/* */

Approche sémantique multiniveaux minimaliste pour le partage de données dans les *dataspaces*

Laurent Bossu — Éric Leclercq

Laboratoire d'Électronique, Informatique et Image (LE2I) - UMR 5158
Universitéé de Bourgogne
BP 47870 - F-21078 Dijon Cedex
{Laurent.Bossu,Eric.Leclercq}@u-bourgogne.fr

RÉSUMÉ. *Dans cet article, nous développons une approche sémantique pour permettre les accès à des sources de données réparties, autonomes et hétérogènes. L'approche des dataspaces introduite récemment permet de proposer une abstraction de différentes sources. Nous proposons d'utiliser les principes développés dans le cadre du web social et du web sémantique à travers une sémantique multiniveaux afin de permettre un partage simple des données dans un dataspace. Les niveaux de coopération reposent sur les notions de communauté d'usage et de sémantique de domaine. Un exemple dans le domaine médical est développé.*

ABSTRACT. *In this paper, we develop an semantic approach to the abstractions on the data sources distributed (data integration system). Introduced recently, approach of dataspaces makes it possible to propose an abstraction on these different sources. We suggest to use principles developed into semantic web and social web through a multilevel semantics in order to allow a sharing of data in a dataspace. Levels are based on community of use and Semantic Domain. An medical example is developed.*

MOTS-CLÉS : *dataspaces, gestion de données, environnement distribué, sémantique multiniveaux, web socio-sémantique*

KEYWORDS: *dataspaces, data management, distributed environment, multilevel semantics, social web, semantic web*

1. Introduction

Avec l'expansion d'Internet, l'arrivée des nouvelles générations de réseaux, de terminaux mobiles, et de dispositifs automatiques de production de données, de nombreuses sources de données sont rendues accessibles. La diversité des applications et des dispositifs d'acquisition a multiplié les formats de données. De plus en plus, les applications requièrent de pouvoir accéder de manière conjointe à plusieurs sources de données hétérogènes et autonomes. Ainsi, le partage de données est devenu un enjeu crucial pour les applications et ceci, à grande échelle que ce soit pour des systèmes d'information supportant des activités classiques, des projets scientifiques ou encore des structures médicales comme les hôpitaux. Bien que les SGBD intègrent des composants pour gérer les données complexes, ils ne permettent de traiter les nouveaux besoins de partage de données.

Le but de l'intégration de données est d'unifier de multiples sources en fournissant un accès uniforme et transparent au travers d'une vue globale. Plusieurs solutions ont été proposées comme les systèmes à base de médiateur (Ullman, 2000), l'approche pair à pair qui répond spécifiquement à la problématique de la mise à l'échelle comme par exemple les systèmes Hyperion (Arenas *et al.,* 2003) ou Piazza (Tatarinov *et al.,* 2003).

Récemment, Franklin et al. (Franklin *et al.,* 2005) ont proposé la notion de *dataspace* qui est une approche d'intégration incrémentale accumulant une connaissance du domaine durant le processus d'intégration . Cette approche propose une architecture d'intégration où les sources de données cohabitent et les mappings s'effectuent au fur et à mesure. Cette approche suscite encore de nombreux défis à réaliser tels que la recherche d'information, la prise en compte d'une sémantique explicite, la localisation des sources, la réutilisation des attentions humaines, etc. Les premières solutions proposent des outils de gestion de données personnelles. Notre vision de cette approche s'oriente plus dans l'axe des intergiciels (*middlewares*) en fournissant une couche d'abstraction sur la couche de persistance des données afin de permettre un partage de données complexes et multi-format à grande échelle.

L'objectif de l'approche développée dans cet article est de proposer une approche pour enclencher le processus d'intégration des *dataspaces*. Les approches web social et web sémantique permettent de traiter de nombreuses sources de données, l'une est basée sur une communauté d'usage et l'autre est basée sur une communauté d'experts. Nous proposons, dans le cadre des *dataspaces* de combiner ces deux approches et d'établir une sémantique à plusieurs niveaux : un niveau local qui exploite l'aspect communautaire et un niveau domaine métier qui permet d'apporter la formalisation métier à l'aspect communautaire.

Cet article est organisé comme suit. Dans la section 2, nous plaçons ce travail selon deux axes : d'une part, dans le contexte des approches récentes d'intégration, et d'autre part dans le contexte des approches du web social et du web sémantique. La section 3 présente un exemple d'utilisation des *dataspaces* dans le domaine médical. Dans la section 4, nous développons notre approche sémantique pour les dataspaces. La section suivante s'attache à illustrer avec l'exemple évoqué auparavant dans la sec-

tion 3, puis nous en discutons. Enfin, nous concluons en résumant notre proposition, et en présentant nos travaux futurs.

2. État de l'art : aspect données et aspect sémantique

Dans cette section nous développons un panorama des méthodes d'accès au données distribuées selon deux axes : le premier concerne les évolutions des architectures d'intégration, le second présente les aspects accès aux données dans le cadre des approches web sémantique et web social.

2.1. *Évolutions récentes des systèmes d'intégrations de données*

Les systèmes d'intégration de données permettent l'accès aux données et le partage de données mais différents problèmes se sont posés comme le passage à l'échelle. En effet, la plupart des systèmes d'intégration se basent sur une architecture avec un schéma global, des schémas locaux, et doivent permettre de répondre au nombre toujours plus important de sources de données. Ainsi, le processus d'intégration de données ne peut pas être figé mais il doit évoluer continuellement. Des approches récentes proposent des solutions pour l'intégration de données à grande échelle comme par exemple les grilles spécialisées pour la gestion de données, les PDMS (*Peer Data Management System*), et les systèmes de gestion de *dataspace* (*DataSpace Management Systems*).

Les PDMS sont formés d'un ensemble de pairs et chaque pair possède son propre schéma représentant son domaine d'intérêt. Les *mappings* dans les PDMS ne sont pas construits par rapport à un schéma global afin d'éviter des problèmes de couplage, mais ils sont construits directement entre pairs et stockés de façon locale. L'absence de schéma médiateur permet de rendre la gestion des *mappings* plus flexible et plus évolutive. Les avantages d'une telle architecture sont l'aspect décentralisé et la scalabilité, mais elle présente des limites telles que la qualité imprévisible et l'incertitude sur les données. De nombreux problèmes sont à résoudre et des travaux ont été publiés concernant les modèles conceptuels (Tzitzikas *et al.,* 2003), *mappings* entre les pairs (Li *et al.,* 2007), définition de schéma (Tatarinov *et al.,* 2003), algorithme de requêtes (Tatarinov *et al.,* 2004).

Les grilles fournissent un accès consistant et coordonné pour des ressources de stockage et de calcul distribuées et hétérogènes (Jagatheesan *et al.,* 2003). Les grilles de données sont destinées à manipuler des données pour partager l'accès aux données et aux systèmes de stockage (Risch *et al.,* 2002). Ainsi, elles fournissent un partage coordonné du stockage de l'information, un espace logique de nommage pour la localisation indépendante d'identifiants et des APIs d'accès. Les systèmes reposant sur cette architecture fournissent également des services de base pour gérer l'état de l'information sur les collections de la grille, la connaissance des évènements et services. Cette approche même si elle permet une intégration à grande échelle reste statique.

Dans l'approche des *dataspaces* (Franklin *et al.,* 2005)(Halevy *et al.,* 2006), l'inté-

gration évolue au cours du temps et seulement où cela est nécessaire contrairement aux approches traditionnelles d'intégration de données. Les *mappings* entre les sources peuvent être soit générés automatiquement soit définis par les utilisateurs. Le processus d'intégration des données exploite différentes informations contextuelles comme les métadonnées sémantiques, des regroupements de fichier dans des répertoires, les requêtes utilisateur. Dans le domaine de la gestion de données personnelles, deux prototypes ont été développés : Semex (Dong *et al.,* 2005) et iMemex (Dittrich *et al.,* 2005). Ces deux systèmes permettent de gérer des données structurées et non structurées provenant de sources réparties mais seulement à un niveau local, c'est-à-dire qu'ils permettent d'intégrer des sources de données personnelles.

Dans les grilles, le processus d'intégration est fixe et ne permet pas de flexibilité. Dans l'approche des PDMS, le processus d'intégration est simple mais la qualité des données est incertaine. En revanche, l'approche *dataspace* propose un processus d'intégration incrémental qui doit simplement être enclenché afin de mettre à disposition les services aux applications s'appuyant sur le *dataspace*.

2.2. *Étude des concepts liés au web sémantique et au web social*

Le web sémantique et le web social sont deux visions du web qui se distinguent sur les connaissances exploitées : les unes sont issues d'une communauté d'expert (domaine métier spécifique) et les autres sont issues d'une communauté d'usage (les utilisateurs). Le web social met l'utilisateur au centre du processus de publication et d'échange d'informations et utilise le *tagging* pour annoter le contenu. Au contraire, le web sémantique repose sur une sémantique délivrée par des métadonnées et des ontologies.

Le web sémantique se base sur l'instauration de marqueurs sémantiques sur les ressources du web qui serviront à expliciter le contenu de ses ressources. Les ontologies serviront à définir le vocabulaire de ces marqueurs. Une ontologie possède donc une taxinomie et un ensemble de règles d'inférence conceptualisant un domaine de connaissances particulier. Les ontologies sont souvent conçues dans une optique particulière et sont souvent appréhendables que par leurs géniteurs (Mikroyannidis, 2007). De plus, les ontologies requièrent des mises à jour constantes afin de correspondre au mieux aux entités qu'elles représentent. De nombreuses ontologies relativement massives ont été élaborées et sont disponibles pour une grande variété de domaines, leurs exploitations requièrent souvent une connaissance très précise du domaine. Le principal obstacle des ontologies réside dans le fait qu'elles représentent un point de vue particulier d'experts sur un domaine et que leur utilisation est trop restrictive pour un utilisateur quelconque. L'intéropérabilité dans les ontologies semble un enjeu crucial et des travaux vont dans ce sens comme le mapping inter-ontologies ou encore la fusion d'ontologies.

Le web social est initié avec l'arrivée des nouvelles applications du web et qui permet d'accroître les moyens expressifs des utilisateurs, notamment grâce aux outils collaboratifs. Le web social a pour but d'augmenter la richesse sémantique du web actuel sans la complexité de mise en œuvre du web sémantique, mais cette approche

reste très limitée. L'apparition récente des folksonomies (Mathes, 2004) offre la possibilité aux utilisateurs d'indexer du contenu sur le web, librement, à l'inverse des taxinomies et montre l'intérêt d'une approche d'indexation personnalisée et sans véritables contraintes pour les utilisateurs. Cependant, même si cette approche présente des avantages au niveau de la souplesse et de l'adaptabilité, elle met en évidence des faiblesses au niveau de la cohérence des tags utilisés. En effet, les tags peuvent être mal interprétés en l'absence d'information contextuelle ou être sans signification par absence de connaissances sur le domaine. Ainsi, le processus d'indexation ne repose plus sur des concepts valides, mais sur des tags au sens incertain. Malgré cette limite indéniable, les folksonomies mettent en évidence la possibilité de concevoir une approche d'un web participatif.

En combinant les bénéfices qu'apportent chacune de ces deux approches, nous proposons un cadre, d'une part, qui favorise la constitution de partage de données ainsi que l'utilisation aisée et libre d'annotation de contenu et, d'autre part, qui apporte une formalisation à un domaine particulier.

3. Exemple illustrant l'intérêt de notre approche

Pour montrer l'intérêt de notre approche, considérons un exemple dans le domaine médical avec le système d'information d'un hôpital. Nous décrivons le système d'information de l'hôpital de façon simplifiée, et nous illustrons notre démarche avec deux situations de coopérations.

L'infrastructure de l'intranet d'un hôpital se compose d'un SGBD destiné à la gestion des patients (partie administrative) et un ensemble de serveurs de stockage de données qui sauvegardent les données des stations et des terminaux des différents services. Nous nous intéressons plus particulièrement à ceux dédiés à la radiologie et aux médecins pratiquants dans l'hôpital (partie données cliniques). Les différents serveurs sont installés dans plusieurs endroits. Un médecin génère et stocke des données différentes et variées sur ses patients, des examens et des interventions qu'il réalise, ses activités de recherche, des données personnelles. Les informations concernant ses patients peuvent être soit classées selon une organisation définie ou soit réparties dans des répertoires quelconques.

Dans la première situation, nous nous plaçons dans un groupe restreint de participants, et dans la deuxième situtation, nous sommes dans le cadre d'une coopération à plus grande échelle avec une intéraction entre plusieurs services.

À l'échelle d'un groupe de médecins travaillant au sein d'un même hôpital et pouvant être affectés dans des services différents, nous considérons que chaque médecin dispose et utilise une application qui s'appuie un système de gestion de *dataspaces*. Cette application permet au médecin d'annoter en définissant ses propres tags dans une folksonomie, de gérer et de rechercher des données. Un médecin X et un médecin Y veulent collaborer pour traiter des patients communs, en partageant leurs documents. L'annotation de documents utilisera des tags assez similaires relatifs à la pathologie, au traitement, etc., donc ces tags pourront être regroupés au sein d'un même folksonomie pour un usage collectif de façon restreinte. Le service de radiologie dispose d'une

application similaire mais dans le but d'annoter les clichés et les comptes-rendus établis par les médecins radiogues et les techniciens, ainsi que pour le service d'analyses médicales, les données des analyses sont stockées dans un SGBD.

À l'échelle de l'hôpital, nous simulons un médecin cancérologue qui veut constituer un dossier médical de patients souffrant de certaines maladies spécifiques. Ce dossier médical nécessite la mise en place d'une application qui requière le regroupement de données issues de plusieurs sources, avec comme contrainte des restrictions de droits d'accès selon le statut de la personne qui le consulte. Pour illustrer la coopération mise en place à un niveau domaine métier, nous intégrons des coopérations locales évoquées auparavant, à savoir un groupe médecin, le service de radiologie et le service d'analyses médicales. L'application gérant le dossier patient utilise une folksonomie permettant une annotation des données regroupées au sein de la coopération et repose sur un système de gestion de *dataspaces*. Chaque source utilise une folksonomie différente qui regroupent les tags qui auront servis à l'annotation. Nous pouvons supposer que nous pouvons établir certains *mappings* basé sur la synonymie, antonymie,etc. entre les tags des folksonomies locales et la folksonomie globale. Ces *mappings* pourront être invalidés s'ils sont incohérents en se référant à une ontologie de domaine qui modélisera la connaissance du domaine.

Au travers de cet exemple, nous avons montré l'intérêt d'une approche sémantique basée sur la définition d'une taxinomie libre établie par des utilisateurs pour une coopération spécifique et l'utilisation d'une ontologie de domaine pour une coopération à plus grande échelle dans un domaine donné. Nous mettons en évidence l'aspect multiniveaux de cette approche : l'un dans un cadre restreint et l'autre à un domaine métier. Dans la section suivante, nous présentons notre approche de façon plus formalisée.

4. Une approche sémantique multiniveaux pour les *dataspaces*

Les *dataspaces* sont une solution de gestion uniforme de données distribuées et ils se caractérisent par leur flexibilité et leur évolutivité. Le système de gestion d'un *dataspace* fournit des services destinés à assurer le contrôle, l'organisation, le stockage et la recherche de données. Les fonctionnalités remplies par ce système sont extensibles selon les nécessités des applications. Étant une méthode d'intégration, ils nécessitent la mise en place d'un schéma virtuel et d'un schéma de *mappings* entre les sources de données. Traditionnellement deux approches d'intégrations ont été proposées l'approche GaV (Global-as-Views), qui consiste à définir le schéma global en fonction des schémas locaux de chaque source de données et l'approche LaV (Local-as-View), qui se base sur la définition des schémas locaux des sources de données en fonction du schéma global prédéfini. L'approche des *dataspaces* repose sur une intégration incrémentale au niveau des sources, par conséquent l'approche LaV répond pleinement aux besoins d'évolutivité des dataspaces au niveau de la gestion des sources de données.

4.1. *Description de notre approche*

Dans une conception du web rapprochant web sémantique et web social, Mikroyannidis (Mikroyannidis, 2007) suggère d'utiliser les folksonomies élaborées par les utilisateurs du web pour construire et faire évoluer les ontologies et les métadonnées utiles au web sémantique. Dans (Rousset, 2004), l'auteur propose de faire évoluer le web actuel par étapes successives en effectuant une annotation manuelle de document basée sur une taxinomie, et en opérant des *mappings* entre les taxinomies afin de pouvoir supporter un processus de recherche d'information. Les ontologies avec leur formalisation métier peuvent apporter des mécanismes de contrôle pour des applications collaboratives. Ainsi, ces réflexions montrent que le web social et le web sémantique sont complémentaires et que nous pouvons exploiter leurs qualités afin de les transposer aux *dataspaces*.

Notre approche repose sur une ontologie, la coexistence de schémas (LaV), et des mappings entre schéma et folksonomies. En nous basant sur ces principes pour les *dataspaces*, nous envisageons des coopérations à deux niveaux, un **niveau local** qui s'articule autour des folksonomies et un **niveau du domaine** qui s'appuie sur les ontologies de domaine afin de fournir une meilleure intéropérabilité sémantique. Chaque source de données ou ensemble de sources, par l'intermédiaire de définition de mappings, est associée à une folksonomie vu comme une méthode d'indexation de contenu établie par l'utilisateur. Pour palier au mieux les inconvénients lié à leur utilisation, les tags de la folksonomie seront classifiés en extrayant des concepts généraux d'une ontologie. Donc tout nouveau tag sera classé selon des concepts préexistants.

Au niveau local, c'est-à-dire dans le cadre restreint d'une petite communauté

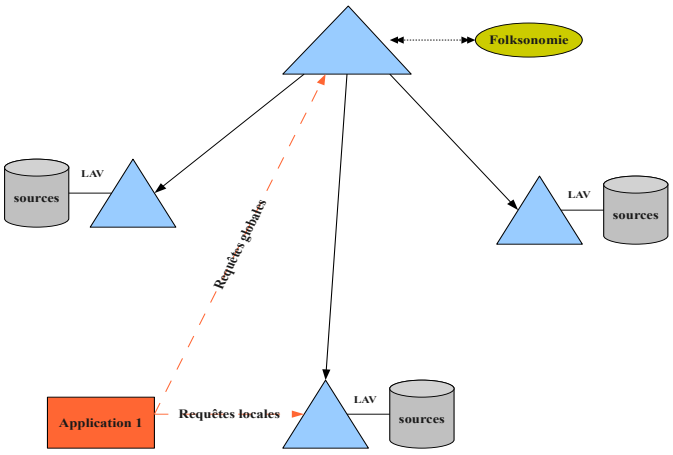

Figure 1. *Configuration pour une approche d'une communauté d'usage*

d'usage collaborant, la terminologie utilisée n'aura pas de variations très marquées, ce qui aura l'avantage d'avoir une simplicité et une uniformité. Ainsi, il est possible d'intégrer les nouvelles sources partagées sans avoir de problèmes d'hétérogénéité sémantique importants à combler. La figure 1 illustre l'approche locale avec trois sources intégrées au moyen d'une vue globale. Une application interroge localement ses sources au travers du schéma local et sur la vue globale pour une requête portant sur des sources réparties.

Au niveau domaine, c'est-à-dire dans le cadre d'un domaine métier (figure 2),

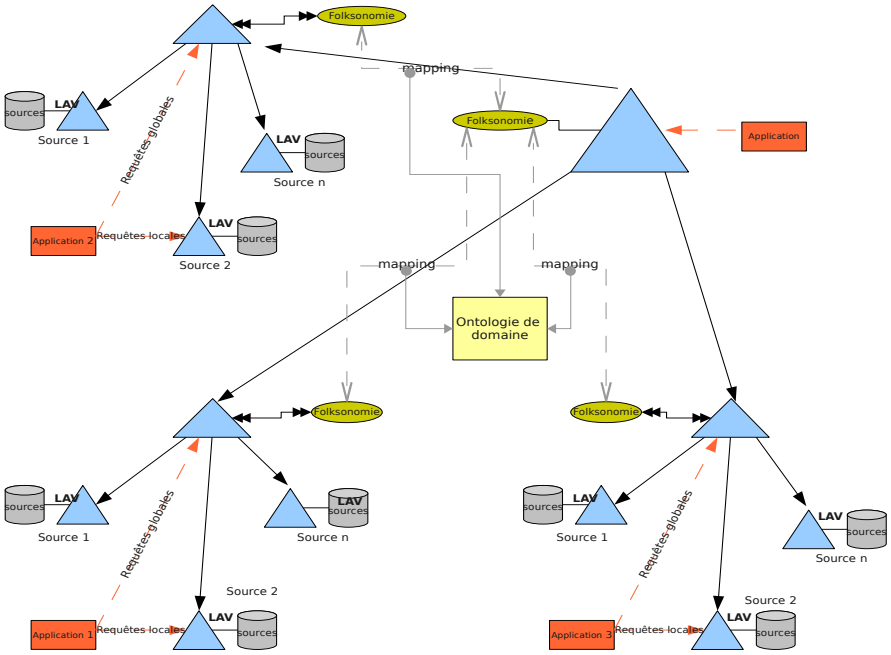

Figure 2. *Configuration pour une approche dans le cadre d'un domaine métier*

nous proposons d'utiliser des *mappings* inter-folksonomies et une ontologie de domaine. Les *mappings* définissent les relations sémantiques entre les folksonomies qui représentent d'une certaine façon la connaissance des sources. Une folksonomie source aura une base de *mappings*, ou correspondances, avec d'autres folksonomies, et l'ensemble des *mappings* est un ensemble de relations entre les termes de cette folksonomie et ceux d'une autre. L'introduction d'une ontologie dans notre système vise à réduire les incohérences terminologiques sur ces *mappings*. Se référer à l'ontologie pour invalider certains *mappings* apportera une sémantique plus rigoureuse et plus cadrée. Ainsi, une application s'appuyant sur cette architecture pourra obtenir des réponses à des requêtes plus précises et de meilleure qualité.

Nous avons montré l'intérêt de l'utilisation des folksonomies dans un cadre restreint permettant d'apporter de la flexibilité dans le processus d'intégration de données de façon collaborative. La folksonomie apporte une sémantique basé sur une communauté d'usage, définie par un groupe restreint et permet de résoudre des conflits sémantiques. À plus grande échelle, les mappings inter-folksonomies conforme à une ontologie de fournir une sémantique plus riche pour un domaine donné.

4.2. *Aspects formels de la sémantique proposée*

Une source de données sera représenté par un schéma local et une folksonomie. Le schéma local est le modèle de données orienté objet. Chaque objet est unique et est défini par un ensemble d'opérations et de propriétés (attributs ou relations entre les objets). Les termes de la folksonomie peuvent être soit des valeurs d'attributs d'un objet, le nom d'un objet ou d'une relation entre deux objets.

Définition 1. *Soit un ensemble de folksonomies $F = \{f_1, ...f_n\}$. Une folksonomie pseudo-classifiée $f \in F$ peut-être représentée sous la forme d'un graphe arborescent, et est définie comme un triplet $\langle R, T, TS \rangle$ où R est la racine, T un ensemble de tags généraux et TS un ensemble de tags spécifiques. Un tag spécifique $ts \in TS$ est associé par une relation d'appartenance à un tag général $t \in T$, et sera noté $\{t : ts\}$*

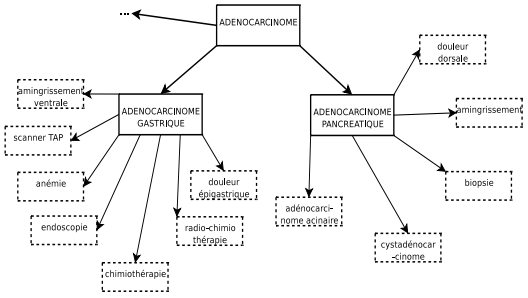

Figure 3. *Exemple d'un graphe d'une folksonomie construite pour les médecins (communauté d'usage)*

La figure 3 montre un exemple d'un graphe construit sur la base de la folksonomie au niveau local. La racine est le label "adénocarcinome" et les tags hiérarchiques sont les autres labels "adénocarcinome gastrique" et "adénocarcinome pancréatique". Les tags spécifiques sont les autres labels rattachés à ses tags.

Nous distinguons trois types de mappings dans lequels les termes de la folksonomie peuvent intervenir. Les termes de la folksonomie peuvent être intervenir au niveau du schéma, mapping schéma-folksonomie, au niveau des données d'une source pour annoter le document, mapping donnée-folksonomie, et au niveau de deux folksonomies pour pouvoir interroger les sources lors d'une coopération à plus grande échelle, mappings inter-folksonomies.

Définition 2. *Un mapping schéma-folksonomie* msf *est une paire* $\left\langle e_{Sc}, \{t:ts\}_{i,f} \right\rangle$, *où* e_{Sc} *est un élément d'un schéma* Sc *et un tag spécifique* $\{t:ts\}_f$ *d'une folksonomie f. Un élément* e_{Sc} *peut être soit une valeur d'attributs d'un objet, soit le nom d'un objet ou d'une relation entre deux objets.*

Définition 3. *Un mapping donnée-folksonomie* mdf *est une paire* $\left\langle e_s, \{t:ts\}_{i,f} \right\rangle$, *où* e_s *est un élément d'une source* s *et un tag spécifique* $\{t:ts\}_f$ *d'une folksonomie f.*

Définition 4. *Un ensemble de mappings inter-folksonomies est défini entre une folksonomie source fs et une folksonomie cible fc, noté* $M = \langle m_{Fs}, m_{Fc} \rangle$. *Un mapping inter-folksonomies* $m \in M$ *représente une correspondance sémantique inter-folksonomies entre un tag spécifique* $\{t:ts\}_{i,Fs}$ *de Fs et un tag spécifique* $\{t:ts\}_{j,Fc}$ *de Fc, noté* $\left\langle \{t:ts\}_{i,Fs}, \{t:ts\}_{j,Fc}, \propto \right\rangle$ *où* $\propto$ *est la relation sémantique entre les deus tags spécifiques* $\{t:ts\}_{i,Fs}$ *et* $\{t:ts\}_{j,Fc}$

Définition 5. *Une relation sémantique* $\propto$ *entre un tag spécifique* $\{t:ts\}_{i,Fs}$ *d'une folksonomie source fs et un tag spécifique* $\{t:ts\}_{j,Fc}$ *d'une folksonomie cible fc est de différente nature : une relation hiérarchique, qui est réciproque, se basant sur un rapport de spécialisation entre un tag général (spécificité (est plus spécifique que) (*$\sqsubseteq$*), généralité (est plus général que) (*$\sqsupseteq$*)) et des tags spécifiques et une relation lexicale classique (synonymie (équivalence) (*$\equiv$*), antonymie (*$\neq$*), similarité terminologique (acronyme, abréviation, accord, etc.) (=)). Ainsi,* $\propto \in \{\sqsubseteq, \sqsupseteq, \equiv, \neq, =\}$

5. Scénario de coopération illustrant notre approche

Nous illustrons notre approche sémantique multiniveaux en reprenant l'exemple précédemment décrit. Nous nous plaçons dans le cas de médecins qui traitent des patients atteints de cancers du pancréas (adénocarcinome pancréatique), de l'estomac (adénocarcinome gastrique), etc. Les médecins sont amenés à rédiger des documents relatifs aux patients, à faire de la recherche sur des traitements, etc. Des examens sont nécessaires pour décrire les stades d'évolution de la maladie. Ils sont de natures biologiques (analyse) ou de nature visuelle (imagerie : radiographie, scintigraphie, scanner, IRM, etc.). Nous considérons un niveau de coopération local, des médecins qui traitent des patients communs, et un autre plus global, les données issues du service de radiologie, de l'analyse médicale et du service clinique (médecins) devront être accessibles en partie.

Un médecin annote ses documents relatifs à ses patients. L'application qui s'appuie sur une couche *dataspace* fournit la possibilité d'annoter tout document texte, image en leur associant des tags grâce à XML. Pour les documents texte produits, des balises XML supplémentaires peuvent être incluses dans l'en-tête du fichier comportant les tags de la folksonomie à l'image des documents OpenOffice.org. De même, les images peuvent être enregistrées dans un nouveau format incluant ses tags. Dans le service de radiologie, cette annotation peut se révéler très importante, du fait que certaines ressources visuelles peuvent être difficilement exploitables et difficiles à re-

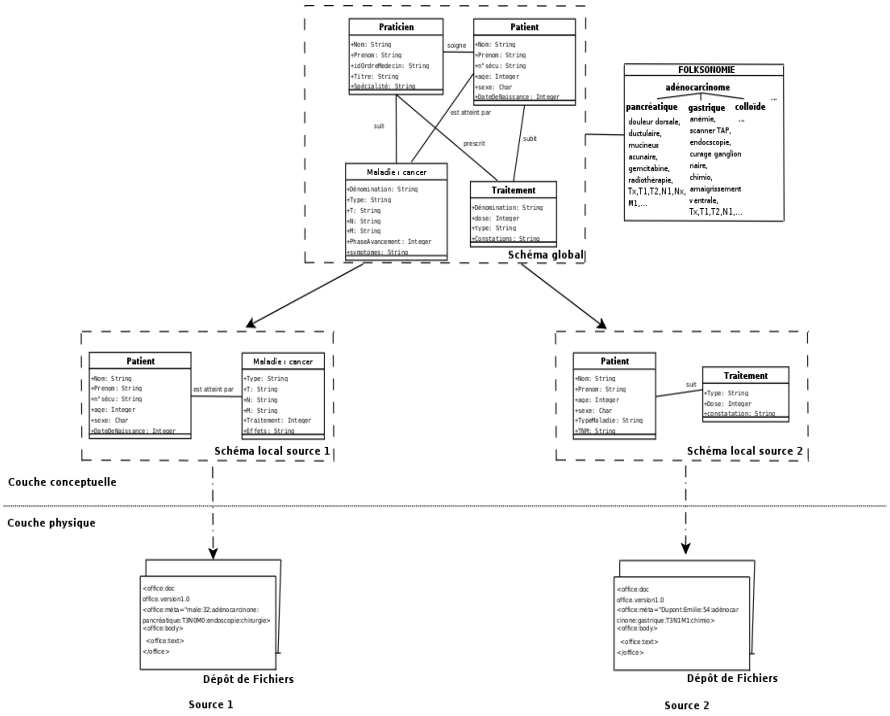

Figure 4. *Illustration d'une coopération en environnement local*

trouver. Pour effectuer des recherches pertinentes et précises, les informations telles que le nom du patient, la type d'examen pratiqué, des mots-clefs du diagnostique du radiologue ou encore les parties du corps observées, etc. Par exemple, l'utilisation du format EXIF (*EXchangeable Image File*) ou les métadonnées IPTC (*International Press and Telecommunications Council*) plus orientées sémantique peuvent être employées. Le service d'analyse médicale sauvegarde ses données dans une base de données. Nous pouvons inclure les tags au niveau du tuple en associant un tag à une valeur particulière du tuple considéré. Par exemple, chaque analyse mesure un certain taux dans une unité particulière. Avec le taux enregistré, on incorpore l'unité de mesure en spécifiant par exemple *35 :mg/L*. Le premier champ correspondrait à la donnée et le suivant au champ du tag. Les figures 4 et 5 montrent la représentation de ces différentes sources de données (partie couche physique).

Des experts du domaine définissent une ontologie de domaine sur les différents cancers au niveau abdominal. Cette ontologie sera limitée à la classification TNM du cancer pancréatique. Trois concepts généraux tumeur (T), adénopathies régionales (N) et métastases à distance (M) sont définis et des termes plus spécifiques sont identifiés pour chaque concept selon cetains critères qualitatifs :

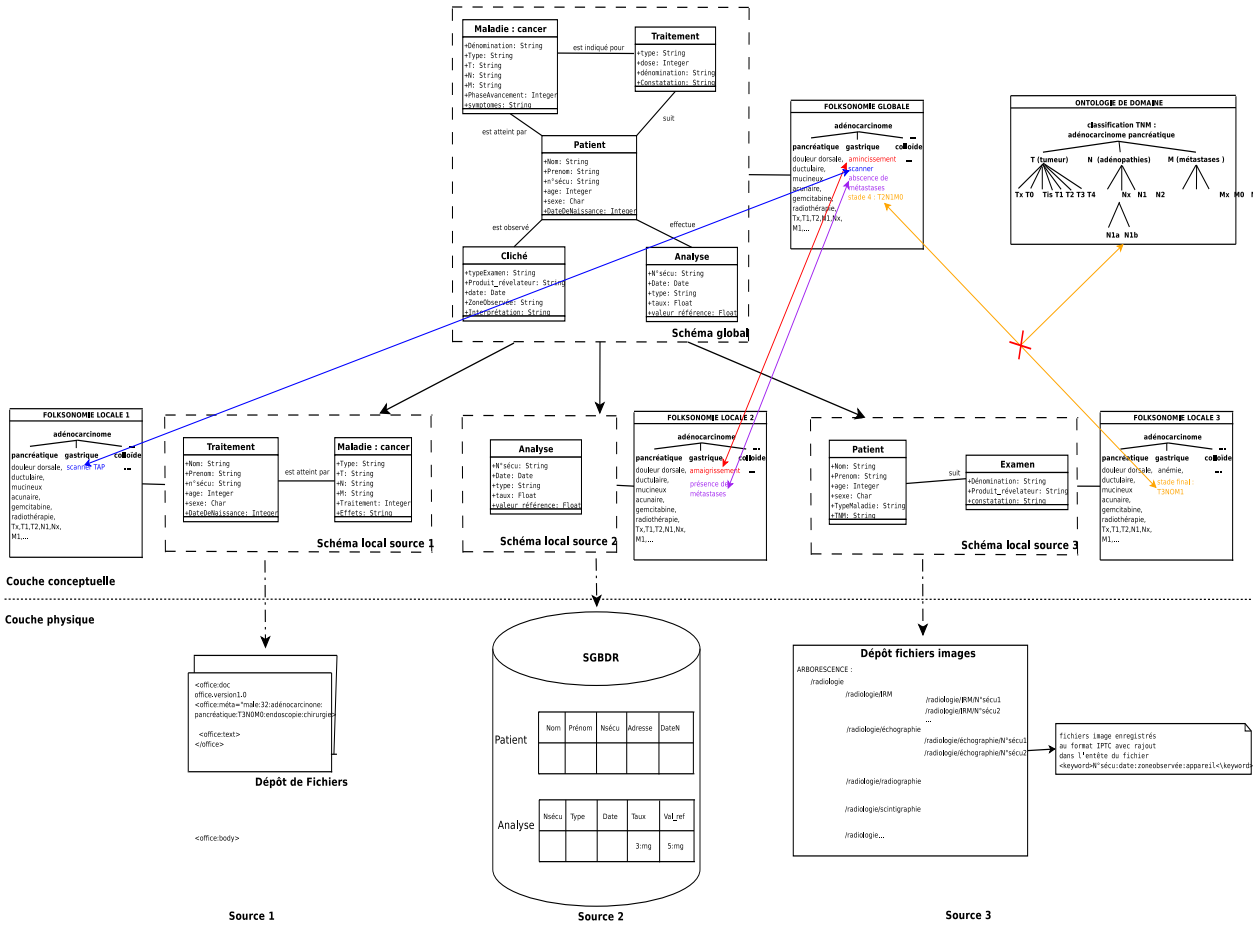

Figure 5. *Illustration d'une coopération en environnement global*

T : Tx (insuffisance des renseignements), T0 (absence de signes de tumeur primitive), Tis (carcinome in situ), T1 (tumeur limitée au pancréas inférieur à 2cm), T2 (tumeur limitée au pancréas supérieur à 2cm), T3 (tumeur touchant un ou plusieurs des organes : duodénum, canal biliaire, tissu péripancréatique), T4 (tumeur touchant un ou plusieurs des organes : estomac, rate, côlon, vaisseaux adjacents) ;

N : Nx (insuffisance des renseignements), N0 (absence de métastase ganglionnaire régionale), N1 (prolifération des ganglions lymphatiques régionaux) ;

M : Mx (insuffisance des renseignements), M0 (absence de métastases à distance), M1 (présence de métastases à distance).

D'autres critères peuvent référencer les stades d'évolution de la maladie selon cette classification : stade I : T1-T2 N0 M0, stade II : T3-T4 N0 M0, stade III : quelquesoit T, N1-N2 M0 et quelquesoient T et N, M1. Ces critères nous permettent d'apporter les règles pour invalider les *mappings* incohérents par rapport à l'ontologie.

Au niveau local, deux médecins cancérologues collaborent pour traiter des patients communs (figure 4) en partageant des données annotées avec des termes d'une folksonomie commune. Ce système permet de partager une sémantique commune.

Au niveau domaine constitué de différents services de l'hôpital, un médecin effectue des travaux sur le cancer pancréatique et il recherche toutes les données en rapport avec cette maladie spécifique pour faire un état des traitements appliqués. L'interrogation des données, annotées par une folksonomie locale, se fait par l'intermédiaire d'un schéma global défini auquel est rattaché une folksonomie globale.

Illustrons les différentes formes de *mappings* évoquées dans la section 3.2 avec notre exemple de la figure 5. Soient Fg la folksonomie source du schéma global et Fsi, avec i=1..3, la folksonomie cible du schéma local de l source Si. La relation $\left\langle \{gastrique : scanner\}_{Fg}, \{gastrique : scannerTAP\}_{Fs1}, \sqsubseteq \right\rangle$ indique que le tag spécifique *scanner* de Fg inclut tout type de scanner (ou tomo-densitométrie). La scanner thoraco-abdomino-pelvienne est un examen plus spécifique que la scanner plus généraliste. Une autre relation $\left\langle \{gastrique : amincissement\}_{Fg}, \{gastrique : amaigrissement\}_{Fs2}, \equiv \right\rangle$ fait référence à un *mapping* basé sur la synonymie, alors qu'une autre relation $\left\langle \{gastrique : absence\}_{Fg}, \{gastrique : presence\}_{Fs3}, \neq \right\rangle$ symbolise un *mapping* basé sur l'antonymie.

Contrairement à d'autres approches qui utilisent l'ontologie pour opérer des *mappings*, ou encore de classifier les termes d'une taxinomie, notre ontologie de domaine commune à l'ensemble d'un domaine métier, dans notre cas l'hôpital, permettra d'invalider les *mappings* qui seront incohérents avec le domaine. Par exemple, la folksonomie du schéma global classe le tag "scanner TAP" dans le concept "gastrique" alors que la folksonomie du schéma de la source S1 spécifie un tag équivalent "tomo-densitométrie thoraco-abdomino-pelvienne" dans le concept "colorectal". L'ontologie pourrait spécifier que ce type d'examen est réservé dans le cas de la maladie "adénocarcinome gastrique", le *mapping* serait alors invalidé. La relation $\left\langle \{gastrique : stadeIVT0N1M0\}_{Fg}, \{gastrique : stadefinalT3N0M1\}_{Fs1}, \equiv \right\rangle$

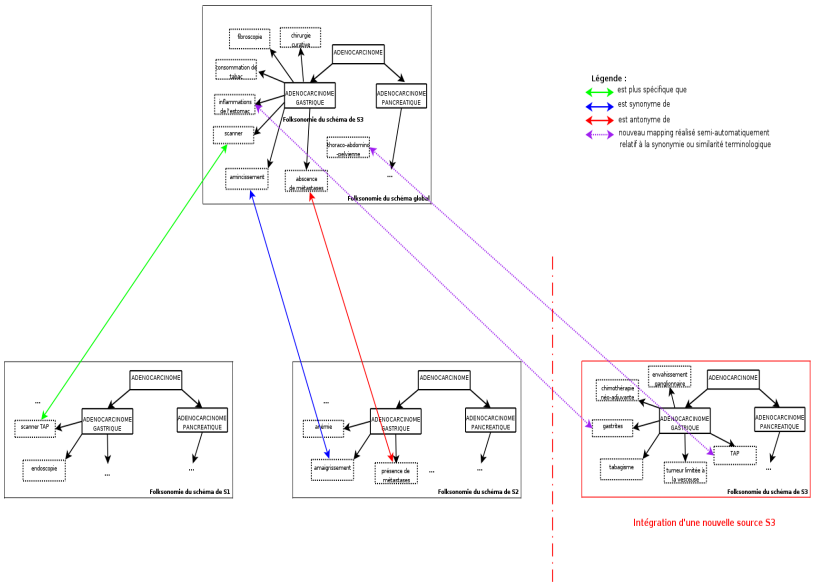

Figure 6. *Illustration de mappings avec l'intégration d'une nouvelle source*

est contraire à la règle de l'ontologie qui spécifie que le dernier stade quelquesoient T, N pour tout M1. On pourra le spécifier avec l'opérateur $\neg$ et la relation sera notée : $\left\langle \{gastrique : stadeIVT0N1M0\}_{Fg}, \{gastrique : stadefinalT3N0M1\}_{Fs1}, \equiv: \neg \right\rangle$

Quand une nouvelle source requiert à être intégrée dans un *dataspace*, des *mappings* inter-folksonomies sont réalisés semi-automatiquement entre la folksonomie source et la folksonomie cible basées sur des relations de synonymie ou de similarité terminologique. Des *mappings* supplémentaires seront établis incrémentalement selon les besoins d'intégration ultérieurement. La figure 6 illustre l'arrivée d'une nouvelle source à intégrer. Sa folksonomie est mise à disposition du schéma global pour que des *mappings* puissent être réalisés avec sa folksonomie tels que $\left\langle \{gastrique : inflammationsestomac\}_{Fg}, \{gastrique : gastrites\}_{Fs2}, \equiv \right\rangle$.
Une requête posée sur le schéma global spécifiée en SQL par exemple telle que `SELECT * FROM Maladie :cancer Mc WHERE Mc.symptomes='gastrites'` n'aurait aucun résultat pertinent sur les deux seules sources déjà intégrées si les *mappings* n'avaient pas été générés au préalable sur la dernière source.

6. Conclusion et perspectives

Les *dataspaces* soulèvent des enjeux majeurs en terme de sémantique, et de processus d'intégration. Le processus d'intégration étant incrémental, nous avons proposé

une méthode pour fournir un socle sémantique qui permettra d'interroger les sources de données et de fournir des services sur ces sources. Notre approche sémantique pour les *dataspaces* incluant plusieurs niveaux et ayant une folksonomie qui repose sur une communauté d'usage et une formalisation d'un domaine métier pour réfuter des *mappings* sémantiques en contradiction avec le domaine. Nous avons montré l'intérêt d'une telle approche dans un domaine spécialisé. Une sémantique enrichie par les acteurs du domaine permet d'accroître la précision de la terminologie employée et ainsi converger vers un ensemble commun de termes dans une utilisation locale. L'ontologie de domaine, fruit d'une communauté d'experts, va permettre d'apporter une mise en conformité des *mappings* par rapport au domaine métier, à plus grande échelle.

Dans notre approche, nous devons tenir compte de l'évolution des constituants du système. La folksonomie, fruit d'un processus collaboratif, aura une évolution inéductable et aura des impacts sur le système. L'évolution de l'ontologie aura aussi des conséquences sur la cohérence des *mappings* entre les termes de la folksonomie. Certains *mappings* seront liés à une version de l'ontologie tandis que d'autres en fonction d'une version ultérieure. Nous pourrions envisager de vérifier automatiquement les *mappings* existants avec la nouvelle version de l'ontologie, ou de proposer des vues globales avec des versions de l'ontologie et de la folksonomie.

Les changements dans les sources peuvent intervenir au niveau du schéma et au niveau des données. Comme nous sommes dans une approche d'intégration de type LAV, les changements de schéma auront peu d'impact sur le schéma global. Dans un cadre scientifique avec des appareils d'acquisition ou des capteurs, des changements au niveau de l'enregistrement des données peuvent intervenir. La prise en compte des versions des données est impérative et doit être opérée.

Dans nos travaux futurs, notre but est de concevoir et implémenter un modèle de *middleware* basé sur le principe des *dataspaces*. Nous nous attachons dans un premier temps à caractériser les types de *mappings* et d'étudier l'approche de réfutation des mappings. Il s'agit de concevoir un framework intermédiaire permettant l'accès et la manipulation de données à un niveau global, sur un ensemble de sources disponibles à un niveau local plus particulièrement au niveau de l'autonomie des sources et des droits d'accès en se basant sur le principe des *homeviews* (Geambasu *et al.,* 2007).

7. Bibliographie

An Y., Borgida A., Miller R. J., Mylopoulos J., « A Semantic Approach to Discovering Schema Mapping Expressions », *ICDE*, p. 206-215, 2007.

Arenas M., Kantere V., Kementsietsidis A., Kiringa I., Miller R., Mylopoulos J., « The Hyperion Project : From Data Integration to Data Coordination », 2003.

Dittrich J.-P., Salles M. A. V., Kossmann D., Blunschi L., « iMeMex : escapes from the personal information jungle », *VLDB '05 : Proceedings of the 31st international conference on Very large data bases*, VLDB Endowment, p. 1306-1309, 2005.

Dong X. L., Cai Y., , Halevy A., Liu J. M., Madhavan J., « Personal information management with SEMEX », *SIGMOD '05 : Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ACM Press, New York, NY, USA, p. 921-923, 2005.

Franklin M., Halevy A., Maier D., « From databases to dataspaces : a new abstraction for information management », *SIGMOD Rec.*, vol. 34, n° 4, p. 27-33, December, 2005.

Geambasu R., Balazinska M., Gribble S. D., Levy H. M., « Homeviews : peer-to-peer middleware for personal data sharing applications. », *in* , C. Y. Chan, , B. C. Ooi, , A. Zhou (eds), *SIGMOD Conference*, ACM, p. 235-246, 2007.

Halevy A., Franklin M., Maier D., « Principles of dataspace systems », *PODS '06 : Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM Press, New York, NY, USA, p. 1-9, 2006.

Jagatheesan A., Moore R., Paton N. W., Watson P., « Grid data management systems & services », *vldb'2003 : Proceedings of the 29th international conference on Very large data bases*, VLDB Endowment, p. 1150-1150, 2003.

Le Deuff O., « Folksonomies : Les usagers indexent le web », *BBF*, n° 4, p. 66-70, 2006.

Li G., Ooi B. C., Yu B., Zhou L., « Schema Mapping in P2P Networks Based on Classification and Probing. », *in* , K. Ramamohanarao, , P. R. Krishna, , M. K. Mohania, , E. Nantajeewarawat (eds), *DASFAA*, vol. 4443 of *Lecture Notes in Computer Science*, Springer, p. 688-702, 2007.

Mathes A., « Folksonomies — Cooperative Classification and Communication Through Shared Metadata », *Computer Mediated Communication*, 2004.

Mikroyannidis A., « Toward a Social Semantic Web », *Computer*, vol. 40, n° 11, p. 113-115, 2007.

Risch T., Koparanova M., Thide B., « Efficient query reformulation in peer data management systems », *Workshop on Distributed Data et Structures - WDAS-2002, March 20-23, 2002*, University Paris 9 Dauphine, 2002.

Rousset M.-C., « Small Can Be Beautiful in the Semantic Web », *in* , F. v. H. S. McIlraith, D. Plexousakis (ed.), *Third International Semantic Web Conference*, vol. 3298, Springer (LNCS), p. 6-16, 2004.

Tatarinov I., Halevy A., « Efficient query reformulation in peer data management systems », *SIGMOD '04 : Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, ACM, New York, NY, USA, p. 539-550, 2004.

Tatarinov I., Ives Z., amd J., Halevy A., Suciu D., Dalvi N., Dong X., Kadiyaska Y., Miklau G., Mork P., « The Piazza Peer Data Management Project », 2003.

Tzitzikas Y., Meghini C., Spyratos N., « Taxonomy-Based Conceptual Modeling for Peer-to-Peer Networks. », *in* , I.-Y. Song, , S. W. Liddle, , T. W. Ling, , P. Scheuermann (eds), *ER*, vol. 2813 of *Lecture Notes in Computer Science*, Springer, p. 446-460, 2003.

Ullman J. D., « Information integration using logical views », *Theor. Comput. Sci.*, vol. 239, n° 2, p. 189-210, 2000.

Yao K.-T., Wagenbreth G., « Simulation Data Grid : Joint Experimentation Data Management and Analysis », *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), November 28 to December 1*, at the Orange County Convention Center, Orlando, FL, USA., 2005.