Interface multi-facettes d'accès au capital documentaire de l'organisation

Guillaume Cabanac Jeune chercheur

Université de Toulouse, IRIT UMR 5505 CNRS
118 route de Narbonne, F-31062 Toulouse cedex 9
Guillaume.Cabanac@irit.fr

RÉSUMÉ. *Les individus au sein des organisations modernes sont couramment amenés à rechercher de l'information, pour réaliser leurs activités. À partir des multiples sources interrogées, ils extraient des documents après en avoir le plus souvent évalué la pertinence. Chacun peut ensuite conserver dans son Espace Documentaire Personnel (EDP) les documents qu'il juge utiles pour ses activités. La structure communément hiérarchique d'un EDP résulte d'efforts cognitifs importants, reflétant la perception de son propriétaire par rapport à ses activités. De ce fait, les EDP d'une organisation représentent un capital documentaire interne à haute valeur ajoutée. Malgré tout, les individus exploitent peu cette source car ils privilégient une source externe telle que le Web, pourtant moins adaptée* a priori. *Afin de limiter cette problématique, nous définissons une interface pour explorer le capital organisationnel, en naviguant dans des « facettes » qui représentent les thématiques des documents et l'usage que les individus en font. Basée sur l'exploitation du SI organisationnel, cette proposition vise à mieux rentabiliser les efforts requis pour la gestion des EDP au travers d'une interface sur le principe du donnant-donnant.*

ABSTRACT. *Knowledge workers are used to search for information in order to carry out their activities. They assess the relevance of the documents extracted from various sources. Then each knowledge worker may store in his Personal Document Space (PDS) the most useful ones. The common hierarchical organization for the PDS results from significant cognitive efforts related to individuals' perceptions. Therefore, the organizational PDSs constitute an in-house value-added capital. Nevertheless, people tend not to use this information source while giving priority an external source, e.g. the Web although it may seem less appropriate. In order to overcome this issue, we define an interface aiming to explore the organizational capital by browsing "facets." They provide content-based and usage-based access to documents and people from the organization. This approach, which exploits the organizational IS, intends to make PDS management-related efforts profitable, thanks to the proposed mutual benefit interface.*

MOTS-CLÉS : *Organisation, système d'information, hiérarchie, contenu et usage des documents.*

KEYWORDS: *Organization, information system, hierarchy, document content and usage.*

1. Introduction et motivations

À l'heure actuelle, le Système d'Information (SI) représente un centre névralgique majeur pour toute organisation : entreprise, industrie, laboratoire de recherche, etc. En effet, les membres organisationnels reposent sur leur SI pour réaliser leurs activités. Ils l'utilisent notamment pour accéder à de l'information contenue dans des documents, qu'ils soient internes ou externes à l'organisation. Par la suite, chaque membre peut stocker de manière organisée les documents utiles à son activité au sein d'Espaces Documentaires Personnels (EDP) : systèmes de gestion de fichiers, favoris Internet... La structure de ces espaces est majoritairement arborescente, elle reflète les efforts cognitifs mis en jeu par l'usager alors qu'il gère cet espace en insérant, supprimant ou déplaçant des documents. En réalité, les arborescences des membres organisationnels sont de véritables mines d'informations en relation directe avec leurs activités. Malgré cela, ces mêmes sources d'information internes sont très souvent délaissées, principalement au profit d'une source externe : le Web. Ce constat résulte en partie de la difficulté à accéder au capital documentaire de l'organisation : à moins qu'un individu ne modifie les droits d'accès de son EDP et en informe ses collègues, les documents organisés dans cet EDP restent confidentiels. Cette situation est d'autant plus paradoxale pour les documents obtenus à partir de sources publiques, au prix de recherches longues et fastidieuses. Bien qu'introduits dans l'organisation, de tels documents publics ne sont pas atteignables *via* les EDP, qui sont pourtant des sources internes structurées. Par conséquent, ce capital documentaire sommeille dans les EDP, alors qu'il serait utile à chaque membre organisationnel s'il était mis en valeur par un système de type donnant-donnant.

Afin d'apporter une réponse à cette problématique, cet article définit une interface multi-facettes d'accès au capital documentaire de l'organisation. Cette proposition ne vise pas à constituer une nouvelle base documentaire ou à enrichir une mémoire d'entreprise ; elle se veut plutôt complémentaire à ces approches. Afin de rendre possible la visualisation et l'exploration du capital documentaire, notre proposition exploite le SI organisationnel tel quel, sans nécessiter des modifications spécifiques. Ainsi, un avantage de l'interface multi-facettes proposée réside dans son aspect non-intrusif : la façon de travailler des membres organisationnels — en termes de pratiques documentaires — n'est pas remise en cause. Concrètement, l'interface proposée permet l'observation des documents et les personnes selon deux axes : thématiquement (en fonction du contenu des documents possédés) ou selon leur usage (mesurant la régularité d'organisation des documents les uns par rapport aux autres dans les EDP).

L'article est structuré comme suit : la section 2 expose plus en détail la problématique relative à la rentabilisation limitée des documents introduits dans le SI. Nous postulons que ces derniers peuvent répondre à des besoins qu'éprouve l'ensemble des membres organisationnels. Afin d'accroître le retour sur investissement du SI en exploitant les EDP structurés, nous définissons une interface multi-facettes à destination des membres organisationnels et du pilotage de l'organisation dans la section 3. Enfin, nous discutons cette proposition dans la section 4.

2. Hiérarchies de documents du SI organisationnel : un capital en sommeil

La proportion des « travailleurs du savoir » (*knowledge workers* en anglais) était estimée à 31 % de la main d'œuvre aux États-Unis en 1995, elle « continuera à croître de façon significative pendant le nouveau millénaire » (Sellen *et al.*, 2003, p. 51). Cette expression fait référence aux personnes qui travaillent avec de l'information et qui en produisent, elle correspond à de nombreuses professions : ingénieur, scientifique, chef de projet, journaliste… Dans les organisations modernes les individus travaillent de plus en plus au contact de l'information, si bien que nous sommes tous des travailleurs du savoir selon (Ballay, 2002). Au sein de l'organisation, ces individus consacrent 15 à 35 % de leur temps à la recherche d'information, sans pour autant trouver les documents recherchés : plus d'une recherche sur deux n'aboutit pas (Feldman, 2004). Dans le cas contraire, l'usager peut conserver les documents trouvés dans son EDP organisé hiérarchiquement, au sein d'une arborescence de répertoires, par exemple. Les classements par projet et thématique sont courants (Jones *et al.*, 2005; Khoo *et al.*, 2007), ils nécessitent un effort cognitif de la part de l'usager et reflètent sa perception des documents en rapport avec ses activités. Ainsi, une hiérarchie personnelle au sein du SI contient les documents utiles à un membre organisationnel, organisés de façon à réaliser au mieux sa réflexion autour de ses activités. À cause du faible rendement de la recherche d'information mentionné précédemment, 90 % du temps nécessaire à la rédaction d'un nouveau rapport serait consacré à la recréation d'informations préexistantes (Feldman, 2004). À la lumière de cette observation, le capital documentaire constitué par chaque membre organisationnel semble être sous-exploité, alors qu'il représente une réelle mine d'information. Nous estimons que ce constat est essentiellement dû aux trois causes suivantes.

– **Caractère personnel des EDP.** À cause de cette caractéristique, les documents trouvés après un effort de recherche important sommeillent dans les hiérarchies personnelles, sans qu'aucun autre membre ne puisse les exploiter. Par conséquent, les documents trouvés par un individu et recherchés ultérieurement par d'autres feront l'objet d'efforts de recherche répétés, parfois en vain. Pour limiter ce problème, les individus peuvent partager tout ou partie de leurs documents avec leurs collègues…

– **Partage manuel limité.** Le partage manuel des documents est limité car il demande un effort cognitif important. Par exemple, l'envoi d'un document par courrier électronique nécessite de connaître les autres membres, de sélectionner les destinataires potentiellement intéressés, de résumer l'intérêt du document… tout en prenant garde à ne pas surcharger ses collègues par de trop nombreux envois. La publication de documents sur l'intranet de l'organisation est une alternative au partage interpersonnel : celui d'IBM comprendrait au moins 5,5 millions de pages (Dmitriev *et al.*, 2006). Cette solution possède également des limites car 40 % des recherches sur l'intranet de grands comptes échoue (Feldman, 2004). Étant données les limites du partage manuel, des approches à base de partage automatique ont été proposées…

– **Partage automatique limité.** Une alternative à la recherche et au partage manuels consiste à mettre en place des processus de recommandation basés sur la description des usagers et de leurs besoins par des profils utilisateur (Montaner *et al.*, 2003).

Les limites de cette approche concernent la difficulté à modéliser les profils et à les faire évoluer afin qu'ils représentent au mieux les attentes réelles de l'usager. De plus l'appariement profil-document souffre également de limites telles que la nécessité d'une masse critique d'usagers, le frein du démarrage à froid (difficulté d'émettre des recommandations à un nouvel usager) et le problème du vocabulaire qui est courant en RI : identification et prise en compte de la synonymie, de l'homonymie, des figures de style, etc.

Afin de pallier ces problématiques, la section 3 expose notre proposition en définissant une interface multi-facettes d'accès au capital de l'organisation. Ce dernier correspond à une partie du SI : les EDP gérés par les membres organisationnels. L'interface proposée est constituée de facettes qui reflètent le contenu des documents ainsi que l'usage qui en est fait par les membres organisationnels. Notre approche est originale à plusieurs égards. 1) Elle vise à rentabiliser le SI organisationnel sur le principe du donnant-donnant, en exploitant les EDP qui ne bénéficient actuellement qu'à leurs propriétaires respectifs. 2) L'interface proposée est destinée aux membres ainsi qu'au pilotage de l'organisation. 3) Enfin, en plus de l'exploitation du contenu des documents (approche classique issue de la RI) nous proposons d'identifier et d'exploiter les relations d'usage qui lient les documents employés par les individus.

3. Interface multi-facettes d'accès au capital organisationnel

L'interface proposée vise à répondre aussi bien à des besoins opérationnels que stratégiques. L'adjectif « opérationnel » fait référence à la réalisation des tâches affectées aux membres organisationnels. Dans ce cadre, l'interface offre une vue globale des documents de l'organisation et en permet l'exploration, par thématique ou par usage. Ces deux mesures de similarité sont complémentaires. La Figure 1 illustre ce fait à partir des douze documents provenant des deux EDP représentés dans la Figure 6. On remarque la présence de deux groupes dans (1.a) contre un seul groupe dans (1.b). De plus, d_1 et d_{11} sont proches thématiquement alors qu'ils ne sont pas utilisés ensemble. Enfin, d_4 et d_5 sont les documents les plus proches par l'usage alors que leurs thématiques sont relativement éloignées. Grâce à l'interface basée sur ces deux mesures, chaque individu peut identifier, à partir de l'ensemble des EDP, les documents connexes ou complémentaires à ses propres documents. Comme il peut être utile de chercher des documents pour trouver les individus associés, et vice versa (Hertzum *et al.*, 2000), l'interface permet de basculer de l'une à l'autre de ces deux dimensions. Par ailleurs, l'adjectif « stratégique » concerne les activités propres au pilotage de l'organisation, notamment au service des ressources humaines. Dans ce contexte, notre proposition permet de visualiser, au travers des EDP, les activités de tout ou partie des membres organisationnels. Une application directe consiste à identifier les documents utilisés pour réaliser les activités associées à un poste donné. La prise en compte de cette information peut aider à trouver des personnes-ressources dans un domaine donné, à composer un groupe de travail adapté aux besoins d'un projet, à identifier les centres d'intérêts émergeants, à lutter contre le *turnover* en anti-

cipant les compétences à renouveler (Boyer *et al.*, 2007). Dans ces travaux, les auteurs proposent de cartographier une organisation en fonction du contenu des documents et des relations établies entre les différents acteurs. Par rapport à cette approche, notre proposition introduit la notion d'usage, permettant d'identifier des liens complémentaires entre les documents en fonction de leur organisation dans les EDP.

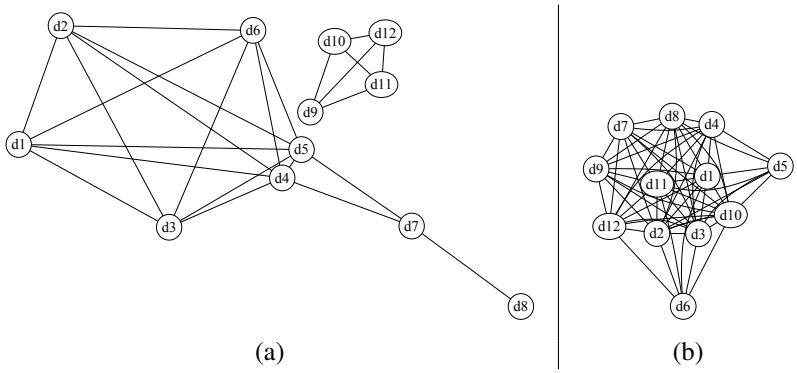

Figure 1. *Comparaison de la similarité des documents sur l'usage (a) par rapport à leur similarité sur le contenu (b). La longueur des arcs est inversement proportionnelle à la similarité entre les nœuds associés qui représentent les documents* d_1 *à* d_{12}.

Nous exposons dans cette section l'interface multi-facettes proposée, en décrivant en premier lieu ses différentes composantes complémentaires qui donnent accès au capital organisationnel (aspect statique). Dans un deuxième temps, nous formalisons les diverses actions que l'usager peut réaliser sur l'interface (aspect dynamique) afin de pouvoir explorer le capital organisationnel au travers des facettes. Enfin, nous présentons dans un troisième temps les aspects relatifs à la mise en œuvre de cette interface : notions de similarité de contenu et d'usage, techniques de visualisation. Ces différents points ont fait l'objet de travaux et de développements antérieurs.

3.1. *Aspect statique de l'interface : représentation du capital organisationnel*

L'interface proposée permet de visualiser les deux dimensions du capital organisationnel : les documents et les personnes. Pour une dimension donnée, l'usager peut explorer un ensemble d'éléments (un groupe de documents ou de personnes) ou un seul élément (un document ou une personne). La combinaison de ces deux paramètres représente quatre cas, matérialisés par des « vues » dans la Figure 2 qui schématise l'architecture globale de l'interface proposée. Une vue peut être assimilée à une fenêtre dans une interface graphique. De plus, chaque vue comprend quatre « facettes » où figurent des informations relatives à la vue choisie. Elles permettent également l'exploration du capital organisationnel car c'est par leur intermédiaire que l'usager passe d'une vue à l'autre (aspect dynamique). Pour chacune des quatre vues, le Tableau 1 recense les facettes disponibles. On distingue trois types de facettes : visualisation,

liste et fiche. Chaque nom de facette est préfixé par l'initiale de son type, soit « v », « l » ou « f ». Nous détaillons dans les sections suivantes chacune des quatre vues, en décrivant les informations accessibles par l'intermédiaire de chaque facette et en donnant un scénario d'utilisation.

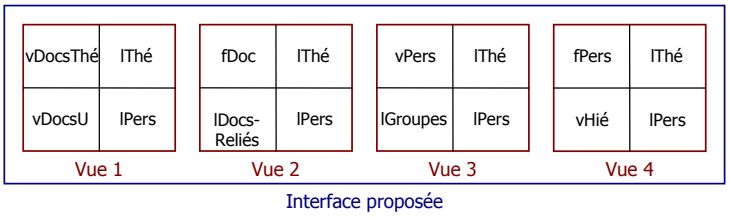

Figure 2. *Architecture générale de l'interface comprenant des vues et des facettes.*

3.1.1. *Vue 1 : représentation d'un groupe de documents*

Les facettes de cette vue permettent de visualiser les documents de l'organisation regroupés par thématique (vDocsThé) et par usage (vDocsU). Ces deux modalités sont respectivement basées sur le contenu des documents et sur leur organisation au sein des EDP. De façon intuitive, deux documents sont d'autant plus proches par le *contenu* qu'ils partagent un nombre important de termes. D'autre part, deux documents sont d'autant plus proches par l'*usage* qu'ils sont organisés ensemble dans les EDP. Le détail des mesures de similarité sur le contenu et sur l'usage est présenté en section 3.3.2. L'usager interagit avec ces facettes en sélectionnant un ou plusieurs documents, il peut alors se focaliser sur ce(s) dernier(s). La facette lPers contient la liste des propriétaires des documents sélectionnés, triable par nombre de documents. Enfin, la facette lThé liste les thématiques concernant les documents sélectionnés, par ordre d'importance.

Facettes		Dimensions			
		Document		Personne	
Nom	Description	Groupe *Vue 1*	Unité *Vue 2*	Groupe *Vue 3*	Unité *Vue 4*
fDoc	fiche d'un document		✓		
fPers	fiche d'une personne				✓
lDocsReliés	liste des documents liés		✓		
lThé	liste de thématiques	✓	✓	✓	✓
lGroupes	liste de groupes			✓	
lPers	liste de personnes	✓	✓	✓	✓
vDocsThé	vue thématique des documents	✓			
vDocsU	vue de l'usage des documents	✓			
vHié	hiérarchie d'une personne				✓
vPers	représentation de personnes			✓	

Tableau 1. *Description des facettes associées aux quatre vues composant l'interface.*

Grâce à cette vue, un individu obtient une représentation des thématiques du fonds documentaire constitué à partir des documents extraits des EDP. En se focalisant sur une thématique particulière, il visualise les personnes qui possèdent ces documents. Il voit également quels documents sont classés avec les documents sélectionnés. Ces documents connexes, issus des EDP de l'organisation, apportent des informations complémentaires par rapport à la sélection originale de l'usager. En reflétant les associations d'idées des membres organisationnels, la facette vDocsU offre un véritable retour sur investissement qui rentabilise l'effort de chaque membre.

3.1.2. *Vue 2 : représentation d'un seul document*

Cette vue présente la fiche d'un document (fDoc) qui donne accès à son titre, à son contenu et aux chemins absolus des EDP qui le contiennent (par exemple, `/home/userX/informatique/bdr/indexation/arbreBalancé/cours.pdf` et `/home/userY/inventeurs/science/info/Rudolf_Bayer/bio.html`). La date de création du document dans chacun de ces chemins est précisée. Les thématiques du document sont listées dans la facette lThé. Les individus qui le possèdent sont recensés dans la facette lPers. Enfin, les documents connexes (utilisés avec le document visualisé) sont listés dans la facette lDocsReliés, ordonnés par similarité d'usage.

Au travers des facettes de cette vue, l'usager identifie les thématiques traitées dans un document. Il connaît également les autres individus qui l'ont rangé dans leur EDP ; les noms des chemins absolus associés fournissent des indications complémentaires sur l'utilisation qui est faite de ce document. Comme l'usager identifie les personnes intéressées par le document, il peut par la suite explorer leurs EDP pour trouver d'autres documents intéressants et éventuellement prendre contact avec eux. Cette fonctionnalité répond aux besoins identifiés dans (Hertzum *et al.*, 2000).

3.1.3. *Vue 3 : représentation d'un groupe de personnes*

Cette troisième vue représente dans la facette vPers un ensemble de personnes et les liens qui les unissent, qu'ils soient d'usage ou de thématique. Cette facette privilégie la visualisation des liens, elle est complétée par la facette lPers qui liste les personnes visualisées. Au sein de l'organisation, chaque individu fait partie de groupes explicites (équipes, groupes de travail, commissions…). Ces derniers sont représentés dans la facette lGroupes : elle contient les noms et le nombre de représentants des groupes distincts correspondant aux personnes visualisées dans vPers, par exemple « Service des ventes (12) ». Enfin, la facette lThé recense les thématiques associées aux EDP des personnes visualisées, triées par nombre de documents associés.

Cette vue permet à un usager d'identifier les intérêts thématiques caractérisant tout groupe de personnes, qu'il soit explicite (une équipe mentionnée dans l'organigramme) ou tacite (des personnes qui ont des affinités, qui déjeunent ensemble, etc.). De ce fait, un membre organisationnel peut identifier et explorer par la suite les thématiques de son équipe. Cette fonctionnalité est très utile en phase d'intégration d'un nouveau collaborateur, lorsque ce dernier doit s'adapter et se former en assimilant les thématiques manipulées par son équipe d'accueil (Boyer *et al.*, 2007). De la même fa-

çon, l'identification des thématiques principales d'une équipe, à partir des EDP, peut aider le service des ressources humaines à établir des fiches de poste. Ces dernières pourront notamment être utilisées pour la création ou le renouvellement d'un poste.

3.1.4. *Vue 4 : représentation d'une seule personne*

Une personne est représentée par sa fiche (fPers) qui contient les informations suivantes : identité (nom, prénom) et groupes d'appartenance. Une représentation hiérarchique des documents structurés dans son EDP est accessible au travers de la facette vHié. La liste des thématiques relatives à son EDP est présentée dans la facette lThé, elles sont classées par importance décroissante. Enfin, la facette lPers recense les personnes qui partagent les mêmes thématiques ou qui utilisent les documents de la même façon que la personne étudiée dans cette quatrième vue.

Un scénario concret d'utilisation consiste, pour un usager donné, à visualiser sa propre fiche pour identifier les personnes proches de lui (par thématique ou par usage). Par la suite, la visualisation de leurs fiches lui permet de connaître les thématiques qui les caractérisent. Il peut alors explorer le contenu de leurs EDP en fonction des thématiques qui l'intéressent et de leur structure.

3.2. *Aspect dynamique de l'interface : exploration du capital organisationnel*

L'interface proposée permet de visualiser le capital organisationnel selon quatre vues spécifiques. Afin de permettre l'exploration et la navigation dans ce capital, nous définissons dans cette section deux types d'interaction entre l'usager et l'interface : l'interaction « intra-vue » et l'interaction « inter-vues ».

Au sein d'une vue quelconque, l'interaction intra-vue consiste à automatiquement répercuter la sélection d'un ou de plusieurs éléments d'une facette sur les trois autres facettes de la vue. Par exemple, la sélection d'un ensemble de thématiques associées à une personne (dans la facette lThé de la vue 4) permet d'identifier, au même moment, ces thématiques dans l'EDP de la personne (facette vHié) et de voir les personnes qui partagent ces mêmes thématiques (facette lPers). Concrètement, chaque facette met en évidence les éléments associés à la sélection grâce à une mise en forme adaptée (couleur différente, graisse de la police, etc.). En fait, l'interaction intra-vue permet de localiser un même élément dans toutes les facettes qui constituent une vue, ces facettes proposant des représentations complémentaires de l'information extraite des EDP.

Le second type d'interaction introduit, appelé inter-vues, permet la navigation d'une vue à l'autre. Concrètement, en fonction d'une action réalisée sur une facette, l'interface remplace la vue actuelle par une autre vue répondant plus précisément au besoin exprimé. La sélection d'une personne au sein de la facette vPers (vue 3) permet par exemple de basculer sur la vue 4, car elle apporte davantage d'informations sur cette personne. De cette façon, les différentes actions réalisées sur les facettes permettent d'explorer le capital organisationnel. Nous avons modélisé la dynamique de l'interface à l'aide du diagramme états-transition de la Figure 3.

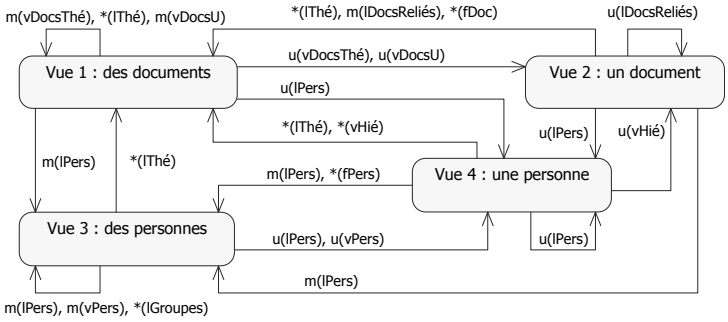

Figure 3. *Diagramme états-transitions décrivant la dynamique de l'interface.*

Les quatre états représentent les vues explicitées dans le Tableau 1. Une transition d'un état e_1 vers un état e_2 est déclenchée par des actions sur une facette de la vue correspondant à e_1. Le détail de ces actions figure sur l'étiquette de la flèche reliant les deux états. Plusieurs actions possibles sont séparées par une virgule. La notation d'une action est du type $s(f)$ où s représente une sélection et f une facette. Plus précisément la sélection multiple est notée « m », la sélection d'un seul élément est notée « u », et « * » désigne une sélection multiple ou unique. Par exemple, l'étiquette « m(lPers), *(fPers) » entre la vue 4 et la vue 3 signifie qu'au travers de la vue 4, une sélection multiple dans la liste des personnes lPers ou une sélection quelconque dans la fiche de la personne fPers mènent à la vue 3.

Afin de donner une vision d'ensemble de l'interface proposée ainsi que des interactions possibles entre les vues, la Figure 4 synthétise les aspects statique (Tableau 1) et dynamique (Figure 3). Les nombreux liens entre les vues montrent le caractère interactif de l'interface, qui facilite la navigation dans le capital organisationnel. Le calcul des facettes incorporées dans les quatre vues fait l'objet de la section suivante.

3.3. *Mise en œuvre de l'interface proposée*

La mise en œuvre de l'interface multi-facettes d'accès au capital organisationnel nécessite de modéliser les données sources à partir desquelles les similarités sur le contenu et sur l'usage sont calculées. Ces dernières peuvent alors être représentées au sein des facettes composant les quatre vues.

3.3.1. *Modélisation des composants du SI nécessaires à notre approche*

L'interface proposée ne vise pas à constituer une nouvelle source d'information, mais plutôt à explorer les EDP gérés par les membres organisationnels. Notre approche est de ce fait basée sur l'exploitation du SI pour extraire des données relatives aux personnes et aux documents de l'organisation. Comme recommandé par la CNIL, nous ne tenons pas compte des répertoires et fichiers personnels afin de respecter la

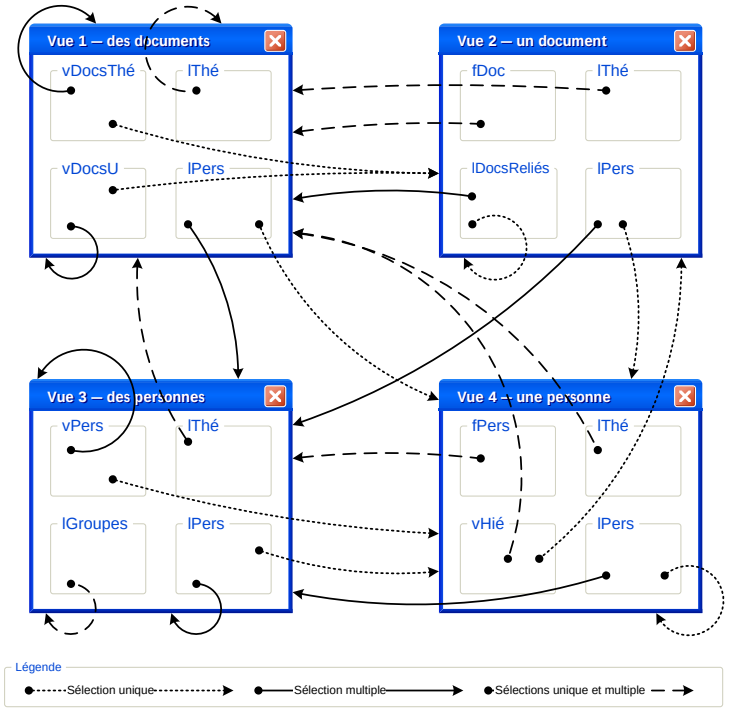

Figure 4. *Synthèse des aspects statique et dynamique de l'interface proposée.*

vie privée des membres organisationnels [1]. Ces éléments sont identifiables grâce à leur nom, qui contient une chaîne de caractères spécifique telle que « perso ». De telles chaînes peuvent être définies au niveau de l'organisation. Ainsi, seuls les répertoires et fichiers non personnels sont exploités. La Figure 5 représente la modélisation conceptuelle UML des données nécessaires au calcul des facettes identifiées dans le Tableau 1. Une fois extraites, ces données sont mises à jour périodiquement pour refléter les activités courantes de l'organisation. Chaque membre est modélisé par la classe Personne, il est caractérisé par son login, son identité (nom et prénom) et le chemin absolu de son EDP (cheminEDP). Il fait éventuellement partie de Groupes, généralement explicités dans l'organigramme de l'organisation ou bien constitués pour des activités spécifiques, telles que des projets. Une personne possède et gère une hiérarchie de Répertoires, ces derniers étant issus de son système de gestion de fichiers, de son arborescence de favoris Internet, etc. Chaque répertoire peut contenir des sous-répertoires et des Exemplaires de Document : plusieurs exemplaires du même document peuvent

1. « un message envoyé ou reçu depuis le poste du travail [...] revêt un caractère professionnel, sauf indication manifeste dans l'objet du message ou dans le nom du répertoire où il pourrait avoir été archivé par son destinataire qui lui conférerait alors le caractère et la nature d'une correspondance privée protégée par le secret des correspondances. » (Bouchet, 2004)

exister dans l'organisation. Un Exemplaire est caractérisé par sa date de création et le nom attribué par son propriétaire. Lors de l'alimentation et des mises à jour incrémentales de la base de données, nous détectons les nouveaux exemplaires correspondant à des documents déjà indexés en appliquant une fonction de hachage sur leur contenu. En complément de l'attribut hachage, la donnée de la taille des documents permet de limiter le problème des collisions de hachage (deux contenus différents possédant une valeur de hachage identique). Pour identifier les thématiques des documents, nous modélisons les Termes qui les composent ; ils sont extraits grâce à un processus d'indexation, classique dans le domaine de la RI (Baeza-Yates *et al.*, 1999, ch. 2). Ce processus comprend généralement les quatre étapes suivantes : segmentation, élimination des « mots vides », lemmatisation et pondération des termes. Le résultat du processus est un ensemble de couples (terme, fréquence relative tf). Cette valeur tf pour un document et un terme donnés (classe Index) représente la fréquence du terme dans le document. Par ailleurs, la fonction getIdf() correspond au pouvoir discriminant du terme dans la collection de documents, appelée « corpus ». Cette valeur est d'autant plus élevée que le terme est rare dans le corpus car, dans ce cas, il a un fort pouvoir discriminant pour les documents qui le contiennent.

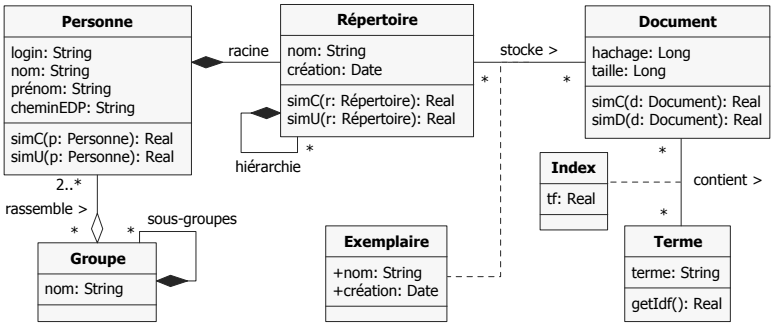

Figure 5. *Diagramme des classes représentant les données exploitées par l'interface.*

La section suivante décrit l'exploitation du contenu des documents représenté par les classes Terme et Index (resp. de l'organisation des documents représenté par la classe Répertoire) et le calcul d'une mesure de similarité thématique (resp. liée à l'usage des documents) correspondant à la méthode simC (resp. simU).

3.3.2. *Mesures de similarité sur le contenu et sur l'usage des documents*

Les informations présentées dans diverses facettes de l'interface sont basées sur le calcul de similarités thématique et d'usage. C'est pourquoi nous détaillons ces similarités avant d'en montrer l'exploitation par des techniques de visualisation adaptées.

3.3.2.1. Similarité basée sur le contenu des documents indexés

Évaluer la similarité entre deux documents est une opération fondamentale dans le domaine de la RI (Baeza-Yates *et al.*, 1999, ch. 2). Une telle similarité est classique-

ment fonction des contenus textuels des documents. Plusieurs modèles mathématiques ont été proposés, le plus répandu étant le modèle vectoriel (Salton *et al.*, 1975) où chaque document est représenté par un vecteur dans l'espace vectoriel des termes distincts du corpus. Ainsi, un document d_i aura pour représentation $\vec{d_i} = (w_i^1, \ldots, w_i^n)$ où un $w_i^j \in \mathbb{R}_+$ correspond au poids du j^{e} terme dans le document d_i, sachant que le corpus comprend n termes. Une pondération courante consiste à fixer $w_i^j = tf_i^j \cdot idf^j$, de ce fait $w_i^j \in [0; 1]$. La similarité entre deux documents d_1 et d_2 est le résultat d'une fonction appliquée aux deux vecteurs qui les représentent, par exemple $\cos(\vec{d_1}, \vec{d_2})$.

Concernant le calcul de la similarité entre deux personnes, nous exploitons l'approche du « méga-document » consistant à représenter une personne comme un document unique, créé en concaténant les contenus des documents issus de son EDP.

Dans la vue 1, la facette vDocsThé est construite à partir des valeurs de similarité calculées pour les documents pris deux à deux. Quant aux thématiques listées dans la facette lThé, elles correspondent aux termes issus de l'indexation, classés par fréquence décroissante. Enfin, la facette vPers de la vue 3 repose sur le calcul des similarités entre personnes prises deux à deux.

3.3.2.2. Similarité basée sur l'usage des documents classés dans les EDP

Contrairement à la similarité de contenu basée sur le résultat de l'indexation, la similarité d'usage définie dans (Cabanac *et al.*, 2007) repose uniquement sur la structure des EDP. Cette mesure n'évalue pas à quel point deux documents contiennent des termes identiques, mais plutôt à quel point ils sont utilisés ensemble par les individus. Les deux similarités (contenu et usage) sont complémentaires : deux documents peuvent être utilisés ensemble sans pour autant contenir les mêmes termes, et vice versa. La similarité sur l'usage repose sur l'observation que les individus regroupent, au sein de leurs EDP, les documents qu'ils estiment similaires selon des critères personnels : par domaine, par objectif, etc. Le calcul de cette similarité repose sur la modélisation des EDP dans un « multi-arbre » qui factorise les documents de l'organisation. La Figure 6 représente un multi-arbre construit à partir des EDP de deux utilisateurs notés u_1 et u_2. Pour deux documents d_x et d_y, la similarité d'usage $s(d_x, d_y)$ dépend de deux facteurs : de leur proximité dans le multi-arbre et du nombre de personnes qui les ont rangés ensemble. Premièrement, pour chaque EDP dont le rang est noté i, nous évaluons la proximité p_i entre d_x et d_y : elle est inversement proportionnelle au nombre d'arcs qui les relient dans le multi-arbre. Puis, nous calculons le nombre n de personnes possédant d_x et d_y dans une branche de leur EDP. Enfin $s(d_x, d_y)$ est proportionnelle à la moyenne des proximités p_i, amplifiée par n.

Par extension, la similarité d'usage $s'(u_x, u_y) = \sum_{1 \leqslant i < j \leqslant m} s(d_i, d_j)$ entre deux personnes u_x et u_y est fonction des documents qu'ils ont en commun — cet ensemble est noté $D = \{d_1, \ldots, d_m\}$. Cette expression traduit le fait que deux personnes sont d'autant plus proches par l'usage qu'elles possèdent les mêmes documents et qu'elles les organisent de la même façon. La condition de la somme prend en compte le fait que la fonction s' est symétrique ; elle génère $N = \sum_{i=1}^{m-1} i = \frac{m(m-1)}{2}$ couples à

évaluer. De ce fait, la valeur de similarité inter-personnes peut être normalisée telle que $\frac{s'(u_x,u_y)}{N \cdot \sup(s)} \in [0;1]$ où $\sup(s)$ représente la borne supérieure de la fonction s.

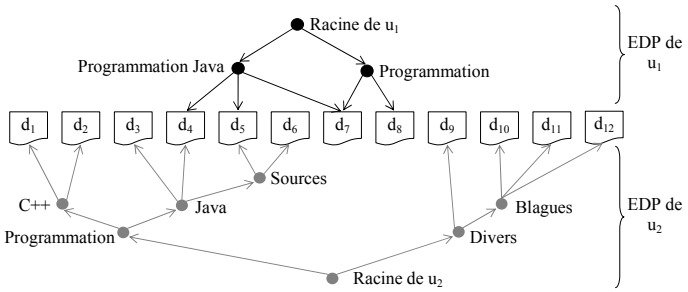

Figure 6. *Exemple d'un multi-arbre construit à partir de deux EDP.*

Le calcul de similarités inter-documents sur l'usage est restitué dans les facettes vDocsU de la vue 1 et lDocsReliés de la vue 2. Concernant la similarité inter-personnes, elle figure dans la facette lPers.

3.3.3. *Techniques de visualisation utilisées*

Il existe une kyrielle de techniques de visualisation dans la littérature (Chen, 2006, ch. 4). Les choix que nous présentons dans cette section peuvent être remis en question en fonction de critères propres à l'organisation ou aux utilisateurs. Nous avons retenu les cartes auto-organisatrices (Kohonen, 2001) pour restituer les thématiques des documents, en se basant sur leurs similarités de contenu. Cette approche est adaptée car elle met visuellement en exergue la notion de proximité thématique, chaque thématique étant représentée par une étiquette (Figure 7). Une telle carte est divisée en zones. La densité des zones, c'est-à-dire le nombre de documents qui y sont associés, est représentée par un dégradé de couleurs. Un avantage de cette technique de visualisation est que l'utilisateur peut consulter les détails d'une zone en la sélectionnant, il obtient alors une nouvelle carte représentant uniquement la zone sélectionnée.

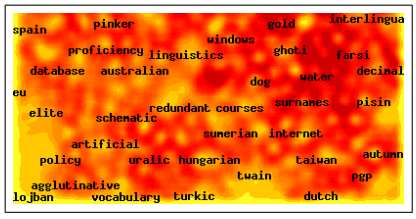

Figure 7. *Carte auto-organisatrice générée par WEBSOM (Lagus* et al.*, 2004).*

D'autre part, les liens d'usage entre les documents et, par extension, entre les personnes sont représentés sous forme de graphes. La notion de proximité d'usage est obtenue par l'application d'un algorithme de placement dirigé par les forces d'attraction-

répulsion (Fruchterman *et al.*, 1991). Le graphe obtenu (Figure 8) représente les documents sous la forme de nœuds, ils sont reliés par des arcs dont la longueur est inversement proportionnelle à leur similarité. Cette visualisation permet d'identifier des groupes de documents utilisés ensemble. Les arcs entre les documents sont étiquetés avec les chemins absolus issus des EDP qui les contiennent.

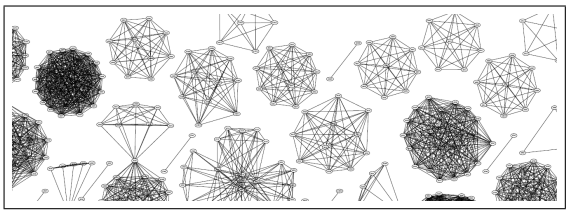

Figure 8. *Graphe de l'usage des documents (Cabanac* et al.*, 2007).*

4. Discussion

Concernant l'analyse des EDP, il convient de souligner un point important : l'interface proposée ne permet pas l'identification d'expertise à proprement parler. En effet, le fait qu'un individu conserve de nombreux documents sur une thématique donnée ne fait pas de lui un expert. Cette observation permet tout au plus de savoir qu'il s'intéresse à cette thématique-là. À l'opposé, un réel expert peut ne pas avoir besoin de stocker dans son EDP des documents qu'il aurait assimilés ou bien qu'il pourrait facilement retrouver par un autre moyen.

Par ailleurs, une étude des motivations d'archivage de documents papier montre que la construction d'un « héritage documentaire » est la seconde motivation après le fait de pouvoir retrouver un document (Kaye *et al.*, 2006). Si les mêmes motivations s'appliquent aux EDP, l'interface proposée permet effectivement le partage des documents et donc la mise en commun de l'héritage documentaire de chacun, sans pour autant demander aux individus de modifier leur façon de travailler. Il est probable que cette faculté sur le principe du donnant-donnant motive les différents membres organisationnels, qui sauront par la suite que leurs efforts d'organisation de leurs EDP bénéficient également à l'organisation dans sa globalité.

5. Conclusion et perspectives

Les « travailleurs du savoir » des organisation modernes disposent d'Espaces Documentaires Personnels (EDP) où ils organisent les documents utiles à la réalisation de leurs activités. La structure hiérarchique est couramment retenue, elle reflète les associations d'idées et plus généralement l'usage des documents qui sont jugés utiles pour les activités de leurs propriétaires. Ainsi, les EDP forment des mines d'informations structurées de façon incrémentale, au fur et à mesure des « découvertes » des membres organisationnels. Bien que le contenu d'un EDP puisse correspondre au besoin de nombreux membres organisationnels (appartenant à une même équipe, par

exemple) il n'est accessible qu'à son propriétaire. De plus, le partage des documents, aussi bien manuellement qu'automatiquement au travers de processus de recommandation, souffre de diverses limites : surcharge cognitive, pertinence des profils usagers dont les centres d'intérêts évoluent sans cesse... C'est en partie pour ces raisons que les individus privilégient des sources externes telles que le Web pour leurs recherches d'information. Pourtant, les EDP semblent plus adaptés aux activités de l'organisation car leur contenu a déjà fait l'objet d'un jugement de pertinence.

Afin de davantage valoriser l'investissement des individus qui gèrent leurs EDP, notre proposition vise à donner accès à ce véritable capital organisationnel. À cet effet, nous avons défini une interface multi-facettes permettant de visualiser les documents et les personnes de l'organisation, à partir des données extraites du SI, plus précisément des EDP organisationnels. Notre proposition repose sur le principe du donnant-donnant : les efforts cognitifs d'un individu sont rentabilisés au niveau de l'organisation qui en bénéficie ; en retour, tout individu peut explorer le capital organisationnel et trouver des documents pertinents eu égard à ses activités. Nous avons souligné l'utilité de l'exploration par thématique et par usage, au travers des facettes de l'interface, pour les membres ainsi que pour le pilotage de l'organisation.

Une perspective à court terme consiste à intégrer au sein de l'interface spécifiée dans cet article les différents composants développés dans nos travaux précédents, notamment dans (Cabanac *et al.*, 2007). Ceci nous permettra d'évaluer l'apport de notre proposition dans une organisation afin de valider l'approche proposée. Dans un premier temps, nous envisageons d'expérimenter cette interface avec une équipe de recherche du laboratoire. Le retour d'expérience des enseignants-chercheurs spécialistes de leurs domaines, ainsi que des nouveaux arrivants néophytes (étudiants en stage de master 2 notamment) fournira une première évaluation qualitative. Des résultats encourageants pourront être approfondis par des évaluations quantitatives. Nous envisageons également d'expérimenter des techniques adaptées à la visualisation de grands graphes (Boutin *et al.*, 2004; Huang *et al.*, 2007). Une autre perspective consiste à prendre en compte la dimension temporelle dans notre approche. En effet, certains besoins requièrent une connaissance actualisée d'un domaine (conseil, veille technologique), alors que d'autres nécessitent une connaissance sur le long terme (recul sur une technologie, rétrospective d'un domaine). Par conséquent, la mise en évidence de l'utilisation réelle des documents sauvegardés dans les EDP permettrait de distinguer les ressources et thématiques qui (ré)émergent par rapport à celles qui sont progressivement abandonnées.

6. Bibliographie

Baeza-Yates R. A., Ribeiro-Neto B. A., *Modern Information Retrieval*, ACM Press/Addison-Wesley, 1999.

Ballay J.-F., "Nous sommes tous des travailleurs du savoir", *L'Expansion Management Review*, vol. 107, p. 94–101, 2002.

Bouchet H., La cybersurveillance sur les lieux de travail, Technical report, CNIL, Paris, France, 2004. available online `http://www.ladocumentationfrancaise.fr/rapports-publics/044000175/`.

Boutin F., Hascoët M., "Focus dependent multi-level graph clustering", *AVI'04: Proceedings of the working conference on Advanced visual interfaces*, ACM, New York, NY, USA, p. 167–170, 2004.

Boyer M., Canut M.-F., Chevalier M., Péninou A., Sèdes F., "Cartographie de l'organisation : une approche topologique des connaissances", *EGC'07 : actes des 7[e] journées Extraction et Gestion des Connaissances*, vol. RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, Cépaduès, p. 557–568, 2007.

Cabanac G., Chevalier M., Chrisment C., Julien C., "An Original Usage-based Metrics for Building a Unified View of Corporate Documents", *DEXA'07: Proceedings of the 18[th] International Conference on Database and Expert Systems Applications*, vol. 4653 of *LNCS*, Springer, p. 202–212, 2007.

Chen C., *Information Visualization: Beyond the Horizon*, 2[nd] edn, Springer, 2006.

Dmitriev P. A., Eiron N., Fontoura M., Shekita E., "Using Annotations in Enterprise Search", *WWW'06: Proceedings of the 15[th] international conference on World Wide Web*, ACM Press, New York, NY, USA, p. 811–817, 2006.

Feldman S., "The high cost of not finding information", *KM World magazine*, vol. 13, n° 3, p. available online `http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=9534`, 2004.

Fruchterman T. M. J., Reingold E. M., "Graph Drawing by Force-directed Placement", *Softw. Pract. Exper.*, vol. 21, n° 11, p. 1129–1164, 1991.

Hertzum M., Pejtersen A. M., "The information-seeking practices of engineers: searching for documents as well as for people", *Inf. Process. Manage.*, vol. 36, n° 5, p. 761–778, 2000.

Huang M. L., Nguyen Q. V., "A Space Efficient Clustered Visualization of Large Graphs", *ICIG'07: Proceedings of the 4[th] International Conference on Image and Graphics*, IEEE Computer Society, Washington, DC, USA, p. 920–927, 2007.

Jones W., Phuwanartnurak A. J., Gill R., Bruce H., "Don't Take My Folders Away!: Organizing Personal Information to Get Things Done", *CHI'05 extended abstracts on Human factors in computing systems*, ACM Press, New York, NY, USA, p. 1505–1508, 2005.

Kaye J. J., Vertesi J., Avery S., Dafoe A., David S., Onaga L., Rosero I., Pinch T., "To Have and to Hold: Exploring the Personal Archive", *CHI'06: Proceedings of the conference on Human Factors in computing systems*, ACM Press, New York, NY, USA, p. 275–284, 2006.

Khoo C. S., Luyt B., Ee C., Osman J., Lim H.-H., Yong S., "How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour", *Information Research*, vol. 11, n° 2, p. available online `http://informationr.net/ir/12-2/paper293.html`, 2007.

Kohonen T., *Self-Organizing Maps*, 3[rd] edn, Springer-Verlag, Secaucus, NJ, USA, 2001.

Lagus K., Kaski S., Kohonen T., "Mining massive document collections by the WEBSOM method", *Inf. Sci.*, vol. 163, n° 1-3, p. 135–156, 2004.

Montaner M., López B., de la Rosa J. L., "A Taxonomy of Recommender Agents on the Internet", *Artif. Intell. Rev.*, vol. 19, n° 4, p. 285-330, 2003.

Salton G., Wong A., Yang C. S., "A Vector Space Model for Automatic Indexing", *Commun. ACM*, vol. 18, n° 11, p. 613–620, 1975.

Sellen A. J., Harper R. H., *The Myth of the Paperless Office*, MIT Press, Cambridge, USA, 2003.