Miipa-Doc : gestion de l'hétérogénéité des classifications documentaires en entreprise

Manuel Zacklad, Orélie Desfriches-Doria, Gilles Bertin, Sylvain Mahé, Benoit Ricard, Noémie Musnik, Jean-Pierre Cahier, Aurélien Bénel, Emmanuel Lewkowicz

Manuel Zacklad, Orélie Desfriches-Doria, Gilles Bertin, Sylvain Mahe, Benoit Ricard, Noémie Musnik, Jean-Pierre Cahier, Aurélien Bénel, Emmanuel Lewkowicz, "Miipa-Doc : application, infrastructure et méthode de gestion de l'hétérogénéité des classifications documentaires en entreprise", Conférence H2TPM'2011, 12-14 octobre 2011, Metz (France), Hermes, 2011

Miipa-Doc : application, infrastructure et méthode de gestion de l'hétérogénéité des classifications documentaires en entreprise[1]

Manuel Zacklad[1,3], Orélie Desfriches-Doria[1], Gilles Bertin[1], Sylvain Mahe[2], Benoit Ricard[2], Noémie Musnik[2, 1], Jean-Pierre Cahier[3], Aurélien Bénel[3], Emmanuel Lewkowicz[3,4 (2)]

[1] CNAM, laboratoire DICEN, EA 4420,
{manuel.zacklad, orelie.desfriches_doria, bertin}@cnam.fr
[2] EDF, DER, STEP,
{sylvain.mahe, benoit.ricard, noemie.musnik}@edf.fr
[3] UTT, ICD, UMR 6279 CNRS, Equipe Tech-CICO
{Jean-Pierre.cahier, aurelien.benel, emmanuel.lewkowicz}@utt.fr
[4]Cogniva Europe SAS

Résumé : Dans le cadre du projet ANR Miipa-Doc « Méthodes et Services Intégrés Institutionnels et Participatifs pour la Classification à Facettes des Contenus Documentaires Complexes », nous visons à développer de nouvelles approches méthodologiques et logicielles pour permettre une gestion plus intégrée de la documentation d'entreprise qui facilite la gestion de l'hétérogénéité des modalités de classification et de classement du patrimoine documentaire. Dans cet article nous présenterons les trois principales contributions techniques du projet : (1) l'application gestion multidimensionnelle et ascendante des SOC, Hypertagging (2) l'infrastructure à base de Systèmes d'Organisation des Connaissances Hétérogènes, SOC-H[3] et (3) la méthode d'Analyse Sémiotique des Transactions Documentaires, ASSET-Doc. Nous présenterons également un premier résultat des études de terrain effectuées qui ont contribué à permettre l'élaboration de la méthode ASSET-Doc et la spécification fonctionnelle d'Hypertagging.

Mots-clés : pratiques collectives et distribuées, ingénierie documentaire, ingénierie des connaissances pour l'entreprise, web socio-sémantique.

---

[1] Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet Miipa-Doc n°2008 CORD 014 03

[2] Les auteurs remercient également les autres contributeurs passés au projet ou non impliqués dans cette rédaction H. Zaher, C. Zhou, C. Vaucelle, M. Diagne, P. Vasante, V. Vijayaraj et ceux qui l'ont rejoint récemment, G. Salzanno, M. Ankoud, M. Hmimida.

[3] SOC-H et Hypertagging sont des prototypes de recherche destiné à être commercialisés dans le produit SémioTag du partenaire du projet Cogniva Europe.

1. Introduction

La croissance exponentielle du nombre de documents numériques au sein des organisations, corollaire du développement du système d'information semi-structuré et des Documents pour l'Action qu'il contient (fichiers bureautiques, messagerie, CMS de type blog et wiki, etc.), et le caractère à la fois anarchique et foisonnant des environnements informatiques de sauvegarde (postes individuels, disques réseaux, serveurs de GED, serveurs de CMS, serveurs en mode cloud, etc.) entrainent des difficultés cognitives et organisationnelles croissantes de gestion de l'information au sein des entreprises. Ces difficultés sont susceptibles d'avoir des conséquences lourdes sur la performance individuelle et collective, sur le bien-être au travail des collaborateurs, mais aussi sur le respect des obligations contractuelles et légales auxquelles les entreprises sont soumises.
Dans le cadre du projet ANR Miipa-Doc « Méthodes et Services Intégrés Institutionnels et Participatifs pour la Classification à Facettes des Contenus Documentaires Complexes », nous visons à développer de nouvelles approches méthodologiques et logicielles pour permettre une gestion plus intégrée de la documentation d'entreprise qui facilite la gestion de l'hétérogénéité des modalités de classification et de classement du patrimoine documentaire. En particulier, nous visons à tirer profit de trois innovations importantes qui transforment les usages du SI et son architecture : la généralisation des moteurs de recherche qui se présentent de plus en plus comme une interface d'accès unifiée à l'information semi-structurée (voire structurée), l'utilisation des métadonnées dont la généralisation apparait de plus en plus comme une solution non pas opposée mais complémentaire à celle des moteurs de recherche et enfin, la généralisation de la participation des utilisateurs au classement de leur documentation qui est une réalité depuis l'avènement de l'informatique personnelle, encore renforcée par les usages du Web 2.0.
Dans cet article nous présenterons les trois principales contributions techniques du projet : (1) l'application de gestion multidimensionnelle et ascendante des SOC, Hypertagging (2) l'infrastructure à base de Systèmes d'Organisation des Connaissances Hétérogènes, SOC-H[4] et (3) la méthode d'Analyse Sémiotique des Transactions Documentaires, ASSET-Doc. Nous présenterons également un premier résultat des études de terrain effectuées qui ont contribué à permettre l'élaboration de la méthode ASSET-Doc et la spécification fonctionnelle d'Hypertagging. Nous cherchons à développer

[4] SOC-H et Hypertagging sont des prototypes de recherche destiné à être commercialisés dans le produit SémioTag du partenaire du projet Cogniva Europe.

une conception orientée infrastructure (Star & Bowker 2002, cf. infra), ce qui implique que nos logiciels doivent s'intégrer dans les infrastructures existantes et en particulier réaliser un compromis entre les classifications existantes, qu'elles soient institutionnelles ou personnelles et les classifications émergentes répondant aux besoins des nouvelles activités individuelles ou collectives.

1 Eléments de théorie du document et définitions

Nous commençons par présenter un certain nombre de notions théoriques sur lesquelles est basée notre recherche sous la forme de définitions.

1.1 Sémiotique des transactions coopératives (et transaction communicationnelle)

Un cadre d'analyse de l'activité en partie complémentaire d'autres cadres comme l'analyse de l'activité ou les approches interactionalistes visant à rendre compte de celle-ci comme étant toujours de nature relationnelle (transactions coopératives) et tout à la fois impliquant et produisant des artefacts médiateurs relevant d'une analyse sémiotique[5]. L'analyse sémiotique des artefacts médiateurs suggère de les analyser selon trois points de vue complémentaires : physique, expressif, agentif (cf. Tab 1)[6]. Quand les artefacts médiateurs ont une dominante expressive nous parlons de transaction communicationnelle (pour la transaction) et de production sémiotique (pour l'artefact médiateur).

Table 1. Différentes dimensions des artefacts médiateurs dans la sémiotique des transactions coopératives

Dimensions génériques	**Physique**	**Expressif**	**Agentif**
Dimensions des artefacts	*Tangible*	*Représentation*	*Symbolique*
Dimensions des personnes	*Corporel*	*Psychique*	*Socio-Relationnel*

[5] Par rapport à l'analyse de l'activité la sémiotique des transactions coopératives insiste sur la dimension intrinsèquement relationnelle de toute action. Par rapport à l'approche interactionaliste elle insiste sur la dimension « productive », y compris au sens économique, des transactions et sur le caractère pérenne des artefacts médiateurs mobilisés.

[6] Par exemple selon M. Coyaud « On considère souvent le signal comme une donnée brute (support physique de l'information), le signe comme quelque chose de mental, et le symbole comme un signe encore plus abstrait. » Coyaud, M., 1966, p.13 cité par le Trésor de la Langue Française (2011).

1.2 Situation transactionnelle

Les transactions coopératives se déroulent dans des situations transactionnelles qui sont caractérisées par différents paramètres (au sens où l'on parle de situation d'énonciation en analyse pragmatique des conversations). Parmi ceux-ci : les réalisateurs et les bénéficiaires en présence, le projet d'action (ou but de l'activité), les relations sociales (statuts, rôles, etc.), le cadre spatio-temporel (setting), les équipements techno-informationnels, etc.

1.3 Production sémiotique

Les productions sémiotiques sont des artefacts médiateurs dans lesquels la dimension expressive prédomine. Ce sont ces artefacts qui sont les objets principaux des sciences de l'information et de la communication comme de l'informatique. La production sémiotique peut être analysée selon les trois points de vue présentés précédemment. Le point de vue physique, ou du médium, peut lui-même se décomposer en point de vue du support matériel et point de vue de la forme d'expression. Etant donné un support matériel (gestes, gestes oraux, supports d'écriture ou de dessin papier, support d'enregistrement numérique, etc.) la forme d'expression matérielle correspond aux moyens physiques qui permettent de matérialiser des signes. Les points de vue expressif et agentif, qui constituent le contenu sémiotique correspondent d'une part, au pouvoir d'évocation de l'artefact médiateur, le contenu représenté, et d'autre part à ses effets potentiel, son contenu performatif au sens de la théorie des actes de langage, ou effets potentiels.

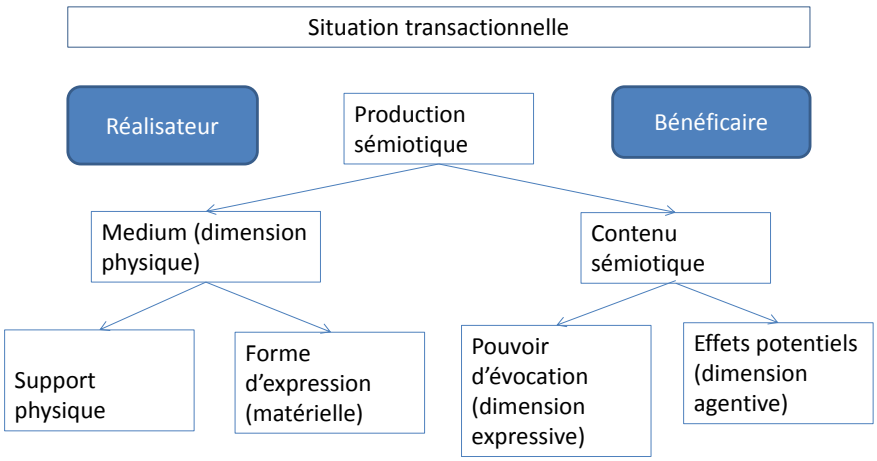

Fig.1. La situation transactionnelle et les composants de la production sémiotique

1.4 Documentarisation et document

Quand le support physique de la production sémiotique est pérenne celui-ci va pouvoir faire l'objet d'un investissement spécifique visant à en faciliter la remémoration ultérieure, le travail de documentarisation. La documentarisation consiste à doter le support « *d'attributs spécifiques permettant de faciliter (i) leur gestion parmi d'autres supports, (ii) leur manipulation physique, condition d'une navigation sémantique à l'intérieur du contenu sémiotique et enfin, (iii) l'orientation des récepteurs, mais également de plus en plus des réalisateurs eux-mêmes* » (Zacklad 2004). Corollairement, un document est défini comme « *une production sémiotique transcrite ou enregistrée sur un support pérenne qui est équipée d'attributs spécifiques visant à faciliter les pratiques liées à son exploitation ultérieure dans le cadre de la préservation de transactions communicationnelles distribuées. Ces attributs doivent permettre au document de circuler à travers l'espace, le temps, les communautés d'interprétation, pour tenter de prolonger les transactions communicationnelles initiées par ses réalisateurs.* ». La documentarisation permet une articulation des fragments du document selon une dimension interne (documentarisation interne) liée aux opérations de lecture, et externe, liée aux opérations de classification du document à l'intérieur d'une collection ou d'un dossier (documentarisation externe).

1.5 Activité de classification

La documentarisation considérée comme une activité cognitive, correspond le plus souvent à une activité de classification.

Dans l'univers documentaire il s'agit d'une opération « *qui consiste à organiser des sujets ou des objets en classes, de sorte que les sujets/objets semblables ou parents soient regroupés et séparés des sujets/objets non semblables. Une classe est un groupe d'entités qui présentent une ou plusieurs caractéristiques communes.* » (Hudon, 1999-2000). Cette définition associe deux activités qui relèvent de problématiques différentes : l'assignation d'un objet documentaire dans une classification existante et l'assignation d'un sujet à un groupe d'autres sujets ce qui revient à créer ou à modifier un système de classification. A l'heure de l'indexation participative, ces deux opérations peuvent être effectivement réalisées de concert. Les activités de classification sont principalement conduites selon deux finalités en partie interdépendantes : une activité de prise de décision ou de résolution de problème et une activité de remémoration. C'est la raison pour laquelle elles interviennent de façon centrale dans la documentarisation. Le plus souvent, ces activités de classification s'appuient

sur des structures existantes représentées à l'aide de Systèmes d'Organisation des Connaissances (SOC)

1.6 Activité de classement

Avant la numérisation, la classification et le classement faisaient l'objet d'opérations bien distinctes, mais cette distinction tend à se dissoudre en partie en environnement numérique. Alors que la classification relevait de la documentarisation externe, le classement tirait les conséquences de cette documentarisation pour placer un exemplaire physique du document dans un rayonnage. Mais si en environnement numérique, sur le disque dur de l'utilisateur, l'affectation d'une métadonnée et la définition d'un chemin d'accès dans une hiérarchie de répertoires correspondent bien à des opérations matérielles différentes, celles-ci tendent à se confondre dans les systèmes de GED et de CMS dans lesquels l'indexation d'un document peut se traduire pour l'utilisateur par un « classement virtuel » dans un espace donné.

1.7 Activité d'indexation

Comme pour le classement, avant la numérisation, classification et indexation correspondaient à des activités différentes. La classification impliquait la sélection d'une classe de destination le plus souvent unique tandis que l'indexation consistait à « *décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse.* » (UNISIST, dans Chaumier, 1988, p. 22). L'indexation s'appuie classiquement sur un langage documentaire ou un système d'organisation des connaissances (cf. infra). Avec la numérisation, cette distinction paraît là encore moins tranchée. Dans les systèmes d'indexation participative mouvants du Web 2.0, l'ajout d'un nouveau « tag »peut vouloir signifier la création d'une nouvelle classe. Le fait que l'indexation puisse faire appel à différents champs sémantiques en partie disjoints peut correspondre, dans ces environnements, à la problématique d'une classification multidimensionnelle ou à héritage multiple.

1.8 Systèmes d'Organisation des Connaissances et technologies associées

Dès que les opérations de documentarisation ont un caractère un peu systématique, elles s'appuient sur des structures représentées à l'aide de langages que nous dénommons à la suite de Hodge (2000) des Systèmes

d'Organisation des Connaissances (SOC). Ce terme 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 2011). Nous y ajoutons également les index de moteurs de recherche que nous considérons comme des SOC automatiques basés sur des algorithmes informatiques de fouille de texte (ou d'autres types de ressources numériques). Les SOC sont conçus pour servir de support aux opérations de classification et de classement mais aussi de nommage des documents et peuvent être intégrés à des technologies très diverses selon les environnements matériels considérés. Indépendamment des index des moteurs de recherche que nous avons déjà mentionnés et qui constituent des SOC automatiques, les principales modalités de mise en œuvre des SOC que nous considérerons sont les métadonnées, les répertoires (chemins d'accès aux documents) et le nommage qui peut relever d'une codification spécifique.

1.9 Les métadonnées

Les métadonnées sont les auxiliaires à la fois les plus classiques et les plus en vogues pour la documentarisation externe des documents en environnement numérique et elles portent sur des dimensions très variées des productions sémiotiques (cf. infra). Si l'on adopte une vision large des SOC et des activités de classification, on peut considérer que tout ensemble de métadonnées relève d'un SOC. Ainsi le fait d'utiliser un champ de métadonnées pour décrire la date de création d'un document à l'aide d'un système de calendrier relève d'une opération de classification du temps et de sa périodisation. La prise en compte du nom de l'auteur d'un document s'appuie sur une classification de l'identité des personnes basée sur le système des patronymes, etc. Pour une large part, la documentarisation s'appuie donc sur des métadonnées, mais elle fait également appel à d'autres opérations liées au classement et au nommage.

1.10 Définition du chemin d'accès à un répertoire

Dans les systèmes d'exploitation de la famille Windows, qui sont les plus utilisés en environnement bureautique, l'enregistrement d'un document implique la sélection d'un répertoire de destination, c'est-à-dire d'un chemin d'accès pour l'enregistrement. Bien que les répertoires relèvent de la problématique du classement, les hiérarchies de répertoires sont les principaux outils dont disposent les utilisateurs pour réaliser une classification conceptuelle de leurs fichiers sur le poste de travail. D'une

certaine manière, l'organisation hiérarchique d'un répertoire correspond à un Système d'Organisation des Connaissances, bien que les répertoires ne fournissent pas de métadonnées directement manipulables par les utilisateurs. Une des difficultés associée à l'usage des répertoires est leur organisation hiérarchique et le fait que les fichiers ne peuvent être placés que dans un seul répertoire de destination dans l'environnement Windows. Grace au protocole WebDav, ces chemins d'accès sont étendus de manière transparente aux disques externes « montés » sur le poste de travail de l'utilisateur.

1.11 Le nommage du document

Le titre du document est un attribut essentiel de sa documentarisation. Il est souvent composé de manière structurée en fonction d'une logique de classification particulière. Des utilisateurs expérimentés utilisent une codification particulière pour leurs noms de fichiers qui correspond à la concaténation de différents paramètres : thème, auteur, date, version, etc.

1.12 Activité d'annotation

La documentarisation, qu'elle soit interne ou externe, met en jeu des activités d'annotation que nous définissons comme « *toute forme d'ajout visant enrichir une inscription ou un enregistrement pour attirer l'attention du récepteur sur un passage ou pour compléter le contenu sémiotique par la mise en relation avec d'autres contenus sémiotiques préexistants ou par une contribution originale* » (Zacklad 2007a). Si à un niveau d'abstraction élevé, la documentarisation relève largement des activités de classification, celles-ci se traduisent du point de vue de la matérialité documentaire et de l'écriture, par la rédaction d'une annotation. Quand elle contribue à la documentarisation externe du document, l'annotation relève de l'annotation associative et puise généralement dans un SOC. L'annotation associative correspond aussi aux activités de « taggage » caractéristiques des applications de type Web 2.0 (Zacklad 2007a).

1.13 Dimension, facette, point de vue

Une discussion approfondie de ces notions excède le cadre de cet article (cf. Marleau et al. 2008 et Zacklad 2011, pour des éléments de discussion). La méthodologie de classification et de classement ascendante ASSET-Doc vise à produire des classifications documentaires ad hoc tout en intégrant les classifications plus institutionnelles ayant cours dans l'entreprise. Le caractère hétérogène des classifications considérées nous conduit à parler de classification multidimensionnelle sans préciser toujours si ces

classifications relèvent de facettes ou de points de vue. Dans un précédent texte (Zacklad 2011), nous établissons une distinction entre des approches à facettes universelles introduites par Ranganathan et des approches à facettes locales dans lesquelles le jeu de facettes peut être particularisé dans le cadre d'une application spécifique, comme dans la méthodologie ISIS (Marleau et al. 2008, Mas & Marleau 2009), vis-à-vis de laquelle le projet Miipa-Doc se positionne en complémentarité.

Les approches à base de point de vue, que nous promouvons dans le cadre du web socio-sémantique et qui correspondent notamment aux « ontologies sémiotiques » (Zacklad 2005, Zacklad et al. 2007), visent à caractériser un ensemble d'items selon différents points de vue « *défendus par un ou plusieurs acteurs et pouvant être socialement et/ou cognitivement conflictuels avec un autre.* » (Zacklad 2011). Outre le fait que, dans les approches par facettes, la sélection d'une valeur est potentiellement exclusive d'une autre, la différence entre facettes et points de vue vient de ce que, dans les facettes, chaque dimension s'inscrit dans une relation de complémentarité avec les autres, tandis que dans les points de vue, il peut y avoir redondance et conflictualité. Dans des versions plus avancées de la méthodologie encore en travaux, les usages respectifs des facettes et des points de vue seront d'avantage développés, notamment dans le cadre de la prise en compte de processus de documentarisation collectifs. Dans le cadre de cet article, nous parlerons de point de vue pour distinguer les différentes approches de l'analyse sémiotique des artefacts médiateurs (situation, support physique, forme d'expression matérielle, contenu représenté, contenu pragmatique). Nous parlerons ensuite de dimensions pour désigner les différentes listes de tags appartenant à des champs sémantiques proches utilisés par les utilisateurs à l'intérieur de chacun de ces points de vue sans préciser, à ce stade de notre recherche, si elles constituent des facettes ou des sous-points de vue.

1.14 Infrastructure et urbanisation durable

Le terme d'infrastructure peut être utilisé comme un quasi-synonyme d'environnement informatique ou conformément à notre approche, de manière plus spécifique, en référence aux travaux dans le domaine du « social informatics » (traduits en français par « pratiques collectives distribuées » par W. Turner, 2007). Dans un article récent Karasti, Baker et Millerand (2010) revisitent cette expression dans la perspective de la conception de plateformes de collaboration entre scientifiques pour l'échange de données dans le domaine de l'écologie. Par rapport à la définition initiale du terme par Star and Ruhleder (1996) qui faisait surtout

référence à l'infrastructure en tant que dispositif, en partie technologique, visant à résoudre la tension existant entre le local et le distant, comme peuvent l'être le respect de standards techniques, largement adoptés sur la planète, ils introduisent les dimensions suivantes : la prise en compte de la dimension temporelle, l'identification de deux orientations temporelles distinctes, celles du « temps du projet » et du « temps de l'infrastructure » et l'adoption d'une orientation spécifique dans le processus de développement, celle du « développement en continu » susceptible de correspondre aux enjeux temporels des infrastructures qu'ils mettent en évidence (Karasti et al. 2010).

De par son objectif central de prise en compte de l'hétérogénéité des classifications documentaires en intégrant à la fois la perspective institutionnelle porteuse des enjeux de la longue durée et la perspective des classifications émergentes issues de nouvelles configurations transactionnelles, le projet Miipa-Doc s'inscrit largement dans cette approche d'une conception centrée infrastructure (Star &Bowker 2002). Les composants d'architecture qui visent à intégrer les classifications hétérogènes et qui sont exploités par l'application Hypertagging relèvent spécifiquement d'une problématique d'infrastructure. En faisant référence à ce concept, nous indiquons également, comme le soulignent aussi Karasti et alii, que notre démarche de conception doit s'intégrer dans les pratiques ancrées des utilisateurs, que celles-ci correspondent à des habitudes de classement ou à l'usage d'un logiciel familier. Cette perspective relève aussi pour nous d'une stratégie d'urbanisation durable selon laquelle, les nouveaux composants applicatifs doivent permettre des innovations sans provoquer de ruptures, respecter une certaine hétérogénéité de l'environnement applicatif tout en facilitant sa gestion, ne pas enfermer les utilisateurs dans des environnements propriétaires empêchant de transférer des données et des paramètres et posséder une ergonomie renforcée offrant des IHM cohérentes et limitant les doubles tâches de saisie.

Dans le cadre de notre projet, nous distinguons le composant applicatif et le composant d'infrastructure. Ce dernier est dédié à la gestion de l'hétérogénéité des Systèmes d'Organisation des Connaissance et particulièrement à leur représentation sous la forme de métadonnées. Dans la suite du texte, nous commencerons par présenter l'application de classification intégrée Hypertagging puis nous présenterons le composant d'infrastructure SOC-H et la méthodologie.

2 Hypertagging : application de classification et de recherche multidimensionnelle

L'application Hypertagging vise à traiter de manière intégrée la problématique de la classification et du classement des documents numériques en environnement hétérogène. Elle comporte deux modules : un module de classification/sauvegarde et un module de recherche. Dans le premier module, représenté Figure 2, l'utilisateur sélectionne un jeu de métadonnées offrant une classification multidimensionnelle des documents. Cette représentation doit être construite au préalable en tenant compte à la fois de l'organisation de l'entreprise, dans ses dimensions instituées et émergentes, et en fonction des besoins de gestion de l'information personnelle de l'utilisateur.
La sélection de tags est utilisée pour déterminer le nom du fichier et le répertoire de classement à l'aide d'un système de règles, regroupant ainsi dans une interface unique les trois principales méthodes de classification existante en environnement bureautique[7]. Le module se lance directement à partir des applications Office pour une bonne intégration dans le processus de création du fichier ou à partir de l'explorateur de fichier pour la classification de documents existants n'impliquant pas leur modification. Pour la recherche, l'utilisateur dispose de la même interface qui lui permet de filtrer progressivement les documents pertinents au fur et à mesure qu'il sélectionne des tags appartenant à différentes dimensions.
Un des caractéristiques originales d'Hypertagging est la représentation des tags par des icônes qui constituent des SémioTags. L'interface met ainsi en œuvre un principe de codage redondant de l'information puisque le tag est représenté à la fois par une expression linguistique et par une image plus ou moins arbitraire. Notre hypothèse, qui devrait faire l'objet d'expérimentations ergonomiques, est que ce « surcodage » sémiotique devrait faciliter la remémoration et la discrimination perceptive et donc accélérer les opérations de sélection lors de la classification et de la recherche.

[7] Par manque de place ces aspects ne sont pas présentés ici.

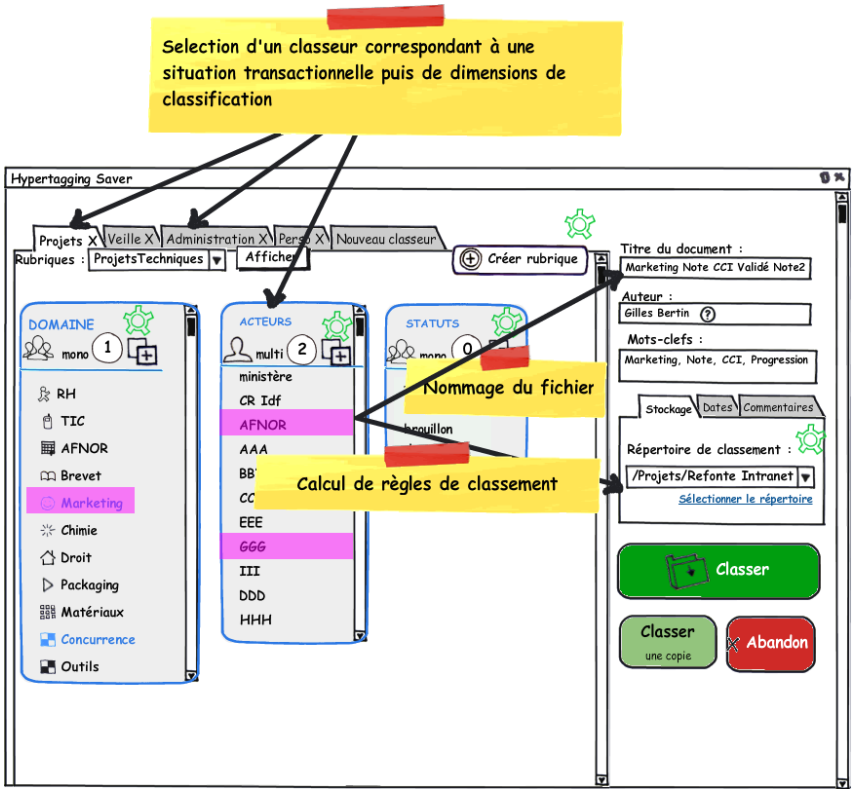

Fig.2 - Interface de classement de l'application Hypertagging

3 Composant d'infrastructure à base de SOC Hétérogène (SOC-H)

Du fait de l'extrême hétérogénéité de la localisation des métadonnées, l'application Hypertagging ne pourrait pas fonctionner sans des composants d'infrastructure assurant l'interface entre les espaces de stockage sur un plan physique et « logique ». Les métadonnées peuvent être situées soit (1) dans le fichier, soit (2) dans le système de gestion de fichiers (intégré au système d'exploitation) sur le poste local ou sur des disques réseau, ou encore dans des entrepôts distants, (3) index des moteurs de recherche (intégrés ou non au système d'exploitation), (4) GED et CMS propres à l'organisation ou (5) liés à des communautés dans l'environnement du Web (Tab 2).

Chaque jeu de métadonnées est géré par des environnements applicatifs différents qui ne possèdent généralement pas d'interface : application d'édition de contenu pour les métadonnées contenues dans les fichiers, explorateur de fichier reflétant des hiérarchies de fichiers diversement structurées, différents types d'index associés aux systèmes de GED et aux

CMS. Cette complexité est en partie réduite par la définition de standards, impliquant la sélection de sous-ensembles de métadonnées communes, permettant de disposer de composants logiciels génériques en mesure d'adresser différents type de fichiers et différents entrepôts (cf. Table 2). Cependant, ces standards ne rencontrent généralement le succès que dans le cadre de propriétés relevant d'une sémantique référentielle liée aux caractéristiques matérielles des fichiers. Les hiérarchies de répertoires ou les métadonnées interprétatives relevant du contenu sont généralement mal prises en compte.

Table 2. Différents espaces de localisation des métadonnées

Localisation des métadonnées	**Standard offrant des accès normalisés**
Fichier et environnement direct	
Fichier (via son éditeur de contenu - application)	XMP
Système de gestion de fichiers (via l'explorateur de fichiers)	WEBDAV
Entrepôts ou index de moteur de recherche	
Index des moteurs de recherche (via interface de requête)	Pas de standard
Bases documentaires de type GED et CMS (via le navigateur)	CMIS
Bases documentaires de type GED et CMS dans l'environnement public ou semi-public du web (via le navigateur)	Certains standards légers existants, par exemple, dans des archives ouvertes.

La figure 3, présente la situation actuelle dans la majorité des écosystèmes documentaires actuels au sein des organisations. L'utilisateur est amené à répéter de manière non coordonnée les opérations de documentarisation externe de ses documents à travers diverses applications. La recherche de ces mêmes documents implique symétriquement de mobiliser sa connaissance des SOC, des métadonnées, des chemins de sauvegarde et des titres des documents pour retrouver les informations correspondant à son besoin.

L'orientation de conception orientée infrastructure que nous adoptons va viser précisément à fournir un composant se situant à l'intersection de ces différents entrepôts et briques applicatives de manière à automatiser

l'écriture des métadonnées dans les différents environnements. SOC-H s'appuie aujourd'hui sur deux choix d'architecture principaux qui sont similaires à ceux mis en œuvre dans l'architecture d'ISIS (Marleau et al. 2008), que nous prolongeons et systématisons. Ceux-ci ne sont pas inamovibles et ils peuvent présenter des inconvénients, mais ils sont particulièrement en phase avec notre principe d'urbanisation durable. Le premier consiste à utiliser les fichiers comme le principal véhicule des métadonnées caractérisant le document. Pour traiter un plus grand nombre de formats, nous nous appuyons sur le standard XMP. Cette option s'oppose à celle suivie dans la plupart des systèmes de gestion documentaire actuels qui possèdent des index spécifiques et qui associent dans leurs tables de métadonnées l'URI du document[8].

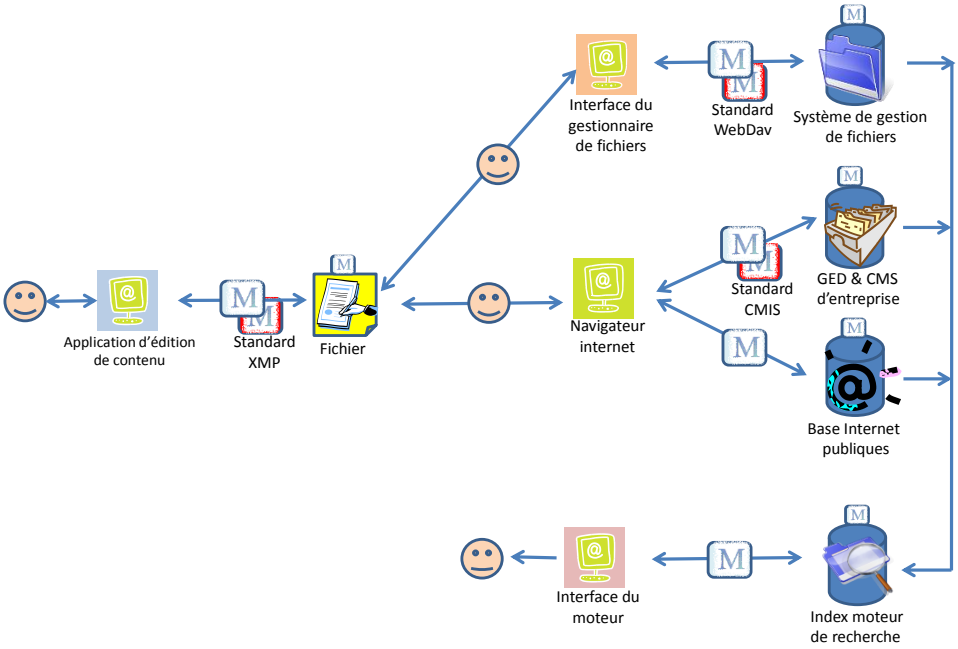

Fig. 3 - Situation actuelle de distribution des métadonnées

L'inconvénient de l'approche classique est que si le document est déplacé, celui-ci perd ses attributs de documentarisation externe (comme c'est aujourd'hui le cas dans les systèmes de gestion de fichiers). L'inconvénient de l'approche que nous avons choisie est que toute évolution de la

[8] Pour des raisons que nous ne pouvons développer ici, nous disposons également d'un entrepôt de métadonnées intégré à notre composant d'architecture qui réplique les métadonnées intégrées dans les fichiers.

classification implique une réécriture physique d'un des champs du fichier ce qui est contradictoire avec certains principes d'archivage et difficile à contrôler dans les systèmes de gestion de fichier actuels[9]. Notons que si le document est classé dans un entrepôt de donnée de type GED ou CMS, SOC-Doc peut dupliquer les métadonnées du fichier à l'intérieur de l'entrepôt en utilisant le standard CMIS (Fig. 3).

Le second principe d'architecture concerne le module de recherche d'information qui utilise les moteurs de recherche déployés dans l'environnement informatique des utilisateurs pour retrouver les fichiers répartis dans les différents espaces de stockage (Fig. 4.). Du fait de l'hétérogénéité des espaces de stockage des documents et de la multiplication non coordonnée des systèmes utilisés pour leur classification, les moteurs de recherche sont devenus des composants d'infrastructure majeurs dans de nombreuses entreprises. Ils possèdent un certain nombre de caractéristiques qui tendent à les constituer en interface unifiée d'accès à l'information : ils peuvent indexer des documents répartis dans plusieurs espaces de stockage, accéder aussi bien à l'information structurée qu'à l'information semi-structurée, effectuer un traitement différencié de certaine métadonnées pour répondre à des requêtes évoluées.

Cependant, les moteurs de recherche, ne parviennent toujours pas à hiérarchiser l'information de manière entièrement pertinente si celle-ci n'a pas été classifiée en amont, notamment selon la dimension interprétative, la plus porteuse de valeur ajoutée pour l'organisation. C'est la raison pour laquelle, malgré l'hétérogénéité et la fragmentation actuelle des systèmes de classification, ceux-ci resteront indispensables à la gestion des documents d'entreprises. Pour l'accès à l'information, l'infrastructure SOC-H se présente donc comme complémentaire de l'infrastructure que constitue le moteur de recherche d'entreprise, qui stocke dans son index la description des documents quel que soit leur localisation et qui actualise cet index de manière continue. Le moteur est interrogé grâce à l'application Hypertagging qui permet à l'utilisateur de lancer des requêtes sélectionnant des métadonnées pertinentes.

[9] L'ajout ou la modification d'une métadonnée entraîne une modification du champ « modifié le » qui est utilisé par défaut dans l'explorateur Windows sans qu'il soit possible d'indiquer si la modification porte sur le contenu principal du document ou sur une annotation « externe » à celui-ci, intervenant dans la classification sans en altérer le fond.

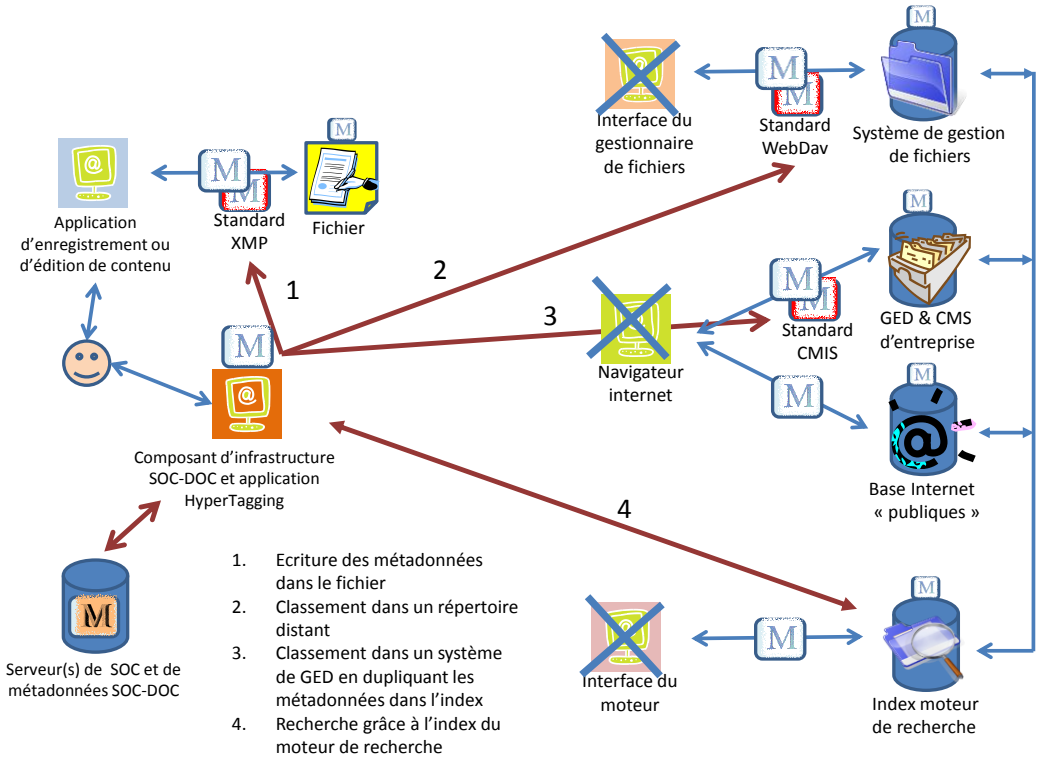

Fig. 4 - Positionnement du composant d'infrastructure SOC-H

4 La méthode d'analyse sémiotique des transactions documentaires (ASSET-Doc)

Le fonctionnement d'Hypertagging repose sur une bonne conception du système de représentation multidimensionnel qui organise les métadonnées qui seront nécessaires à la classification et à la recherche. C'est le rôle de la méthode d'analyse sémiotique des transactions documentaires que de fournir un guide pour cette conception. Nous ne présenterons ici que les principes qui sous-tendent la classification multidimensionnelle et pas les aspects dynamiques de la démarche qui doit articuler point de vue institutionnels, personnels et émergence de nouvelles normes. La classification des Tags est basée sur les quatre points de vue qui organisent l'analyse sémiotique des artefacts médiateurs présentés plus haut auxquels se rajoute le point de vue de la situation transactionnelle.

Le point de vue de la situation transactionnelle[10]correspond à une documentarisation basée sur des informations qui ne sont pas explicitement présentes dans le contenu du document, mais qui interviennent de manière importante dans les activités de remémoration et de classement parce qu'elles décrivent la dynamique de la transaction qui lui a donné naissance: finalité principale du document ou contexte socio-spatio-temporel dans lequel il a été produit, par exemple. Alors que dans la méthodologie ISIS le contexte correspond toujours à la représentation formelle d'un processus métier (fonction, activité, structure organisationnelle, poste, rôle…), la situation transactionnelle dans laquelle le document est produit peut renvoyer à des types d'actions collectives différentes dans ASSET-Doc : animation d'une communauté, veille informationnelle, activité personnelle, activité de formation, etc.

Par ailleurs, même dans le cadre d'activités d'entreprise, la situation transactionnelle est considérée comme relevant de mondes sociaux (Strauss 1993, Fitzpatrick et al. 1995) assez différents, auxquels sont associées des productions sémiotiques spécifiques relevant soit d'une activité métier soit d'une activité de management. Concernant les activités métiers, les situations transactionnelle sont extrêmement différentes selon le domaine, mais toujours bien identifiée : situation d'audit, de diagnostic, de spécification, d'étude, de conception, d'entretien, de maintenance, etc. Les activités de management sont souvent plus génériques et transverses à différents domaines : situation de gestion de projet, de gestion des connaissances, de suivi budgétaire, de suivi d'activité, d'entretien d'évaluation, etc.

A chaque situation transactionnelle relevant d'un métier, d'une activité de management, d'une activité de loisir ou encore d'une activité de formation, pour ne prendre que quelques exemples, correspond un ensemble de documents. Ceux-ci seront classés en utilisant les quatre points de vue génériques du support physique, de la forme d'expression matérielle, du contenu représenté et du contenu pragmatique. A chacun de ces points de vue peut correspondre une ou plusieurs dimensions de classification. Par exemple, le contenu représenté peut-être appréhendé à partir du nom du projet décrit, des « objets » concernés, des actions entreprises, des personnes impliquées par les tâches, etc. Le point de vue pragmatique peut-être décrit par des dimensions liées à l'état de la rédaction, à son statut plus ou moins officiel, au fait qu'il s'agit d'un document émis ou reçu, etc.

[10] Sous un certain angle, il correspond au modèle du contexte dans la méthodologie ISIS, avec cependant des différences d'approches assez marquées.

A ces dimensions relevant de la classification du document peuvent s'ajouter d'autres dimensions de classification relevant de critères propres à la situation transactionnelle qui, comme nous l'avons dit plus haut, ne sont pas représentés explicitement dans le document. Ces critères situationnels externes peuvent, par exemple, relever d'effets performatifs[11] propres aux usages que le bénéficiaire du document souhaite en faire (veille personnelle, attribution à tel collègue, à utiliser dans tel contexte, etc.). Ils pourraient aussi correspondre à des évaluations du contenu en vue de telle ou telle action future.

Le point de vue du support ne donne généralement pas lieu à la création d'une dimension explicite, parce qu'il est porté par l'indication graphique du format dans la plupart des systèmes de gestion de fichiers. Le point de vue de la forme d'expression correspond le plus souvent à la dimension « genre du document ». Selon notre approche méthodologique, le genre correspond à la rencontre d'une forme scripturale et d'un type de contenu représenté ou pragmatique. Les genres de document sont plus ou moins fortement marqués dans leur forme et le type de contenu qui leur est associé. Dans les organisations très avancées d'un pont de vue documentaire, ils correspondent à des modèles de fichiers spécifiques. Mais la plupart du temps, c'est l'utilisateur qui doit, à la création, déterminer le genre dont va relever son écrit.

Le point de vue du contenu représenté correspond aux thématiques évoquées dans le document. Il faut bien sûr s'en tenir à des classifications générales, l'objectif de ces classifications de haut niveau n'étant pas de se substituer au document d'origine en reproduisant, par exemple, des formulaires mais de faciliter son classement en vue d'une réutilisation ultérieure. Enfin, le contenu pragmatique correspond à la dimension performative du contenu sémiotique, c'est-à-dire aux effets potentiel du document dans un flux transactionnel au sein de l'organisation. Cette dimension correspond, par exemple, au statut du document eu égard à l'état de la rédaction, à l'état du processus de décision qu'il relate, à son importance en tant que témoignant de tel ou tel engagement (voir Tab 3 et l'étude de terrain au paragraphe suivant)

[11] Dans (Zacklad 2007b) nous proposons une typologie des actes de langages perlocutoires selon leurs effets : immédiats, médias, différés. Si les effets potentiels immédiats attendus sont souvent représentés dans le contenu du document (explicitation d'une demande, remerciement, etc.) ce n'est pas le cas des effets différés.

Table 3. Exemple de dimensions de classification pouvant être utilisées dans la méthode ASSET-Doc.

Point de vue de la situation transactionnelle	**Exemple de dimensions**
Situation d'animation de communauté d'intérêt (botanique)	1. Cycle de vie de la communauté : situation de recrutement de membres, évènements conviviaux, vœux, accueil de nouveaux membres, rapport de cueillette, bulletin d'actualité, etc. 2. Evaluation des documents selon leur « qualité » 3. Statuts des rédacteurs et destinataires (non nécessairement représentés dans le document)
Point de vue de la production sémiotique	
Dimension(s) du support	4. Format physique de document : ppt, doc, pdf, etc.
Dimension(s) de la forme d'expression	5. Genre du document : affiche, compte-rendu, support de présentation, lettre interne, lettre externe, fiche de description de plante…
Dimension(s) du contenu représenté	6. Grandes familles de plantes concernées dans une fiche de description 7. Thèmes évoqués dans la manifestation 8. Partenaires mentionnés 9. Etc.
Dimension(s) du contenu pragmatique	10. Statut du document au regard de l'état de la rédaction 11. Statut du document au regard de la demande de fonds 12. Statut du document au regard de sa conservation

5 Eléments d'illustration en environnement professionnel

Dans cette dernière partie nous présenterons les premiers résultats de l'analyse effectuée chez le partenaire industriel du projet. La table 4 est issue d'une analyse des répertoires d'ingénieurs experts avec la grille ASSET-Doc. Cette analyse a pour but de réaliser un premier déploiement de l'application Hypertagging en effectuant une reprise partielle de l'existant documentaire sur les disques locaux et réseau.

Table 4. Dimensions de classification mise en évidence par l'analyse des répertoires d'ingénieurs experts

Situation transactionnelle	Activité Amont, Activité de production, Suivi de réalisation, Communication, Administratif, Encadrement
Dimensions de la Situation transactionnelle	Grandes activités : observation, recherche documentaire, accompagnement, suivi, conception, restitution, bilan, évaluation Modalité de partage : personnel, métier, projet, prestataire
Dimension du support physique	Format généré automatiquement par l'application : .doc, .ppt, .pdf, etc.
Dimension de la forme d'expression matérielle	Type de document : rapport, CR, Présentation, Contrat, Lettre, Fiche projet
Dimension du contenu représenté	Type de contenu : documentation, proposition, synthèse, analyse, spécification, contribution, description Thème : urbanisme, fonctionnalités, résultats Nom du projet : castor, pollux, achille, etc.
Dimension du contenu pragmatique	Niveau de diffusion : Pré-diffusion, Validation, Approuvé, Diffusion, Archivé Etat de la rédaction : Document de travail, Version intermédiaire, Final, Archive, Ancien

6 Conclusion

DOC-H reprend un certain nombre de principes développés dans lSIS (Mas& Marleau, 2009), comme l'écriture des métadonnées dans les fichiers bureautiques pour assurer leur classement et l'utilisation du moteur de recherche pour procéder à leur recherche. Tout en systématisant les principes de gestion de l'hétérogénéité documentaire (i.e utilisation des standard XMP et CMIS, utilisation d'une interface de classification indépendante de la suite Office[12], etc.).Hypertagging se différencie surtout d'ISIS par la prise en compte d'une diversité de modèles de situation (ou d'action) et en offrant la possibilité d'une modélisation ascendante et progressive conforme aux méthodes popularisées par le Web 2.0. Les deux approches sont complémentaires sur le plan méthodologique et technique, l'approche ISIS étant particulièrement adaptée à la gestion des processus métiers institués des grandes organisations.

D'autres projets de recherche se sont attelés à la problématique de la gestion sémantique intégrée des documents des utilisateurs en référence directe au programme du web sémantique. C'est le cas du projet Semantic Desktop (voir, par exemple, Sauermann 2005), qui relève d'une ambition en partie

[12] Dans ISIS la classification des fichiers s'effectue à l'intérieur de l'application Office et pas au moment de la sauvegarde ou via le « clic droit ».

similaire mais qui procède de manière différente à la fois du point de vue du modèle classificatoire et du point de vue de l'architecture. Dans le Semantic Desktop le modèle est basé sur des ontologies qui impliquent une formalisation poussée des modèles qui n'est pas compatible avec la dynamique de conception ascendante et évolutive à base d'indexation individuelle et participative défendue dans Miipa-Doc[13]. Par ailleurs, les choix d'architecture sont différents, en particulier du point de vue de l'intégration de la solution logicielle dans le système de gestion de fichiers et du point de vue de la gestion des métadonnées. Dans le Semantic Desktop les métadonnées sont uniquement gérées dans un serveur Web externe et ne sont pas intégrées aux documents. Les documents ne peuvent pas être stockés dans un répertoire quelconque de l'environnement informatique des utilisateurs.

Le projet Miipa-Doc va entrer dans sa dernière phase qui doit permettre le déroulement d'observations visant à évaluer en grandeur réelle un certain nombre d'options de conception. Nous envisageons de mettre en œuvre plusieurs démarches d'évaluation allant d'un mode très guidé à un mode libre. Dans les démarches fortement guidées par la méthodologie, des analyses documentaires et des interviews approfondies seront conduites auprès des utilisateurs de manière à leur proposer des modèles classificatoires les plus adaptés possibles aux situations transactionnelles dans lesquelles ils sont engagés.

Dans les démarches libres, les utilisateurs définiront eux-mêmes les classeurs et les dimensions de classifications adaptées à leurs besoins. Nous n'excluons pas de reproduire de manière semi-automatique certaines hiérarchies de répertoires et certains modèles de classement existant au sein de l'organisation, mais ils seront invités à compléter ces schémas par d'autres dimensions correspondant à leurs modalités de classification personnelles ou collectives émergentes en tenant compte des opportunités de classification multidimensionnelle offertes par HyperTagging. Enfin, une dernière option consistera à proposer une série de dimensions génériques correspondant à différents modèles d'actions (processus métier, projet, communauté, veille…) à particulariser par les utilisateurs en fonction de leur situation spécifique.

[13] Même si certains chercheurs du projet (Decker 2006) rendent compte de l'intégration du semantic desktop avec les technologies du Web 2.0, l'esprit de ces recherches est toujours de considérer que les ontologies formelles doivent finir par désambiguïser les wiki ou les folksonomies, une approche qui est en contradiction avec celles que nous défendons dans le courant du web socio-sémantique et notre vision de l'hétérogénéité intrinsèque des classifications (Zacklad 2005, Zacklad et al. 2007).

Références.

Adobe Systems Inc. Extensible Metadata Platform (XMP) Specification: Part 1, Data Model, Serialization, and Core Properties. Adobe, 2010.
http://www.adobe.com/content/dam/Adobe/en/devnet/xmp/pdfs/XMPSpecificationPart1.pdf

Chaumier, J. (1988). *Le traitement linguistique de l'information.* 3e édition mise à jour et augmentée. Paris, Entreprise moderne d'édition. 186 p. (Systèmes d'information et nouvelles technologies)

Choy D., et al. Content Management Interoperability Services (CMIS) Version 1.0. OASIS Standard. OASIS, 2010.
http://docs.oasis-open.org/cmis/CMIS/v1.0/cmis-spec-v1.0.html

Coyaud, M. (1966). *Introduction à l'étude des langages documentaires*, Paris-Librairie C.Klincksieck, (Coll. publiée sous le patronnage de l'ATLA, CNRS), 1966, 148p.

Fitzpatrick, et al., (1995). Work, locales and distributed social worlds. In *Proceedings of the fourth conference on European Conference on Computer-SupportedCooperativeWork* (ECSCW'95), Hans Marmolin, YngveSundblad, and Kjeld Schmidt (Eds.). Kluwer AcademicPublishers, Norwell, MA, USA, 1-16.

Goland Y., et al., HTTP Extensions for Distributed Authoring -- WEBDAV. RFC 2518. IETF, 1999.
http://tools.ietf.org/html/rfc2518

Hudon, M. (1999-2000), *COURS BLT 6054 - Analyse et représentation documentaires 1; COURS BLT 6055 - Analyse et représentation documentaires 2*. Université de Montréal, École de bibliothéconomie et des sciences de l'information.

Karasti, et al., (2010), "Infrastructure Time: Long-term Matters in Collaborative Development", *Computer Supported Cooperative Work*, 19:377–415

Decker, S. (2006). "The social semantic desktop: Next generation collaboration infrastructure", *Information Services and Use*, vol. 26/2, 139-144.

Mas S., Marleau Y., (2009). "Proposition of a Faceted Classification Model to Support Corporate Information Organization and Digital Records Managemen"t, *Hawaii International Conference on System Sciences*, pp. 1-10, 42nd Hawaii International Conference on System Sciences, 2009

Marleau Y., et al. (2008). Exploitation des facettes et des ontologies sémiotiques pour la gestion documentaire dans Broudoux, E. & Chartron, G. (Eds). *Traitements et pratiques documentaires : vers un changement de paradigme ?* Paris : ADBS Editions, p. 91-110.

Sauermann, L. (2005), "The Semantic Desktop - a Basis for Personal Knowledge Management", in *Proceedings of I-KNOW '05* Graz, Austria, June 29 - July 1, 2005 [en ligne] http://www.dfki.uni-kl.de/~sauermann/papers/Sauermann2005b.pdf

Star, S. L., & Ruhleder, K. (1996). "Steps toward an ecology of infrastructure: borderlands of design and access for large information spaces". *Information SystemsResearch*, 7(1), 111–134.
Star, S. L., & Bowker, G. C. (2002). How to infrastructure? In L. A. Lievrouw& S. L. Livingstone (Eds.), *The handbook of new media. Social shaping and consequences of ICTs* (pp. 151–162). London: Sage Publications.
Strauss, A. (1993): *Continual Permutations of Action*, Aldine de Gruyter, New York.
Trésor de la langue française, http://atilf.atilf.fr/tlf.htm [consulté le 10/02/2011].
Turner, W. (2007), Eléments pour une socio-informatique, dans Reber, B., Brossaud, C., Humanités numériques. *Nouvelles technologies cognitives et concepts des sciences sociales*, HermesPublishing, Londres-Paris (à paraître).
Zacklad, M. (2004). Processus de documentarisation dans les Documents pour l'Action (DopA). *Le numérique : impact sur le cycle de vie du document*. ENSSIB,Colloques de l'ENSSIB[en ligne] http://www.enssib.fr/bibliotheque-numerique/document-1223
Zacklad, M. (2005), Introduction aux ontologies sémiotiques dans le Web Socio Sémantique. Dans : Jaulent, M.-C. *16èmes journées francophones d'Ingénierie des Connaissances*, 30-03 Avril 2005, Nice. Grenoble: PUG, 12 p.
Zacklad, M. (2007a). Annotation : attention, association, contribution. In P. Salembier et M. Zacklad eds, *Annotations dans les Documents pour l'Action*. Lavoisier, Paris : 29-46.
Zacklad, M. (2007b). Transactions communicationnelles et actes de langage dans l'économie de services. In: Chabrol, C., Orly-Louis, I., et Najab, F. *Interactions communicatives et psychologies*. Paris: Sorbonne Nouvelle, 2007.
Zacklad, M., et al., (2007), Hypertopic : une métasémiotique et un protocole pour le Web socio-sémantique, in *Actes des 18èmes journées francophones d'Ingénierie des Connaissances*, 2-6 juillet 2007, Grenoble (à paraitre).
Zacklad, M. (2011). Cinq critères d'évaluation des Systèmes d'Organisation des Connaissances, *Les cahiers du numérique*, à paraitre.