification [DORIA 2013]	Un savoir complet du domaine n'est pas nécessaire [DORIA 2013]	La nécessité d'avoir des connaissances professionnelles sur l'environnement d'organisation sauf les systèmes d'exploitation qui nécessitent des connaissances minimes	La nécessité d'avoir des connaissances complètes

TABLE 1.3: Les différences entre les structures organisationnelles

A partir de ce tableau qui résume les différences entre les structures organisationnelles présentées dans des différentes sections de ce chapitre, nous constatons qu'il existe une typologie implicite des utilisateurs ayant des objectifs et besoins différents :

— Un simple utilisateur : est une personne n'ayant pas de connaissances avancées et professionnelles sur les méthodologies, les technologies et les structures utilisées pour organiser ses documents. Il représente la première cible du marché des technologies d'organisation des documents. Souvent, il cherche à organiser des documents non codifiés, documents ressources et documents pour l'action [ZACKLAD 2007], par la structure la moins complexe et le système ayant la plus simple interface de recherche ;

— Un documentaliste utilise la structure bibliothécaire pour organiser ces documents non codifiés. Il définit les sujets représentés par les documents, les classes et les facettes associées aux sujets ;

— Un informaticien, par exemple un développeur, organise ses documents codifiés comme les fichiers de programme, dans des environnements spécifiques : les environnements de développements, environnements de versioning, les bases des données (eux-mêmes considérées comme un document), etc.

1.7 Conclusion

L'organisation, la représentation et la recherche des connaissances n'est pas en soi quelque chose de nouveau. Elles sont quotidiennement employées pour communiquer, comprendre et expliquer des informations efficacement. L'innovation se présente au niveau des méthodologies et des technologies conçues pour gérer les connaissances. Par exemple dans une entreprise, l'organisation des différents types du contenu : les documents ressources (fichiers texte, fichiers audiovisuels, images...), les fichiers de programmation, les outils logiciels, les bases des données, les sites web et les blogs, etc., facilite le travaille collaboratif et le partage de l'information au sein de l'organisation à l'aide des systèmes d'information innovants comme les GED (Gestion Electronique des Documents) et les CMS (Content Management System).

Deuxième partie

Problématique et contributions

Chapitre 2

Problématique et méthodologies

2.1 Introduction

Dans ce chapitre nous présentons le cadre dans lequel se situent nos travaux de recherche ainsi que la problématique en expliquant quels sont nos objectifs. Nous décrivons par la suite les solutions méthodologiques que nous proposons pour réaliser les objectifs fixés.

2.2 Cadre de recherche

Nos recherches dans cette thèse se situent dans le cadre du projet de l'Agence Nationale de la Recherche (ANR) intitulé Méthodes et services intégrés institutionnels et participatifs pour la classification à facettes des contenus documentaires complexes (Miipa-Doc). Ce projet rassemble différents contributeurs appartenant à des disciplines et établissements différents :

- Des chercheurs en science de l'ingénieur plus précisément en ingénierie des connaissances et informatique de l'UTT (laboratoire Tech-CICO) ;
- Des chercheurs en science de l'information et de la communication du Cnam (laboratoire DICEN) ;
- Des ingénieurs en ingénieries des connaissances de EDF-R&D ;
- Des ingénieurs concepteurs du Cogniva-europe.

L'objectif de Miipa-Doc est de concevoir de nouvelles méthodes d'ingénierie sociocognitives ascendantes pour l'organisation des contenus documentaires complexes de l'entreprise et de développer l'architecture logicielle correspondante tout en gérant les problèmes d'hétérogénéité rencontrés à la gestion de ces contenus.

2.3 Problématique et objectifs

La classification documentaire en tant que tâche informatisée est réalisée à travers les systèmes d'information qui utilisent des SOC. Elle se matérialise par des opérations fortement interdépendantes aujourd'hui qui sont : le classement, l'indexation, le nommage du document dans des espaces de stockage partagés ou locaux.
Avant le numérique, ces opérations étaient des activités différentes. La classification représentait une activité intellectuelle permettant la sélection d'une classe de destination. Le classement correspondait à la conséquence de cette activité en plaçant physiquement un document dans une étagère et l'indexation correspondait à la description d'un document par les représentations des sujets qu'il présente dans son contenu en utilisant nécessairement un langage documentaire.
Avec le numérique, la classification et l'indexation sont devenues corrélées. En effet, en indexation participative où il n'y a pas un langage de documentaire prédéfini, l'utilisateur peut ajouter un nouveau 'tag' à fin de créer une nouvelle 'classe' de rangement des documents similaires. Le classement garde la même finalité qui est l'emplacement du document. Il représente son enregistrement dans un espace de stockage, un disque dur par exemple, en sélectionnant le chemin d'accès. Le document a essentiellement un nom qui lui est attribué par l'utilisateur en respectant une logique de classification par exemple la concaténation des attributs du document (auteur, date, version, etc.). La documentarisation classificatoire, présentée par Zacklad, « *correspond à un sous-ensemble des activités de documentarisation qui regroupe les trois opérations de nommage, d'indexation de classement* » [ZACKLAD et al. 2011a]. Elle représente « *une documentarisation externe portant sur l'ensemble du document et permettant une « articulation sémantique externe avec d'autres documents renvoyant à des projets transactionnels proches » (Zacklad 2004) opposée à la documentarisation interne basée sur des métadonnées qui orientent la lecture à l'intérieur*

du document » [ZACKLAD et al. 2011a].

L'un des principaux objectifs des SOC est de faciliter la communication efficiente d'une connaissance dans une communauté d'utilisateur en fournissant une conceptualisation partagée d'une partie du monde. En effet, la classification documentaire dépend de la communauté d'utilisateurs concernée, la structure du SOC utilisé et la façon dont il est représenté. L'hétérogénéité des SOC tient à la diversité de structuration, qui dépend de la nature d'un SOC et de son utilisation, et l'hétérogénéité des formats et langages de leur représentation. Ces diversités et hétérogénéités entraînent un problème d'interopérabilité entre les outils, les systèmes et les solutions basés sur les SOC.
Nos travaux ont pour but de montrer la faisabilité d'un système d'information de *la documentarisation classificatoire* qui facilite la gestion de l'hétérogénéité des modalités de la classification documentaire en répondant aux objectifs suivants :

— Utilisation et création facile des concepts du SOC pour un simple utilisateur ;
— Une gestion multidimensionnelle de la classification ;
— Une gestion plus intégrée de la documentation de l'entreprise ;
— La réalisation de compromis entre des classifications existantes et les classifications émergentes ;
— La position de ce système d'information dans l'environnement des outils de gestion du contenu dans les entreprises.

Ce dernier objectif sera détaillé dans le quatrième chapitre.

2.4 Propositions et méthodologies

Nous proposons une approche orientée modèle et incrémentale qui gère les quatre préoccupations conceptuelles pour mettre en place un système d'information de documentarisation classificatoire adapté aux objectifs cités précédemment. Elles sont :

— Construction du modèle de l'Interface Homme Machine (IHM) et l'ergonomie du SI en assurant une interaction entre l'utilisateur et le SI pour une réalisation la plus simplifiée et la moins coûteuse des tâches utilisateur ;

— Construction du modèle de SOC permettant une indexation collaborative ;
— Conception orientée infrastructure permettant la représentation de la communication interne du SI ;
— Conception de l'interopérabilité avec autres SI.

Ces quatre préoccupations représentent des axes de modélisation nécessaires pour réaliser la solution la plus adaptée à la représentation, au stockage et à la mise à disposition de SOC à une communauté d'utilisateurs au sein de l'entreprise.
En effet, une tâche de classification d'un document s'étale sur les différents niveaux conceptuels des différentes profondeurs du SI, de l'interaction avec l'utilisateur à travers une IHM et son évolution en participant à la construction du SOC (construction participative) jusqu'aux techniques et protocoles de communications inter-composants du SI et avec autres SI comme le montre le schéma suivant :

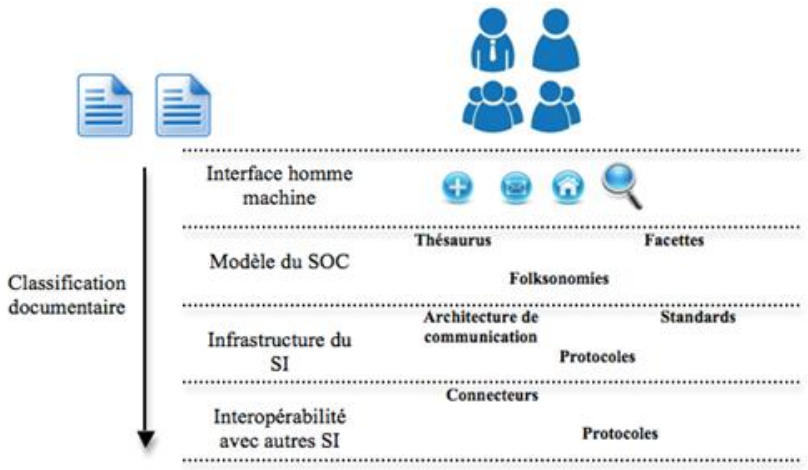

FIGURE 2.1 – Les niveaux conceptuels intervenant dans la classification documentaire

Notre démarche est pluridisciplinaire : elle fait appel aux Sciences de l'Information et de la Communication (SIC) et à l'Informatique, plus particulièrement à une approche incrémentale dirigée par les modèles, pour apporter des éléments conceptuels, méthodologiques et technologiques de support aux aspects collaboratifs et évolutifs du système d'informa-

tion basé sur un système d'organisation des connaissances. En effet, nous nous basons sur l'ingénierie dirigée par les modèles (IDM) pour proposer une conception progressive en décomposant l'objectif global de cette conception en une suite d'objectifs conceptuels intermédiaires. Elle est aussi incrémentale en ce qu'elle gère l'hétérogénéité des éléments des différents niveaux d'abstraction (IHM, SOC, l'infrastructure ou l'architecture, l'interopérabilité) du SI et inspirée d'UGC (User Generated Content) pour assurer un rôle d' « acteur » à l'utilisateur. L'UGC représente le principe fondateur du web 2.0 (Read-Write Web), il est à la base des services proposés par les plateformes de partage de contenus numériques et les « réseaux sociaux », qu'ils soient généralistes (Facebook [25],Twitter [26]), ou spécialisés, comme Youtube [27] ou Dailymotion [28] en vidéo, ou encore Delicious [29] ou Diigo [30] pour la gestion des favoris.

2.4.1 Le rôle des usagers

2.4.1.1 Valorisation des rôles des usagers

Dans cette démarche, les acteurs, ils peuvent participer par l'aval à faire reconnaître leurs propres points de vue, dans une approche de participation et d'émancipation. Ce qui valorise les « savoirs d'usage » et renforce le nouveau rôle et les apports des usagers. Cet aspect est particulièrement important dans les activités de veille préalables à la « mise sur le marché » de produits innovants.

Comme l'indique VULBEAU [2010], « *depuis que l'usager est considéré comme un acteur capable d'intervenir dans les questions qui le concerne, tout un vocabulaire a émergé pour mieux rendre compte de ce nouveau statut. Avec les « savoirs d'usage », on suppose que les pratiques des usagers forment une expérience sociale, urbaine, culturelle, etc. et que cette expérience, cumulée et réfléchie, permet de déduire des compétences inédites. Ces savoir-faire efficaces portent sur les choses du quotidien, souvent mal identifiées ou considérées comme mineures. Ainsi, un piéton empruntant tous les jours le même itiné-*

25. http ://www.facebook.com/
26. http ://twitter.com/
27. http ://www.youtube.com/
28. http ://www.dailymotion.com/fr
29. http ://www.delicious.com/
30. http ://www.diigo.com/

raire connaît d'expérience les passages sûrs, les espaces rendus dangereux par la circulation automobile et peut témoigner des problèmes d'autres catégories de piétons. Tout aménagement qui viserait à sécuriser les déplacements urbains devrait ainsi tenir compte de cette expérience qui se fonde sur un savoir de première main, marquée par les épreuves répétitives de la quotidienneté. ».

2.4.1.2 Responsabilisation des usagers dans un processus d'indexation collaborative

CHAKER et al. [2010] proposent un système de recherche d'information (SRI) pour améliorer les performances des usagers dans la résolution de leurs tâches métier. Ce système est basé sur un modèle de contexte métier, qui repose sur un triptyque : la modélisation à long terme des usagers, la modélisation des tâches métier (associées aux tâches informationnelles) ainsi que la modélisation de l'environnement. SRI élabore aussi un principe de « stabilisation » des situations de travail, basé sur ces modèles. Dans un objectif similaire au « principe de stabilisation », notre système d'information de la documentarisation classificatoire vise à établir des règles d'apprentissage (appelées règles de stockage) pour classer des documents en fonction du contexte et faciliter ensuite la collecte d'informations. Dans cette perspective, pour évaluer la qualité globale de ce type du système, il nous semble important de pouvoir évaluer finement les activités de nommage et de classification des documents opérées par les usagers au cours du temps. [SALZANO et ANKOUD 2012]

2.4.2 L'approche IDM

Dans le monde d'ingénierie du logiciel, l'ingénierie dirigée par les modèles (IDM) est une approche de conception basée sur le principe « tout est modèle ». Cette approche propose d'utiliser des représentations simplifiées des différents aspects du système et définir des solutions pour chacune des préoccupations conceptuelles (fonctionnelle, structurelle, ...).

L'IDM est basé sur un ensemble de concepts et des relations qui sont :

— La notion de modèle : elle représente le concept central de cette approche. Il est défini par COMBEMALE [2009] comme suit : « *Un modèle est une abstraction d'un système, modélisé sous la forme d'un ensemble de faits construits dans une intention*

particulière. Un modèle doit pouvoir être utilisé pour répondre à des questions sur le système modélisé » ;

— Le principe de la substituabilité : une relation appelée *représentationDe* est définie entre le modèle et le système qu'il représente. Un modèle doit être une abstraction pertinente de ce système, c'est-à-dire : « *Un modèle doit être suffisant et nécessaire pour permettre de répondre à certaines questions en lieu et place du système qu'il est censé représenté, exactement de la même façon que le système aurait répondu lui-même* » [COMBEMALE 2009] ;

— La notion du méta-modèle : un modèle, une représentation de la réalité, a besoin d'un langage de modélisation qui le définit clairement. La définition de ce langage a pris lui-même la forme d'un modèle appelé méta-modèle et défini par COMBEMALE [2009] comme suit : « *Un méta-modèle est un modèle qui dénit le langage d'expression d'un modèle, c.-à-d. le langage de modélisation* ». Donc une deuxième relation est identifiée entre modèle et méta-modèle, appelée *conformeA*.

L'OMG [31] se base sur ces principes pour définir des standards. Cet organisme a développé cette approche en proposant une extension avec l'initiative de MDA (Model Driven Architecture). En admettant la notion du méta-modèle, plusieurs méta-modèles ont émergé pour définir chacun leurs spécifications dans un domaine particulier. Cette grande variété des méta-modèles donne lieu à des problèmes d'indépendance et d'incompatibilité, ce qui nécessite la définition d'un cadre général pour la description des méta-modèles. Le méta-méta-modèle est la réponse à ces problèmes et l'un des principes de base de la modélisation d'OMG. D'après COMBEMALE [2009], il représente « *un modèle qui décrit un langage de méta-modélisation, c.-à-d. les éléments de modélisation nécessaires à la dénition des langages de modélisation. Il a de plus la capacité de se décrire lui-même* ».
Les principes de modélisation d'OMG sont représentés sous la forme d'une pyramide comme le montre la figure suivante :

Les quatre niveaux M0, M1, M2 et M4 représentent respectivement :

— Le monde réel, le système modélisé avec les relations et les différents cas d'utilisa-

31. The Object Management Group (OMG), http ://www.omg.org/

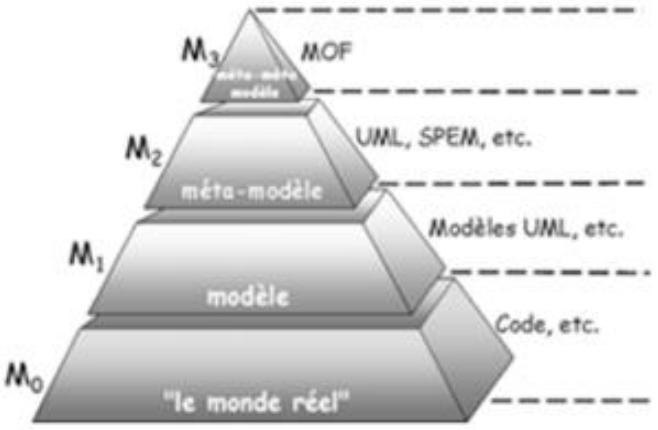

FIGURE 2.2 – Pyramide de modélisation d'OMG [BEZIVIN 2003]

tions de ce système ;

— Le modèle du système réel définit dans un langage choisi, appelé aussi modèle d'instance ;
— Le méta-modèle définissant ces modèles (par exemple UML) ;
— Le méta-méta-modèle définissant le méta-modèle.

Chaque modèle du M0 à M2 doit être conforme à celui qui le précède en respectant l'ordre descendant sur la pyramide de modélisation [SCHREIER 2011].
La modélisation et la méta-modélisation utilisent des normes de modélisation, les principales normes d'OMG sont :

— MOF (Meta-Object Facilities) : est un langage de définition des méta-modèles, il représente le méta-méta-modèle.
— Ecore : est une implémentation de MOF, utilisée avec l'approche EMF (Eclipse Modeling Framework).
— UML (Unified Modelling Language) : est un langage de modélisation des modèles et des méta-modèles.
— OCL (Object Constraint Language) : est un langage de description des contraintes sur le modèle. Il est utilisé dans la définition des méta-modèles.

— CWM (Common Warehouse Metamodel) : est un langage de modélisation des ressources, des données et de la gestion de l'entreprise [SCHREIER 2011].

Le schéma suivant montre un exemple de la hiérarchie entre les quatre niveaux :

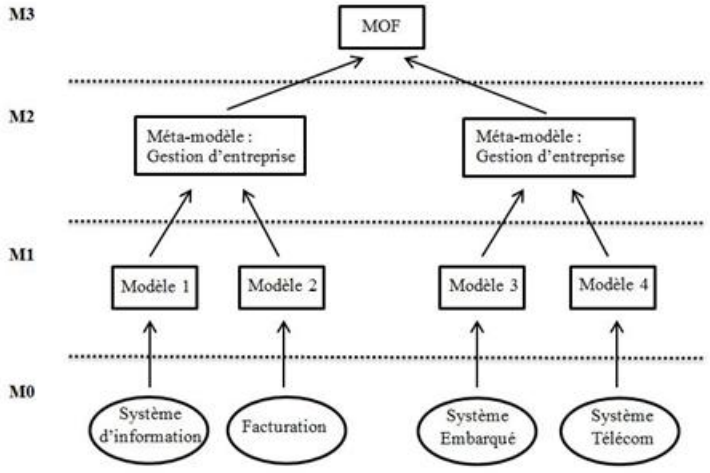

FIGURE 2.3 – Exemple du modèle à 4 niveaux

2.4.3 Construction du modèle de l'IHM

2.4.3.1 Interface Homme-Machine

L'interface homme-machine (IHM), considérée comme un composant du système d'information interactif, englobe tous les aspects de la solution informatique, tels que l'environnement de travail et les diverses façons de manipuler et de stocker des informations, qui influencent la participation de l'utilisateur aux tâches informatisées de la documentation classificatoire (nommage, classement, indexation) et de la recherche des documents classés. En général, l'interaction Homme-Machine s'introduit dans la problématique d'usage pour répondre à la question :

— Comment les humains utilisent-ils les systèmes d'information afin de concevoir des systèmes qui satisfassent aux mieux les besoins d'utilisateur ?

Plus particulièrement, nous cherchons dans cette solution à établir une interaction participative et collaborative à travers une interface unique, de qualité, qui permet la participation d'utilisateur à l'enrichissement progressif des SOC et leur partage dans un groupe de travail. Cette interface unique permet une semi-automatisation des opérations de la documentarisation classificatoire (nommage, classement, indexation, enregistrement), la recherche des documents et l'exploration des structures classificatoires des SOC utilisés. Une IHM de qualité est celle qui facilite l'apprentissage, offre une utilisation efficace, facile et sécurisée à l'utilisateur pour atteindre la satisfaction des usagers, l'acceptabilité du logiciel et la rentabilité pour l'entreprise. En effet, une IHM doit s'adapter aux processus cognitifs des utilisateurs en se basant sur des méthodes de l'ergonomie afin d'améliorer la communication entre l'utilisateur et la machine (le logiciel plus précisément). [SATTEL et al. 2005] [GRISLIN et KOLSKI 1996]

2.4.3.2 L'ergonomie

L'ergonomie consiste à adapter le travail, les outils et l'environnement à l'être humain en caractérisant la relation entre eux. Depuis des dizaines d'années, l'ergonomie des logiciels est devenue une discipline essentielle qui contribue à la conception et l'évaluation des systèmes critiques comme les applications militaires, les centrales nucléaires, le contrôle du trafic aérien et les systèmes d'information dans des nombreuses organisations et entreprises.

La particularité de l'ergonomie des logiciels est que la conception logicielle établit le contenu des informations disponibles aux utilisateurs et les relations visuelles entre eux, ainsi que la séquence des actions que l'utilisateur doit effectuer et les résultats de ces actions. [SCAPIN 1986]

2.4.3.3 Le modèle de l'IHM

La construction du modèle de l'IHM dépend du contexte. En effet, la sélection des propriétés de la qualité ergonomique dépend des caractéristiques de la communauté et des exigences des tâches. Ces propriétés sont :

— L'utilité : elle correspond aux capacités fonctionnelles, aux performances et à la qualité de l'assistance technique fournie à l'utilisateur par le SI pour atteindre ses

objectifs de travail

— L'utilisabilité : elle dépend de la qualité de l'interaction Homme-Machine. Elle représente une frontière entre l'utilité potentielle d'un système et son utilité réelle.

La conception et la construction d'une IHM doivent respecter les principes ergonomiques suivants :

— La compatibilité : l'accord entre les caractéristiques de l'utilisateur, les tâches et l'organisation des entrées, sorties et du dialogue de l'application des données ;
— L'homogénéité et la cohérence : la façon de conserver les choix de conception de l'IHM pour des contextes identiques ou différents ;
— L'adaptabilité et la flexibilité : la capacité du système à réagir selon le contexte, les exigences des tâches, les stratégies, les habitudes de travail de l'utilisateur ;
— Le guidage : l'orientation, le conseil, l'information et la conduite de l'utilisateur lors de ses interactions avec le SI ;
— La charge de travail : l'augmentation de l'efficacité du dialogue et la réduction de la charge de travail ;
— Le contrôle explicite : la prise en compte du contrôle de traitement des actions explicites de l'utilisateur ;
— La gestion des erreurs : la réduction des erreurs et leur correction. [SATTEL et al. 2005][GRISLIN et KOLSKI 1996]

A ce niveau conceptuel, nous utilisons l'IDM pour modéliser le modèle d'IHM et les tâches effectuées par l'usager pour avoir à la fin une interface graphique qui répond au mieux aux exigences de la documentarisation classificatoire. Nous détaillons les modèles conceptuels réalisés dans le chapitre suivant.

2.4.3.4 L'interface graphique

Une interface graphique est un dispositif de dialogue homme-machine, dans lequel les objets à manipuler sont dessinés sous forme d'une représentation graphique à l'écran. L'usager peut utiliser cette interface en manipulant les objets avec un dispositif de pointage

comme la souris ou le doigt pour les écrans tactiles, etc. Elle se base sur des concepts généraux comme :

- Les métaphores : permet de transférer une compréhension supposée existante chez l'utilisateur dans un nouveau contexte. Par exemple la métaphore du bureau, formulaire papier, tableau de bord, etc.
- Les icônes : sont les représentations graphiques des objets sur l'écran. Elles sont déterminées en utilisant généralement la métaphore, par exemple, une feuille de papier pour un document. Elles permettent de déduire intuitivement le fonctionnement des objets et gérer efficacement les données par manipulation directe.
- La manipulation directe : l'utilisateur manipule les objets à l'écran avec l'impression d'être sans autre intermédiaire grâce aux présentations continues des objets et actions physiques (sur des boutons) avec un effet immédiatement visible.
- Le WYSIWYG (What You See Is What You Get) : est un complément logique de la manipulation directe lors que le résultat de l'action est montré à l'écran tel qu'il apparaîtra dans le monde réel.

2.4.4 Construction du modèle du SOC

2.4.4.1 Approche orientée métier, fondée sur un SOC folksonomique à facettes

Une nouvelle méthodologie de classification orientée métier et basée sur la folksonomie à facette est élaborée dans le cadre de ce projet de recherche. Ma contribution consiste à modéliser les composants de cette méthodologie et ses évolutions afin de la mettre en œuvre dans le système d'information de la documentarisation classificatoire.

Aujourd'hui, dans les entreprises, la gestion du contenu dépend du projet, ses phases, du processus métiers, des tâches, et des acteurs du métier. Ces derniers appartenant au même métier n'ont pas forcément les mêmes pratiques mais ont nécessairement des éléments en commun comme les significations contextuelles des termes, la culture, et les représentations du métier. Les acteurs ont donc des activités de gestion du contenu en dimension personnelle et/ou partagée.

L'approche élaborée dans ce projet de recherche (Miipa-Doc) se focalise sur ces concepts

métiers et les activités des acteurs dans les projets d'une organisation. Elle représente une démarche orientée métier de la gestion des connaissances métiers en utilisant un SOC individuel et collectif de classification folksonomique à facettes. Elle permet aux acteurs de :

- Représenter leurs activités dans une dimension individuelle ;
- Partager et diffuser les représentations jugées utiles avec d'autres acteurs exerçant le même métier ;
- Améliorer et développer les pratiques métiers en profitant de l'expérience métier des autres acteurs.

Cette méthodologie intègre l'utilisateur au processus de création du SOC itérativement. Les activités de construction, d'alimentation et de gestion de ce dispositif sont considérées comme des activités méta-fonctionnelles permettant la préparation du SOC aux activités fonctionnelles. En effet, « *l'activité fonctionnelle correspond à la gestion du flux permanent de l'information, tandis que l'activité méta-fonctionnelle se rapporte à la gestion des connaissances, et correspond concrètement à l'anticipation des usages possibles des documents gérés ainsi qu'à une vision prospective de l'activité fondée sur l'expérience* » [DORIA 2013].

Ce SOC est une folksonomie à facette qui permet la contribution des utilisateurs à la classification et indexation des ressources grâce aux mots-clés sous la forme des groupes de tags afin de représenter les significations contextuelles souhaitées.

Une folksonomie formelle est basée sur des liaisons entre les utilisateurs, les tags et les ressources. Elle est définie par l'ensemble F : = (U, T, R, Y) dont U, T, R sont des ensembles finis, respectivement des utilisateurs, des tags et des ressources et Y représente la relation d'affectation de tags à l'ensemble (U x T x R). Ce modèle de folksonomie simple n'est pas suffisant pour attribuer un contexte à l'affectation de tags.

La folksonomie à facette est un modèle défini par extension au premier modèle permettant la création des différentes facettes d'une affectation de tags. Elle est représentée par l'ensemble F : = (U, T, R, Y, C, Z) dont les éléments d'extension C et Z définissent respectivement le contexte d'information dans lequel s'élabore la folksonomie et la relation d'affectation du contexte à l'ensemble (Y x C) [ABEL et al. 2011].

2.4.4.2 Les composants du SOC

2.4.4.2.1 Les vues

Les vues correspondent à une sélection de facettes pertinentes dans un contexte d'activité donné pour fournir la structuration globale de la classification. Par exemple, dans une application de gestion de projet, on pourrait organiser les vues selon les étapes du projet (étude préalable, réalisation, déploiement...) ou avoir deux vues, l'une relative à la temporalité (étape du projet, type de document, cycle de vie du document), l'autre aux sujets traités dans le projet (composant de l'artefact, compétences impliquées, clients associés).

2.4.4.2.2 Les facettes

Les facettes représentent la catégorisation selon plusieurs critères de l'ensemble des documents à classer. Chaque facette décrit un ensemble de propriétés dont le type de leurs valeurs représente le titre de la facette. Les propriétés sont représentées par les tags, associées aux facettes sous la forme des attributs, par exemple, pour le type de document, « rapport », « compte-rendu », « courrier », « présentation », etc. La modélisation de ce concept doit respecter les règles suivantes :

— La même facette peut se retrouver dans plusieurs vues pour faciliter le travail d'analyse de l'utilisateur ;
— Deux facettes ne peuvent pas avoir le même titre si elles n'ont pas les mêmes attributs ;
— Du point de vue sémantique de la classification à facette, il n'est pas possible d'avoir deux facettes ayant deux titres différents et les mêmes attributs ;
— Une facette peut être multivaluée (attribuer à un document plusieurs tags de la même facette à la fois) ou, dans le cas contraire, monovaluée ;
— Une facette peut être personnelle ou partagée avec d'autres utilisateurs (groupes de partage).

2.4.4.2.3 Les Tags

Les tags sont les valeurs inscrites dans les facettes. Elles représentent des mots clés permettant de décrire les documents à classer en langage naturel. La figure suivante montre les relations entre les éléments du SOC et les documents.

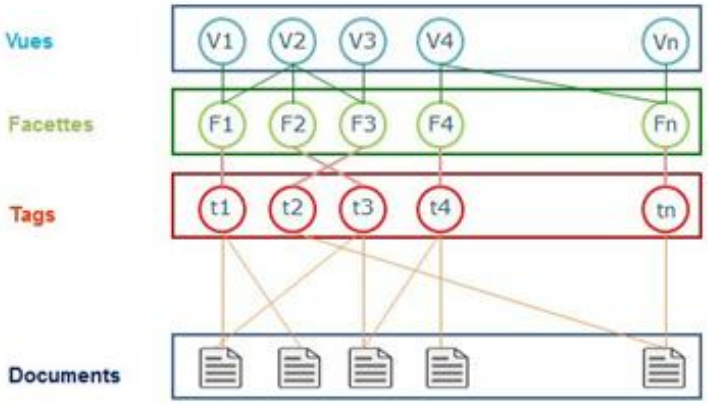

FIGURE 2.4 – Relations entre les éléments du SOC et les documents classés

Afin d'avoir une représentation graphique significative, chaque tag est représenté par une icône qui représente visuellement le sens du tag et une description textuelle. Le sens d'un tag provient du couple : titre de la facette et tag. En effet, un tag peut appartenir à plus d'une facette ayant des titres différents. Dans la même facette, les valeurs des tags sont homogènes au sens contextuel et sémantique (le fond) mais variables en terme de forme, par exemple une facette générique « Lieu » qui contient comme valeurs toutes entités représentant un lieu (nom de ville, pays, capitale, nom d'une rue, etc.). La figure suivante est une illustration de cet exemple :

Une facette peut contenir un type exact des valeurs, par exemple une facette qui décrit un lieu de type «Pays » dont les noms de pays sont les valeurs des tags. La figure suivante représente la facette « Pays » :

Ces composants et leurs caractéristiques offrent aux utilisateurs une indexation collaborative en permettant :

— une implémentation d'une logique métier ;
— une classification de documents par le système de tagging en injectant les tags

FIGURE 2.5 – Exemple de la facette « Lieu »

FIGURE 2.6 – Exemple de la facette « Pays »

comme des métadonnées dans le document ;
— la recherche de documents suivant un filtre progressif et selon les facettes et les tags prédéfinis.

Pour mettre en place cette méthode orientée métier et alimenter les composants du SOC dans une organisation, une procédure de collecte des informations sur le processus métier, les activités des acteurs, les fonctionnements et les habitudes de classement individuel et collectif est nécessaire. Cette collecte est réalisée à travers des réunions en collectif global, des réunions en collectif métier, des entretiens individuels et des entretiens d'élaboration individuelle de l'approche.

2.4.4.3 Démarche méthodologique pour la gestion d'évolution du SOC

L'objectif de notre démarche consiste à gérer les évolutions des composants du SOC en identifiant des éléments structurants notre SOC d'indexation collaborative. Nous constituons notre feuille de route comme suit :

— se positionner dans le cadre d'ingénierie des évolutions des systèmes d'information, identifier des recommandations concernant le support de l'évolution des ressources et apporter des éléments méthodologiques et technologiques de support aux aspects collaboratifs et évolutifs ;
— apporter des éléments de modélisation conceptuelle concernant les éléments du SOC.

Cette démarche traite la problématique d'évolution en deux étapes comme nous l'avons présenté sur la feuille de route et nous le détaillerons dans le chapitre suivant.

2.4.5 Construction du modèle orienté infrastructure

2.4.5.1 Approche orientée infrastructure

Le terme «infrastructure» évoque de vastes ensembles d'équipements collectifs nécessaires aux activités humaines, comme les bâtiments, les routes, les ponts, les voies ferrées, canaux, ports et réseaux de communication. Au-delà de briques, de mortier, des tuyaux

ou des fils, l'infrastructure comprend également des entités plus abstraites, telles que des protocoles (humains et informatiques), des normes et de la mémoire.
Pour les systèmes d'information, l'infrastructure se réfère à des installations numériques, des services informatiques, services d'assistance, des référentiels de données et des protocoles de communication et de transport des données à l'intérieur et à l'extérieur du SI. [BOWKER et al. 2010]
Depuis les années 80, ce concept représente un domaine d'émergence dans les études scientifiques et technologiques (STS : Science and Technology Studies). Dans différents domaines comme l'informatique, la science de la bibliothèque et la science de la collaboration à long terme, la conception de l'infrastructure joue un rôle crucial pour informatiser les spécifications fonctionnelles du SI, matérialiser les tâches des utilisateurs et établir une communication et une cohérence entre les différents composants hétérogènes du SI.

2.4.5.2 Objectif de l'approche orientée infrastructure

L'objectif de notre démarche conceptuelle consiste à élaborer une architecture système pour mettre en œuvre les différentes fonctionnalités et activités de la documentarisation classificatoire collaborative. Cette architecture est basée sur un protocole de communication entre ses composants structuraux.

Cette démarche est réalisée en respectant les étapes suivantes :

1. définir le type de l'architecture et ses composants qui conviennent aux attentes méthodologiques et technologiques de l'approche orientée métiers, basée sur l'indexation collaborative et le partage des éléments du SOC et des documents ;

2. choisir le protocole de communication sur le quel se base cette architecture ;

3. donner un modèle conceptuel générique de ce protocole et l'instancier par la suite pour l'appliquer sur le prototype réalisé dans le cadre de ce projet.

Ces étapes sont détaillées dans le chapitre qui suit.

2.4.6 Vers l'interopérabilité du SI

Nous continuons dans la même perspective orientée infrastructure, mais nous nous intéressons à la caractérisation d'ouverture de l'infrastructure et à son interopérabilité avec autres systèmes d'information.

2.4.6.1 Définition de l'interopérabilité

Suite au développement des systèmes d'information de gestion de contenu d'entreprise, plusieurs types de matériels et de logiciels sont apparus. Pour assurer la communication entre eux, ils doivent posséder une perception commune des informations, ce concept est appelé interopérabilité. Deux systèmes ou plus interopèrent s'ils dialoguent sans ambiguïté.

2.4.6.2 Aspects méthodologiques et techniques de l'interopérabilité du SI

L'interopérabilité des systèmes d'information est limitée aux aspects d'hétérogénéité et de normalisation. Elle est assurée au niveau des couches basses de l'architecture système en s'appuyant sur des standards pour la normalisation des interfaces physiques. Comme exemples, nous citons :

— Les protocoles du réseau Internet (TCP/IP, http,...) qui permettent l'échange des informations entre les différents systèmes d'exploitation et les applications systèmes sans ambiguïté ;
— Le protocole CMIS (Content management interoperability services) qui assure l'interopérabilité entre les systèmes de gestion de contenu en utilisant des services de type SOAP ou REST.

Dans le projet Miipa-Doc, pour répondre à la problématique d'hétérogénéité des documents traités par les différents systèmes de gestion du contenu, nous intégrons le protocole CMIS dans l'infrastructure de notre SI afin d'assurer :

— La recherche et l'exploration de documents classés par un autre SI au sein de notre SI ;
— La recherche et l'exploration de documents classés par notre SI au sein d'autre SI.

Dans la même approche conceptuelle, nous étudions les exigences applicatives d'usabilité de données issues de systèmes hétérogènes, nous détaillons le fonctionnement du CMIS en se basant sur un modèle conceptuel qui décrit ses composants et la communication entre eux et nous enrichissons par la suite le modèle de l'architecture conçu précédemment dans l'approche orientée infrastructure pour qu'il supporte ce protocole.

2.5 Conclusion

Après avoir fixé nos objectifs et présenté notre démarche méthodologique basée sur une approche conceptuelle incrémentale qui gère les hétérogénéités des différents niveaux d'abstraction d'un système d'information de gestion du contenu documentaire. Nous détaillons nos contributions fondées sur des aspects méthodologiques et conceptuels dans le chapitre suivant.

Chapitre 3

Approche dirigée par les modèles

3.1 Introduction

Dans ce chapitre, nous présentons le prototype HyperTagging, réalisé dans le projet Miipa-Doc. Par la suite, nous montrons comment une approche dirigée par les modèles permet de gérer les hétérogénéités entre les différents niveau d'abstraction du système d'information de la documentarisation classificatoire en proposant une modélisation incrémentale des composants de ce système, par les méta-modèles que nous instancions pour modéliser les composants du prototype HyperTagging.
L'aspect générique des méta-modèles enrichit notre approche conceptuelle. En effet, il donne une caractérisation d'évolutivité à long terme aux modèles conceptuels que nous proposons dans ce chapitre et à l'outil HyperTagging.

3.2 Le prototype HyperTagging

3.2.1 Description et fonctionnement

Le prototype HyperTagging modélisé dans le cadre du projet Miipa-Doc est un outil de gestion documentaire permettant de gérer des Systèmes d'Organisation des Connaissances hétérogènes, de manière ascendante, multifacette et distribuée. Il est représenté à l'utilisateur sous la forme de deux fenêtres : une fenêtre de classification et sauvegarde du document et une deuxième pour la recherche d'un document. Elles ont un module en commun qui représente graphiquement le SOC sur lequel se base la classification des docu-

ments dans ce prototype. Il est composé d'onglets pour représenter les vues, de facettes et de tags. Ces composants offrent une classification multidimensionnelle lorsque l'utilisateur sélectionne les métadonnées représentées par les tags associés aux facettes dont chacune appartient à un ou plusieurs contextes (représentés par les vues) à la fois. La sélection des tags permet aussi de déterminer l'emplacement de classement d'un document à l'aide des règles de stockage formées par le couple « chemin d'accès du répertoire de classement et ensemble de métadonnées ». La construction de ce module dépend de l'organisation de l'entreprise et des besoins de gestion de l'information personnelle de l'utilisateur.

Une unité de stockage peut être locale ou partagée, par exemple, les disques partagés ou les espaces gérés par des logiciels de stockages comme Dropbox [32] basé sur la notion du « Cloud Computing ». L'originalité de cette interface est qu'elle représente une expérimentation ergonomique par la mise en œuvre du codage redondant de l'information en représentant un tag par une icône et une expression linguistique. Cette représentation sémiotique facilite la mémorisation et accélère l'opération de sélection des tags au moment de la classification ou de la recherche d'un document.
Il fonctionne en mode :

— Multiposte : plusieurs postes pour un seul utilisateur avec des versions synchronisées des ressources pour chacun ;
— Monoposte : un poste par utilisateur ;
— Multi-utilisateur : plusieurs utilisateurs utilisent une ressource partagée en même temps ;
— Mono-utilisateur : un seul utilisateur avec une seule version des ressources.

3.2.2 Document pour l'action

Dans notre outil HyperTagging, le document traité ne désigne pas seulement les documents numériques bureautiques ou destinés à la publication, mais tous les supports d'écriture y compris ceux utilisés dans les applications de messagerie ou de web social, qui supportent toutes activités collectives collaboratives comme dans la définition des Docu-

32. Dropbox est un logiciel qui permet de stocker, synchroniser et partager des fichiers en ligne.

ments Pour l'Action (DOPA) fournie par Zacklad [ZACKLAD 2005b].

En science de l'information, le document est une notion fondamentale. Sa définition a connu une évolution progressive et influencée par les innovations technologiques de l'audiovisuel et du Web. Roger le définit comme suit : « Un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs » [ROGER 2003]. En suivant cette définition, nous considérons les fichiers (Word, Excel, pdf, text, les images, les vidéos...), les mails, les blogs et les pages HTML comme des documents. Parmi les documents nous associons une place spécifique aux documents évolutifs par M. Zacklad nommés des documents pour l'action (DOPA). Cet auteur s'est appuyé sur l'analyse des pratiques documentaires et sur les opportunités de collaboration offertes par l'Internet et Intranet pour définir le DOPA comme un objet de transaction et d'échange entre partenaires, indispensable aux activités collectives coopératives [ZACKLAD 2005b]. Ce type « désigne une classe de documents, numériques rassemblant, au fur et à mesure de leur élaboration, les contributions d'un collectif engagé dans une activité commune » [ZACKLAD 2005b]. Ses caractéristiques sont les suivantes :

— « *soutien à l'activité collective pendant sa rédaction ;*
— *preuve et témoin des décisions ou des analyses dont il acte l'existence ;*
— *mémoire des transactions associées aux processus de résolution de problème et de délibération dont il est le résultat ;*
— *offrant des informations exploitables par les processus visant à l'indexer, le rechercher et l'exploiter.* » [ZACKLAD 2005b]

3.3 Modélisation de l'IHM

Dans cette partie, nous modélisons l'IHM du prototype HyperTaging en nous basant sur l'ingénierie d'interaction homme machiné dirigée par les modèles [SATTEL et al. 2005]. La modélisation de l'IHM s'articule en différents niveaux d'abstraction qui sont :

— la spécification du domaine : spécifie les tâches utilisateur et les concepts du domaine manipulés dans les tâches. Dans ce niveau, nous nous intéressons à deux modèles : le modèle des concepts et le modèle des tâches.
— l'espace de travail : représente la zone d'interaction et la réalisation des tâches utilisateur. Ce niveau donne lieu au modèle d'espace.
— les interacteurs : représentent les composants de l'interface qui sont en interaction direct avec l'utilisateur (les fenêtres, les champs de texte, les boutons, etc.). Ce niveau donne lieu au modèle des interacteurs.
— l'interface finale programmée : la maquette conçue est implémentée dans un ou plusieurs langages de programmation. Un modèle de programme montre cette implémentation.

3.3.1 Modèle des tâches

Le modèle des tâches décrit les tâches utilisateur, les sous-tâches et les tâches élémentaires. Il est représenté graphiquement par les arbres. Une tâche est un couple « but, procédure » qui correspond respectivement à « l'objectif à atteindre par l'utilisateur, la façon d'atteindre cet objectif ». Les tâches sont liées entre elles avec des opérateurs binaires de type logique (OU, ET) ou de type temporel (PUIS, AVEC). Elles peuvent être décorées par des opérateurs unitaires pour exprimer une valeur par défaut, la complexité, la fréquence, le caractère optionnel ou itératif [SATTEL et al. 2005]. Le méta-modèle suivant définit les relations entre les tâches et les opérateurs.

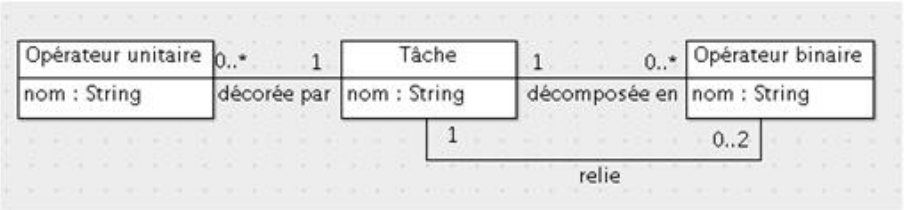

FIGURE 3.1 – Méta-modèle de tâches

Nous modélisons les tâches utilisateur d'HypeTagging par des instances du méta-modèle de tâches. Les tâches sont :

— Créer un jeu de facettes qui correspond à la structure de la classification des documents dans ce prototype. Cette tâche est réalisée en créant les vues et leur associe des facettes existantes ou nouvellement créées. Pour créer une facette, l'utilisateur commence par son squelette puis ses tags. La figure suivante représente cette tâche et ses sous-tâches :

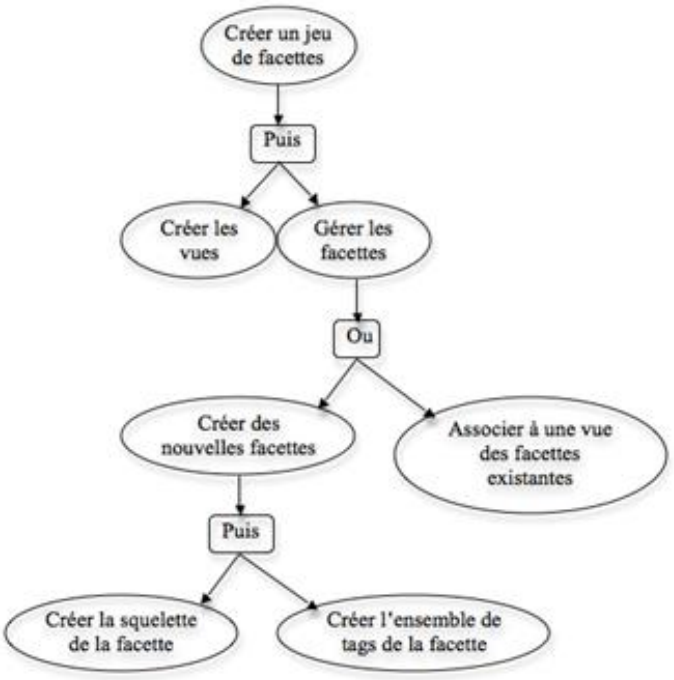

FIGURE 3.2 – Modèle de tâches « Créer un jeu de facettes »

— Classer un document est une tâche itérative (décorée par une étoile dans la figure suivante) composée de quatre sous-râces : choisir le mode de classement (mode document ouvert ou mode clic droit sur le document), sélectionner les tags avec les quels le documents sera indexé (les tags sont injectés comme des métadonnées du document en vue des activités de recherche), gérer l'emplacement du classement (en sélectionnant une unité de stockage locale ou partagée, une règle de stockage ou l'emplacement actuel du document), gérer le nommage du document (soit garder le nom actuel du document soit saisir un nouveau) et gérer le type du classement (classer le document original ou une copie). La figure suivante représente le modèle

de cette tâche et ses sous-tâches :

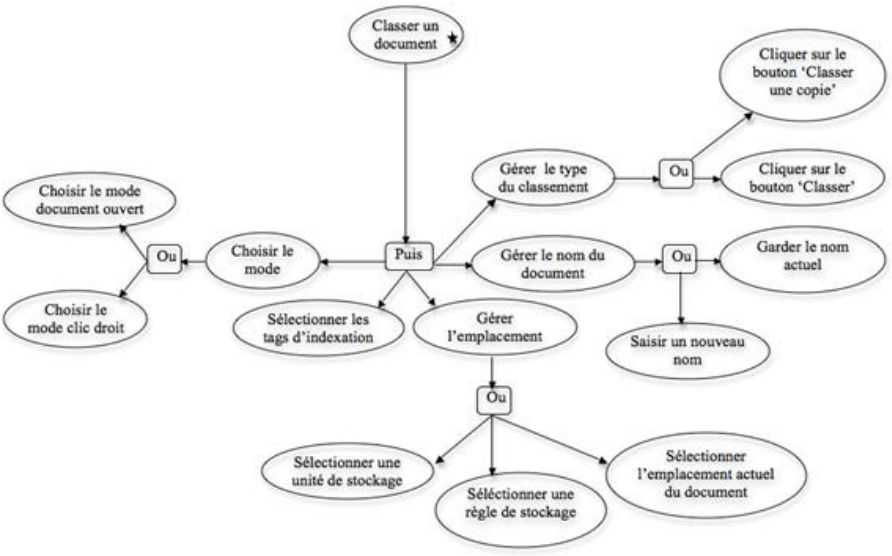

FIGURE 3.3 – Modèle de tâches « Classer un document »

— Chercher un document est une tâche itérative (décorée par une étoile) composée de trois sous-tâches réalisées successivement : Cliquer sur le bouton de recherche à partir du menu de démarrage du système d'exploitation, sélectionner les tags qui représentent les annotations (forment des métadonnées du document) avec les quelles le document était classé et choisir finalement le document cherché à partir de la liste des documents ayant les mêmes tags. La figure 3.4 représente le modèle de cette tâche.

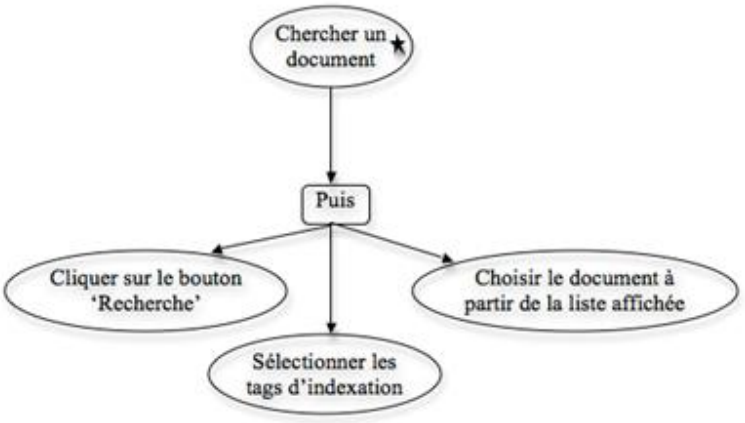

FIGURE 3.4 – Modèle de tâches « Chercher un document »

3.3.2 Modèle des concepts

Le modèle des concepts représente les entités du domaine manipulées dans les tâches, comme le modèle des classes dans le langage de modélisation UML, il s'appuie sur les relations d'héritage et d'association. Une entité est représentée par une classe qui contient des attributs. Elle peut être héritée ou associée à une ou plusieurs classes comme le montre le méta-modèle dans la figure 3.5.

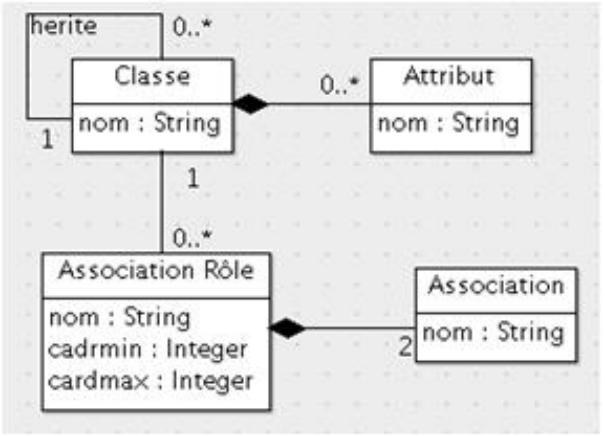

FIGURE 3.5 – Méta-modèle des concepts [SATTEL et al. 2005]

Nous instancions le méta-modèle de concepts pour avoir un modèle des concepts ma-

nipulés dans le prototype HyperTagging. Comme le montre la figure 3.6, ce modèle est composé de classes Vue, Facette, Tag, Document, Unité de stockage et Règle de stockage en respectant ces contraintes :

— une vue regroupe un ensemble (éventuellement vide) de facettes ;
— une facette est constituée d'un ensemble (éventuellement vide) de tags ;
— un tag est associé à un ensemble (éventuellement vide) de documents ;
— une règle de stockage est composée par un ensemble (au minium un élément) des tags et une unité de stockage ;
— un document peut être classé directement dans une ou plusieurs unités du stockage ou par une ou plusieurs règles de stockage.

Inversement,

— un tag appartient à une et une seule facette ;
— une facette peut ne pas être associée à une vue spécifique, ou être associée à une ou plusieurs vues ;
— un tag peut appartenir à une ou plusieurs règles de stockage ;
— une unité de stockage peut appartenir à une ou plusieurs règles de stockage.

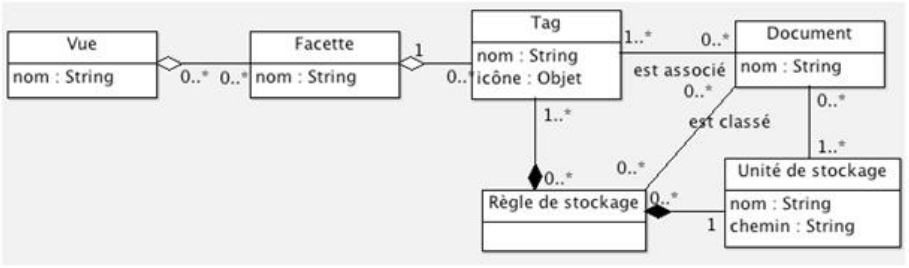

FIGURE 3.6 – Modèle des concepts du prototype HyperTagging

3.3.3 Modèle des espaces

Dans le modèle des espaces qui dérive du modèle de tâches, les espaces sont identifiés sur la base des tâches d'utilisateur. Ils contiennent les concepts manipulés dans ces tâches et s'enchainent conformément aux tâches représentées. Un espace peut représenter l'espace de commencement ou/et de l'achèvement des tâches. Le méta-modèle suivant représente

les relations entre les deux composants « espace » et « enchainement » de ce modèle.

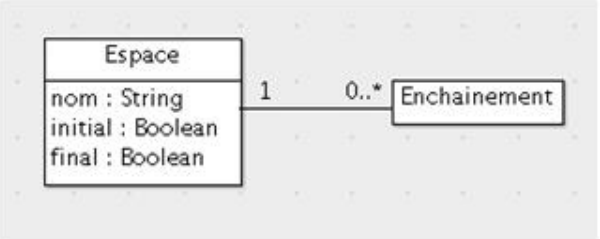

FIGURE 3.7 – Méta-modèle des espaces de travail [SATTEL et al. 2005]

Nous instancions le méta-modèle des espaces pour modéliser les espaces de réalisation des tâches et sous-tâches identifiées précédemment. Un espace de travail est associé par tâche et l'enchaînement entre espaces calque fidèlement les opérateurs entre tâches. En effet :

- pour le cas d'un opérateur binaire de type 'OU' entre tâche mère et ses sous-tâches, les sous-espaces redonnent accès à l'espace père une fois la sous-tâche accomplie ;
- pour le cas d'un opérateur binaire de type 'ET' entre tâche mère et ses sous-tâches, les sous-espaces redonnent accès à l'espace père une fois toutes les sous-tâche accomplies ;
- pour le cas d'un opérateur temporel de type 'Puis' entre tâche mère et ses sous-tâches, l'espace père donne accès au premier fils qui donne accès au espace fils suivant, récursivement jusqu'au dernier fils qui donne accès à l'espace père ;
- pour le cas d'un opérateur temporel de type 'Avec' entre tâche mère et ses sous-tâches, l'espace père donne accès aux sous-espaces qui redonnent accès à l'espace père une fois toutes les tâches accomplies.

Les figures suivantes représentent les modèles des espaces associés aux tâches mères ainsi que leurs sous-tâches :

- Ce modèle est composé d'une tâche mère « Créer un jeu de facettes » et deux sous-tâches, « Créer les vues » et « Gérer les facette », successives ;

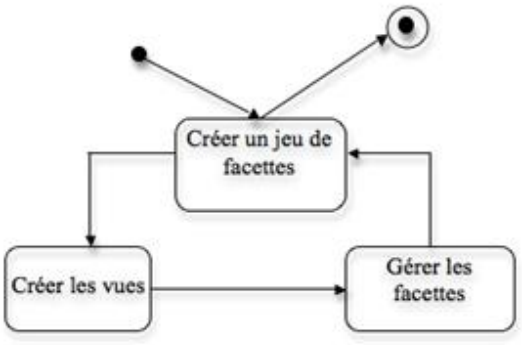

FIGURE 3.8 – Modèle d'espace de la tâche « Créer un jeu de facettes »

— La sous-tâche « Gérer les facettes » est effectuée par la réalisation de l'une des sous-tâches « Créer des nouvelles facettes » et « Associer à une vue des facettes existantes » ;

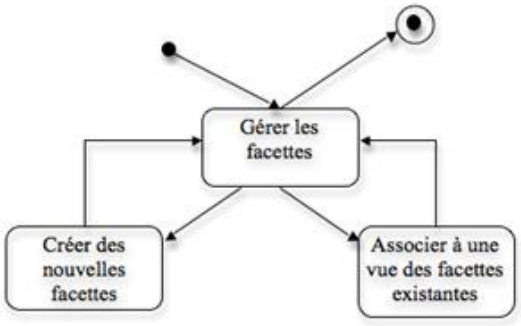

FIGURE 3.9 – Modèle d'espace de la sous-tâche « Gérer les facettes »

— Le modèle suivant représente la sous-tâche « Créer des nouvelles facettes » composée de deux sous-tâches successives « Créer le squelette de la facette » et « Créer l'ensemble de tags de la facette » ;

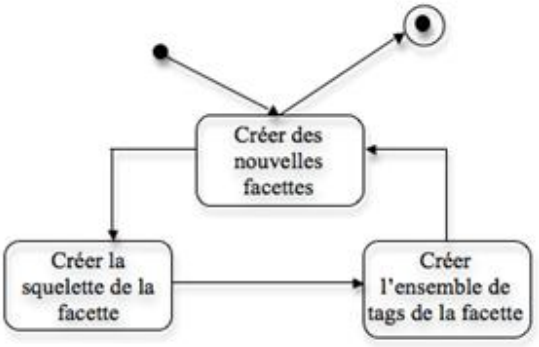

FIGURE 3.10 – Modèle d'espace de la sous-tâche « Créer des nouvelles facettes »

— La tâche « Classer un document » est composée d'une succession des sous-tâches « Choisir le mode », « Sélectionner les tags d'indexation », « Gérer l'emplacement », « Gérer le nom du document » et « Gérer le type du classement » ;

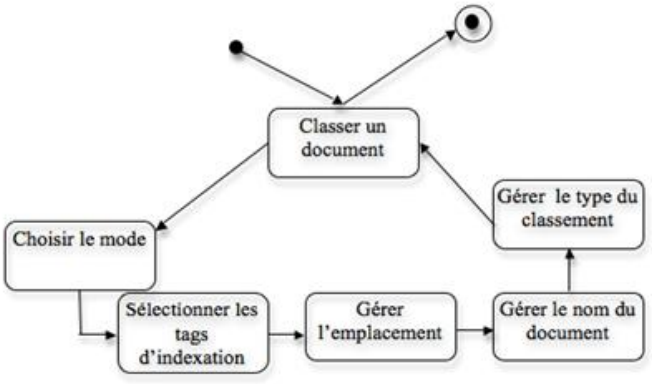

FIGURE 3.11 – Modèle d'espace de la tâche « Classer un document »

— « Choisir le mode » est une sous-tâche du modèle précédent, elle est réalisée si la sous-tâche « Choisir le mode document ouvert » ou « Choisir le mode de clic droit » est réalisée ;

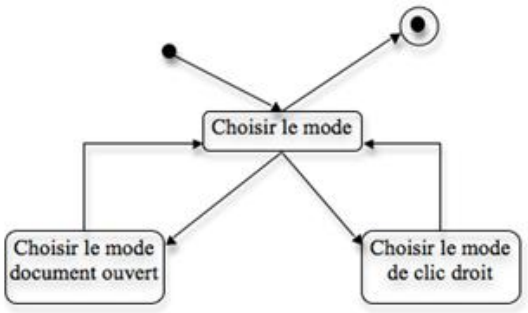

FIGURE 3.12 – Modèle d'espace de la sous-tâche « Choisir le mode »

— « Gérer le type du classement » est réalisée par l'achèvement de l'une des sous-tâches suivantes : « Cliquer sur le bouton 'Classer une copie' » ou « Cliquer sur le bouton 'Classer' » ;

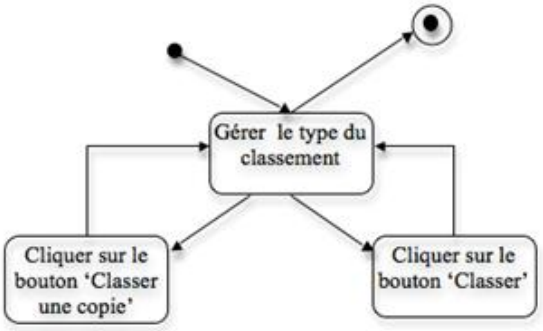

FIGURE 3.13 – Modèle d'espace de la sous-tâche « Gérer le type de classement »

— « Gérer l'emplacement » est effectuée par la réalisation de la sous-tâche «Sélectionner une unité de stockage », « Sélectionner une règle de stockage » ou « Sélectionner l'emplacement actuel du document » ;

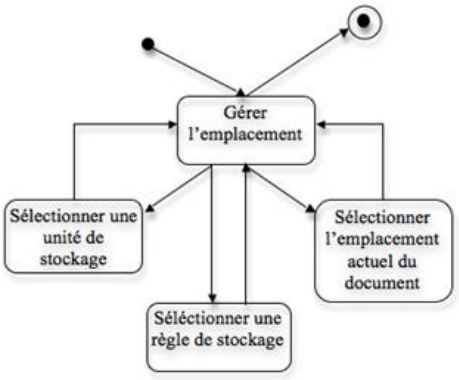

FIGURE 3.14 – Modèle d'espace de la sous-tâche « Gérer l'emplacement »

— Deux façons permettent de « Gérer le nom du document », qui sont « Garder le nom actuel » ou « Saisir un nouveau nom » ;

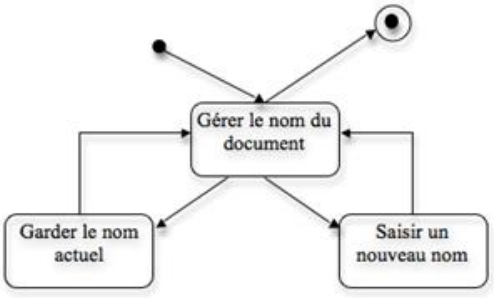

FIGURE 3.15 – Modèle d'espace de la sous-tâche « Gérer le nom du document »

— Pour « Chercher un document» il faut « Cliquer sur le bouton Recherche », par la suite « Sélectionner les tags d'indexation » et finalement « Choisir le document à partir de la liste affichée ».

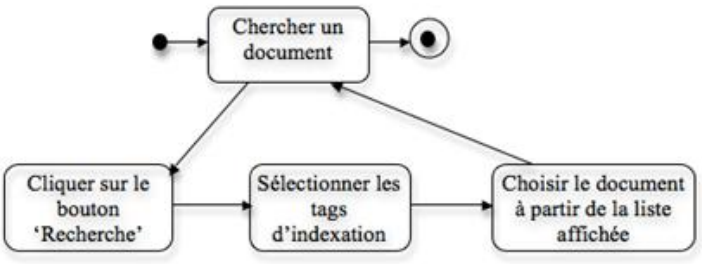

FIGURE 3.16 – Modèle d'espace de la tâche « Chercher un document »

3.3.4 Modèle des interacteurs

Le modèle des interacteurs détermine la représentation graphique des espaces de travail (fenêtres ou canevas), les concepts du domaine (champs du texte, listes, etc.) et les enchainements entre espaces (onglets, liens hypertexte, boutons, etc.).
Les espaces de travail deviennent des conteneurs, les concepts et les opérations applicables deviennent des interacteurs et l'enchainement entre les espaces devient une suite des interacteurs de navigation. Les interacteurs de navigation et les conteneurs sont des interacteurs comme le montre le méta-modèle suivant :

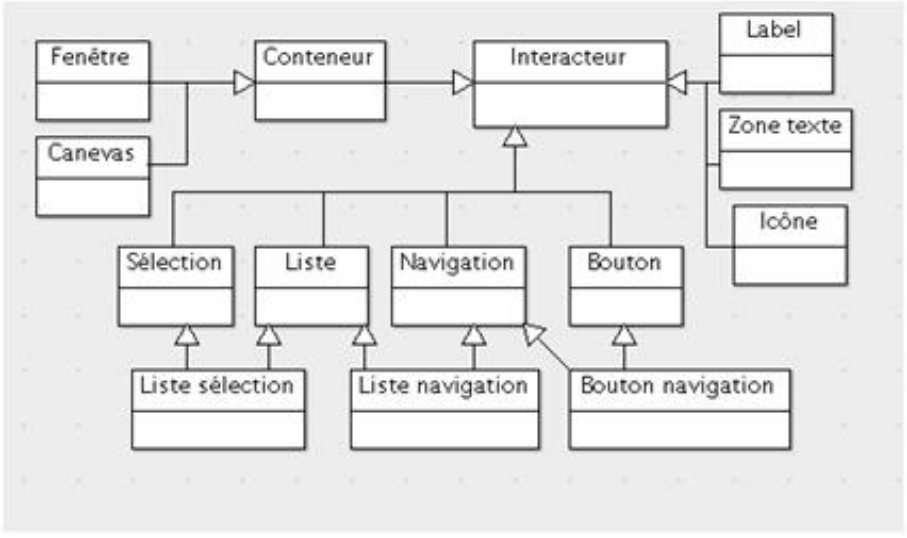

FIGURE 3.17 – Méta-modèle des interacteurs

Nous instancions ce méta-modèle pour avoir deux modèles des interacteurs d'Hyper-Tagging. Chaque modèle est composé principalement par une fenêtre qui contient deux

conteneurs dont le premier est commun aux deux modèles. Ce conteneur en commun contient les interacteurs permettant une création évolutive des vues, des facettes et des tags.

En effet, la structure de classification (le jeu de facettes) évolue au cours du temps puisque l'utilisateur peut à tous moments recréer des nouveaux éléments (vue, facette, tag). Le deuxième conteneur dépend de la tâche souhaitée, pour classer un document, il contient des interacteurs permettant la gestion du nom du document (la zone texte), la sélection du placement de stockage, la sélection de la règle de stockage (liste) et le choix de type de classement (les boutons) comme le montre la figure suivante :

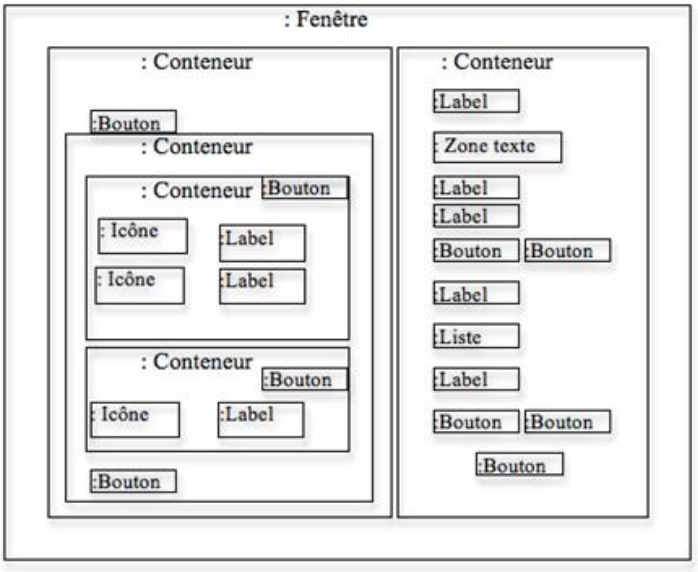

FIGURE 3.18 – Modèle des interacteurs pour la création d'un jeu de facettes et le classement d'un document

Et pour chercher un document, ce deuxième conteneur contient une liste des documents comme représentée dans la figure suivante :

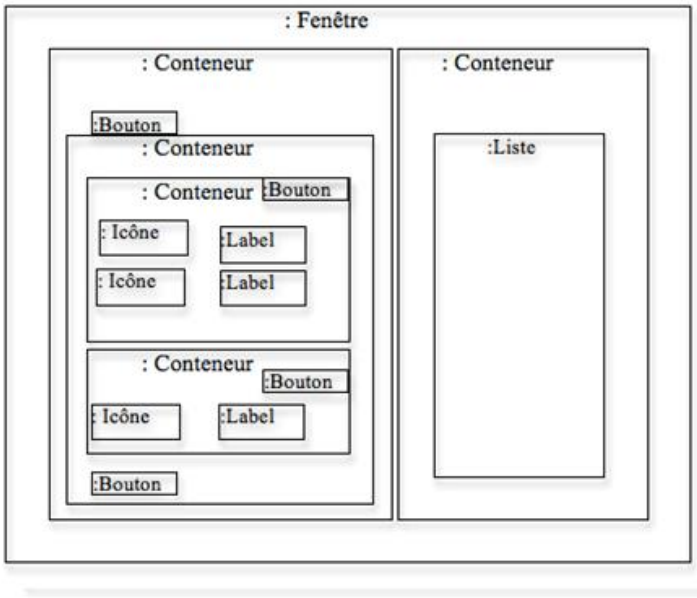

FIGURE 3.19 – Modèle des interacteurs pour la création d'un jeu de facettes et la recherche d'un document

3.3.5 Modèle des programmes

Le modèle des programmes consiste à représenter l'IHM avec un langage de programmation. Les deux modèles suivants représentent l'implémentation des modèles des interacteurs :

— Ce modèle de programmation permet la création du jeu de facettes et le classement des documents :

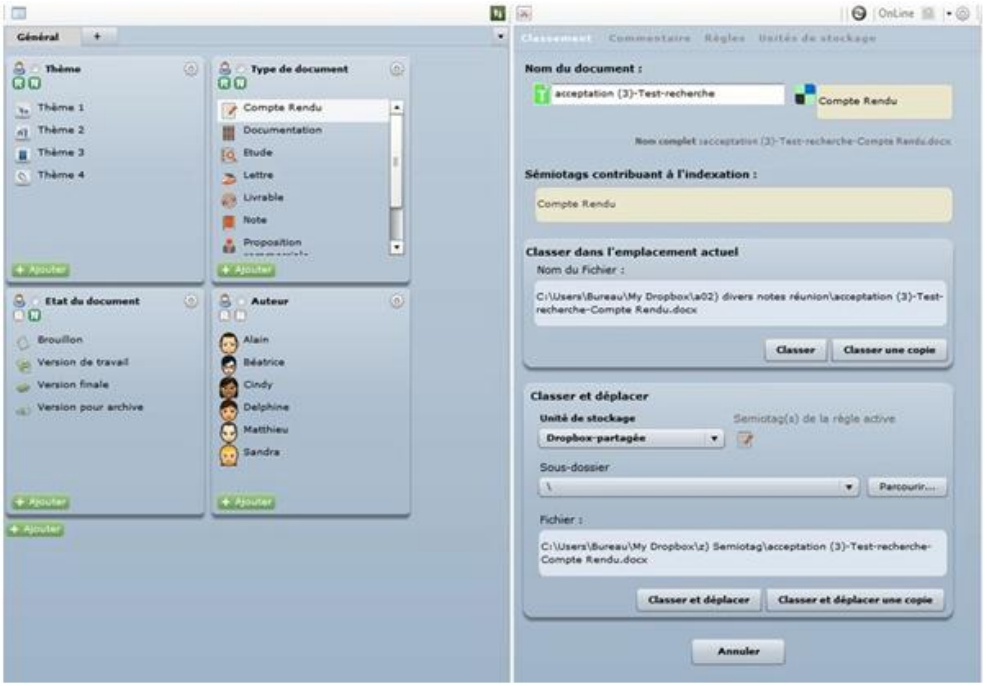

FIGURE 3.20 – Modèle programmé de l'IHM d'HypeTagging : création du jeu de facettes et classement des documents

— Ce modèle de programmation permet la création du jeu de facettes et la recherche des documents :

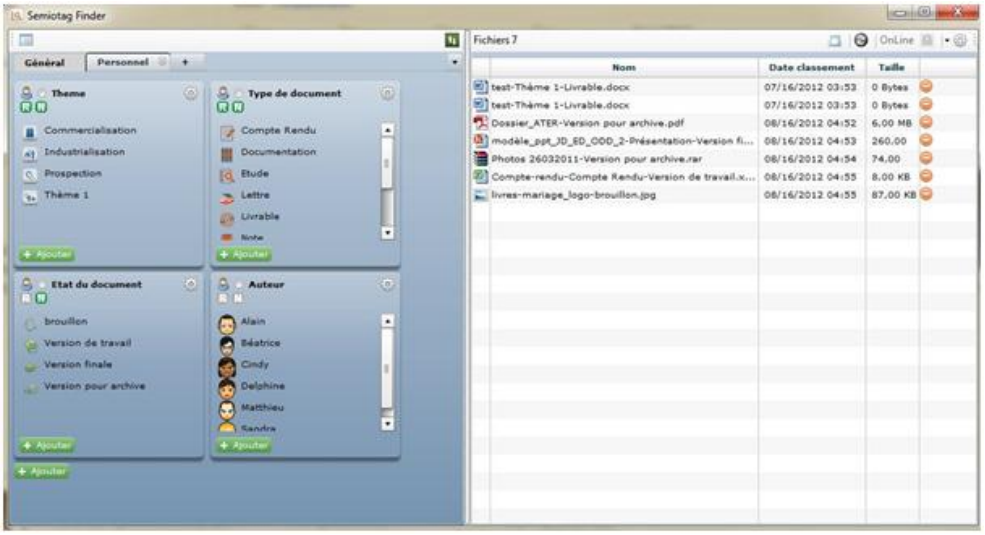

FIGURE 3.21 – Modèle programmé de l'IHM d'HyperTagging : création du jeu de facettes et recherche des documents

3.4 Modélisation du SOC et son évolution

3.4.1 Les principes méthodiques et technologiques de la gestion de l'évolution du SOC

La modélisation que nous proposons est orientée infrastructure avec différentes perspectives. En effet les composants de notre SOC multidimensionnel sont aussi des composants de l'infrastructure du SI de la documentarisation classificatoire.
Nous nous positionnons dans le thème émergent de l'ingénierie des évolutions des systèmes d'information. Ce thème de recherche englobe l'analyse et la conception de systèmes d'information devant faire face à des évolutions concernant les besoins fonctionnels et non fonctionnels, les principes organisationnels, les processus métiers et les plateformes techniques [AKOKA et COMYN-WATTIAU 2009]. Dans cette thématique, nous retenons des principes méthodiques et technologiques pour la gestion des évolutions au sein d'un SOC en nous focalisant sur trois perspectives inter-corrélées : le travail collaboratif, la classification et les métadonnées, et le partage des ressources.

Dans la perspective du travail collaboratif, Karasti et ses collègues se focalisent [KARASTI et al. 2010] sur les exigences de durabilité et évolutivité des composants du SI. Ils introduisent la notion du « long now » qui apporte une vision organisationnelle de la planification et de l'évolution des infrastructures collaboratives et distribuées. Cette vision corrèle plusieurs dimensions (institutionnelle, organisationnelle et technologique), dans lesquelles la perspective à long terme est intégrée dès les premières phases de la conception. Ces principes nous semblent particulièrement pertinents pour les SOC de grandes entreprises, dont le rôle institutionnel a une portée à large échelle (régionale, nationale ou internationale).
Afin de formaliser les problèmes d'hétérogénéité sémantique liées à des données dont les descriptions évoluent dans le temps, nous intégrons la notion du long terme dans notre conception par une perspective de classification et des métadonnées focalisées sur le concept de « cycle de vie ». Ce concept est approprié pour gérer dans le temps les spécifications de vues, facettes et tags évolutifs, dans des SOC d'entreprise distribués et hétérogènes.

Pour assurer le partage des composants du SOC et de documents qu'il gère, nous nous focalisons sur la notion d'URI (Uniform Resource Identifier), un identificateur qui définit l'emplacement d'une ressource par une adresse URL (Uniform Resource Locator) ou son nom par une adresse URN (Uniform Resource Name). Nous les considérons comme une donnée descriptive des composants du SOC et des documents qu'ils gèrent dont les historiques doivent aussi être gérés explicitement. La propriété de « persistance des URI » garantit que chaque URI identifiant une ressource continue dans le temps à identifier cette ressource pour fournir une représentation consistante de la ressource. La persistance d'un URI est d'autant plus cruciale que la valeur de l'URI est élevée c'est-à-dire que les références à cet URI sont nombreuses. Pour le W3C « La gestion des représentations par les propriétaires d'URI est source d'attentes sociales » [W3C-ARC 2006].
Face aux différentes raisons poussant les gestionnaires de ressources à changer d'URI (par exemple dans le cas de fusions de sociétés), plusieurs recommandations permettent de limiter ces changements ou, le cas échéant, de les gérer. Elles comportent des aspects organisationnels, conceptuels et techniques, nécessairement corrélés.
Les recommandations méthodologiques rentrent dans le cadre d'adoption d'une perspective de conception sur le très long terme, avec des règles de nommage les plus stables possibles et indépendants des systèmes de classifications ; spécifications de métadonnées décrivant les URI, notamment en liaison aux dates de création et d'expiration de l'URI et de la ressource identifiée.
Sur le plan technologique, HTTP a été conçu pour renforcer la consistance des représentations, notamment par des mécanismes de redirection et de négociation de contenus [W3C-ARC 2006]. Cependant, ces mécanismes doivent eux-mêmes s'inscrire dans la durée et s'accompagner de dispositifs de gestion de correspondances entre URI.

3.4.2 Modèle conceptuel pour la gestion des évolutions du SOC

A partir de ces principes méthodologiques et technologiques, nous nous basons sur le modèle de concepts du prototype HyperTaging modélisé précédemment dans l'approche conceptuelle de l'IHM de ce prototype, qui spécifie les relations entre les composants du SOC : vue, facette et tag. Pour détecter les éléments d'innovation issus des individus et

du collectif, nous utilisons les métadonnées pour décrire les éléments structurant le SOC et les documents qu'il gère.

Ce modèle comprend les concepts (les classes en UML) Document et ElémentSOC. Ce dernier se spécialise dans les concepts (les classes en UML) Vue, Facette et Tag reliées par des agrégations décrites précédemment.

Le concept MétadonnéeEvol se compose de deux concepts : l'URI plus précisément un URL et Métadonnées. Il se spécialise en deux sous-classes, MétadonnéesDoc et MétadonnéesEltSoc. La figure suivante représente ce modèle :

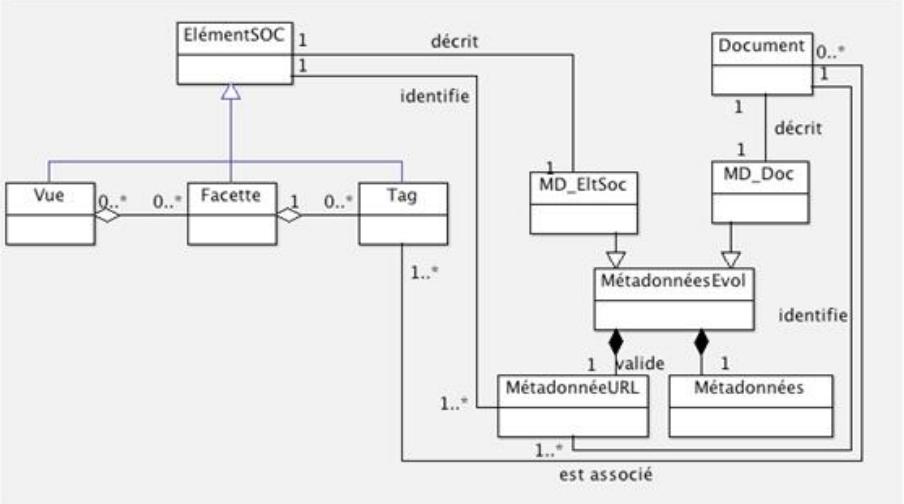

FIGURE 3.22 – Modèle de concepts : « Document » et « Elément du SOC »

3.4.3 Représentation des métadonnées

Pour la gestion des évolutions, les métadonnées s'appuient sur les concepts « URL » et « cycle de vie » qui permettent respectivement de :

— gérer l'historique d'identification de la ressource à travers les variables « validité » qui décrivent pour chaque URL son état (valide ou invalide), « date d'identification » qui contient la date d'association de l'URL à une ressource et « date fin de validité» qui correspond à la date de fin de validité de l'URL identificateur d'une ressource. Une ressource est associée à un ensemble des URL (au minimum un élément) dont un URL est obligatoirement valide et fait partie des métadonnées de cette ressource ;

— tracer la gestion des documents comme des éléments du SOC, en enregistrant les actions successives opérées sur eux, en fonction de leur type (par exemple, création, classement, modification ou suppression de documents, création/modification d'éléments de classification, ajout/suppression de tags à des facettes...).

Ainsi les métadonnées principales des documents et des éléments du SOC (vue, facette, tag) sont respectivement :

— pour les documents : un URL valide, le titre, un ensemble de mots-clés et, pour chaque action sur le document (création, classement, suppression, modification, ajout de commentaire), la date à laquelle cette action est exécutée, le (ou les) auteur(s) de l'action et le libellé de l'action ;

— pour les éléments du SOC : un URL valide, le nom et le type de l'élément, ainsi que, pour chaque action sur l'élément (création, suppression, modification), la date à laquelle cette action est exécutée, l'auteur de l'action et le libellé de l'action.

Ces métadonnées sont résumées dans les deux figures suivantes :

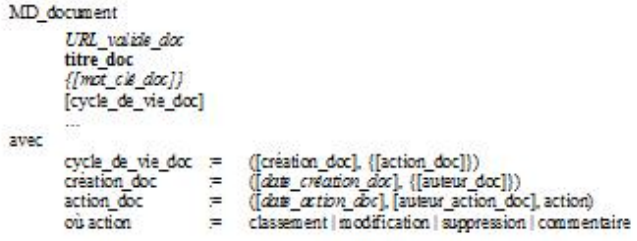

FIGURE 3.23 – Métadonnées des ressources « documents » [SALZANO et ANKOUD 2012]

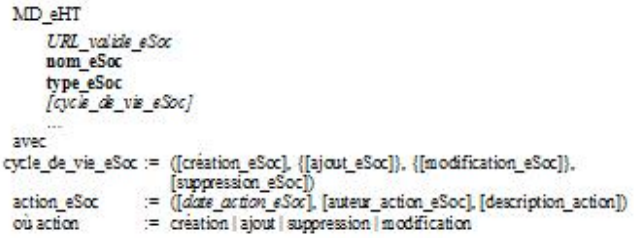

FIGURE 3.24 – Métadonnées des éléments du SOC [SALZANO et ANKOUD 2012]

Dans ces figures, on a représenté en gras, les métadonnées obligatoires et en italique, les métadonnées recommandées pour gérer les évolutions. Les notations ., [.] et (., ., .) représentent respectivement : la répétition d'un élément, un élément optionnel et une structure formée de plusieurs éléments.

Une certaine dynamique entre les classifications adoptées au niveau d'équipes ou de collectifs et les classifications adoptées au niveau global, d'entreprise, est introduite par les apports informationnels de ces métadonnées. Ces dernier consistent à :

— déterminer les contextes temporels de validité des documents et des Vues, Facettes et Tags. Par l'analyse des messages stockés dans la base de données des métadonnées, on peut déterminer, par exemple, les changements opérés dans le temps par des utilisateurs sur la configuration d'une vue, par la mise à jour (insertion, suppression) de tags dans une facette donnée ;

— favoriser à travers la gestion de leur cycle de vie l'évaluation des aspects collaboratifs du système d'information : on peut analyser les contributions des acteurs à la création, au partage et à la diffusion de documents et éléments de classification, en détectant les « leaders » de ces activités ;

— mieux manager l'intégrité des données gérées dans le SOC : par exemple, lorsqu'une facette est supprimée, une date de suppression est notifiée dans les métadonnées. Elle détermine le contexte de validité des tags associés à cette facette, qui ne seront pas supprimés.

3.5 Représentation architecturale et modélisation de la communication

3.5.1 Cadre méthodologique de l'architecture

Pour mettre en place l'approche orientée métiers, le SOC à facette d'indexation collaborative et la gestion des évolutions par les métadonnées, l'architecture système doit permettre l'accès multiples des ressources dans des environnements collaboratifs (au sein d'un service ou d'un département), organisationnels (pour toute l'entreprise) et inter-

organisationnels (entreprise étendue à ses partenaires). Dans ce cadre, l'interaction entre les usagers et la ressource se présente en trois phases, comme le montre la figure ci dessous d' « Interaction usagers-ressources » : la création de la ressource sous forme des données structurées ou non-structurées, la diffusion de la ressource entre les usagers (en utilisant un système de partage) et la gestion (modification, lecture, suppression) de la ressource.

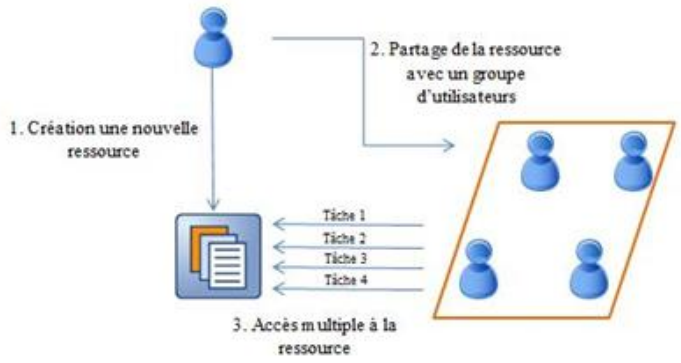

FIGURE 3.25 – Interaction usagers-ressources

L'exploitation des ressources partagées peut se faire dans l'une des situations suivantes, de la plus simple à la plus complexe :

- Utilisation des plusieurs ressources par un seul utilisateur dans des intervalles du temps différents ;
- Utilisation des plusieurs ressources par un seul utilisateur dans un seul intervalle de temps ;
- Utilisation d'une ressource par plusieurs utilisateurs dans des intervalles de temps différents ;
- Utilisation d'une ressource par plusieurs utilisateurs dans un seul intervalle de temps.

La figure suivante présente ces quatre cas dont la dernière représente le cas problématique, d'accès multiple, traité dans notre contribution :

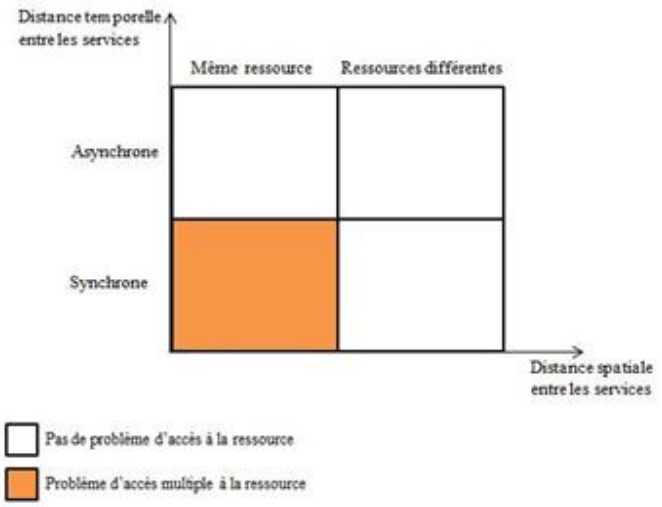

FIGURE 3.26 – Les cas d'accès à une ressource

A fin de gérer ce concept nous nous basons sur le mécanisme des services dans une architecture SOA que nous présentons dans la section suivante.

3.5.2 Choix de l'architecture

3.5.2.1 L'architecture orientée services SOA

SOA n'est pas un standard mais un style d'architecture qui s'appuie sur des standards. Il représente un modèle de conception qui définit la structure, les composants et les connecteurs des systèmes logiciels en spécifiant le vocabulaire de communication, les règles d'interconnexion et les propriétés des composants et des connecteurs. Il existe trois styles architecturaux : architectures orientées objets, architectures orientées ressources et architectures orientées services. Nous nous intéressons dans cette étude au dernier type qui peut s'appliquer à toutes les technologies et pour tout type de réalisation. Face au succès du déploiement d'Internet dans les entreprises et dans le public, SOA (Service-Oriented Architecture) est devenu plus populaire en 2005. C'est une architecture centralisée, qui permet au client de rechercher, consulter et consommer des services. Ces derniers sont des fonctions logicielles autonomes, qui acceptent, traitent des requêtes et fournissent des résultats. Ils utilisent un protocole normalisé pour communiquer entre eux [COUTURIER

2011][FIGER 2006].

Cette architecture se caractérise par :

— Un couplage lâche entre les services pour ne pas dépendre de l'état d'autres services ;

— L'utilisation de standards ouverts ;

— L'interopérabilité inter-services ;

— Les services sont sans état pour faciliter l'extensibilité et l'orchestration ;

— Les services sont exposés et accessibles par d'autre services et applications ;

— Les services sont distribués pour que l'on puisse les installer à l'intérieur et à l'extérieur des organisations [COUTURIER 2011][FIGER 2006].

D'après COUTURIER [2011], la SOA est composée de trois couches : une couche d'affaires, une couche des services et une couche applicative comme le montre la figure suivante :

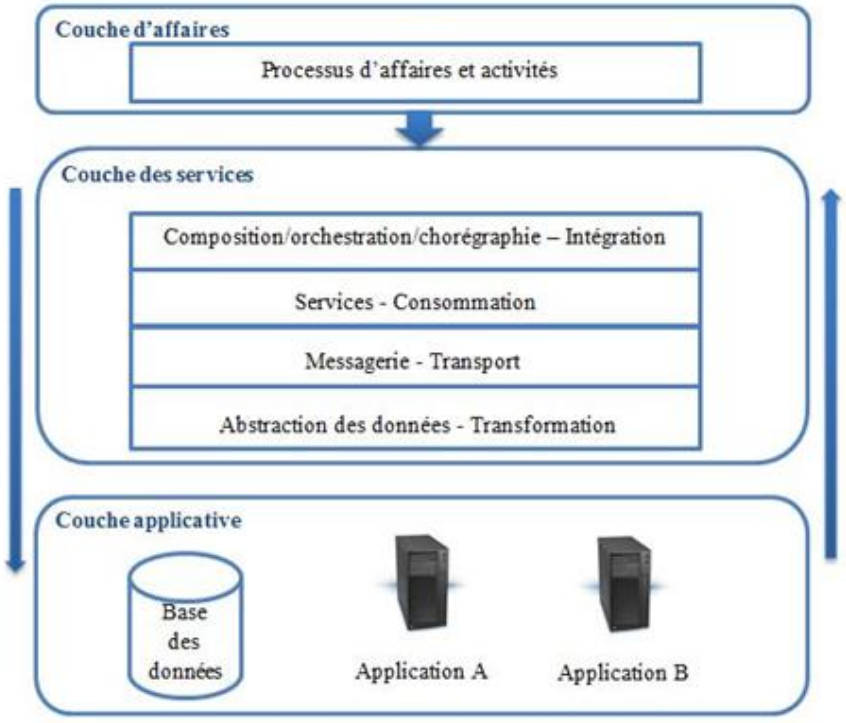

FIGURE 3.27 – Les couches d'une architecture SOA

— Un couplage lâche entre les services pour ne pas dépendre de l'état d'autres services ;

— La couche d'affaires : représente les processus d'affaires et les activités de l'organisation qui utilisent cette architecture ;

— La couche des services : est subdivisée en sous-couches. La première est la couche de transformation qui utilise des services applicatifs pour interagir avec la couche applicative, extraire les données et les transformer en un format standardisé. La deuxième couche est la messagerie, elle assure le transport des données. La troisième est la couche qui contient les services d'utilisation des données. Ces deux dernières couches forment les services d'affaires qui réalisent les activités élues de la couche d'affaires. La dernière permet de structurer, regrouper et communiquer les services entre eux, par composition, orchestration (coordination entre les services) ou chorégraphie (la gestion d'échange entre les acteurs : partenaires, clients ou fournisseurs) ;

— La couche applicative : contient la base des données, les applications et les systèmes auxquels on souhaite donner accès. Elle est située en dessous de la couche « transformation ».

3.5.2.2 Les services

Un service correspond à un périmètre fonctionnel que l'on souhaite exposer à un certain type de consommateurs (ensemble de fonctionnalités qui font sens). Il expose un petit nombre d'opérations offrant un traitement de bout en bout. Nous étudions deux types différents de service : service basé sur SOAP et service basé sur REST.

3.5.2.2.1 Les services du style d'architecture SOAP

Ce type de service repose sur un ensemble de protocoles et de standards pour échanger les données entre différentes applications : WSDL, SOAP et UDDI.

— WSDL (Web Service Description Langage) : est un standard qui offre un langage de description des services, utilise le format XML. Il définit l'interface du service ainsi que les paramètres d'appel de service ;

— SOAP (Simple Object Access Protocol) : est un protocole, défini par W3C (World Wide Web Consortium), assure la communication interservices entre les applications par l'échange des messages XML en utilisant le protocole http, SMTP ou FTP. Ce protocole contient trois parties :
  — « L'enveloppe qui définit ce que contient le message et comment il doit être traité ;
  — Un ensemble de règles d'encodage permettant de représenter des formats de données définis par des applications tierces ;
  — Une convention pour se représenter les appels et les réponses distants (à l'aide de message RPC (Remote Procedure Call) ou de message de type document). » [COUTURIER 2011]

— UDDI (Universal Description, Discovery and Integration) : est un annuaire de services qui contient la description de chaque service. Un fournisseur enregistre et publie son service dans l'UDDI. Le client peut chercher, par une description fournie grâce à WSDL, le service qui répond à ses besoins, il peut le consommer par la suite via le protocole SOAP [COUTURIER 2011][FIGER 2006].

Une architecture fondée sur ce type de services appartient au style d'architecture orienté services SOAP.

3.5.2.2.2 Les services du style d'architecture REST

Les services REST sont des services fondés sur le protocole de communication http pour échanger des messages XML, sur le standard URI comme syntaxe universelle pour adresser les ressources et sur le type MIME (text/xml, text/html, image/jpeg...) pour présenter les ressources [COUTURIER 2011][FIGER 2006].
Le style d'architecture orientée ressources et services, REST, inventée par Roy Thomas Feilding en 2000, se caractérise par :

— Un modèle de type client/serveur, en séparant l'interface utilisateur et les données ;

— La communication entre le client et le serveur est sans état, ce qui donne trois propriétés importantes : la visibilité (pour exécuter une requête, le système n'utilise que des informations simples), la fiabilité (il est plus facile de gérer les échecs partiels) et la faculté de monter en charge (comme il n'y a pas de stockage des requêtes, le serveur libère rapidement des ressources) ;

— L'utilisation de cache est la performance majeure de cette architecture. Dans cette technologie les données de réponses doivent être marquées par la possibilité de les mettre en cache ou non. Le cache contient les données de réponses qui sont réutilisables pour des requêtes équivalentes, ce qui élimine certaines interactions, améliore l'efficacité et la montée en charge ;

— L'utilisation d'une interface unique (un connecteur) entre les composants qui permet de simplifier l'architecture globale du système et améliore les interactions en optimisant le transfert et normalisant les données. L'interface unique est le point fondamental qui distingue l'architecture REST des autres architectures, mais elle n'est pas efficace pour toutes les interactions (efficace pour le transfert des données hypermédia et le web classique) ;

— Une structure en couches hiérarchiques, chaque couche n'est visible qu'aux composants immédiats avec lesquels elle interagit. Les couches sont indépendantes et la complexité globale du système est limitée. Chaque couche peut encapsuler des services, faciliter l'exécution des anciens et protéger des nouveaux en créant des espaces intermédiaires partagés contenant les fonctionnalités rarement utilisées ;

— Le support de la contrainte facultative «Code à la demande », cette architecture « permet l'extension des fonctionnalités d'un client par le biais de téléchargement et d'exécution de code sous forme d'applet ou de scripts ». Le téléchargement des fonctionnalités simplifie les clients, réduit le nombre des interactions et améliore l'efficacité. [FIELDING 2000]

Indépendamment de la mise en œuvre et de la description des protocoles utilisés, REST définit trois éléments architecturaux fondamentaux en englobant les contraintes d'interaction entre eux [FIELDING 2000] :

— Les éléments des données : La communication dans cette architecture est basée sur le transfert d'une représentation d'une ressource dans un format standard, en passant par les interfaces normalisées, et le partage de type des données transmises par les métadonnées. « *Le format de données d'une représentation est connu comme étant un type de média* » dont le type est simple ou composé. Une ressource (document, image) est identifiée par un identifiant (URI) qui sera par la suite mis en relation avec la représentation. Cette dernière est composée des données, métadonnées sur les données et/ou métadonnées sur la représentation. REST utilise aussi des données de contrôle pour spécifier le type de communication entre les composants, le type des messages transférés, l'action demandée et la signification de la réponse.
— Les connecteurs : REST utilise plusieurs types de connecteurs, ces derniers offrent une interface de communication entre les composants, c'est-à-dire que chaque composant possède au minimum un connecteur pour arriver à communiquer avec des autres :
  — Connecteur client : le client envoie une demande à travers ce connecteur ;
  — Connecteur serveur : le serveur est à l'écoute des demandes. Un composant peut avoir les deux connecteurs à la fois (connecteur client et connecteur serveur ;
  — Connecteur serveur-cache ou client-cache : le cache contient les réponses réutilisables pour un composant ;
  — Connecteur résolveur : « *le résolveur traduit les identifiants partiels ou complets de ressource en information d'adresse réseau nécessaire pour établir une connexion entre composants* » ;
  — Connecteur tunnel : « *il relaie simplement la communication à travers une connexion ayant des limites, comme un pare-feu ou une passerelle réseau de bas niveau* ».

— Les composants :
  — Un agent utilisateur : comme le navigateur Web, il utilise le connecteur client pour envoyer une demande et recevoir la réponse ;
  — Serveur d'origine : utilise le connecteur serveur, il reçoit les requêtes et envoie les réponses aux demandeurs, il est la source définitive des représentations des

ressources ;

— Passerelle : est un intermédiaire imposé par le réseau ou par le serveur d'origine ;

— Serveur mandataire : est un intermédiaire choisi par le client.

3.5.2.2.3 SOAP vs REST

Sur le niveau de performance et de fiabilité, il n'y a pas une différence entre les deux architectures puisque toutes les deux transfèrent les données à travers le protocole http et sous la forme XML. Au niveau du déploiement, développement et test, REST est le plus simple, pour tester une requête. Il suffit d'utiliser la barre d'adresse du navigateur contrairement à SOAP qui demande la mise en place d'un langage de programmation et d'une infrastructure convenables.
La différence majeure se manifeste dans la façon d'accéder à une ressource demandée par une requête. Dans REST, le nom d'une ressource est précisé dans l'URI, donc l'action demandée est directement dirigée vers le service concerné (voir la figure 3.28). Concernant le style SOAP, les actions demandées sont cachées dans l'enveloppe (voir la figure 3.29) pour des raisons de sécurité et de confidentialité qui sont le point fort de cette architecture mais qui entrainent des inconvénients en même temps puis qu'on se trouve dans un couplage fort entre la sécurité et l'application. Tous les messages sont envoyés vers un point d'entrée unique, ce qui complique la gestion des performances et des priorités en amenant à un couplage fort entre les services au niveau des priorités d'accès. La simplicité du mécanisme du REST permet de mettre en place un système de contrôle d'accès aux ressources facilement, avec une indépendance totale entre l'autorisation et l'application [COUTURIER 2011][FIGER 2006].

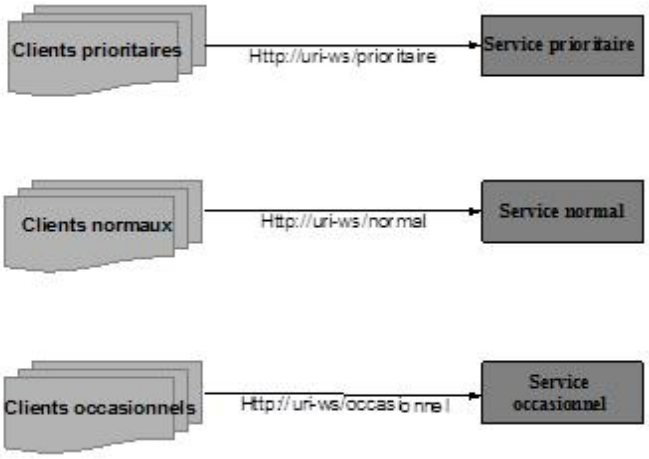

FIGURE 3.28 – Nommage de la ressource par REST

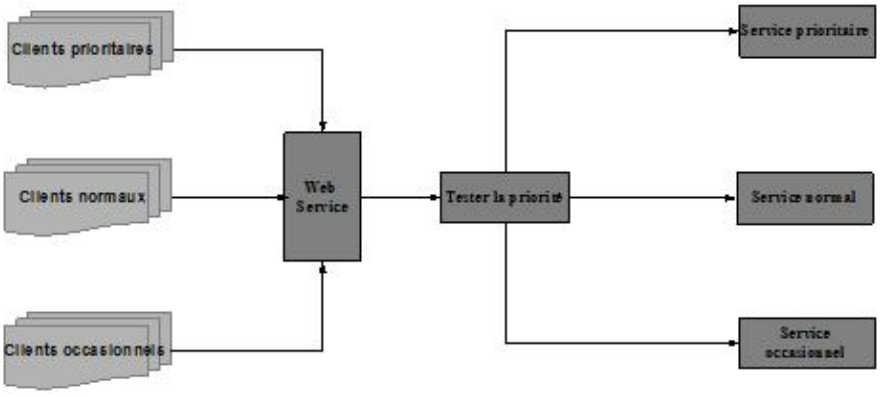

FIGURE 3.29 – Nommage d'une ressource par SOAP

Dans le tableau suivant nous résumons les différences et les similarités entre REST et SOAP sur lesquelles nous nous appuyons pour choisir l'architecture convenable à notre projet.

Après cette comparaison et pour des raisons de simplicité, fiabilité et performance, nous choisissons d'étudier l'architecture REST pour trouver les meilleures approches et modèles applicables dans notre projet.

3.5.3 Cadre technologique de l'architecture

Du point de vue technologique, cette démarche architecturale orientée services (dite SOA : Service Oriented Architecture) est basée sur des standards informatiques qui as-

	REST	SOAP
Performance	Il n'y a pas de différence, les deux sont performantes	
Fiabilité	Il n'y a pas de différence, les deux sont fiables	
Déploiement	Simple	Complexe
Développement	Simple	Complexe
Test	A travers la barre d'adresse du navigateur (simple)	A travers un langage de programmation (complexe)
L'accès à une ressource demandée	Chaque demande est dirigée directement vers le service concerné => la simplicité de la gestion des demandes	Toutes les demandes sont envoyées vers un point d'entrée unique => la complexité de la gestion des demandes
Transfert des données	Par le Protocol http et en format XML	

TABLE 3.1 – Comparaison entre REST et SOAP

surent l'interfaçage entre nos composants systèmes.

L'architecture de notre SI s'appuie sur trois composants fonctionnels majeurs. Les composants Saver et Finder, dédiés respectivement aux fonctions de nommage, indexation et classement (Saver) et de recherche (Finder), nécessitent un composant de gestion des métadonnées (vues, facettes et tags). Chaque composant est géré par un service spécialisé.

Pour pallier le problème de l'hétérogénéité des métadonnées, cette architecture s'appuie sur des métadonnées externes. Deux types de métadonnées sont stockées dans la base de données : des métadonnées relatives aux documents classés et des métadonnées, en format XML, d'éléments du SOC (vues, facettes, tags), dites aussi éléments système dans la figure suivante qui présente une vue des composants fonctionnels de cette architecture.

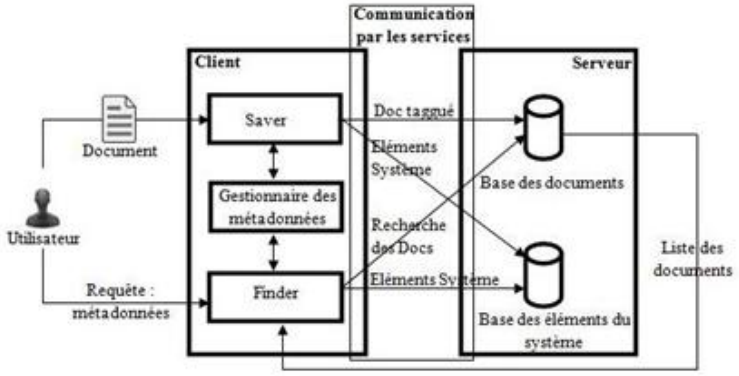

FIGURE 3.30 – Composants fonctionnels de l'architecture

Elle est basée sur les services REST. Ces derniers assurent une communication entre les composants et les bases de données du système. Les services gèrent aussi les métadonnées. Pour des raisons d'efficience et de réutilisation des requêtes d'interrogation des données, ils ont besoin de s'appuyer sur un protocole qui assure la communication et le transfert des données entre les composants.

3.5.4 Modélisation du style architectural REST

Nous continuons dans la même perspective conceptuelle incrémentale. Notre objectif dans cette section est de représenter un modèle conceptuel du REST qui décrit les composants architecturaux et la communication entre eux. Pour construire ce modèle, nous nous basons sur des travaux existants. En effet, notre démarche est composée de trois phases :

— nous décrivons les modèles du REST existants ;
— à partir des lacunes des autres modèles, nous améliorons l'existant pour concevoir le concept d'accès multiples aux ressources ;
— finalement nous instancions un modèle sur lequel l'architecture d'HyperTagging sera fondée.

3.5.4.1 Approche d'Alarcon et Wilde

3.5.4.1.1 Cadre méthodologique de l'approche

Les services REST sont très étendus dans le domaine industriel, grâce à leur évolutivité élevée garantie par une conception à couplage lâche, leur simplicité d'implémentation et leur ouverture aux différentes normes et standards. En utilisant les principes fondamentaux de l'architecture REST (la priorité de la ressource, l'utilisation d'URI, la représentation des ressources avec des métadonnées), Alarcon et Wide ont introduit le concept de la représentation du service REST dans un méta-modèle.
Ce concept représente les composantes fonctionnelles, les différents états des interactions du client avec les différentes ressources et les liens nécessaires pour ces interdépendances. Il contient aussi une description du service associé et des contraintes d'exécution : les exigences et les restrictions. Les descriptions sont utilisées par la suite dans la documentation, la publication des fonctionnalités et/ou guide utilisateur [ALARCON et WILDE 2010].

3.5.4.1.2 Le méta-modèle REST d'Alarcon et Wilde

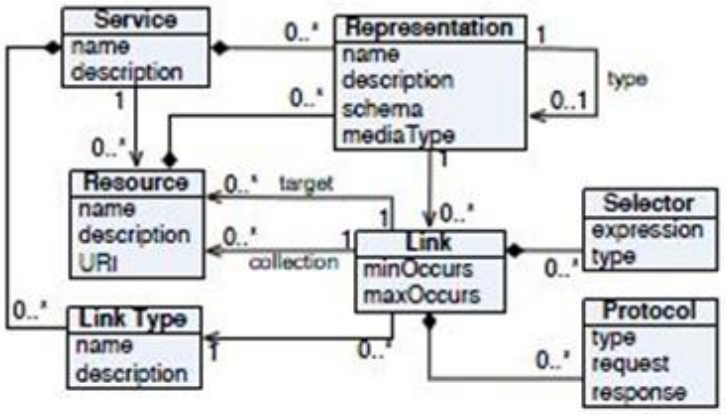

FIGURE 3.31 – Méta-modèle d'Alarcon et Wilde : Description du service REST [ALARCON et WILDE 2010]

Comme le montre la figure ci-dessus, ce méta-modèle est un modèle composé de sept concepts (classes en UML) : Service, Ressource, Representation, Link, Selector, Protocol et Link Type. Ils sont liés entre eux tout en respectant les règles suivantes :

— Un service (Service) peut fournir une ou plusieurs ressources dont chacune est identifiée par un URI ;
— Une ressource (Resource) peut avoir une ou plusieurs représentations (Representation) dont chacune peut contenir des liens (Link) vers d'autres ressources ;
— Chaque lien (Link) a un type (Link Type). Il peut être récupéré de la représentation en utilisant le sélecteur (Selector) ;
— La spécification des sélecteurs dépend du format de la représentation ;
— Les liens respectent les règles du protocole de la communication (Protocol) utilisé durant la demande et la réponse.

3.5.4.2 Approche de Schreier

3.5.4.2.1 Cadre méthodologique de l'approche

Pour décrire son approche de modélisation de l'architecture REST, Schreier s'est basé sur le choix de Stahl et ses collègues d'Ecore, choix fondé sur les éléments décrits dans le tableau 3.2.

Cette approche est composée de trois méta-modèles qui représentent séparément les types de ressources, la structure et le fonctionnement de l'architecture.

3.5.4.2.2 Méta-modèle de différents types de ressources

Comme il est expliqué précédemment, la communication REST est fondée sur le transfert de la représentation (données et métadonnées) d'une ressource identifiée par un URI. Elle utilise aussi les données de contrôle qui précisent le type de la communication et des messages transférés ainsi que les actions demandées. En se basant sur ce concept, Schreier présente une typologie des ressources et propose un méta-modèle qui englobe les relations entre les différents types.

Elément	Description
Eclass	Identifié par un nom ; Peut contenir un ensemble d'EAttributes, qui décrit ses propriétés ; Peut contenir un ensemble d'EReferences, qui décrit les relations avec les autres EClass ; Hérite tous les attributs et les références d'un classe mère ; Peut-être « abstract » pour ne pas créer une instance.
EdataType	Identifié par un nom ; Représente les types basiques comme « String » et « Integer ».
Eattribute	Identifié par un nom ; Son type est défini par EDataType ; Définit un nombre minimum et un nombre maximum de valeurs d'attribut.
Ereference	Identifié par un nom ; Représente les relations entre les EClass ; Définit le type et les multiplicités des relations.

TABLE 3.2 – Les éléments d'Ecore

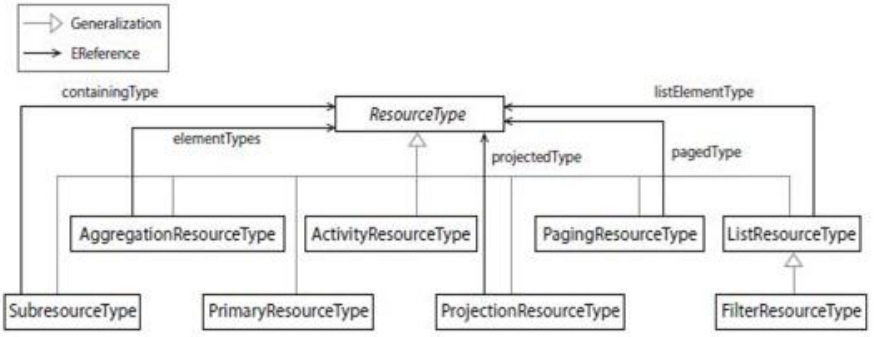

FIGURE 3.32 – Méta-modèle de Schreier : La hiérarchie et les relations des différents types de la ressource [SCHREIER 2011]

RessourceType est une EClass abstraite qui modélise le type d'une ressource de l'architecture REST. Elle a un attribut maxResources qui représente le nombre des ressources autorisées, variant entre 1 et l'infini.
SubressourceType, AggregationRessourceType, PrimaryRessourceType, ActivityRessourceType, ProjectionRessourceType, PagingRessourceType, ListRessourceType et FilterRessourceType sont des EClass héritant de RessourceType, comme il est explicité dans le tableau 3.3. Chacune représente un type de ressource particulier.

3.5.4.2.3 Méta-modèle de la structure de l'architecture REST

Schreier présente la structure de l'architecture REST et les relations entre ses composantes dans le méta-modèle suivant :

Eclass	Type de ressource	Description
SubressourceType	Sous-ressource	Une partie d'une ressource accessible directement par une autre ressource.
AggregationRessourceType	Ressource d'agrégation	L'assemblage des attributs des différentes ressources.
PrimaryRessourceType	Ressource primaire	Le concept de base d'un domaine modélisé. Exemple : les photos et les albums.
ActivityRessourceType	Ressource d'activité	Un sous-ensemble des attributs d'une autre ressource.
PagingRessourceType	Ressource divisée	La division de larges ressources en pages. Exemple : les pages de l'album.
ListRessourceType	Liste des ressources	Une liste des ressources primaires. Exemple : toutes les photos de l'album.
FilterResourceType	Des ressources filtrées	Une liste des ressources primaires, filtrées par les propriétés désirées. Exemple : les photos d'album avec le profil d'une seule personne.

TABLE 3.3 – Les différents types de ressources [SCHREIER 2011]

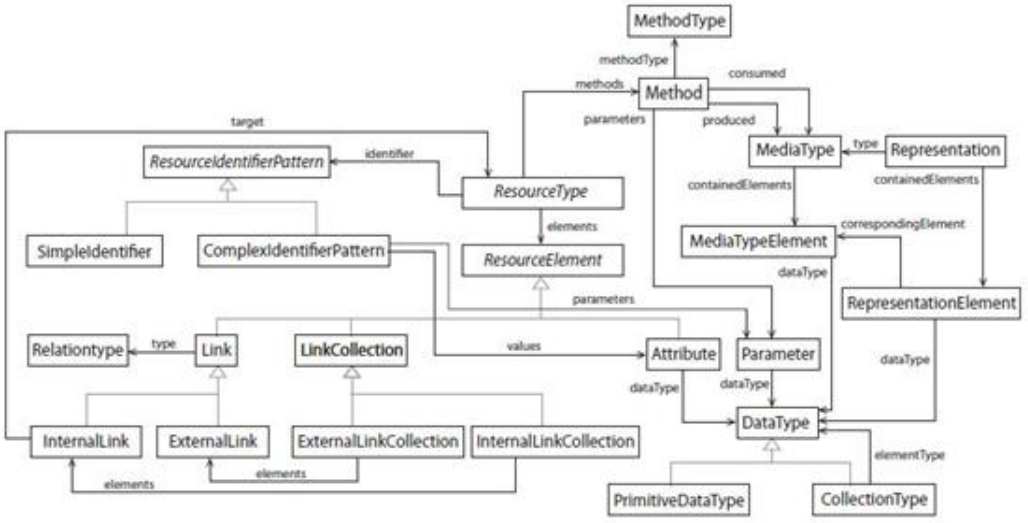

FIGURE 3.33 – Méta-modèle de Schreier : La structure du REST [SCHREIER 2011]

Chaque ResourceType (ressource est une instance) est identifiée par une seule ResourceIdentifierPattern (URI est une instance). Cette dernière EClass abstraite peut être SimpleIdentier (pour les cas d'identifiants simples) ou ComplexdentifierPattern (pour les cas d'identifiants composés des valeurs des attributs de la ResourceType et/ou des paramètres de méthodes utilisant cette ressource).

ResourceType contient une liste des éléments comme les attributs (Attribute) et les liens (Link), cette liste est modélisée par ResourceElement.

Les attributs sont des représentations optionnelles des propriétés de la ressource sous la forme des données conformes à la DataType. Cette dernière EClass abstraite peut être PrimitiveDataType (représente les données primitives comme 'String' et 'Integer') ou CollectionType (comme une liste ordonnée des données).

Une MethodeType, identifiée par son nom, définit toutes les méthodes utilisées dans REST comme les méthodes du http (GET, PUT ...). Une liste des méthodes du type MethodeType est associée à chaque ResourceType. Toute méthode produit ou/et utilise des MediaType chez le serveur en utilisant des paramètres optionnels : Parameter. Ces Parameters contiennent des données sous la formes de DataType, ils sont aussi utiles pour

l'identifiant de la ressource.

Une MediaType est identifiée par son nom, elle modélise les données du type média envoyées par le serveur. Elle contient des éléments du type MediaTypeElement qui sont des informations représentées par DataType. Une MediaType est définie par une Representation contenant des éléments RepresentationElement qui sont des données conformes à DataType.

La relation entre deux ressources est définie par Link, RelationType précise le type de cette liaison. Le client connait le type de la relation et la méthode exécutée pour répondre à une requête comme les informations de navigation 'précèdent' et 'suivant' (RelationType) entre deux pages web (ResourceType). Il y a deux types de relation : interne et externe, représentées respectivement par InternalLink et ExternalLink. Elles peuvent être regroupées sous la forme des collections LinkCollections : ExternalLinkCollection et InternalLinkCollection.

3.5.4.2.4 Méta-modèle du fonctionnement de l'architecture REST

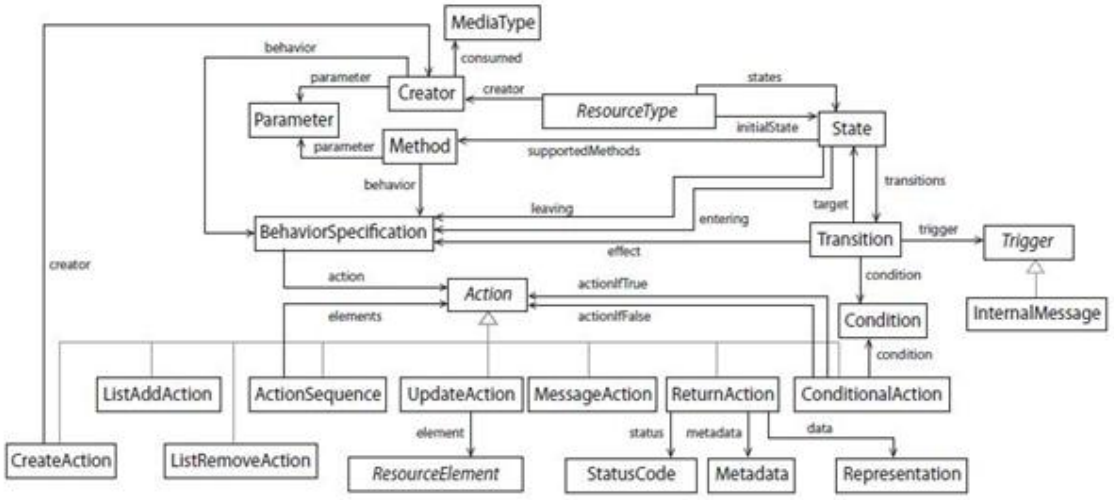

FIGURE 3.34 – Méta-modèle de Schreier : Le fonctionnement du REST [SCHREIER 2011]

Le changement d'état d'une ResourceType est géré par la machine à états finis qui permet de représenter l'état actuel de la ressource et ses différents états possibles après une transition ou demande précise. Comme le décrit la figure ci dessus, la ressource définit

une liste d'état States dont une seule est initiale. Une EClass State contient une liste des transitions sortantes appelées Transitions, chacun de ses éléments n'a qu'une seul cible State.
Une transition peut être déclenchée par une ou plusieurs déclencheurs (Trigger), comme les messages internes (InternalMessage), si la condition (Condition) associée est vérifiée.
Pour décrire le comportement des différentes composantes fonctionnelles, Schreier a introduit une EClass BehaviorSpecification. Elle définit le comportement et les contraintes de la transition (Transition), l'état (State) en entrant et en sortant, les méthodes (Method), les actions (Action) et les créateurs (Creator) des ressources, des actions et des méthodes (ils définissent les paramètres des méthodes et le type de données à consommer).
Les EClass CreateAction, ListAddAction, ListRemoveAction, ActionSequence, UpdateAction, MessageAction, ReturnAction et ConditionalAction héritent de la EClass abstraite Action et représentent respectivement la création d'une nouvelle ressource, l'ajout d'un élément à la liste des actions, la suppression d'un élément de cette liste, une séquence des actions, le changement des attributs et des liens d'une ressource, l'envoi des requêtes vers d'autres ressources, la réponse d'une demande (composée de la représentation, les données et l'état d'une ressource) et les contraintes sur le comportement d'un élément.

3.5.4.3 REST et l'accès multiples aux ressources

3.5.4.3.1 Cadre méthodologique de l'approche

Pour gérer l'accès multiples aux ressources dans une architecture client/serveur basée sur REST, nous attribuons à chaque ressource une boite de réception de messages. Ces derniers sont les requêtes envoyées par les différents clients utilisant cette ressource. Les messages sont rangés en respectant le temps d'arriver et la priorité.
Tout en respectant les règles du protocole de communication utilisé, un client envoie une requête pour afficher, créer, mettre à jour ou supprimer une ressource (par exemple les méthodes GET, PUT, POST, DELETE du protocole HTTP). Suite à la modification d'une ressource, des demandes de synchronisation sous la forme de messages sont déclenchées par des événements (envoyés par le serveur). Les messages qui concernent les services accédant en lecture à la ressource sont les plus prioritaires. La figure suivante décrit ce mécanisme.

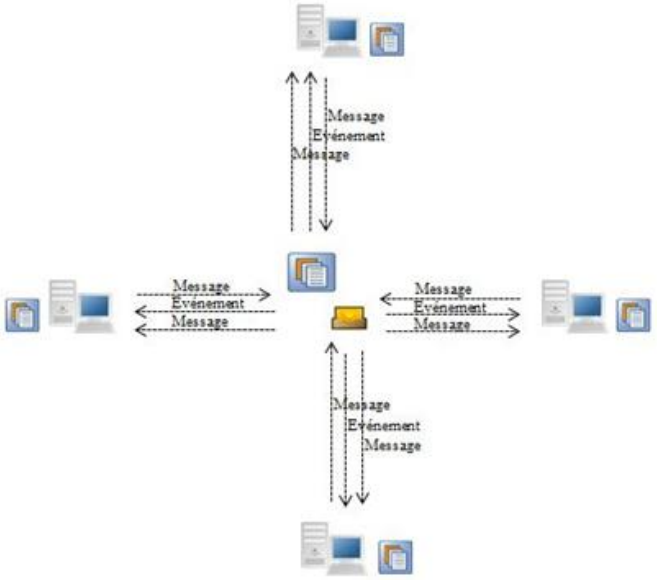

FIGURE 3.35 – Vue de l'architecture d'accès multiple

Les messages d'une ressource sont gérés par priorité selon l'état de la ressource, deux cas sont possibles :

— *Sans modification de la ressource* : tant que la ressource n'est pas modifiée, chaque message arrive à la boîte de réception avec une priorité égale à 1.
— *Avec modification de la ressource* : après chaque modification, les services utilisant cette ressource sont informés à travers les alertes propagées par le serveur et ils génèrent automatiquement des requêtes de synchronisation demandant les nouvelles mises à jour. Dans ce cas la priorité dépend du service émetteur :
  — S'il utilise la ressource par affichage ou par référencement, la priorité de message de la demande de synchronisation est égale à 0.
  — S'il a des demandes de modification qui ne sont pas encore traitées, s'il s'agit d'une suppression, la priorité attribuée est égale à 0 sinon elle est égale à 1.

Les messages ayant des priorités égales sont traités en respectant la règle « le premier

arrivé est le premier servi ».
Nous présentons ces concepts comme une extension à l'approche de Schreier pour assurer la gestion des accès concurrent aux ressources et l'intégration de nouveaux services à cette architecture quand elle est en cours d'utilisation.

3.5.4.3.2 Méta-modèle structurel

Nous ajoutons trois EClasse au méta-modèle structurel de Schreier : Service, ServiceType et Inbox. Elles correspondent respectivement aux : service qui envoie les requêtes utilisant une ou plusieurs méthodes vers la ressource, type du service et boîte de réception des actions sur la ressource. Chaque Service est décrit par une Représentation. Ce méta-modèle est présenté par la figure 3.36.

3.5.4.3.3 Méta-modèle fonctionnel

La EClasse Service spécifie la méhode utilisée dans chaque requête par la EClasse Method. BehaviorSpecification décrit l'action correspondante à une méthode, elle utilise une fonction Manager qui gére les priorités des actions qui seront par la suite gérées par une liste, l'attribut de la EClasse Inbox.
Chaque méthode effectue des modifications sur une ressource, notifie le service émetteur des changements pour qu'il puisse les demander sous la forme des messages internes (InternalMessage) et synchroniser les differentes versions de la ressource. Toute action envoie un retour vers le service émetteur.
Ce méta-modèle est présenté par la figure 3.37.

3.5.4.4 Instance d'HyperTagging

Tout en respectant les modèles conceptuels réalisés dans les sections précédentes, nous décrivons les composants architecturaux du prototype HyperTagging en instanciant les méta-modèles structurel et fonctionnel présentés précédemment.
Une instance est composée de deux type de modèles : un modèle de concepts (classes) et un modèle (ou plusieurs) qui décrit l'état d'une ressource.

Le modèle de concepts décrit par la figure 3.38 est un exemple parmi d'autres dans

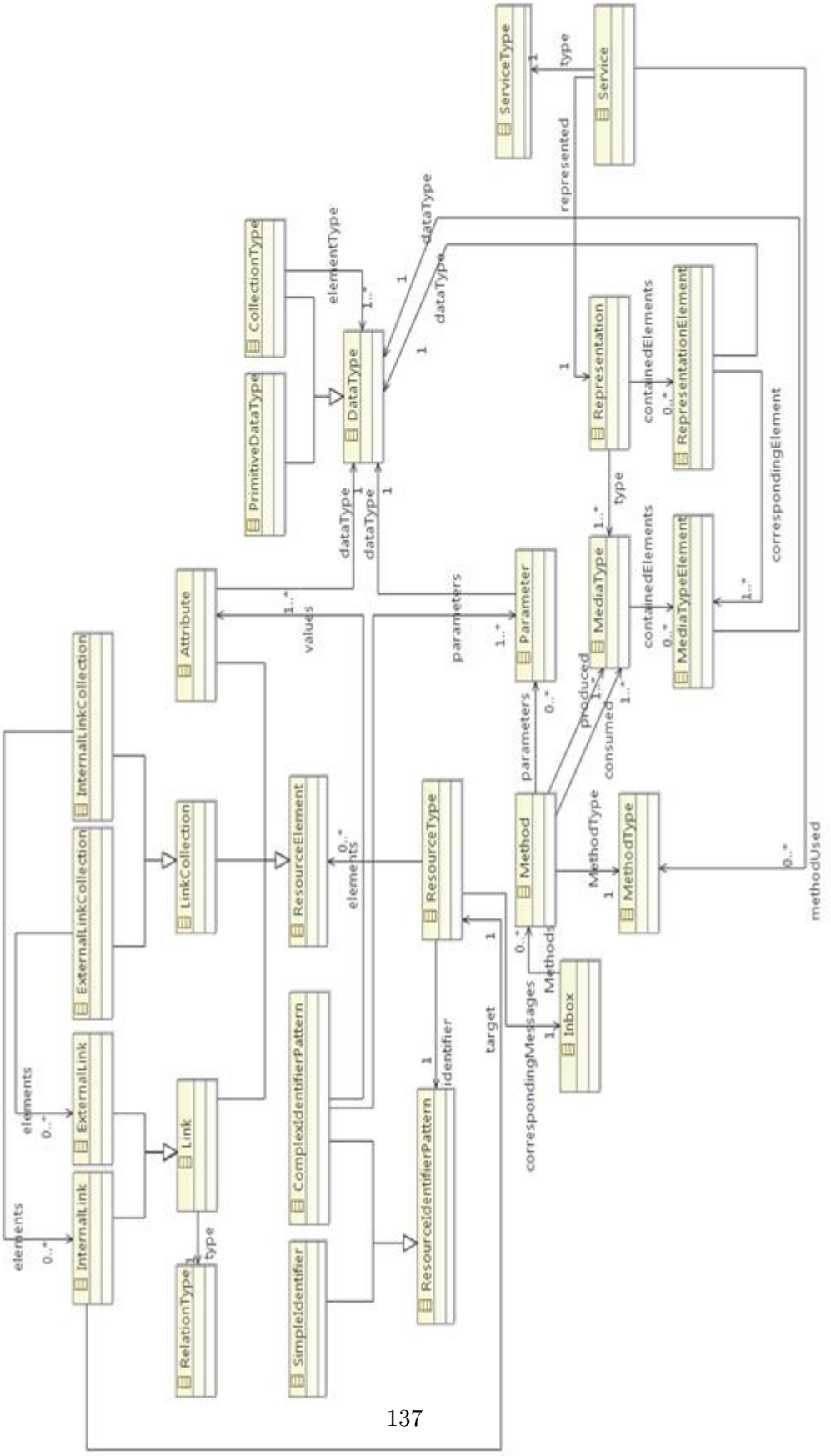

FIGURE 3.36 – Méta-modèle structurel

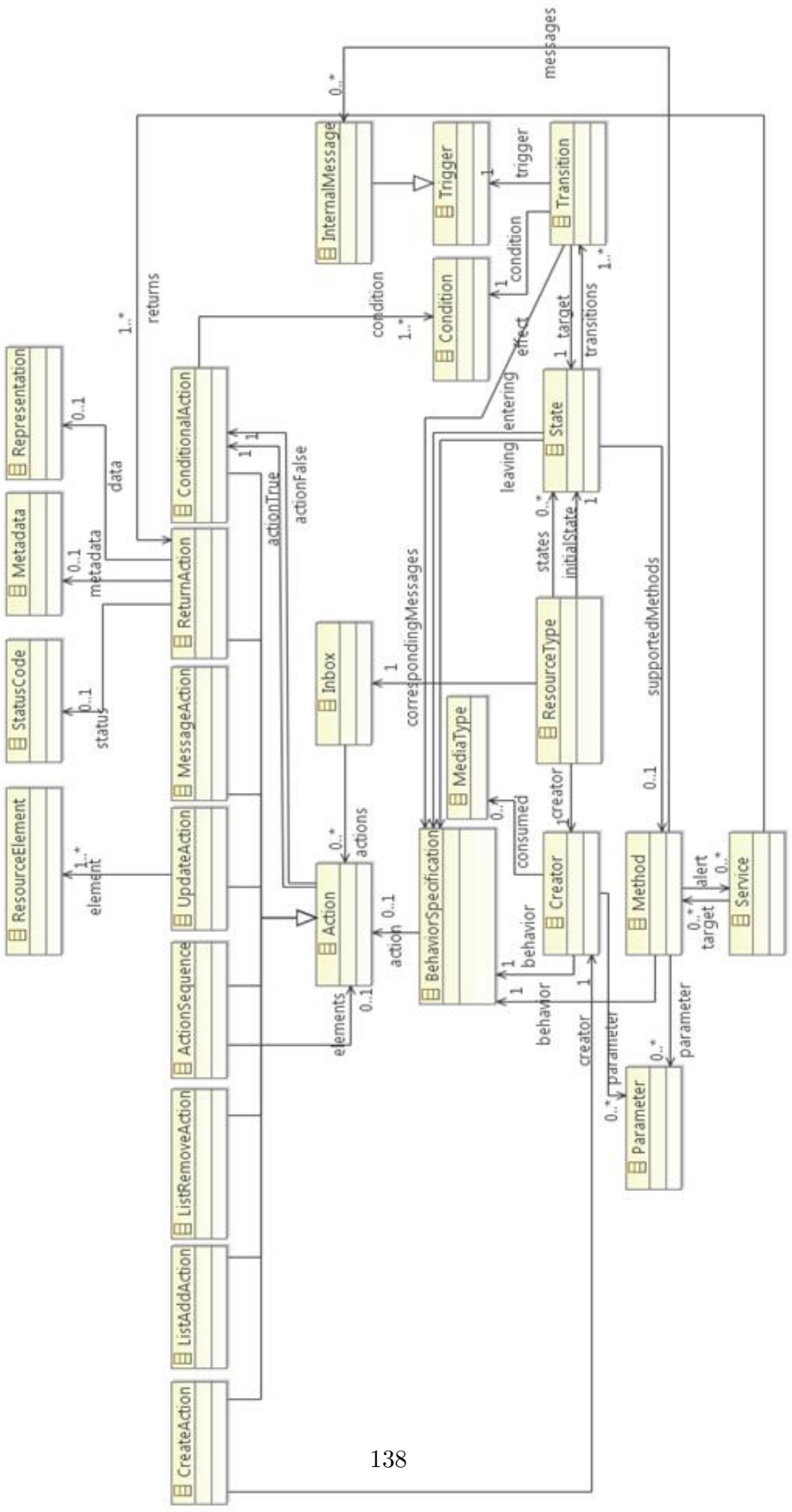

FIGURE 3.37 – Méta-modèle fonctionnel

lequel nous considérons les éléments du SOC, les règles du stockage, les unités du stockage et les documents comme des ressources primaires « PrimaryRessourceType ». Vues, Tags de Facette et Tags de règle, qui représentent respectivement toutes les vues créées dans un projet, les tags d'une facette et les tags utilisés pour construire une règle du stockage, sont des ressources de type liste « ListRessourceType ». Données_Tag, représentant l'icône associée à un tag, est une sous ressource de type « SubressourceType ». Chaque ressource primaire est associée à une boite de message, cette dernière contient la liste (de type pile) des actions à effectuer sur cette ressource. Elles sont gérées par deux opérations : empiler pour ajouter une nouvelle action à la liste et dépiler pour exécuter une action tout en respectant les règles de priorité.

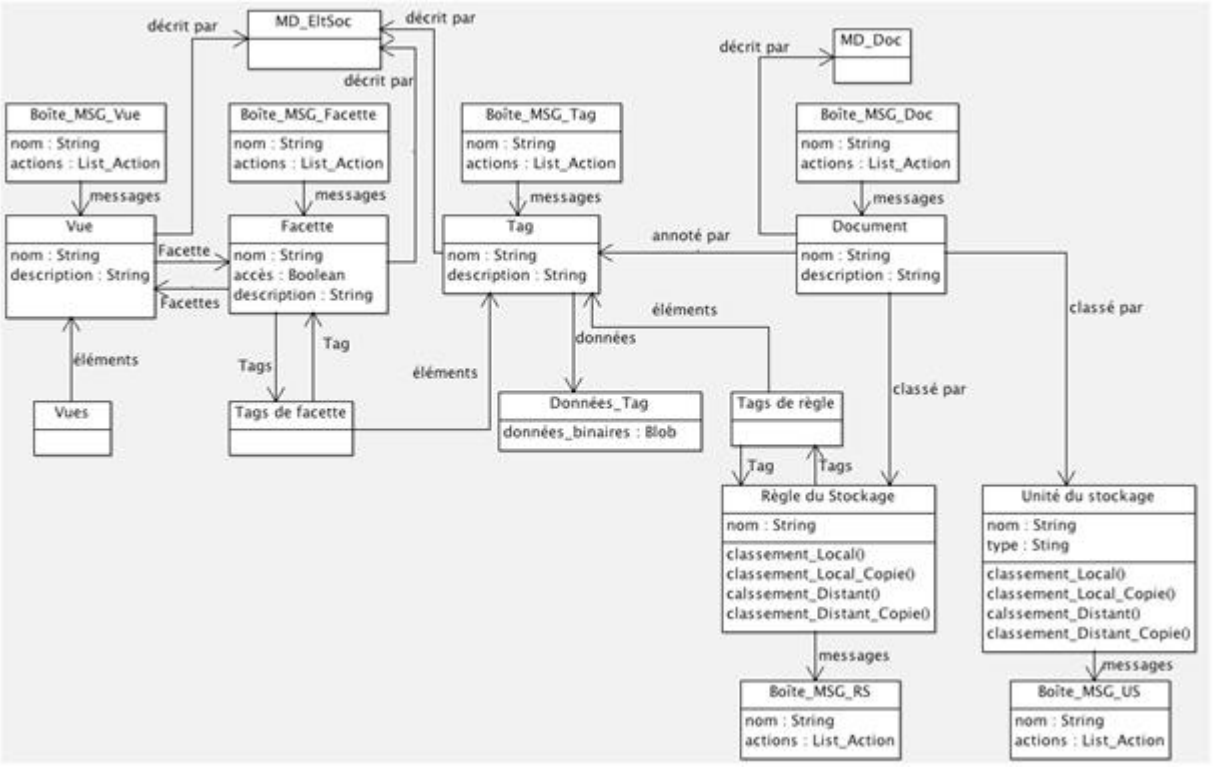

FIGURE 3.38 – Exemple d'instance : modèle de concepts architecturaux d'HyperTagging

Nous avons choisi la ressource « Règle du stockage » pour donner un exemple du modèle d'états d'une ressource. « Règle du stockage » digère quatre états : sa création (créée), sa suppression (supprimée), son utilisation pour classifier les documents (en cours

de classement) et sa libération (libérée). Elle supporte les méthodes suivantes : POST, GET, PUT et DELETE, ce qui n'est pas le cas pour toutes les ressources. La figure suivante représente le modèle des états de cette ressource :

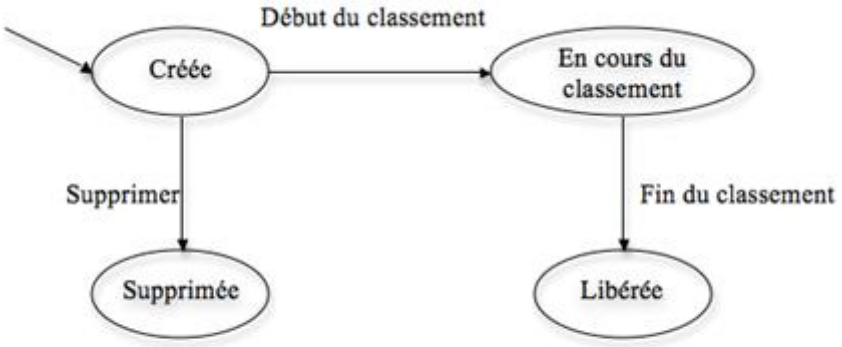

FIGURE 3.39 – Exemple d'instance : modèle d'état de la ressource « Règle du stockage »

La communication entre les services du prototype HyperTagging est basée sur les méthodes du protocole http. Nous citons quelques exemples de communication :

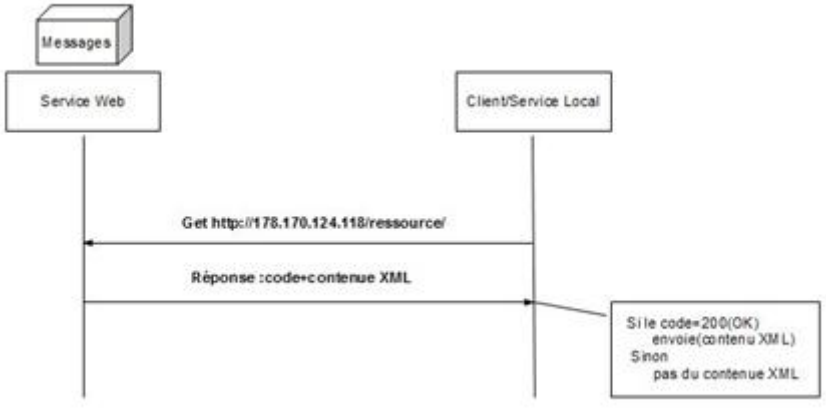

FIGURE 3.40 – Demande d'une ressource

Le client (service local) demande une ressource avec la méthode GET du protocole http en précisant l'URL de la ressource comme le montre la figure 3.40. La réponse est

composée principalement d'un code d'état et le contenu de la ressource sous format XML.

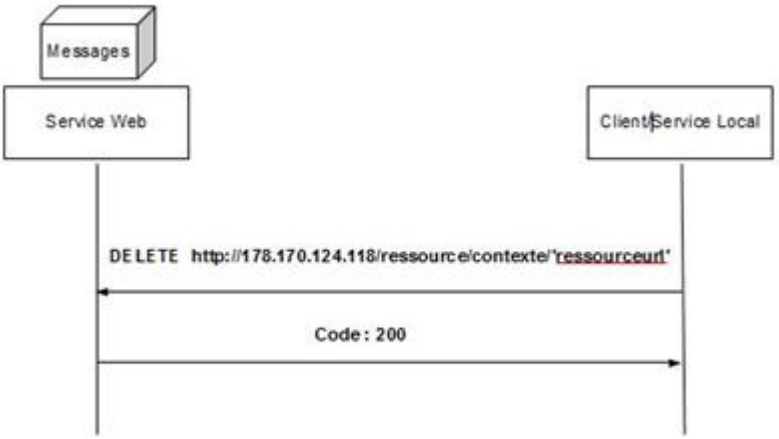

FIGURE 3.41 – Suppression d'une Vue

Cette figure représente la suppression d'une ressource, vue, avec la méthode DELETE. Le serveur à travers le service web envoie le code 200 au client pour l'informer que la requête est traitée avec succès.

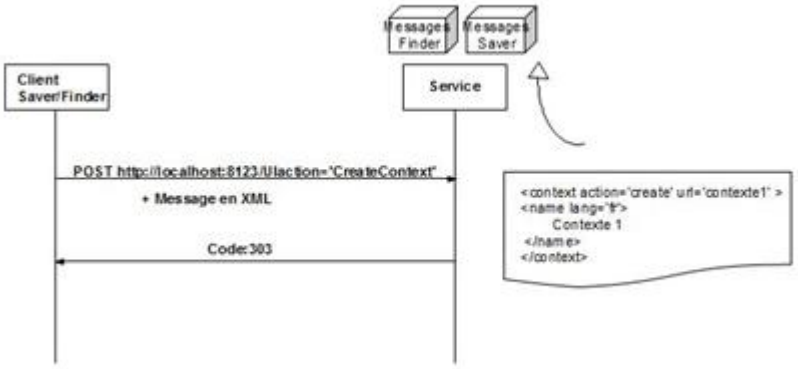

FIGURE 3.42 – Suppression d'une Vue

La création d'une ressource est faite par la méthode POST comme il est décrit par la figure 3.42. Nous présentons dans cet exemple le contenu d'un message XML d'une vue. Les messages envoyés par le client peuvent être de la part du Saver ou du Finder. Le

serveur répond par un code http 303 (See other), c'est-à-dire que l'URL n'est pas valide (par exemple utilisé par une autre ressource).

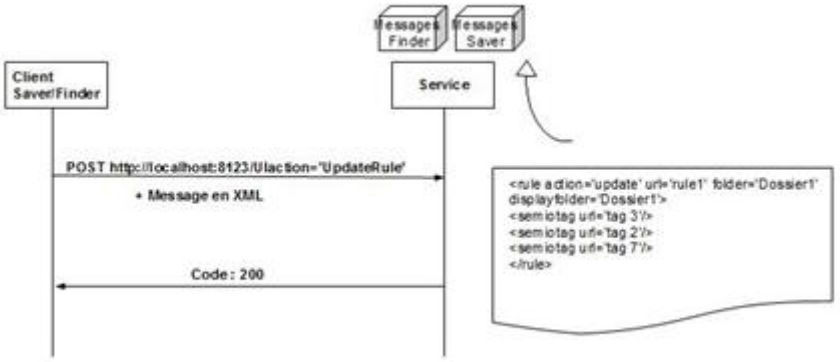

FIGURE 3.43 – Modification d'une Facette

Cette figure représente la modification d'une ressource de type Facette en utilisant la méthode POST.

3.6 Modélisation de l'interopérabilité

3.6.1 Exigences applicatives

Afin d'atteindre le niveau de performance nécessaire aux nouvelles applications interopérables, nous pensons que conception et interopérabilité doivent être corrélées. Nous continuons dans la même démarche conceptuelle orientée infrastructure tout en intégrant la notion d'interopérabilité conformément aux exigences applicatives suivantes :

— Ouverture de l'architecture à travers la normalisation et l'utilisation des standards communs entre la plupart des outils de gestion de documents ;

— Performance de l'architecture en faisant correspondre les évolutions conceptuelles et techniques avec les exigences sémantiques de l'outil (le partage des documents et la gestion d'évolution d'une ressource).

3.6.2 Protocole CMIS

3.6.2.1 Description

CMIS (Content Management Interoperability Services) est un protocole ouvert géré par OASIS (Organization for the Advancement of Structured Information Standards) [33]. Il était fondé par IBM, Microsoft et ECM. La première version est officialisée en 2010. Ce projet est en évolution permanente, les participants à l'OASIS CMIS actuellement sont Adobe, Alfresco, ASG, Booz Allen Hamilton, dotCMS, EMC, fme AG, IBM, Magnolia, Microsoft, Nuxeo, Open Text, Oracle, Quark, SAP, Saperion, Sun Microsystems et d'autres. L'objectif de la spécification du CMIS est de fournir des services pour traiter les différents contenus. Il définit un ensemble de fonctions communes entre les services de type REST et SOAP. Ce protocole propose une spécification ciblée à « *la gestion de documents* », ce qui représente son apport majeur. En effet, deux méthodes d'accès sont définies dans la spécification du CMIS : une interface de service web et une interface basée sur le protocole de publication Atom (protocole qui gère les ressources web avec les méthodes GET, SET, DELETE) qui utilisent respectivement les standards SOAP/WSDL et REST.

3.6.2.2 Méta-modèle du CMIS

Ce modèle définit les concepts et les relations de référentiels CMIS. Il fournit un vocabulaire commun pour tous les utilisateurs de ce protocole. Un référentiel est un conteneur d'objets conforme à CMIS.
Quatre types communs d'objets sont définis dans la spécification du CMIS et peuvent être traités et gérés par des référentiels CMIS (des outils normalisés utilisant un service CMIS) : l'Objet Document, l'Objet Dossier, l'Objet Relation et l'Objet Politique. Ils sont représentés dans la figure suivante :

33. https ://www.oasis-open.org

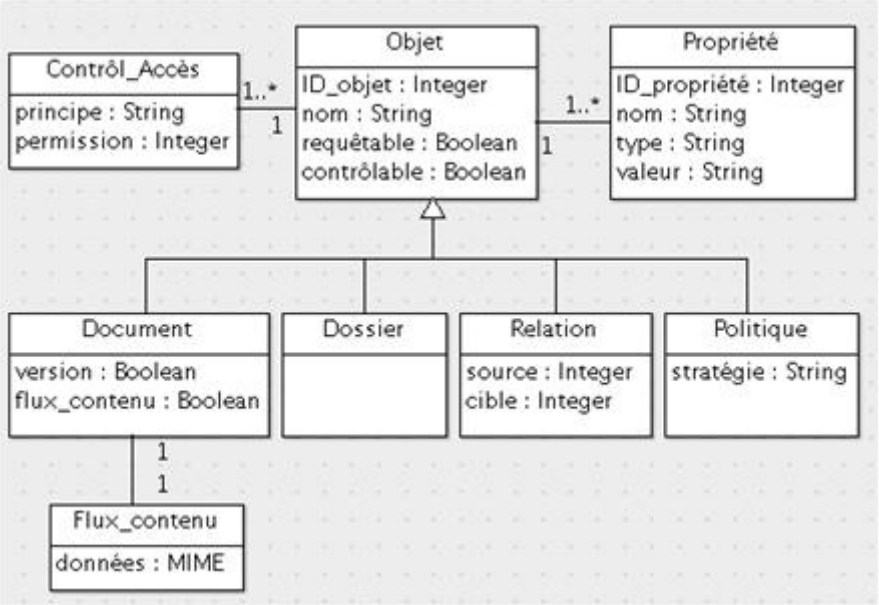

FIGURE 3.44 – Méta-modèle du CMIS

Ce méta-modèle est conforme aux descriptions suivantes :

— Chaque objet a un type et un ensemble de propriétés dont chacune est identifiée par un *ID_propriété* et décrite par un *nom*, son *type*, sa *valeur* et autres ;

— Les objets ont un ensemble d'attributs en commun comme l'identifiant *ID_objet*, son *nom*, *requêtable* (pour préciser si cet objet est requêté ou non), *contrôlable* (pour préciser si cet objet est contrôlé par une politique ou non), etc ;

— Les objets Document et Dossier, comme leurs noms l'indiquent, représentent respectivement les documents et les dossiers gérés (classés, indexés, stockés...) par l'outil digestion du contenu. Ils ont des attributs supplémentaires : version pour indiquer si ce document a des versions ou non et *flux_contenu* pour indiquer si ce document est autorisé à avoir des flux de contenu (flux binaire des données de type MIME) ou non. C'est le seul objet qui peut avoir des versions dont chacune représente un objet avec son propre identifiant ;

— L'objet Relation définit les relations entre les objets, il a deux attributs supplémentaires : *source* et *cible* qui représentent respectivement l'objet source de la relation et son objet cible ;

— Politique est l'objet qui permet de gérer les stratégies d'accès aux contenus (par

exemple : contenu contrôlable) ;

— Une liste de contrôle d'accès est associée à chaque objet en définissant le principe qui peut être une personne, un groupe ou un rôle et les permissions d'accès qui sont lecture, écriture ou tous. [POTTS 2009][WALDT 2009]

3.6.3 Normalisation de l'approche orientée infrastructure

Nous rajoutons aux trois services associés aux composantes de nommage, indexation et classement de documents, de recherche de documents et de gestion des éléments du SOC, un autre service basé sur REST qui gère le dialogue avec les autres outils de gestion de documents en utilisant une interface d'API CMIS. Pour assurer ce fonctionnement, nous adaptons le modèle des concepts conçu précédemment aux spécifications du CMIS. Les figures suivantes montrent les évolutions conceptuelles effectuées dans ce sens :

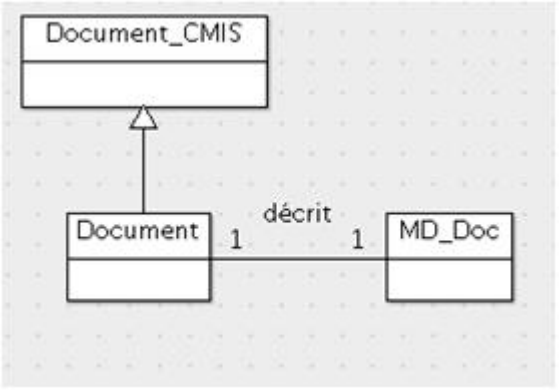

FIGURE 3.45 – Adaptation du Document d'HyperTaging avec CMIS (1)

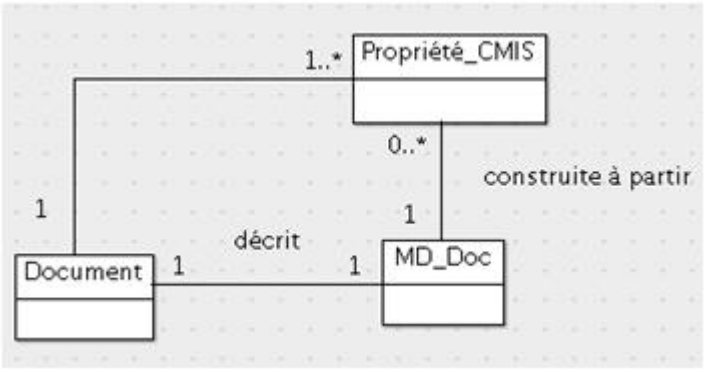

FIGURE 3.46 – Adaptation du Document d'HyperTaging avec CMIS (2)

Le concept Document que nous avons introduit précédemment dans nos modèles conceptuels hérite de Document_CMIS dércit dans l'API de CIMS. Cet héritage permet à notre concept Document de recevoir toutes les caractéristiques du Document_CMIS, y compris la liste des propriétés. Cette dernière est construite à partir des métadonnées du document dont chaque attribut (l'URL_valide_doc, le nom_doc, les mots clés, la création_doc et les actions_doc) représente une propriété. Par exemple, l'attribut URL_valide_doc devient une propriété avec :

— ID_propriété est un entier généré automatiquement ;
— le nom est 'URL_valide_doc' ;
— le type est 'String' (une chaîne de caractères) ;
— la valeur est la valeur de l'URL_valide_doc (un URL).

La normalisation de notre système d'information lui permet de dialoguer avec autres outils de gestion documentaire normalisés (avec CMIS) comme Alfresco, Nuxeo, Documentum et autres. La figure suivante représente ce dialogue dans une architecture orientée services :

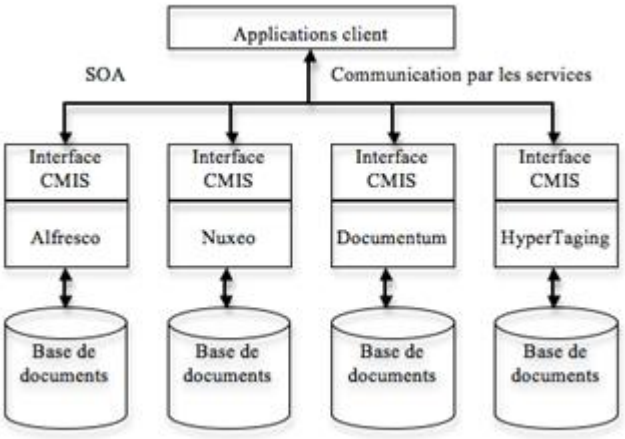

FIGURE 3.47 – Interopérabilité entre différents services utilisant CMIS

3.7 Conclusion

A travers cette démarche conceptuelle incrémentale, nous avons présenté un modèle conceptuel complet évolutif capable de gérer les notions de la documentarisation classificatoire tout en présentant une interface d'interaction avec un usager acteur, un système d'organisation des connaissances multidimensionnel, une gestion d'évolution des composants du SOC et de documents classés, et une infrastructure normalisée permettant l'accès multiples à une ressource et l'interopérabilité avec d'autres systèmes de gestion du contenu documentaire.

Chapitre 4

Evaluer pour positionner !

4.1 Introduction

L'évaluation des approches proposées dans un cadre de recherche est complexe mais néanmoins importante puisqu'elle conditionne la crédibilité de leur caractère scientifique. Pour réaliser une telle étude, il faut suivre un modèle ou un processus analytique à multiple perspectives : organisationnelle, structurelle et technique.

4.2 Objectifs et démarche

Dans ce chapitre, notre objectif consiste à positionner HyperTagging dans l'environnement de gestion de contenu de l'entreprise (ECM : Entreprise Content Management). En effet, nous proposons un modèle d'évaluation du SI de gestion de contenus, du point de vue de la classification documentaire dans l'entreprise. L'originalité de cette contribution réside dans le processus d'évaluation, notamment la détermination des critères fins sur le SOC et le document traité dans ce type du système. Dans cette étude analytique comparative, nous répondons aux questions suivantes :

— Dans quel cadre cette évaluation est-elle proposée ?
— Dans quel environnement d'évaluation se positionne Hypertagging ?
— Sur quels critères faut-il se baser pour implémenter cette évaluation ?

Pour atteindre nos objectifs, nous nous basons sur la démarche méthodologique suivante :

— Etude de l'évaluation des systèmes d'information de gestion de contenu de l'entreprise appelés ECM (entreprise content management) par des grandes entreprises comme Forrester et Gartner ;
— Proposition d'un modèle d'évaluation basé sur des critères spécifiques ;
— Positionnement d'HyperTagging.

4.3 Evaluation des ECM

Plusieurs entreprises dans le monde commercialisent leur expertise et leur savoir professionnel sous forme de rapports d'analyse dans un domaine particulier et pour une clientèle précise.
Nous étudions deux exemples, deux grandes entreprises internationales Gartner et Forrester, connues dans le marché comme des experts dans le domaine d'étude et consulting des nouvelles technologies.

4.3.1 L'évaluation de Gartner

4.3.1.1 Présentation de Gartner

Gartner [34] est une entreprise américaine de conseil et de recherche dans le domaine des technologies avancées, fondée en 1979. Elle mène des recherches, analyses et statistiques sur les nouvelles technologies et stratégies.

4.3.1.2 L'évaluation d'ECM chez Gartner

Les analystes de Gartner considèrent que l'ECM est :
— une approche stratégique de la gestion de l'information, qui permet de contrôler le contenu d'entreprise, encourager la collaboration, faciliter l'accès et le partage de l'information ;
— et un ensemble d'outils logiciels : des capacités et des applications qui interagissent et peuvent travailler séparément pour la gestion de cycle de vie du contenu. [GILBERT et al. 2012]

34. www.gartner.com, visité en septembre 2013

Ils décomposent l'ECM en sept composants, la gestion du document, les applications d'imagerie, workflow et la gestion de processus des affaires (BPM), la gestion d'archive, la gestion du contenu de Web (WCM), le contenu Social et les composants étendus (DAM, e-forms, gestion des e-mails, etc.) et attribuent une degré d'importance pour chacun, respectivement : 15%, 18%, 22%, 13%, 7%, 15% et 10%.

Les experts de Gartner ont effectué une étude analytique, en Octobre 2012, pour évaluer les fournisseurs des ECM du marché mondial comme Alfresco, IBM, Microsoft, Oracle, ECM, OpenText, etc. En première étape, ils ont réparti le marché en quatre catégories selon le type d'acteur : les leaders, les challengers, les visionneurs et les acteurs de niche. Dans une deuxième étape ils ont fixé deux types de critères d'évaluation sur le fournisseur et son outil : les capacités du fournisseur à l'exécution de son produit et l'analyse du marché.À chaque critère est attribué un degré d'importance, élevée, standard ou faible. [GILBERT et al. 2012]

4.3.1.2.1 Les critères de la capacité d'exécution

Nous citons les différents critères fixés par Gartner qui concernent la capacité du fournisseur à l'exécution de son produit, ainsi que la description et le degré d'importance de chacun dans le tableau suivant :

4.3.1.2.2 Les critères d'analyse du marché

Nous citons les critères complémentaires fixés par Gartner dans la table suivante :

En s'appuyant sur ces différents critères, les analystes de Gartner ont montré les points forts et les points faibles de chaque fournisseur actif dans le marché et l'ont placé dans l'une des catégories du marché (leader, challenger, visionneur ou acteur de niche). IBM, Oracle, Microsoft, Opentext, ECM et Hyland Software sont les leaders de marché en 2012. [GILBERT et al. 2012]

Critère	Description	Importance
Produits et services	Les capacités, fonctionnalités et qualité des produits et services offerts par le fournisseur.	Elevée
La fiabilité globale	L'unité opérationnelle des affaires, de finance, de stratégie et d'organisation.	Elevée
La commerciale et la tarification	L'efficacité dans la chaine de vente : l'avant-vente, les contrats, les négociations, et les tarifications.	Standard
La réactivité du marché et l'antécédent	L'aptitude de réponse, la flexibilité, la réussite de la compétitivité, le respect de la concurrence, l'évolutivité de la clientèle et du marché. Ce critère prend en compte l'historique de la réactivité (antécédent).	Standard
Marketing	La clarté, l'efficacité et la qualité d'influence sur le marché ; La création d'une image positive sur le produit/marque et l'organisation, dans l'esprit des acheteurs ; L'utilisation des moyens de Marketing : les publicités, les promotions, le bouche-à-oreille, le leadership éclairé, les activités de vente.	Standard
Expérience client	Les relations, les services, les produits et/ou les programmes qui facilitent l'utilisation du produit aux clients ; La mise en disposition d'un soutien technique à la clientèle.	Elevée
Opérations	Les capacités, les compétences, les expériences, les systèmes et la qualité de la structure organisationnelle et l'efficacité de fonctionnement et d'organisation.	Standard

TABLE 4.1 – Les critères de la capacité d'exécution

Critère	Description	Importance
Compréhension du marché	La capacité d'écouter les clients et comprendre leurs besoins.	Standard
La stratégie du marketing	La communication dans toute l'organisation et à l'extérieur et les programmes proposés pour les clients.	Standard
La stratégie de vente du produit	L'utilisation des réseaux direct et indirect appropriés aux ventes.	Standard
La stratégie des offres	Les méthodologies de développement et de la livraison du produit.	Elevée
Stratégie de l'industrie	La stratégie du fournisseur utilisée pour orienter les offres, les compétences, les ressources pour répondre aux besoins des différentes catégories du marché.	Standard
Innovation	Innovation des ressources (directes, liées, complémentaires), de l'expertise et de la consolidation.	Standard
Stratégie géographique	La stratégie du fournisseur pour gérer les offres, les compétences, les réponses aux besoins du marché sur les différentes zones géographiques, directement ou à travers des intermédiaires.	Faible

TABLE 4.2 – Les critères d'analyse de la vision du marché

4.3.2 L'évaluation de Forrester

4.3.2.1 Présentation de Forrester

Le groupe Forrester [35] est fondé en 1983 aux USA, il effectue des recherches à l'échelle internationale. Il accompagne les entreprises en présentant des conseils professionnels dans le domaine des technologies, le marketing et les stratégies des affaires.

4.3.2.2 L'évaluation d'ECM chez Forrester

Le groupe Forrester divise les ECMs en 4 catégories :

— ECM fondamental, qui fournit les fonctionnalités basiques de la gestion de contenu. Les technologies des contenus fondamentaux comprennent les services de bibliothèque de base, les flux de production, la recherche et la gestion des dossiers.

— ECM d'affaires, les technologies de contenu de l'entreprise poussent l'expérience de travail et fournissent des capacités permettant aux employés de travailler en collaboration et gérer les droits de gestion des documents composés.

— ECM transactionnel, supporte le processus de la gestion des affaires et les processus qui intègrent le contenu avec les applications de back-office (non accessible aux utilisateurs finaux ou clients).

— ECM persuasif, supporte les contenus qui influent sur le comportement du public externe (contenu de marketing). Exemples : le contenu du Web et la gestion des actifs numériques

Les analystes divisent le marché en quatre segments : les leaders, les performants, les concurrents et les pris risqués. Ils se basent dans cette étude, comme les experts de Gartner, sur des critères d'évaluation : l'offre actuelle sur le marché, les stratégies opérationnelles et la présence sur le marché avec des pondérations d'importance qui sont respectivement : 50%, 50% et 0%. [WEINTRAUBN 2011]

35. www.forrester.com, visité en septembre 2013

4.3.2.2.1 Le critère de l'offre dans le marché

Le critère de l'offre dans le marché dépend du type du contenu traité par l'ECM proposé aux clients (contenu fondamental, contenu des affaires, contenu transactionnel et contenu persuasif) et de la capacité technique de son architecture. Ces aspects sont considérés comme des sous-critères et chacun est représenté par un pourcentage selon son importance dans le fonctionnement global d'ECM, fondamental, des affaires et transactionnel. Nous représentons ces différentes pondérations dans la table suivante :

Sous-Critère	Pondération globale	Pondération fondamentale	Pondération des affaires	Pondération transactionnelle
Support du contenu fondamental	20%	80%	30%	5%
Support du contenu des affaires	20%	10%	50%	5%
Support du contenu transactionnel	20%	0%	5%	80%
Support du contenu persuasif	20%	0%	5%	0%
Capacité de la base de l'architecture	20%	10%	10%	10%

TABLE 4.3 – Les sous-critères de l'offre du marché

4.3.2.2.2 Le critère de stratégies opérationnelles

D'après Forrester, les stratégies opérationnelles sont la stratégie de la solution intégrale, la production, l'emballage du produit, les licences et l'établissement du prix. Chaque stratégie est présentée dans la table ci-dessous par un pourcentage d'importance dans le fonctionnement global d'ECM, fondamental, des affaires et transactionnel.

4.3.2.2.3 Le critère de la présence sur le marché

Forrester mesure la présence d'un ECM dans le marché à travers la présence financière du fournisseur, les formations offertes sur l'ECM, la taille de la clientèle et la présence géographique.

Sous-Critère	Pondération globale	Pondération fondamentale	Pondération des affaires	Pondération transactionnelle
Stratégie de la production	70%	65%	60%	70%
Stratégie de la solution intégrale	25%	30%	30%	25%
Stratégie de l'emballage du produit	25%	5%	10%	5%
Licences et l'établissement du prix	0%	0%	0%	0%

TABLE 4.4 – Les sous critères des stratégies opérationnelles

Sous-Critère	Pondération globale	Pondération fondamentale	Pondération des affaires	Pondération transactionnelle
La présence financière de l'entreprise	50%	50%	50%	50%
Formation	5%	5%	5%	5%
Clientèle de l'ECM	45%	45%	45%	45%
Présence géographique	0%	0%	0%	0%

TABLE 4.5 – Les sous critères de la présence sur le marché

4.4 Evaluation des ECM du point de vue de la classification documentaire

Dans les deux études effectuées précédemment par Granter et Forrester, nous constatons que les analystes choisissent les critères d'évaluation selon le domaine d'analyse et les besoins de la clientèle. Ils se basent sur des critères orientés marché concernant la production et la situation des fournisseurs plus que le fonctionnement et la caractérisation du produit.

Dans notre étude, nous nous intéressons à l'aspect orienté classification documentaire des ECM. Pour réaliser cette étude, nous nous inspirons du modèle du processus d'évaluation des Systèmes d'information décrit par Kéfi. H et Kalika. M, qui est formé de trois grandes parties : « Cadre analytique», « Contexte du processus d'évaluation » et « Processus d'évaluation ». [KEFI et KALIKA 2004]

4.4.1 Modèle d'évaluation

La mise en relation d'HperTagging et de l'environnement ECM repose sur un triptyque : la spécialisation de cette étude sur l'organisation des connaissances dans l'entreprise et du classement des documents (le cadre analytique), l'encadrement des environnements d'ECM (le contexte d'évaluation) et une démarche en trois étapes (le processus d'évaluation). Ces étapes sont : le classement des environnements d'ECM, la détermination des critères d'évaluation et le positionnement d'HyperTagging. Dans la figure suivante, nous représentons ces différentes étapes du modèle d'évaluation :

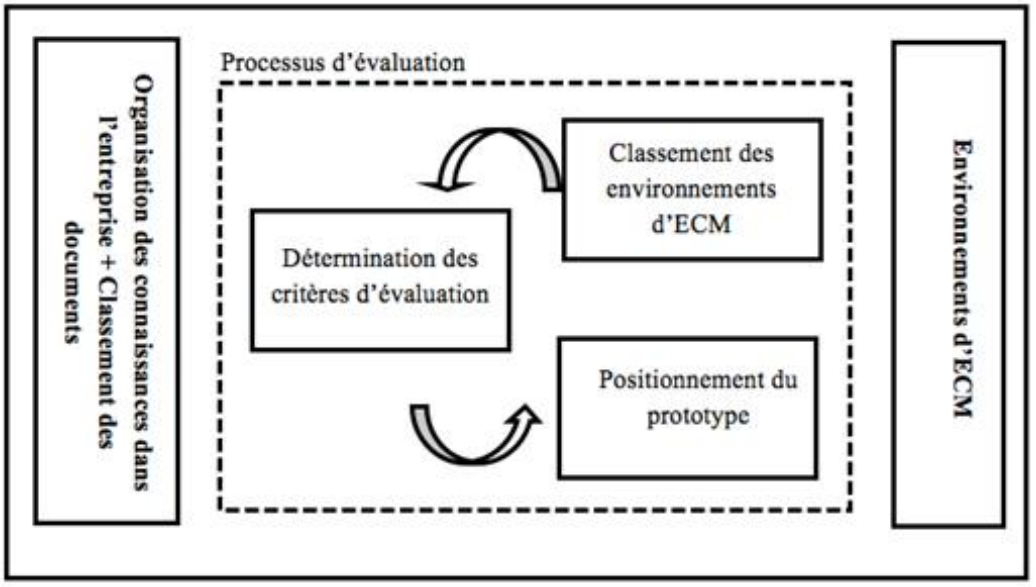

FIGURE 4.1 – Le modèle d'évaluation [ANKOUD et HMIMIDA 2013]

4.4.2 Cadre et contexte d'évaluation

Notre étude s'adresse à une large majorité d'acteurs au sein de l'entreprise, ayant des objectifs individuels et collectifs pour la gestion de contenu dans leurs propres besoins de travail, notamment pour les professionnels du classement et de l'indexation de documents au niveau « privé » et « partagé » en se basant sur les SOC. La gestion de contenu nécessite une convergence entre la gestion des systèmes d'information et la gestion de l'information, qui est assurée par les ECM en construisant la liaison entre les contenus métier de l'entreprise et les fonctions techniques disponibles [DUPIN 2010]. Dans cette

étude, au sein de l'environnement ECM, nous nous concentrons sur les fonctionnalités documentaires comme les règles d'organisation, de classement et de stockage, ainsi que sur le cycle de vie du document

4.4.3 Processus d'évaluation

Ce processus est composé de trois étapes comme le montre la figure précédente : Classement des environnements ECM, Détermination des critères d'évaluation et Positionnement du prototype étudié. L'achèvement d'une étape entraine une passation à la suivante. Dans la démarche suivante, nous traitons le cas du prototype HyperTagging dans la troisième étape.

4.4.3.1 Classement des environnements d'ECM

Plusieurs systèmes de gestion de l'information semi-structurée sont regroupés autour du terme général ECM (Entreprise Content Management, Gestion de contenu d'entreprise), qui utilise des technologies pour capturer, gérer, stocker, classer, préserver et diffuser du contenu semi-structuré en lien avec les processus organisationnels internes et externes. C'est une approche qui permet une meilleure prise de décision au sein de l'organisation et une efficacité de partage de l'information enrichie grâce à la collaboration. Les ECM regroupent aujourd'hui plusieurs types de technologies et associent des stratégies de gestion correspondant aux besoins de l'entreprise, nous proposons de les catégoriser comme suit :

— Des environnements issus de la Gestion Electronique des Documents (Alfresco, OpenText, Filenet, Documentum..) ;
— Des environnements de travail collaboratif ou groupware (Lotus notes, Sharepoint) ;
— Des environnements participatifs issus du web 2.0 (réseaux sociaux d'entreprise, environnement de tagging de ressources partagées, etc.) ;
— Des environnements issus de l'ingénierie des connaissances (approches basées sur les ontologies comme le Semantic Desktop) ;
— Des environnements de publications (les blogs, Wiki ou des Content Management Systems issus de l'univers du Web).

Chaque type du système de gestion du contenu intègre un ou plusieurs systèmes d'organisation des connaissances (SOC), qui permettent de décrire le contenu et le statut organisationnel d'un document (ontologies, thésaurus, facettes, folksonomies, schémas de classification...) sans que les propriétés de ces SOC ne soient toujours clairement explicitées, ce qui rend le choix du meilleur environnement difficile pour les entreprises. Cette complexité justifie le besoin de méthodologies d'évaluation des systèmes d'ECM et le développement de benchmarks scientifiques portant sur les modèles classificatoires intégrés à ces documents.

4.4.3.1.1 GED (Gestion Electronique des Documents)

La GED correspond à l'ensemble des technologies et logiciels permettant l'acquisition, la gestion, l'indexation, la recherche, le stockage, le classement, la consultation, le traitement et la diffusion des informations structurées et non-structurées qui circulent dans une entreprise. Elle peut inclure des composants de type « Workflow » pour la circulation du document. Au sens strict, ces environnements ne permettent pas la publication en ligne et le travail coopératif mais ce sont eux qui offrent les meilleurs outils de gestion des SOC (Classement, indexation, stockage et recherche) [CHAUMETTE et DESBIENS 2008].

4.4.3.1.2 Le travail collaboratif ou Groupware

Le Groupware est « *une combinaison de technologies, de personnes et d'organisation qui facilite la communication et la coordination nécessaire à un groupe pour réaliser son travail de manière collective et efficace, atteindre un but partagé et assurer un gain pour chacun de ses membres* » [DIENG 2001]. Par comparaison avec le GED et le Workflow qui facilitent le travail en groupe, le groupeware utilise ces technologies pour former une structuration globale et échanger les connaissances collectives en interne. Pour classer une fonctionnalité synchrone ou asynchrone (la gestion des documents par création, suppression ou modification, la messagerie électronique...), les applications du travail collaboratif sont réparties en trois groupes orientées « mémoire », « routage » et « échange » qui correspondent respectivement aux principes du Groupeware « coopération », « coordination » et « communication » [CHAUMETTE et DESBIENS 2008].

4.4.3.1.3 Des environnements participatifs issus du Web 2.0

Le Web 2.0 ou le web social est une évolution du Web permettant la participation des usagers. Ceux-ci contribuent à la construction et à la diffusion des informations et des connaissances sur le Web en se basant sur la collaboration (créer, partager, indexer) entre les usagers à travers les sites de réseau social, les blogs, les Wikis, les sites de partage des vidéos et de la musique... [PROULX et al. 2010]. Il a donné lieux à l'apparition de nouveaux SOC basés sur le tagging individuel ou collaboratif dénommés folksonomies.

4.4.3.1.4 Des environnements d'ingénierie des connaissances

L'ingénierie des connaissances est une approche symbolique interdisciplinaire (informatique, linguistique et psychologie cognitive). A travers des techniques, des concepts et des méthodes, elle permet de transmettre, enregistrer les savoir-faire, modéliser, acquérir et gérer les connaissances [AUSSENAC-GILLES et CHARLET 2001]. Puisque les connaissances ne sont pas des objets matériels, Bruno Bachimont a défini l'ingénierie des connaissances comme suit : « l'ingénierie des inscriptions numériques de connaissances ». Il se base sur un support numérique universel et homogène comme une substance d'inscription matérielle des connaissances. Il distingue deux branches dans cette approche :

— « *l'ingénierie des représentations formelles de connaissances* » : ingénierie ontologique ;
— « *l'ingénierie des inscriptions documentaires de connaissances* » : ingénierie documentaire. [BACHIMONT 2006]

4.4.3.1.5 Des environnements de publication

Les systèmes de gestion des contenus sont un groupe des logiciels qui permettent de créer des sites Web et garantissent leurs maintenances. Ils gèrent les différents types des contenus par l'ajout, la modification et la suppression des informations et assurent leur publication sur le Web. Par exemple, Spip, c'est un CMS de publication des articles à partir d'une interface graphique qui facilite la saisie, l'organisation et la gestion des informations. Les CMS jouent un rôle important dans le Web collaboratif en créant un type spécifique

du site Web : le blog. Ce dernier est « *enrichi régulièrement par un ou plusieurs personnes d'articles courts appelés billets ou posts ordonnés* » [GAUTIER 2006]. Il est basé sur le principe de communication entre l'auteur ou les auteurs de l'article et les visiteurs du blog à travers la possibilité du partage des commentaires sous chaque billet. Contrairement au Wiki, un autre type du site Web collaboratif, qui donne la possibilité à tous les visiteurs d'intervenir et changer le contenu des différentes pages pour former un contenu structuré. [GAUTIER 2006][CHAUMETTE et DESBIENS 2008].

4.4.3.2 Détermination des critères

Cette étape consiste à déterminer les critères d'évaluation des Environnements d'ECM du point de vue de l'Organisation des Connaissances. L'objectif de cette évaluation est de mettre en évidence la gestion des connaissances en s'appuyant sur les langages documentaires (comme les thésaurus) ou les langages représentatifs (comme les ontologies) ou une infrastructure de classement implicite. Dans ce qui suit, nous répondons aux questions suivantes :

— Sur quel type de SOC l'environnement est-il basé ?
— Comment le Cycle de vie de document est-il géré ?
— Quel est le degré d'interopérabilité de l'environnement ?
— Quelles sont les finalités collectives visées au sein d'une organisation ?

Ces questions concernent respectivement les critères présentés dans les sections 4.4.3.2.1, 4.4.3.2.2, 4.4.3.2.3 et 4.4.3.2.4.

4.4.3.2.1 Les critères liés au SOC

Ces critères sont :

— Le degré de formalité des SOC utilisés : les Systèmes d'Organisation des Connaissances peuvent appartenir à l'une des familles, présentées précédemment dans le premier chapitre, allant du plus au moins formel :
  — les ontologies formelles et le web sémantique ;

- les classifications épistémiques universelles de la bibliothéconomie et les approches à facettes universelles ;
- les langages documentaires et les thésaurus ;
- les approches multidimensionnelles : les ontologies sémiotiques (web socio-sémantique) et les approches à facettes locales ;
- les annuaires de ressources internet collaboratifs et les folksonomies ;
- les index automatiques des moteurs des recherches.

— L'évolution du SOC en cours d'utilisation par l'indexation basée sur les langages documentaires ou l'indexation collaborative. Par exemple, la contribution de l'utilisateur à l'indexation d'un document sur le web qui permet l'enrichissement de la liste des index ou l'évolution d'une ontologie dans le cadre du Web sémantique.

4.4.3.2.2 Les critères liés au document

Ces critères sont présentés comme suit :

— Le type du document : un fichier, un mail, un blog, une page HTML, DOPA. Il peut être individuel ou partagé.

— Son cycle de vie : il peut être géré soit par des métadonnées sur les actions effectuées sur le document, soit par un outil de maitrise de cycle de vie du document comme le workflow.

— La localisation des schémas de métadonnées : ils représentent un composant conceptuel fondamental de la classification documentaire [CALDERAN et al. 2008]. Ce critère correspond à la manière dont les MD sont stockés à l'intérieur des documents ou dans une base des données externe.

— L'espace de classement du document : il peut être un espace de stockage local, un espace de stockage distant (disque dur partagé) ou un système de gestion (GED ou CMS). [CHOY et al. 2010]

4.4.3.2.3 L'interopérabilité

L'interopérabilité avec les autres environnements d'ECM, qui peut être atteinte en suivant des standards tels que CMIS, http, REST, etc. Par exemple l'intégration d'un CMS à travers des connecteurs.

4.4.3.2.4 La collaboration

Ce critère représente la contribution de l'environnement dans la veille collaborative et l'intégration d'un système d'interopérabilité et du développement du travail collaboratif centré sur l'organisation des connaissances [SALZANO et ANKOUD 2012][HUDON et HEDI 2010]. Ce critère est supporté par d'autres critères à des finalités collaboratives : le type de l'indexation, le type de document indexé et classé et la gestion des activités collectives.

4.4.3.3 Positionnement d'HyperTaging

Le prototype d'Hypertagging, fondé sur une approche à facettes multidimensionnelles orientée métier, offre une interface de documentarisation classificatoire unique qui permette une semi-automatisation des opérations de nommage, de classement et d'indexation.
La simplicité de cette interface donne une responsabilisation aux usagers dans l'enrichissement du SOC et du document dans un environnement du travail collaboratif, en s'appuyant sur le principe fondateur du Web 2.0 (Read-Write Web), l'UGC (User Generated Contents). Ce principe est à la base des services proposés par les plateformes de partage de contenus numériques et les « réseaux sociaux ». Ce prototype met ainsi en œuvre un système d'indexation collaborative, son fonctionnement s'appuie sur l'élément graphique de classification « Tag », « *permettant de propager les métadonnées issues de l'annotation (tagging) dans l'environnement numérique et de les exploiter durant la recherche* ». [SALZANO et ANKOUD 2012][ZACKLAD et al. 2011a][POMPIDOR 2009]
Le cycle de vie du document indexé et des éléments de classification est géré par une conception évoluée de base des données qui décrit toutes les évolutions temporelles de ces composants.

Hypertagging s'intègre au système de fichiers Windows et à des espaces partagés de stockage des fichiers, gérés par exemple à l'aide d'un logiciel comme Dropbox, logiciel qui permet de stocker, synchroniser et partager des fichiers en ligne, basé sur la notion du « Cloud Computing ». Il se situe en priorité aux niveaux individuel (poste de travail) et collectif (par exemple au sein d'un service ou d'un département d'entreprise). A ces niveaux, les outils de production et échange de documents et informations les plus fréquents sont par exemple les outils bureautiques, messageries, blogs, FAQ et forums, wikis, DOPA et espaces partagés. Hypertagging privilégie deux exigences non fonctionnelles : l'usabilité, en termes d'ergonomie et de prise en compte des perspectifs métiers des utilisateurs ; l'interopérabilité, notamment en termes de cœxistence et échanges avec les autres composants du SI, comme les CMS (Joomla, SPIP, ...), les GED (SharePoint, Alfresco ...) et les serveurs des données. Les apports d'HyperTagging au travail collaboratif comme le support du document de type DOPA, l'indexation collaborative, l'interopérabilité avec d'autres environnements et le classement collectif des documents représentent une contribution à la veille collaborative.
Dans le tableau récapitulatif 4.6, nous présentons les relations directes et les similarités entre HyperTagging et chacune des cinq classes d'ECM.

4.4.3.4 Apport d'HyperTagging à la veille technologique

En 1996, D. Rouach définissait la veille comme « *l'art de repérer, collecter, traiter, stocker des informations et des signaux pertinents (faibles, forts) qui vont irriguer l'entreprise à tous les niveaux de rentabilité, permettre d'orienter le futur (technologique, commercial...) et également de protéger le présent et l'avenir face aux attaques de la concurrence. La veille se pratique dans la légalité et le respect des règles de déontologie. En d'autres termes, la veille conduit à transférer des connaissances de l'extérieur vers l'intérieur de l'entreprise, en suivant des règles de fonctionnement strictes.* » [ROUACH 1996].

La veille s'organise selon un processus itératif, chaque itération comprenant quatre phases :

— Phase 1 : définition des besoins de recherche ;

— Phase 2 : collecte de l'information ;

ECM	HyperTagging
Environnement du GED	Points de similarité : — Tous les deux gèrent le cycle de vie du document mais en deux manières différentes. Dans HT, cette gestion est réalisée par une conception évoluée de base des données (la gestion des métadonnées), et par le Workflow pour la GED ; Relations directes : — Interopérabilité : la possibilité d'échange et de coexistence entre HT et GED ; — GED représente un espace de classement pour HT.
Environnement du Groupeware	Points de similarité : — Coordination : HT permet une création des dimensions, des vues et des Tags collaboratifs au sein de l'entreprise ou d'un groupe de l'organisme ; — Communication : elle est assurée dans HT à travers la recherche des documents collaborés ; — Coordination et communication sont deux principes fondamentaux du Groupeware.
Environnement du Web 2.0	Points de similarité : — L'indexation collaborative ; — Le principe d'UGC.
Environnement d'ingénierie des connaissances	Relations directes : — Le langage de description des connaissances par des facettes dans HT s'inscrit sous la branche d'ingénierie des connaissances.
Environnement de la publication	Relations directes : — Représente une source documentaire pour HyperTagging sous la forme d'un DOPA ; — Interopérabilité : la possibilité d'échange et de coexistence.

TABLE 4.6 – Liaisons et similarités entre HyperTagging et les classes d'ECM

— Phase 3 : analyse et traitement de l'information ;

— Phase 4 : diffusion de l'information.

HyperTagging contribue aux premières phases, il permet de définir les besoins de recherche et la collecte de l'information par la définition des éléments de classification (vues, facettes et tags) et l'indexation collaborative. Son intégration de documents issus d'environnements hétérogènes au sein d'un même environnement facilite la phase d'analyse et de traitement de l'information. La phase de diffusion de l'information est supportée par les fonctionnalités de partage de documents et de plans de classement parmi des collectifs et des utilisateurs individuels. [SALZANO et ANKOUD 2012]

A l'aide d'HyperTagging, le veilleur peut classer certaines ressources pertinentes à sa recherche comme les brevets, les articles scientifiques, les rapports de recherche et les thèses, sur des unités partagées, en leur associant des métadonnées, et ainsi les diffuser à des collaborateurs.
Le fonctionnement d'HyperTagging s'adresse aux différents acteurs au sein de l'entreprise réalisant des activités de veille au quotidien pour :

— des besoins individuels, représentant leurs propres responsabilités, procédures, étapes du classement, indexation et recherche de documents au niveau privé ;

— des besoins collectifs, représentant un domaine d'information partagé qui couvre des activités de veilleurs « professionnels » : identifier des sujets émergents, par la détection de signaux « faibles », faire émerger de nouvelles pratiques et de nouveaux modes de représentation collective des savoirs, relier des personnes avec des centres d'intérêt proches.

Cet aspect de la gestion des activités collaboratives permet à HyperTagging de supporter les processus d'innovation en entreprise. Par exemple, dans un environnement de type projet, il permet de définir une ou plusieurs vues en correspondance de chaque stade du projet, avec les dimensions et tags associés. Ainsi, par exemple :

— une vue « Veille multifacette » peut comprendre les dimensions : « Marchés potentiels», « Concurrence », dans lesquelles on définira par exemple les tags « Client »,

« Produit » ;

— une vue « Veille réglementaire » peut comprendre les dimensions « France », « Union Européenne », « Monde », dans lesquelles on définira par exemple les tags « Energie », « Transport », « Santé ».

FIGURE 4.2 – Exemple des deux vues : « Veille réglementaire » et « Veille multifacette »

	Nom	Date classement	Taille
	transavia-Energie-Transport.pdf	10/09/2012 02:07	83,00 KB
	Partage de facette-Directive Inspire.docx	08/10/2012 08:21	72,00 KB
	Modèle_ppt.ppt	10/10/2012 08:29	260,00
	Image-Energie.JPG	08/10/2012 08:15	2,00 MB
	ETC	10/09/2012 02:12	3,00 MB

FIGURE 4.3 – Exemple des documents classés

L'analyse de tags associés à une veille réglementaire permet de mesurer les impacts et opportunités liés aux règlementations les plus récentes, aussi l'approche bottom-up pour le nommage et le classement de documents, adoptée dans HyperTagging, contribue à faire émerger l'apparition et la consolidation de certaines vues, dimensions, certains tags et donc la génération d'idées nouvelles.

4.4.4 Cas d'étude

Ce modèle d'évaluation peut être utilisé pour tout SI de gestion de contenu. Nous avons choisi d'étudier l'exemple d'Alfresco.

4.4.4.1 Cadre et contexte d'évaluation d'Alfresco

Alfresco est une solution ECM, une plateforme de collaboration et de gestion de contenu d'entreprise. Elle permet de gérer des processus à gros volumes de documents et collaborer en toute sécurité. Cet ECM offre différents services, nous nous intéressons à la gestion de documents et aux différents outils, stratégies et méthodes utilisés pour la classification documentaire.

4.4.4.2 Caractéristiques d'Alfresco

Alfresco est capable de gérer de multiples moyens de classification de documents tout en respectant les différentes natures de documents.
Il gère plusieurs types du document : courriers, factures, contrats, documents techniques, etc., dont le cycle de vie est assuré par un workflow.
Le schéma des métadonnées de chaque document (Nom, titre, description, type, type MIME, date de création, auteur, taille, date de dernière modification, modificateur) ainsi que les métadonnées de classifications (mots clé, étiquettes (tags)) sont stockés dans une base de données (PostgreSql pour la version Alfresco 4.0). Ces données sont stockées suivant un des modèles de données de base d'alfresco : sous la forme des graphes dans des fichiers XML. Chaque graphe est un ensemble des nœuds liées par différentes associations. Un nœud est défini par son type, ce dernier est composé de propriétés (métadonnées utilisées lors de traitement du nœud), aspects (ensemble des propriétés et associations qui peuvent être communes entre plusieurs types de nœuds) et associations (relations entre les nœuds de type source-destination ou parent-enfant). Concernant le stockage de documents, il existe deux possibilités. La première consiste à les stocker sous forme des documents cryptés dans le système des fichiers géré par Alfresco, par défaut dans le dossier alf_data/contentstore. La deuxième possibilité est l'utilisation du nuage avec l'application Alfresco Cloud comme espace du stockage de documents.

Sur cette plateforme, des différents schémas classificatoires sont intégrés :

— Le plan de classement : l'arborescence des dossiers et sous-dossiers dans laquelle se fait le stockage des contenus. Il s'agit d'une organisation hiérarchique ou verticale en espaces et sous-espaces semblable à l'organisation dans un système de fichiers classique. La différence majeure entre les dossiers d'un système de fichier classique et les espaces d'Alfresco appelés « dossiers intelligents », c'est qu'ils se caractérisent par une configuration des règles de gestion du contenu à travers des attributs spécifiques pour appliquer des actions sur les documents qu'ils contiennent.

— Les catégories : Alfresco permet aussi de classer les documents par catégories en formant un vocabulaire hiérarchique des mots-clés. C'est une organisation selon une taxinomie tout en fournissant une recherche de contenu performante avec un filtre de navigation « catégorie ».

— Les tags : comme les catégories, ils permettent aux utilisateurs de naviguer d'une manière sémantique dans l'entrepôt. Par contre, ils représentent des mots-clés non-organisés, créés librement par les utilisateurs finaux. Les tags permettent une classification de contenus par une folksonomie. En effet, tout contenu peut être classé par un ou plusieurs tags déjà existants ou nouvellement créés. Une recherche de contenu avancée est mise en place par un filtre de navigation « Tag », qui permet d'afficher tous les contenus étiquetés par le tag sélectionné.

Le travail collaboratif est assuré par le workfow. En effet ce dernier gère les différents états d'un document : 'Envoyer pour réviser', 'Réviser et approuver', 'Réviser et approuver en mode groupe' et 'Réviser et approuver en mode partagé'.

L'architecture d'Alfresco est une SOA, composée principalement de trois niveaux : Client, Serveur et Stockage (la base des données et le fichier des systèmes). La communication entre les clients et le serveur d'application est assurée par des services utilisant de nombreux protocoles comme http, FTP, RMI et SOAP. Elle est aussi normalisée par le protocole CMIS, ce denier lui permet de communiquer avec d'autres ECM.

Exemple d'utilisation :

Dans un bureau d'études de structure dans le domaine du bâtiment et du génie civil spécialisé dans la construction et le renouvellement des logements, un rapport technique est rédigé pour chaque projet ou chantier. Les logements sont différenciés selon différentes typologies :

— Logement social, public ou privé ;

— Chambre, appartement ou maison ;

— Individuel ou collectif.

La construction de logements doit respecter les normes de qualité d'ISO comme ISO 91.010 pour "Industrie de la construction", ISO 91.080 pour "Structures de construction", ISO 91.200 pour "Techniques de construction", etc., et des normes environnementales sur la performance énergétique, par exemple, la norme "Haute performance énergétique" (HPE 2005)et la norme "Haute performance énergétique Énergies renouvelables" (HPE ENR 2005).

La figure suivante montre l'interface d'Alfresco pour le classement des rapports techniques selon le type du logement, les normes d'ISO et les normes environnementales. Ces derniers sont représentés dans ce système par des tags. Comme le montre la figure ci-dessous, pour chaque document après son nom, la date de sa modification, son auteur et sa taille, une liste d'étiquettes (tags) lui est associée.

Tous les tags créés appartiennent à la même liste Tags qui apparaît à droite de la figure en bas avec le nombre de documents utilisant chaque tag, sans prendre en considération les contextes comme dans HyperTagging.

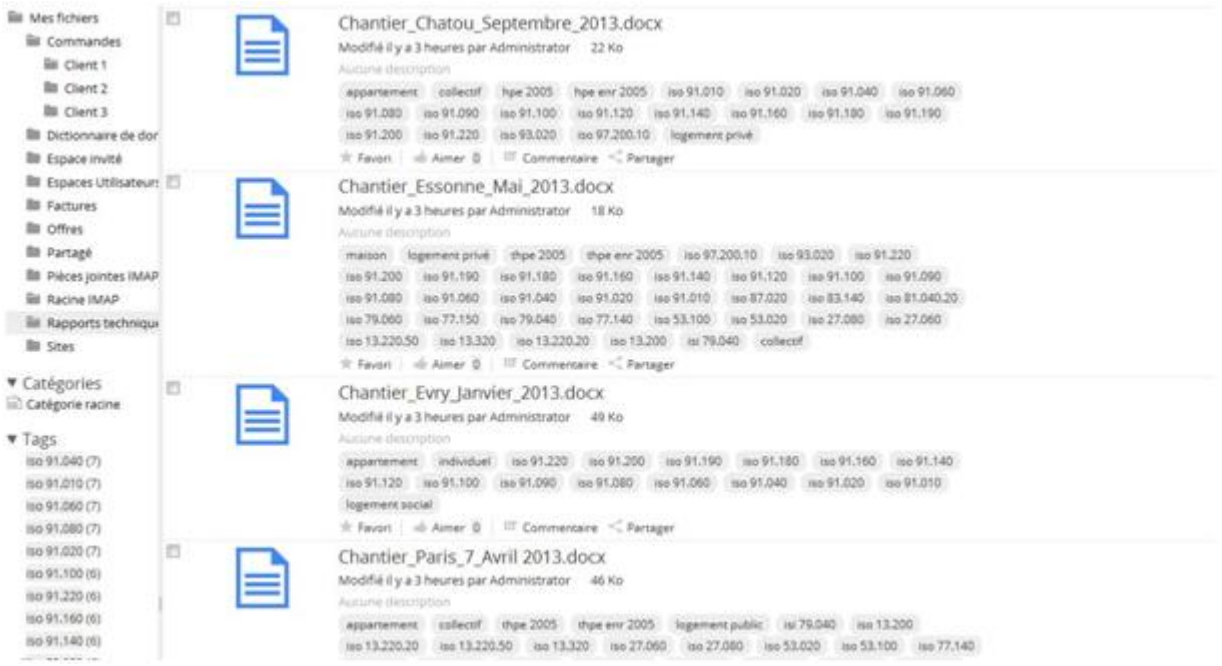

FIGURE 4.4 – Interface d'Alfresco

4.5 Conclusion

Ce chapitre avait comme premier objectif le positionnement d'HyperTagging dans l'environnement des ECM. Après avoir étudié les évaluations des ECM réalisées par des analystes spécialistes utilisant des critères qui ne correspondent pas à nos objectifs dans cette étude, nous avons défini un modèle d'évaluation basé sur des critères particuliers, orientés classification documentaire. Dans ce cadre, nous avons positionné HyperTagging par rapport aux différents environnements d'ECM (les environnements du GED, les environnements du travail collaboratif, les environnements du Web 2.0, les environnements d'ingénierie des connaissances et les environnements de publication), ainsi que par rapport à ses apports aux activités de la veille. Pour conclure, nous avons présenté un exemple d'application de notre modèle d'évaluation sur l'ECM Alfresco.

Conclusion et perspectives

Le travail avec des groupes de chercheurs de différentes disciplines comme les chercheurs (documentalistes) de la science de l'information et communication ainsi que les chercheurs de l'informatique nous a facilité l'accès au monde de la recherche scientifique. Cette expérience nous a permis de consolider et d'enrichir nos connaissances dans plusieurs domaines adjacents : l'organisation des connaissances, l'ingénierie dirigée par les modèles et les architectures Web. Nous avons représenté dans ce mémoire nos connaissances acquises sous formes des chapitres.

Dans l'introduction générale, nous avons présenté le contexte de cette recherche, son cadre, le projet Miipa-doc ; notre contribution sous la forme de deux approches « Top Down » et « Bottom Up » ; et l'organisation du manuscrit.

Le premier chapitre représente une étude de l'organisation des connaissances, les différents systèmes d'organisation des connaissances (SOC), la classification au sens général, la classification documentaire, la classification des documents codifiés, les différents schémas classificatoires et les structures organisationnelles.

La problématique de recherche présentée dans le deuxième chapitre concerne la gestion d'hétérogénéité d'un système d'information de documentarisation classificatoire, basé sur une approche orientée métiers et un SOC folksonomique à facettes, collaboratif, multidimensionnel et synchronisé, ainsi que le positionnement du prototype HyperTagging dans l'environnement des ECM. Nous avons identifié les différents niveaux d'abstraction et les préoccupations conceptuelles de ce type du SI ainsi que les méthodologies et les technologies à mettre en œuvre pour répondre à cette problématique.

Le troisième chapitre décrit notre démarche, progressive et pluridisciplinaire. Nous avons montré comment une approche dirigée par les modèles permet de gérer l'hétérogénéité entre les différents niveaux d'abstraction du SI : les principes de gestion de l'interaction homme machine, l'évolutivité et la temporalité des éléments du SOC, l'évolution architecturale par la synchronisation du protocole REST et l'interopérabilité avec autres SI à travers CMIS. Ces principes, qui s'appuient sur l'analyse de la littérature, concernent aussi bien le niveau technologique que méthodologique.

Les évolutions préconisées étendent les recherches liées aux besoins d'intégrité documentaire dans des environnements collaboratifs. Les cycles de vie associent des informations

temporelles aux actions effectuées (classement, mise à jour, suppression...). Ils permettent ainsi de construire des historiques enrichis des évolutions des documents classés et des plans de classement opérés par les utilisateurs.
Dans un objectif d'indexation collaborative, ce SI couvre les fonctions de nommage, indexation et classification de documents : le suivi dans le temps de ces fonctions «responsabilise» davantage les utilisateurs dans les étapes préalables à la recherche.

Dans le dernier chapitre, nous avons présenté le positionnement d'HyperTagging par rapport aux sous-familles (les classes) d'ECM. L'HyperTagging est un prototype d'application permettant la gestion de Systèmes d'Organisation des Connaissances hétérogènes, de manière ascendante et multifacette. En suivant un modèle d'évaluation nous avons présenté les classes d'ECM qui sont au nombre de cinq (les environnements du GED, les environnements du travail collaboratif, les environnements du Web 2.0, les environnements d'ingénierie des connaissances et les environnements de publication). Nous avons déterminé les critères d'évaluation qui ont permis de positionner et évaluer notre approche.

Avec cette étude, nous espérons contribuer à répondre aux besoins des entreprises pour le choix d'un environnement d'ECM le plus adapté à leurs besoins d'organisation des connaissances mais également offrir des recommandations de conception et de méthode pour l'environnement HyperTagging développé dans le projet ANR Miipa-Doc.
Nous esquissons maintenant quelques perspectives qui nous semblent intéressantes, voire importantes pour élargir le champ d'application de nos contributions. Nous envisageons de :

— formaliser des scénarios significatifs de gestion des évolutions des SOC, en environnement distribué et hétérogène. Ceci permettra de spécifier le service de gestion des évolutions des éléments d'HyperTagging, avec leurs contextes temporels, et d'étendre l'architecture actuelle par les composants informatiques associés à ce service.
— poursuivre le processus d'évaluation par l'évaluation de l'efficacité de notre méthode à travers les retours d'expériences.
— proposer une nouvelle classification multi-vues des environnements ECM, basée sur les critères que nous avons déterminés, et en montrant l'impact de la mise en œuvre

sur le processus métier de l'entreprise

— représenter des relations entre les documents, ceci permettrait de représenter le temps et l'espace par des relations telles que «avant», «dans le même contexte», etc. La prise en compte de relations permettrait de plus d'exprimer les liens de dépendances entre documents.

— présenter une interface en langage naturel. Comme notre navigation se présente sous la forme d'un dialogue homme-machine, où chacun peut poser aussi bien des questions que des affirmations, nous voyons une connexion possible avec le traitement des langues naturelles tel que les graphes conceptuels, dont la navigation donnerait le principe d'interprétation des phrases de l'humain et de génération des phrases de la machine.

— Implémenter une version mobile d'HyperTagging. Puisque devant la croissance exponentielle de la mobilité dans le marché de consommation des Smartphones, les utilisateurs sont attirés de plus en plus par ces nouvelles solutions mobiles.

Bibliographie

F. ABEL, N. HENZE, R. KAWASE, D. KRAUSE, and P. SIEHNDEL. Tagme ! - enhancing social tagging with spatial context. 2011. 57, 86

J. AKOKA and I. COMYN-WATTIAU. Vers l'ingerie des évolutions, dans ingénierie des systèmes dinformation. *RSTI, série ISI*, Vol. 14, n 6/2009 :p9–17, 2009. 111

R. ALARCON and E. WILDE. Restler : Crawling restful services. 2010. 24, 127

M. ANKOUD and M. HMIMIDA. Un modèle d'èvaluation d'un soc dans l'environnement d'ecm prototype : Hypertaging. *ISKO-Maghreb'2013 : Concepts et Outils pour le Management de la Connaissance*, 2013. 25, 156

N. AUSSENAC-GILLES and J. CHARLET. Ingénierie des connaissances : Introduction. *Cours de DEA/DESS Gracq, ingénierie des connaissances*, 2001. 159

B. BACHIMONT. Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *Cours de DEA/DESS Gracq, ingénierie des connaissances*, 2000. 33

B. BACHIMONT. Arts et sciences du numérique : 'ingénierie des connaissances et critique de la raison computationnelle'. *Mémoire d'habilitation à diriger des recherches*, 2004. 33, 34

B. BACHIMONT. Ingénierie des connaissances. 2006. URL `http://www.utc.fr/~bachimon/Recherche.html`. mise en ligne en 2006, visitn 2012. 159

B. BACHIMONT. Ingénierie des connaissances et des contenus : Le numérique entre les ontologies et les documents. 2007. 34

K. LA BARRE. Facet analysis. *Annual Review of Information Science and Technology*, 44(1) :p243–284, 2010. 55

C. BEGHTOL. Relationships in classificatory structure and meaning, dans c.a. bean et r. green (dir.), relationships in the organization of knowledge. norwell, ma : Ed. kluwer academic publishers. 2001. 48, 49

C. BEGHTOL. From the universe of knowledge to the universe of concepts : the structural revolution in classification for information retrieval. *Axiomathes*, 18(2) :p131–144, 2008. 51

J. BEZIVIN. La transformation de modéles. *INRIA-ATLAS Université de Nantes, Ecole dEté dInformatique CEA EDF INRIA 2003*, 18(2), 2003. 22, 81

G-C. BOWKER, K. BAKER, F. MILLERAND, and D. RIBES. Toward information infrastructure studies : Ways of knowing in a networked environment. *International Handbook of Internet Research 2010*, pages p97–117, 2010. 91

A. BURGUN and O. BODENREIDER. Mapping the umls semantic network into general ontologies. *Proceedings of the AMIA Annual Symposium*, pages p81–85, 2001. 37

L. CALDERAN, B. HIDOINE, and J. MILLET. Métadonnées : Mutations et perspectives. *Séminaire INRIA*, pages p81–85, 2008. 161

H. CHAKER, M. CHEVALIER, C. SOULE-DUPUY, and A. TRICOT. Adapter le contexte métier pour améliorer le système de recherche d'information. *Actes de l'Atelier ERATSI (Evolution, Réutilisation, Adaptation et Trailits Systèmes dInformation) : Associé au 28ème Congrès INFORSID 2010*, 2010. 79

J. CHALET, M. ZACKLAD, G. KASSEL, and D. BOURIGAULT. Ingénierie des connaissances : recherches et perspectives. *Ingénierie des connaissances*, 2004. 33

B. CHANDRASEKARAN, J-R. JOSEPHSON, and V-R. BENJAMINS. What are ontologies and why do we need them ? *IEEE Intelligent Systems*, 14(1) :p20–26, 1999. 37

C. CHAUMETTE and D. DESBIENS. Typologie des systèmes d'information. *Universit Lille 3*, 2008. 158, 160

D. CHOY, A. BROWN, R. MCVEIGH, and F. MULLER. Content management interoperability services (cmis), version 1.0. *OASIS Standard, OASIS*, 2010. 161

E-F. CODD. A relational model of data for large shared data banks. *Communications of the ACM*, 13(6) :p377387, 1970. 63

B. COMBEMALE. Ingénierie dirigée par les modèles (idm) : Etat de l'art. 2009. 79, 80

D. COTTE. L'organisation des connaissances entre le formalisme des outils et la complexité des représentations : une illustration par le cas des portails d'entreprise. *Dans Sixiéme Congrés du chapitre frans de l'ISKO, Organisation des connaissances et société des savoirs : concepts, usages, acteurs* , 2007. 35

J-F. COUTURIER. Elaboration et expérimentation d'une méthodologie agile permettant la migration vers une architecture orientervices en pme à l'aide d'openup. 2011. 117, 118, 120, 123

G. DELAUNEY. La place de la bibliothéconomie dans l'organisation des connaissances et les classifications. 2010. 51

W. DENTON. How to make a faceted classification and put it on the web. 2003. 45

O. LE DEUFF. Folksonomies : Les usagers indexent le web. *BBF*, n 4 :p66–70, 2007. 39

R. DIENG. Méthodes et outils pour la gestion des connaissances. *Dunod*, 2001. 158

O. DESFRICHES DORIA. La classification à facettes pour la gestion des connaissances métier : Méthodologie d'élaboration de folkclassification à facettes. *Mémoire de thèse*, 2013. soutenue le 26 Novembre 2013. 45, 46, 48, 49, 50, 54, 56, 70, 71, 86

R. DUCOUNAU, J. EUZENAT, G. MASINI, and A. NAPOLI. Langages et modéles à objet : Etat des recherches et perspectives. *inria-00340768*, version 1, 2008. 22, 65, 66, 67

A. DUDEZERT. Le km au cœur de la stratégie d'entreprise. *Documentaliste-Sciences de l'Information*, 49(2) :p26–43, 2012. 35

C. DUPIN. Gestion de contenu : de nouveaux horizons dans l'entreprise, de nouvelles perspectives pour les professionnels de l'information. *Documentaliste-Sciences de l'Information*, 47(3), 2010. 156

S. FERRE. Systèmes d'information logiques : un paradigme logico-contextuel pour interroger, naviguer et apprendre. *Mémoire de thèse*, 2002. soutenue en 21 octobre 2002. 22, 58, 59, 61, 62, 63, 65

R-T. FIELDING. Architectural styles and the design of network-based software architectures. *Doctoral dissertation*, 2000. 121

J-P. FIGER. Architectures orientées services. 2006. 118, 120, 123

G. GARDARIN. Bases des données. 2003. 5éme tirage. 22, 64

P. GAUTIER. Blogs et wikis : Quand le web s'appuie la société de l'information. *BBF*, 51(3), 2006. 160

M-R. GILBERT, K.M. SHEGA, K. CHIN, G. TAY, and H. KOEHLER-KRUENER. Magic quadrant for enterprise content management. *ID :G0023778.1*, 2012. 149, 150

M. GRISLIN and Ch. KOLSKI. Evaluation des interfaces homme-machine lors du développement des systèmes interactifs. *Technique et Science Informatique (TSI)*, pages p265–296, 1996. 83, 84

N. GUARINO. Some organizing principles for a unified top-level ontology. *Proceedings of the AIII Spring Symposium on Ontological Engineering*, 1997. 36, 37

B. HJORLAND. Semantics and knowledge organization. annual review of information science and technology. 41(1) :p367–405, 2007. 40, 48

B. HJORLAND. What is knowledge organization (ko) ? *Knowledge Organization*, 35(2-3) : p86–101, 2008. 34

B. HJORLAND. Knowledge organization = information organization ? *Knowledge Organization*, 13 :p8–14, 2012. 56

B. HJORLAND. Facet analysis : The logical approach to knowledge organization. *Information Processing & Management*, 49(2) :p545–557, 2013. 55

M. HUDON. Structuration du savoir et organisation des collections dans les répertoires du web. *BBF*, 46(1) :p57–62, 2001. 37

M. HUDON and W. MUSTAFA EL HEDI. Organisation des connaissances et des ressources documentaire : de l'organisation hiérarchique centralisée à l'organisation sociale distribuée. *Organisation des connaissances et de Web 2.0*, LCN(3), 2010. 26, 34, 39, 40, 41, 42, 43, 45, 46, 162

H. KARASTI, K-S. BAKER, and F. MILLERAND. Infrastructure time : Long-term matters in collaborative development. *Computer Supported Cooperative Work*, Vol. 19 : p377–415, 2010. Special Issue on Supporting Scientific Collaboration Through Cyberinfrastructure and e-Science. 111

H. KEFI and M. KALIKA. Evaluation des systs dinformation : une perspective organisationnelle. *ECONOMICA*, 2004. 155

B-H. KWASNIK. The role of classification in knowledge representation and discovery. *Library Trends*, 48(1) :p22–47, 1999. 43, 44, 46

J. MANIEZ. Des classifications aux thésaurus : du bon usage des facettes. *Documentaliste-Sciences de l'information*, 36(4-5) :p249–262, 1999. 38

J. MANIEZ. Actualits langages documentaires : fondements théoriques de la recherche d'information. *ADBS*, 2002. 43

P. POMPIDOR. Evolutions récentes des technologies web du côté client. le web 2.0 au service des applications riches. techniques de l'ingénieur. *Base Documentaire : Technologies logicielles Architectures des systèmes*, 2009. Référence H7440. 162

J. POTTS. Getting started with cmis. 2009. 145

S. PROULX, F. MILLERAND, and J. RUEFF. Web social : Mutation de la communication. *Presses de l'Université de Québec*, 2010. 159

E. QUINTARELLI, A. RESMINI, and L. ROSATI. Information architecture : Facetag : Integrating bottom-up and top-down classification in a social tagging system. *Bulletin of the American Society for Information Science and Technology*, 33(5) :p10–15, 2007. 57

S-R. RANGANATHAN. Classification and communication. *Dehli, Inde : Ess Ess Publication pour Sarada Ranganathan Endowment for Library Science*, 1951. 54

S-R. RANGANATHAN. Prolegomena to library classification. *Dehli, Inde : Ess Ess Publication pour Sarada Ranganathan Endowment for Library Science*, Vol. 1, 1967. 3ème ed. 49, 50, 51, 53

T-P. ROGER. Document : forme, signe et mum. *les reformulations du numque*, 2003. Version 3. 96

D. ROUACH. La veille technologique et lintelligence nomique. 1996. 163

L. SAADANI and S. BERTRAND-GASTALDY. Cartes conceptuelles et thésaurus : essai de comparaison entre deux modèles de représentation issus de différentes traditions disciplinaires. *ACSI-CAIS*, 2000. URL `http://www.cais-acsi.ca/2007call_fr.htm`. In : Kublik, A., ed. Les dimensions d'une science de l'information globale. 36

P. SALVAN. Les progrès de la classification à facettes. *BBF*, pages p5–17, 1962. num. 1. 38

G. SALZANO and M. ANKOUD. Hypertagging : un soc d'entreprise en environnement web 2.0. apports aux activités de veille. *Séminaire Veille Stratégique Scientifique et technologique (VSST) 2012*, 2012. 23, 24, 79, 114, 162, 165

J-S. SATTEL, G. CALVARY, and J-M. FAVRE. Ingénierie de l'interraction homme-machine dirigée par les modèles. 2005. 23, 83, 84, 96, 97, 100, 102

D. SCAPIN. Guide ergonomique de conception des interfaces homme-machine. 1986. 83

S. SCHREIER. Modeling restful applications. *WS-REST '11*, 2011. 21, 24, 81, 82, 130, 131, 132, 133

C. SHIRKY. Ontology is overrated : Categories, links, and tags. 2005. 56

B-C. VICKERY. La classification à facettes, guide pour la construction de schémas spéciaux. 1963. 44, 55

B-C. VICKERY. The structure of subject classification for document retrieval. 2008. 40

A. VULBEAU. Contrepoint - usages, usagers : les mots pour le dires. *Informations sociale*, 2(158) :p99–99, 2010. 78

W3C-ARC. W3c architecture du world wide web. Vol. 1, 2006. Traduction franse. [En ligne] http ://opikanoba.org/tr/w3c/webarch/ Document original en anglais [En ligne] http ://www.w3.org/TR/2004/REC-webarch-20041215/. 112

D. WALDT. Content management interoperability services (cmis). *OASIS CMIS Technical Committee*, 2009. 145

A. WEINTRAUBN. The forrester wave : Enterprise content management, q4 2011. *For Contents and Collaboretion Prefessionels*, 2011. 153

P-J. WILD, M-D. GIESS, and C-A. MC MAHON. Describing engineering documents with faceted approaches : observations and reflections. *Journal of Documentation*, 65(3) : p420– 445, 2009. 51, 52, 53, 56

M. ZACKLAD. Introduction aux ontologies sémiotiques dans le web socio sémantique. *actes de la conférence Ingénierie des Connaissances*, 2005a. 38, 39

M. ZACKLAD. Processus de documentarisation dans les documents pour l'action (dopa) : statut des annotations et technologies de la coopération associées (nouvelle version corrigée). *Dans Actes du colloque Le numérique : Impact sur le cycle de vie du document pour une analyse interdisciplinaire* , 2005b. 96

M. ZACKLAD. Une théorisation communicationnelle et documentaire des tic. *Hermes Science Publications*, 2007. in Reber, B., Brossaud, C., Humanités numériques 2 : socio-

informatique et démocratie cognitive (Traité IC2, série cognition et traitement de l'information). 21, 47, 48, 49, 58, 71

M. ZACKLAD. Evaluation des systémes d'organisation des connaissances. *Les cahiers du numérique*, 6(3) :p133–166, 2011b. 35, 36, 38

M. ZACKLAD, O. DESFRICHES-DORIA, B. BERTIN, S. MAHE, B. Ricard, N. MUSNIK, J-P. CAHIER, A. BENEL, and E. LEWKOWICZ. Miipa-doc : Gestion de l'hétérogénéité des classifications documentaires en entreprise. *Actes de la onzième édition de la conférence internationale H2PTM*, page p323p333, 2011a. 75, 76, 162

Publications

M. ANKOUD, M. HMIMIDA. Etude de l'évaluation des ECM. ISKO-Maghreb'2013 : Concepts et Outils pour le Management de la Connaissance, 8-9 Novembre 2013, Marrakech Maroc.

G. SALZANO, M. ANKOUD, O. DESFRICHES DORIA. Apports du SOC d'entreprise Hypertagging aux activités de veille. Numéro Spécial Document Numérique, ed. Lavoisier à paraître en 2013.

M. ANKOUD, M. HMIMIDA. Un modèle d'évaluation d'un SOC dans l'environnement d'ECM Prototype : HyperTaging. ISKO-Maghreb'2012 : Concepts et Outils pour le Management de la Connaissance, 3-4 Novembre 2012, Yasmine Hammamet Tunisia.

M. HMIMIDA, M. ANKOUD. Recommendation level in faceted classification for documentary classification. ICEELI'2012, 1-3 juillet Sousse, Tunisia, IEEE, pp1-5.

M. HMIMIDA, R. KANAWATI, A. ANKOUD. Nouveau Modèle de recommandation pour la classification à facettes. In Actes du 15ème Colloque International sur le Document numérique, CIDE15, Tunis, Tunisie, Europia, pp.145-157.

G. SALZANO, M. ANKOUD. Hypertagging : un SOC d'entreprise en environnement Web 2.0. Apports aux activités de veille. Séminaire Veille Stratégique Scientifique et technologique (VSST) 2012, Ajaccio (France). 24-25 mai 2012.

G. SALZANO, M. ANKOUD, M. HMIMIDA, M. ZACKLAD. Gestion des évolutions dans un SOC d'entreprise, multidimensionnel et distribué. Conférence Inforsid, 24-26 mai, Lille, 2011.

Annexes

Annexe A

Annexe

Exemples d'utilisation d'HyperTaging HypTaging vise des finalités collectives et des activités professionnelles dans des différents domaines par exemple la réalisation de la ville règlementaire par une collectivité dans le domaine de l'environnement. Nous dérivons les objectifs de chaque facette présentée dans la figure A.1 :

— « Nom du collecteur » et « Nom de l'analyste » servent à indiquer si le document a été collecté et analysé par deux personnes différentes et lesquelles. Les valeurs sont des initiales de noms de personnes.
— « Thématique » indique quel aspect environnemental est principalement concerné.
— « État du texte » est dédié au degré d'officialisation d'un texte de loi.
— « Type de texte de loi » décrit les différents types de textes de loi qui sont susceptibles d'être indexés dans le cadre de cette veille règlementaire.
— « Source » signale quel est l'organisme officiel dont est issu le document à indexer.
— « Périmètre géographique » fournit des informations sur le périmètre géographique concerné par le contenu du document.
— « Type de signal » a pour objectif d'indiquer à quel type de signal le contenu correspond, une fois le document analysé.
— « Applicabilité » et « Conformité » sont des éléments utiles aux veilleurs redans le cadre d'une veille règlementaire, elles participent de l'analyse des documents repérés pendant la collecte.

— « Rubrique ICPE Substance » et « Rubrique ICPE Activités » correspondent au premier niveau de classification de l'Inspection des installations classées du ministère de l'Écologie, du Développement durable et de l'Énergie. ICPE signifie installation classée pour la protection de l'environnement. Ces dimensions permettent une correspondance avec la classification officielle des substances et des activités dans le domaine environnemental et constituent un repère pour des personnes qui travaillent habituellement avec ces catégories.

FIGURE A.1 – Exemple d'une vue de la veille règlementaire environnementale

Nous citons un deuxième exemple, par la figure A.2, qui illustre la capacité d'HyperTaging à supporter la gestion du projet. Les facettes de cette figure représentent les métadonnées suivantes :

— Phase du projet : correspond aux tâches à effectuer successivement pour la mise en œuvre de la directive Inspire.
— Partenaire : indique les interlocuteurs avec lesquels les personnes chargées d'appliquer cette directive vont collaborer.
— Types de documents : ce sont les types documents qui vont constituer des ressources ou être produits pendant la mise en œuvre.
— Jeux de données collectées : il s'agit des types de données géolocalisées sur lesquelles les personnes de la collectivité récoltent des informations.
— Version des documents : il s'agit de l'état de validation des documents relatifs au projet.
— Thèmes : il s'agit de thèmes relatifs à la mise en œuvre de la directive Inspire.

FIGURE A.2 – Exemple d'une vue de gestion du projet

Manel ANKOUD

Gestion de l'hétérogénéité dun SI de classification documentaire multifacette et positionnement dans l'environnement des ECM

Résumé : L'organisation des connaissances est une discipline investie par des bibliothécaires, documentalistes, archivistes spécialistes de l'information, informaticiens et tous professionnels du document. Elle englobe toutes activités, études et recherches qui élaborent et traitent les processus d'organisation et de présentation des ressources documentaires utiles dans une organisation. Dans ce contexte, le projet ANR Miipa-Doc a pour objectifs d'explorer de nouvelles méthodes d'indexation ascendantes, en utilisant des termes descripteurs formulés par les individus plutôt que choisis parmi une liste préétablie, pour l'organisation des contenus documentaires complexes au sein des entreprises de large taille, et concevoir l'architecture logicielle correspondante.
Dans ce projet, notre contribution consiste à gérer l'hétérogénéité d'un système d'information d'organisation des contenus documentaires, basé sur une approche orientée métier et un SOC (système d'organisation des connaissances) folksonomique à facettes. Nous proposons dans cette gestion une approche incrémentale dirigée par les modèles, issue de l'IDM (ingénierie dirigée par les modèles), basée sur des méta-modèles pour garantir l'aspect d'évolutivité. Après l'implémentation du prototype HyperTaging qui met en place ces deux approches, nous proposons un processus d'évaluation permet de positionner ce prototype et tous SI de classification documentaire dans l'environnement des ECM, en se basant sur des critères d'évaluation fins et particuliers.
Mots-clés : organisation des connaissances, indexation, classification documentaire, IDM, SOC, folksonomie à facettes, système d'information, processus d'évaluation, critères d'évaluation, ECM

Abstract : The knowledge organization is invested by librarians, archivists, information specialists, IT professionals and all discipline of document. It includes all activities, studies and research which develop and treat organization process and presentation of relevant information resources in an organization. In this context the Miipa-Doc project aims to explore new ascendants indexing methods, using descriptors made by individuals rather than selected given list for complex contained in the organization document, in large size companies, and design the corresponding software architecture.
Our contribution in this project is to manage the heterogeneity of an information system of document organization, based on a business-oriented approach and a KOS (knowledge organization system) of folksonomy facet. We propose an incremental approach this management model driven, outcome of MDE (Model Driven Engineering), based on meta-models to ensure scalability appearance. After implementing the HyperTaging prototype, that implements both approaches, we propose an evaluation process used to position the prototype and all IS of documentary classification in the environment of ECM based on purposes of delicate and particular evaluation criteria.
Keywords : knowledge organization, indexing, document classification, MDE, KOS, folksonomy facet, information system, evaluation process, evaluation criteria,ECM