Intégrer les LOD dans un cube de données : Transformons une action technique en valeur organisationnelle

Selma Khouri*, Ladjel Bellatreche** Abdessamed Réda Ghomari * Yasmine Aouimer*,***

*École nationale Supérieure d'Informatique, Algérie
(s_khouri, ey_aouimeur, a_ghomari)@esi.dz
http://www.esi.dz
**ISAE/ENSMA, Poitiers, France
bellatreche@ensma.fr
***Université Paris Lumière
Paris, France

Résumé. Avec la multiplication des sources LOD (Linked Open Data ou données ouvertes et liées), les organisations ont vu l'opportunité d'étendre leurs cubes de données internes avec ce type de données externes. Les sources LOD sont certes porteuses d'une réelle valeur ajoutée, mais elles ont apporté leur lot de difficultés techniques liées à la gestion de ces données (flux ETL, volume des données, qualité de la source, etc.). Nous constatons dans la littérature existante, que la décision d'incorporer les LOD dans un cube se concentre sur la gestion de ces difficultés techniques, alors que cette décision est avant tout 'organisationnelle'. Cette omission s'explique par le fait que l'intégration des LOD est étudiée selon une vision orientée sources, alors qu'une vision orientée besoins serait plus adéquate pour étudier efficacement la pertinence de l'intégration de cette nouvelle source. Dans cet article, nous apportons une vision organisationnelle orientée buts dans laquelle les indicateurs de performance clés (key Performance Indicators -KPI-) sont l'élément central liant les sources de données aux objectifs organisationnels. Nous proposons une approche orientée buts permettant d'estimer la valeur organisationnelle (via les valeurs des KPI) des ressources LOD dans le cube. L'approche est basée sur un méta-modèle alignant les trois modèles: le modèle de données, le modèle de buts et le modèle KPI. Des expérimentations sont menées pour valider notre approche.

1 Introduction

Réduire le fossé entre la vision des décideurs des politiques organisationnelles (policy makers) et la vision des ingénieurs de données (préparateur de données, administrateur, concepteur, intégrateur) dans une organisation a toujours été considéré comme un problème central lié à la conception des systèmes de gestion de données. Cette problématique a permis de grandes évolutions. Par exemple, pour la conception des bases de données, le modèle E/A est né de la volonté de son inventeur (P. Chen) de fournir un modèle plus abstrait que les modèles physiques et logiques existants compréhensibles uniquement par les développeurs, qui puissent

être assimilé aussi par les gestionnaires (Chen (2002)). Dans le domaine du génie logiciel, des langages de modélisation tels que UML ont été adoptés pour assurer la communication entre les les gestionnaires, les concepteurs de bases de données et les développeurs d'applications afin de partager une vision unifiée des besoins. Dans le domaine de l'entreposage des données (ED) également, les premières problématiques traitées ont été tournées vers les données en termes de modélisation et aussi d'exploitation (optimisation, tuning, etc.). Ces travaux ont ensuite été complétés par une vision orientée vers les besoins et buts des décideurs, ce qui a donné lieu à l'*approche de conception dirigée par les besoins*. De nombreuses problématiques liées aux besoins des ED ont été développées durant cette dernière décennie et que nous avons longuement étudié comme l'identification des besoins, leur modélisation ou leur intégration Djilani et al. (2017). Dans cette vision orientée besoins, l'évaluation du processus de prise de décision est réalisée par l'évaluation de la réalisation des buts définis. Cette évaluation est cruciale pour mesurer le succès d'une politique organisationnelle, elle est réalisée en analysant des métriques nommées indicateurs de performance clés (KPI) qu'il faut identifier pour chaque but défini (Silva Souza et al. (2012)). Les valeurs de performance des KPI sont calculées à partir des sources de données, faisant ainsi des KPI le maillon liant ente les sources de données à considérer et les buts et objectifs à atteindre.

Du point de vue des sources de données, les ED et leurs cubes associés ont évolué en prenant en compte différentes générations de sources de données (relationnelles, XML, Nosql, etc.). Ces dernières années, les technologies du Web sémantique ont intégré l'espace des ED et une nouvelle architecture de cube sémantique a vu le jour, définie comme un référentiel de données sémantiquement intégrées (Deb Nath et al. (2015)). La popularité croissante des initiatives portant sur l'ouverture des données incite les organisations à publier leurs données sous forme de données liées définies en RDF [1] (Deb Nath et al. (2015)), le langage standard du W3C pour l'échange de données. Ces nouvelles sources appelées LOD ont rapidement été perçues comme porteuse de valeur et pouvant générer des connaissances métiers nécessaires pour la conception des cubes de données d'une organisation. Différentes nouvelles approches sont proposées dans cette optique (Gallinucci et al. (2018); Baldacci et al. (2017); Abelló Gamazo et al. (2016); Matei et al. (2014)). Ces études abordent la question de l'intégration des sources externes LOD selon une vision orientée sources, précisément parce que ces travaux ont eu à traiter les difficultés techniques liées à cette nouvelle source de données. Cette complexité est présente dans différentes tâches de conception telles que l'intégration de la source (par exemple, le formalisme de graphe utilisé dans les sources LOD), la vérification continue de la disponibilité des sources externes, la vérification de leur véracité, le suivi de leur évolution, etc. Cependant, ces études ont négligé la dimension organisationnelle liée à la décision d'intégrer ces sources externes. Notre proposition appelle à un retour à la philosophie de l'approche dirigée par les besoins, visant à établir un lien entre les décideurs des politiques de l'organisation et la vision du gestionnaire des données. Converger ces deux visions peut être bidirectionnel : les gestionnaires de données peuvent éclairer certaines décisions stratégiques en faisant appel aux données externes, et vice versa, les décideurs peuvent orienter la décision d'aller vers les sources externes pour répondre à des besoins précis.

Si la décision d'intégrer des sources externes est connectée au niveau organisationnel, le concepteur aura des indicateurs pertinents pour identifier la valeur ajoutée de ces sources pour l'ED. Nous formulons la problématique traitée dans cet article comme suit : (a) considérant

1. https://www.w3.org/RDF/

une hiérarchie de buts, (b) leurs KPI associés, (c) le schéma d'un cube multidimensionnel sémantique (l'existence de ce schéma est une hypothèse formulée par différents travaux traitant de l'intégration de données internes et externes LOD (Deb Nath et al. (2015))), notre étude vise à fournir une approche orientée besoins permettant d'estimer *la valeur organisationnelle de l'incorporation technique des LOD dans un cube*. Dans notre vision, la valeur organisationnelle est reflétée par la performance estimée de chaque but défini et son impact sur le but global. Nous détaillerons les techniques utilisées pour estimer cette performance.
Cette notion de valeur, popularisée grâce à la mouvance du déluge de données (un des V du big Data) est étudiée intensément dans de nombreuses recherches récentes dans le contexte de système d'informations Princea et al. (2017) ou particulièrement dans le contexte des ED. Par exemple, Berkani et al. (2019) proposent de calculer la valeur ajoutée d'une conception impliquant les LOD mais cette valeur est strictement orientée sources (principalement basée sur le nombre de concepts et le nombre d'instances ajoutées). Nous proposons dans un premier temps, un méta-modèle qui aligne les trois niveaux : modèle de buts, modèle KPI, le modèle global de données qui englobe : les schémas des sources, le schéma cible du cube et les flux de données entre eux. La décision d'intégrer un fragment LOD est exprimée à travers un but défini (les buts sont considérées de façon incrémentale), la performance du KPI associé à ce but est calculée à partir des sources de données (internes et externes LOD), ensuite propagée à travers la hiérarchie des buts pour calculer le degré de satisfaction de l'objectif global établi. Dans cette approche, les KPI deviennent le pivot liant la vision orientée sources de données, et la vision orientée vers les buts et la stratégie de l'organisation. Lorsque les KPI nécessitent des ressources internes et externes, la difficulté revient à calculer leurs performances sont des schémas au formats hétérogènes (par exemple des sources internes relationnelles et des sources LOD ayant le format RDF). Une principale contribution a été d'aligner la spécification des KPI au niveau conceptuel ontologique du cube de données pour gérer l'hétérogénéité des sources. Nous avons ensuite analysé et établit un lien entre les grammaires en langage naturel proposées pour la spécification des KPIs, et les langages du Web sémantique exploitables sur le cube. Des expérimentations sont conduites en utilisant le banc d'essai Berlin SPARQL Benchmark (BSBM [2]) qui construit un cas de business intelligence autour d'un système de commerce électronique. Le document est structuré comme suit : la section 2 présente un exemple de motivation. La section 3 décrit les travaux connexes. La section 4 présente le background nécessaire décrivant les notions de buts et KPI. La section 5 présente l'approche proposée. La section 6 présente les expérimentations menées. La section 7 conclut l'article.

2 Exemple de motivation

L'exemple de motivation est illustré par l'étude de cas utilisée pour nos expérimentations (cf. section 6). Parmi les quelques bancs d'essai sémantiques existants, nous avons opté pour le banc d'essai BSBM qui définit un scénario de Business Intelligence construit autour d'un système de commerce électronique, dans lequel un ensemble de produits est proposé par différents fournisseurs et les clients postent des avis sur les produits. Ce banc d'essai utilise trois ressources LOD : Friend of a Friend (FOAF [3]) qui décrit les personnes et les relations sociales

2. http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/spec/BusinessIntelligenceUseCase/index.html#bimetrics
3. http://xmlns.com/foaf/0.1/

sur le Web et RDF Review Vocabulary (REV [4]) qui fournit un vocabulaire pour exprimer des commentaires et des évaluations. La troisième ressource est géographique et décrit la liste des pays [5], mais comme cette ressource n'est plus accessible, nous l'avons remplacé par un fragment de Dbpedia [6]. Notre étude de cas considère les requêtes analytiques fournies par le banc d'essai [7] comme les besoins des décideurs (les buts), que nous avons complétés par des KPI associés. Nous avons considéré ces buts comme les buts feuilles de l'arbre des buts, nous avons ajouté 6 nouveaux buts (14 buts au total) pour construire l'arbre reflétant la hiérarchie des buts, comme illustré dans la figure 1 (en anglais car les banc d'essai est décrit en anglais). L'arbre des buts présenté dans cette figure est modélisé en utilisant l'outil Constraint Goal Model Tool (CGM-Tool [8]). Des exemples de ces buts identifiés sont les suivants : (1) Augmentez le nombre de catégories de produits ayant le plus de popularité. (2) Diminuer le nombre d'évaluateurs spammeurs (qui donne une évaluation des produits plus élevée que la moyenne).

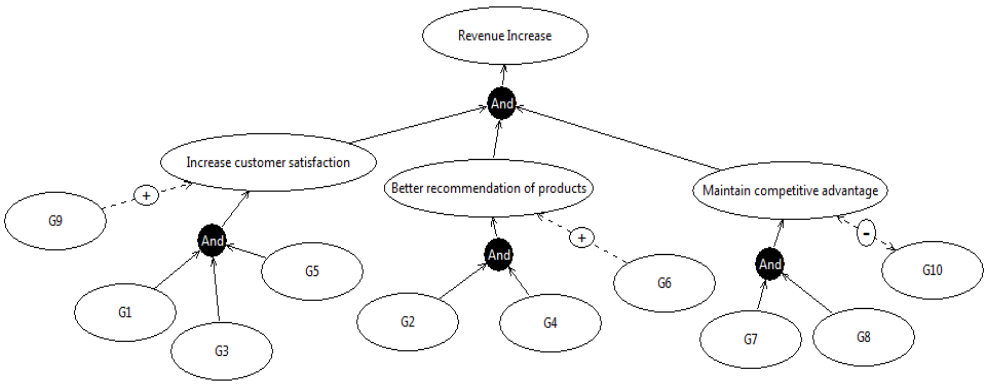

FIG. 1 – *Hierarchies des buts du banc d'essai BSBM*

L'ensemble des buts se trouvant comme feuille dans l'arbre sont considérés de façon incrémentale, reflétant ainsi une situation réelle dans laquelle une organisation doit intégrer de nouveaux besoins pouvant nécessiter des ressources externes. L'objectif de notre approche est de fournir des indicateurs permettant de juger de la pertinence d'intégrer les ressources externes identifiées. L'approche est basée sur deux contributions : (1) un méta-modèle offrant une vision orientée besoins en permettant de lier l'ensemble des buts, concepts internes et externes et flux de données, (2) une approche orientée buts permettant de calculer la valeur organisationnelle indiquant le niveau de performance des buts lors de l'introduction du besoin ayant requis la ressource externe. Le niveau de performance peut prendre différentes valeurs allant du but complètement non performant au but complètement performant en passant par les buts partiellement performants. Cette valeur de performance est calculée à partir du schéma sémantique de BSBM décrivant les sources de données.

4. http://purl.org/stuff/rev#
5. http://downlode.org/rdf/iso-3166/countries
6. http://fr.dbpedia.org/ontology/Country
7. http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/spec/BusinessIntelligenceUseCase/index.html#queriesTriple
8. http://www.cgm-tool.eu/

3 Travaux connexes

Différentes études récentes ont été consacrées à l'intégration des LOD dans des cubes de données. Ces travaux se sont naturellement focalisés sur les problématiques techniques liées à cette nouvelle source de données, que nous pouvons classer comme suit :

(i) La définition du schéma du cube sémantique : les travaux existants se basent sur l'existence d'un schéma commun unifiant les schémas de sources internes et externes, ce schéma est soit supposé existant (Abelló Gamazo et al. (2016)) ou construit par des tables de correspondance utilisant des mesures de similarité (Ravat et al. (2016); Deb Nath et al. (2015)) ou des règles de mapping (Etcheverry et al. (2014)). D'autres études portent sur l'identification du rôle multidimensionnel des nouvelles ressources externes à intégrer (Abelló Gamazo et al. (2016); Gallinucci et al. (2018)).

(ii) Le processus d'intégration des flux internes et externes : d'autres travaux portent sur la définition des processus ETL utilisés pour unifier les flux issus des sources internes et externes. Ces approches proposent des techniques de mappings entre les données internes et externes (Deb Nath et al. (2015)), une extraction et chargement des fragments externes à la demande (Baldacci et al. (2017)), ou une interrogation des données externes par des opérateurs OLAP pour leur chargement dans le cube (Kämpgen et Harth (2011)).

(iii) Le processus d'interrogation : d'autres travaux portent principalement sur la définition d'opérateurs OLAP adaptés aux données LOD (Matei et al. (2014); Saad et al. (2013); Kämpgen et al. (2012)).

Contrairement aux études existantes, notre proposition reconnecte l'incorporation des sources de données externes aux buts et objectifs qui dirigent la conception du cube. Même si la notion de valeur des LOD est implicite dans ces études, elle n'est pas mesurée au niveau organisationnel. Notons d'une part, que quelques études connectent les données aux buts organisationnels et leurs KPI d'un cube de données (Nasiri et al. (2017)) mais ces études sont faites dans un contexte de sources internes et pour un modèle relationnel. D'autre part, nous avons proposé dans Khouri et al. (2019) une approche permettant d'estimer le compromis entre la valeur managériale des LOD et les actions techniques requises. Notre étude dans cet article pose les fondements pour un retour à une approche orienté besoins, elle permet le calcul de la valeur organisationnelle des sources internes et externes en fournissant l'algorithme de transformation des KPI en requêtes Sparql, qui est la pièce maîtresse pour le calcul de la valeur organisationnelle.

4 Notions de base et définitions : Background

Les buts. Les buts représentent le moyen le plus naturel d'exprimer les exigences d'un entrepôt. Un but représente l'état souhaité, défini lors de la planification stratégique et poursuivi lors de l'exploitation du projet (Barone et al. (2011)). Les buts sont présentés sous forme de hiérarchies arborescentes utilisant des relations ET/OU entre un but et ses sous buts afin de refléter la situation stratégique d'une organisation (figure 2). Dans la hiérarchie des buts, la satisfaction d'un but dépend donc de la satisfaction de ses sous-buts. Les relations d'influence (également appelées relations de contribution) peuvent être définies entre les buts, signifiant que la satisfaction d'un but peut être affectée de manière positive ou négative par des buts autres que ses sous-buts.

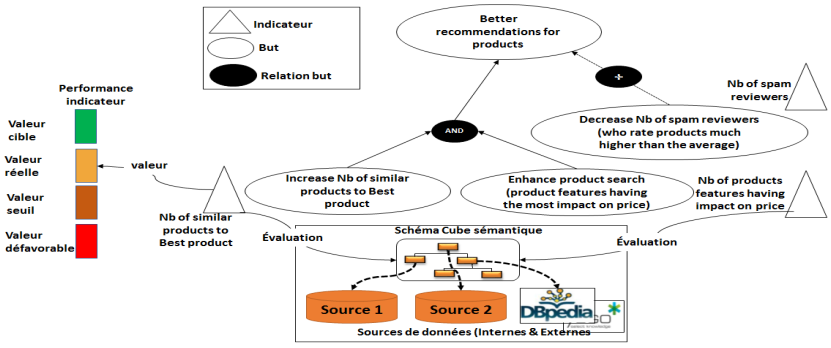

FIG. 2 – *La hierarchie des buts - la hiérarchie des KPI.*

Les indicateurs. Un indicateur clé de performance (KPI) représente une mesure évaluant la performance par rapport à un but établi (Barone et al. (2011)). De la même façon que pour les buts, les KPI sont associés aux buts et sont donc organisés en une hiérarhcie arborescente similaire à celles des buts (figure 2). Par exemple, l'indicateur "Nombre de produits appartenant aux catégories ayant reçu plus de 50 avis positifs" permet de mesurer le but du cas BSBM "Augmenter le nombre de produits pour les catégories les plus populaires". Dans notre étude, nous supposons que les buts et les choix de leurs KPI sont bien définis, en nous basant sur les techniques proposées dans la littérature (Maté et al. (2017b)).

5 Approche proposée

L'approche proposée est illustrée dans la figure 3. Elle considère les buts de façon incrémentale, et est structurée en deux étapes principales : (1) connecter l'incorporation des sources externes LOD (requises par le but) au niveau organisationnel, (2) approche orientée buts pour l'incorporation des LOD dans le cube sémantique. La première étape fournit le modèle global et montre que la décision d'incorporer les LOD est une décision organisationnelle. La deuxième étape détaille l'approche à mener pour mesurer la valeur organisationnelle de l'introduction des ressources externes sur le cube.

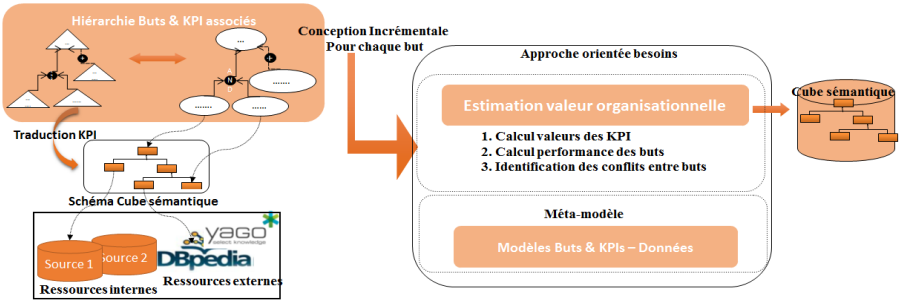

FIG. 3 – *L'approche générale proposée.*

5.1 Connecter l'introduction des LOD au niveau organisationnel

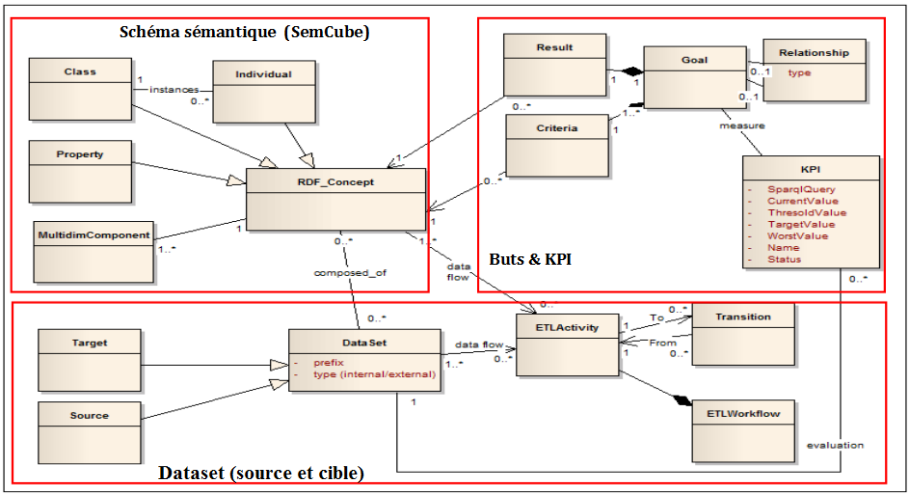

FIG. 4 – *Connecter les LOD au niveau organisationnel.*

Le modèle proposé est illustré dans la figure 4. Notre approche étant orientée besoins, le modèle détaille la représentation des buts. Notre analyse des modèles de buts existants fournit deux principales coordonnées à chaque but : le résultat du but et les critères (i.e. les dimensions d'analyse). Le modèle de buts est composé des classes But, la classe Relationship (relations entre les buts qui sont de type ET/OU ou relation d'influence), ainsi que les classes représentant les coordonnées des buts. Nous avons fourni dans des études précédentes (Khouri et al. (2015)) une analyse détaillée du cycle de conception d'un cube sémantique, ce qui nous permet d'avoir une vue globale des corrélations entre les différents artefacts de conception utilisés, ceux orientés besoins et ceux orientés sources de données. Ainsi, les coordonnées des buts sont exprimés et permettent d'identifier les concepts du schéma cible de l'ED. Ces concepts sont alimentés par les sources de données. Nous obtenons de ce fait, le lien entre les sources (internes et/ou externes) et les buts de l'organisation.

Pour le modèle des sources de données, nous avons considéré le modèle sémantique RDF qui est le langage de partage utilisé pour les LOD. Le modèle de données (sources et cible du cube) est composé du Concept (classe, propriété et instance), une annotation multidimensionnelle (pour le schéma cible), et le dataset auquel appartient le concept. Dans notre étude, nous ne traitons pas l'unification des schémas des sources internes qui n'est pas l'objectif de l'étude, mais plusieurs études démontrent que dans le contexte d'intégration de données sémantiques LOD dans un cube, le schéma pivot considéré est un schéma sémantique (appelé *SemCube*) supposé existant ou construit pour unifier les sources (Abelló Gamazo et al. (2016); Deb Nath et al. (2015); Berkani et al. (2019)). Ce schéma permet d'une part de faciliter la définition des flux de données, i.e les mappings permettant l'intégration des sources de données vers le schéma cube cible du cube (classes ETLActivity et ETLWorflow). D'autre part, c'est ce schéma sémantique que nous utilisons pour évaluer les performances des buts définis. Ces performances sont évalués grâce à l'évaluation des valeurs des KPIs de chaque but. Le

modèle KPI est inspiré de (Maté et al. (2017a)), il est composé de la classe KPI caractérisée par le nom de l'indicateur, son statut (performant ou pas) et ses valeurs (réelle, seuil, cible et défavorable). La classe KPI est également caractérisée par la requête Sparql correspondante, nous verrons dans la suite du document que la valeur réelle du KPI doit être évaluée sur le schéma sémantique du cube, et nécessite par conséquent la traduction de la spécification du KPI en requête Sparql. L'objectif de l'évaluation des KPI au niveau sémantique est d'obtenir des informations sur la pertinence de considérer les ressources internes et externes, avant de les considérer effectivement au niveau physique dans le cube. Un autre avantage est que ces informations permettent au concepteur de décider d'une stratégie de matérialisation adéquate en stockant les ressources externes dans le cube ou en gardant uniquement trace de ces ressources et des mappings nécessaires pour les interroger au besoin.

5.1.1 Considération des buts dans le méta-modèle

Notre approche commence par exploiter le modèle décrit ci-dessus, où la définition des buts sur le schéma sémantique *SemCube* permet d'identifier les fragments de sources internes et externes requis. *SemCube* est ainsi enrichi dès l'introduction d'un nouveau besoin, éventuellement par de nouvelles ressources du schéma cible de l'ED et des ressources internes et/ou externes alimentant la cible. Chaque dataset est décrit dans *SemCube* par un identifiant URI unique. En utilisant les KPI, il devient possible de mesurer pour ce nouveau besoin la valeur organisationnelle de chaque fragment de source nécessaire pour ce but. Ainsi, si les premiers besoins qui ont donné lieu au cube initial (avant le recours au LOD), on peut calculer la valeur initiale du cube. Comme notre approche est incrémentale, elle permet le calcul de la valeur ajoutée organisationnelle pour chaque nouveau besoin nécessitant des LOD, ce qui constitue un aspect essentiel de notre approche.

5.2 Approche orientée buts pour l'incorporation des LOD dans le cube sémantique

Les étapes de notre approche peuvent être résumées dans l'algorithme ci-dessous. L'algorithme prend en entrée : (1) le schéma du cube sémantique (définissant le schéma interne, le schéma externe et le schéma cible), (2) l'ensemble des buts et KPI associés. L'objectif est de pouvoir estimer la valeur organisationnelle des LOD requis par chaque besoin.

Entrées : Hiérarchie de buts (But) et leurs indicateurs (KPI), schéma sémantique du cube (*SemCube*) - Ressources externes
Résultat : Valeur organisationnelle des LOD

for all $B_i \in$ Buts (et son KPI) **do**
  Ajouter le but dans la hiérarchie ▷ Ajout au niveau du méta-modèle, indiqué en section *cf. section* 5.1.1
  Identifier les concepts définissant le but
  Formaliser KPI ▷ Gestion organisationnelle, cf. section 5.2.1
  Traduire KPI en une requête Sparql sur *SemCube* ▷ cf. section 5.2.2
  Calculer la valeur organisationnelle du cube ▷ cf. section 5.2.3
  Identifier les buts conflictuels ▷ cf. section 5.2.4
end for

La valeur organisationnelle est mesurée grâce aux KPI. Cependant, il y a un écart important entre la spécification *naturelle* d'un KPI défini par un décideur et les langages sémantiques exploitables sur le cube, que nous proposons de gérer comme décrit ci-dessous.

5.2.1 Spécification des KPI

Dans (Maté et al. (2017b)), les auteurs ont proposé une spécification en langage naturel structuré, utilisée par les décideurs pour définir les KPI, que nous utilisons dans notre approche. Prenons l'exemple suivant "nombre total des avis sur un produit en 2018" illustrant est une expression d'un KPI. La spécification définit une expression KPI comme étant des valeurs simples ou complexes obtenues en appliquant des opérateurs successifs. Chaque valeur de KPI est calculée en appliquant un opérateur unaire ("nombre" dans l'exemple précédent) à une valeur (par exemple, "Avis") sur des niveaux de dimension ("Produit", "Année"). Les valeurs peuvent être contraintes de respecter certaines conditions booléennes indépendamment de l'ensemble des niveaux (par exemple, "année = 2018").

5.2.2 Traduction des KPIs en requêtes Sparql

La spécification précédente est proposée pour faciliter la spécification des KPI par les décideurs. Pour faciliter la traduction de cette spécification (qui est en langage naturel) en un langage formel tel que Sparql exploitable sur *SemCube*, nous proposons de passer par un langage intermédiaire.
Une revue de litérature nous a permis de choisir Cube Query Language (CQL) comme langage intermédiaire pour la traduction des KPI en requêtes Sparql. CQL est proposé dans (Ciferri et al. (2013)) en tant que langage *conceptuel* pour interroger des cubes de données. Ce langage couvre la spécification précédente et peut être traduit en langage sémantique. Il est à noter que cette proposition a été motivée par l'absence d'une conceptualisation de manipulation d'un cube qui soit proche du langage d'un décideur malgré la popularité des cubes et des opérateurs OLAP, qui sont vus comme des définitions du niveau logique (et non pas conceptuel). Nous commençons par présenter brièvement CQL. CQL fournit un ensemble de cinq (5) principaux opérateurs : Roll-up, Drill-down, Slice, Dice, Map et Drill-across. Chaque opérateur a une signature précise décrite en fonction des ensembles suivants : l'ensemble de tous les cuboïdes dans une instance de cube C, l'ensemble des dimensions D, l'ensemble des mesures M, l'ensemble des niveaux de dimension N et l'ensemble des expressions booléennes sur les mesures et les dimensions. Par exemple, la dimension REVIEW du cube du banc d'essai BSBM utilisé est définie en CQL comme suit :

```
CUBE BSBM {
DIMENSION REVIEW
LEVEL Review ATTRIBUTES {title(String), Text (String), Rating (int)}
LEVEL Category ATTRIBUTES {category (String)}
Review ROLL-UP to Category}
```

Le premier but de BSBM est défini comme suit : "*Augmenter les catégories de produits les plus populaires d'un pays spécifique (sur la base du nombre d'avis, i.e. review)*". Le cas d'utilisation est motivé par le fait que le fournisseur souhaite savoir quelles catégories de produits d'un pays spécifique retiennent le plus l'attention des usagers (le nombre d'avis est supérieur à un seuil) et l'objectif est d'augmenter le nombre de ces catégories de produits. Le KPI correspondant selon la spécification décrite est "le nombre de catégories de produits les plus discutées (nombre avis $>$ 20) provenant d'un pays spécifique". Ce KPI est exprimé en CQL comme suit :

```
C1 <- Slice(SemCube, Vendor)
C2 <- Slice(C1, Offer)
C3 <- Slice(C2, Time)
```

```
C4 <- Dice (C3, country='GB')
C5 <- RollUp(C4, (Product -> ProductType, Person  -> Producer,
Region -> Country, Review -> All), Count(Review))
C6 <- Dice (C5, Count(Review)>20)
```

Dans (Etcheverry et Vaisman (2016)), les auteurs fournissent un processus de traduction d'expressions CQL en Sparql sur un cube QB4OLAP (qui est un vocabulaire destiné à publier des données multidimensionnelles sous forme LOD sur le web). Notre approche a nécessité une réadaptation de ce processus et nous avons proposé un nouvel algorithme de traduction. Afin que l'algorithme ne soit pas dépendant d'un langage particulier (RDF, QB2OLAP, etc), nous avons préféré le définir en fonction du graphe construisant la requête, qui est une représentation générique. Par exemple, La requête Sparql correspondante à l'exemple précédent est la suivante :

```
Select ?productType ?reviewCount
{ { Select ?productType (count(?review) As ?reviewCount)
  {   ?productType a bsbm:ProductType .    ?product a ?productType .
   ?product bsbm:producer ?producer .    ?producer bsbm:country GB.
   ?review bsbm:reviewFor ?product .
  }  Group By ?productType
 } filter (?reviewCount>20)
 }Order By desc(?reviewCount) ?productType
```

Rappelons qu'une requête Sparql est définie comme un triplet <*type* , *Graph Pattern P*, *modificateurs*> (Bonifati et al. (2017)), le *type* dans notre cas est toujours un SELECT, les *modificateurs* représentent les agrégations, les conditions de filtrage des données, etc. Le *Graph Pattern* est le coeur de la requête et représente tous les concepts joints sous forme de triplets (formant un graphe) pour définir la requête. L'algorithme ci-dessous décrit le processus de traduction des KPI.

Inputs : Inputs : requête CQL, *Output* : Requête Sparql Q
1: Type Q := SELECT (fonctions d'agrégation du RollUp)
2: Modificateur Q := ajouter à Q des expressions filter (Opérateurs Dice), ajouter à Q un Group by (première dimension du RollUp)
▷ Construire la graph-pattern de Q
3: Construire un graphe prenant tout les éléments du RollUp et Drill-cross
4: Ajouter tous les arcs possibles entre les noeuds du graphe selon *SemCube*
5: Eliminer du graphe les noeuds indiqués dans les opérateurs Slice
6: Ajouter les graph pattern (rdf :type) pour récupérer les classes des variables du noeuds du Select
7: Ajouter à Q un pattern correspondant à chaque couple de noeuds restants dans le graphe.
8: Ajouter les filtres indiqués dans le Dice à la clause Filter de la requête

L'algorithme repose sur la création d'une template d'une requête Sparql de ype SELECT définissant une fonction d'agrégation et ayant les clauses Filter et Group by. Cette dernière est définie selon la dimension définie dans l'opérateur ROllUp. Toutes les fonctions d'agrégation définies dans la requête CQL sont ajoutées dans la requête Sparql. Le graph pattern de la requête est le plus délicat à construire. Il repose sur l'identification des concepts RDF utilisés dans les opérateurs RollUp et Drill-cross. Le graphe RDF correspondant à ces concepts est extraits à partir du schéma SemCube en utilisant une technique de modularité. Ces statements liés par des jointures vont constituer le graph pattern de la requête Sparql. Les concepts indiqués dans les opérateurs Slice sont éliminés du graphe pattern. Les filtres indiqués dans le Dice sont ajoutés à la clause Filter de la requête.

5.2.3 Calcul de la valeur organisationnelle des sources

Chaque indicateur a une valeur réelle calculée, elle est évaluée (comme performante ou non performante) à l'aide d'un ensemble de paramètres : valeur cible, valeur seuil et valeur défavorable (Barone et al. (2011); Horkoff et al. (2014)). La valeur réelle des indicateurs se trouvant comme feuilles dans l'arbre des indicateurs, est extraite de l'analyse de données (internes et externes). La valeur d'un indicateur de niveau supérieur dépend des valeurs des indicateurs de niveau inférieur dans la hiérarchie (Barone et al. (2011)) ; ceci se fait par une approche de propagation quantitative ou qualitative. Trois principales techniques de propagation sont identifiées (Barone et al. (2011); Horkoff et al. (2014)) : les facteurs de conversion, la normalisation et le raisonnement qualitatif. Les deux premières techniques sont quantitatives et nécessitent la définition d'expressions métriques qui sont définies sur la base de données historiques et/ou des connaissances de domaine. Nous notons que les trois techniques peuvent être utilisées dans notre approche, le choix revient au concepteur selon la disponibilité des informations sur les expressions métriques. Nous optons donc la troisième catégorie de raisonnement qualitatif qui s'applique même en cas d'indisponibilité des expressions métriques.

Le raisonnement qualitatif repose sur deux variables : la performance positive (per+) et la performance négative (per-) attribuées à chaque indicateur, ces deux variables prenent les valeurs suivantes : ("complet", "partiel", "néant"). Ces valeurs sont attribuées conformément aux règles de mappage définies dans (Horkoff et al. (2014)) et détaillées comme suit. En supposant que (vc=valeur cible, vr=valeur réelle, vs=valeur seuil, vd= valeur défavorable) :
(1) Si vr $\geq$ vc Alors les variables de l'indicateur (per+,per-)=(complet, néant) Et l'indicateur est considéré comme *Complètement performant*
(2) Si vs $<$ vr $<$ vc Alors (per+,per-)=(partiel, néant) Et l'indicateur est considéré comme *Partiellement performant*
(3) Si vr $=$ vs Alors (per+,per-)=(partiel, néant) Et l'indicateur est considéré comme *Partiellement performant*
(4) Si vd $<$ vr $<$ vs Alors (per+,per-)=(néant, partiel) Et l'indicateur est considéré comme *Partiellement non-performant*
(5) Si vr $\leq$ vd Alors (per+,per-)=(néant,complet) Et l'indicateur est considéré comme *Complètement non-performant*

La propagation des valeurs des indicateurs de niveau inférieur vers les indicateurs d'un niveau supérieur est basée sur des règles de propagation définies dans dans la table TAB. 1 (Horkoff et al. (2014)). Dans cette table, les valeurs (S) représentent des relations d'influence entre les buts (S+ pour une influence positive et S- pour une influence négative), et le cas des relations (OU) sont duales aux relations (ET). Cette table est utilisée comme suit. Les valeurs (per+) et (per-) des KPI de niveau supérieur dans l'arbre des KPI (première colonne du tableau) sont calculées en fonction des valeurs des KPI de niveaux inférieur (fonction min ou max de leurs valeurs), et aussi selon les relations (ET/OU ou influence) les liant.

5.2.4 Identification des conflits entre les buts

Les conflits peuvent être identifiés lorsqu'un indicateur d'un but B1 prend comme valeur "performant" à partir d'un but B2 dans l'arbre des buts, et prend une valeur contradictoire (eg. "non-performant") à partir d'un autre but B3 aussi en relation avec B1. De telles informations

	$(I_2, I_3) \xrightarrow{ET} I_1$	$I_2 \xrightarrow{+S} I_1$	$I_2 \xrightarrow{-S} I_1$	$I_2 \xrightarrow{++S} I_1$	$I_2 \xrightarrow{--S} I_1$
per+ (I_1)	min(per+(I_2), per+(I_3))	min (per+(I_2), P)	N	per+(I_2)	N
Per- (I_1)	max(per-(I_2), per-(I_3))	N	min (per+(I_2), P)	N	per+(I_2)

TAB. 1 – *Propagation des valeurs de performance des indicateurs (I est un Indicateur, N : Néant, P : Partiel)*

sont importantes pour le concepteur lorsqu'un but conflictuel est détecté par l'introduction d'une ressource externe.

6 Etude de cas et expérimentations

Notre étude de cas utilise Berlin SPARQL Benchmark (BSBM). Comme illustré dans la section 2, nous avons considéré les questions analytiques comme des buts feuille et nous avons construit la hiérarchie des buts illustrée dans la figure 1. Trois sources LOD sont utilisées : FOAF, REV et la liste des pays de Dbpedia. Nos expérimentations ont été menées en considérant l'ensemble des buts de façon incrémentale, afin d'analyser l'impact de l'ajout de nouveaux buts, nécessitant des ressources internes et/ou externes. Le tableau 2 illustre la performance organisationnelle de la prise en compte des buts nécessitant des sources internes/externes. Les lignes représentent le buts feuilles (10 buts feuilles). Le tableau contient pour chaque but feuille les colonnes suivantes : (1) le KPI identifié, (2) les ressources externes auxquelles se réfère le but, (colonnes 3 à 6) l'ensemble des valeurs : la valeur défavorable, valeur seuil, valeur cible (que nous avons estimée aléatoirement) et les valeurs réelles calculées à partir des sources internes et externes (plus précisément sur *SemCube*). (7 & 8) les variable de performance (Per+ et Per-) calculées sur la base des règles de mappage définies ci-dessus, (9) le résultat des indicateurs des buts feuilles, (10-12) les résultats des KPI des trois buts intermédiaires dans la hiérarchie (R1, R2 & R3) (figure 1) calculés à l'aide des règles de propagation (13) le résultat du KPI du but global (*revenue increase* dans la figure 1) également calculé sur la base des règles de propagation énoncées, (14) l'identification d'un conflit. Les acronymes utilisés dans le tableau sont les suivants : PP (Partiellement Performant), PNP (Partiellement Non Performant), CP (Complètement Performant), CNP(Complètement Non Performant).
Le tableau est obtenu par l'exécution des étapes de l'approche décrite dans la section précédente et montre ainsi la faisabilité de notre proposition. Le résultat du tableau montre au concepteur comment l'incorporation progressive des buts influence la performance du but global qui représente la valeur organisationnelle du cube. Le tableau indique que les deux premiers buts et le but global sont performants. Le but global devient partiellement non performant à partir de l'introduction du but 3, ce but tient compte de la source externe (REV). Ce but requiert l'attention du concepteur pour atteindre le but global, et le concepteur peut décider de matérialiser ce fragment de la source externe jusqu'à ce que le but soit atteint. Les KPI des buts 5, 6 et 8 sont également non performants, nous constatons que parmi ces buts, les buts 5 et 6 nécessitent des ressources externes. Ces fragments externes peuvent être considérés comme complément au cube sémantique. L'analyse de la dernière colonne indique la présence d'un conflit global lors de la prise en compte des buts 6 et 9. Ces deux derniers buts nécessitent des ressources externes (REV et FOAF). Dans notre cas, ces conflits se produisent parce que les buts intermédiaires sont performants lors de la propagation des valeurs des buts

KPIs	Res. externes	vd	vs	vc	vr	Per+	Per-	Résultat But	But R1	But R2	But R3	But Global	conflits
1	REV	1	5	10	9	partiel	néant	PP	PP	/	/	PP	Non
2	/	3	10	15	17	Complet	néant	CP	PP	CP	/	PP	no
3	REV- PURL	10	20	30	11	none	partial	PNP	PNP	CP	/	PNP	non
4	/	20	70	80	82	Complet	néant	CP	PNP	CP	/	PNP	non
5	REV -PURL	10	15	20	11	néant	partiel	PNP	PNP	PNP	/	PNP	no
6	REV	20	5	0	24	néant	Complet	CNP	PNP	conflit	/	PNP	oui
7	REV-COUNTRY	20	60	100	100	Complet	néant	CP	PNP	conflit	CP	PNP	no
8	/	1	5	10	2	néant	partiel	PNP	PNP	conflit	PNP	PNP	no
9	REV -FOAF	10	7	0	7	partiel	néant	PP	conflit	conflit	PNP	PNP	Oui
10	/	50	70	100	75	partiel	néant	PP	conflit	conflit	PNP	PNP	no

TAB. 2 – *Performances des buts de BSBM, vr : valeur rélle, vd : valeur défavorable, vs : valeur seuil, vc : valeur cible*

feuilles et non performants à partir des buts d'influence. Ces conflits indiquent au concepteur de considérer les buts et le fragment externe qui provoque la non performance des buts et leurs conflits jusqu'à la satisfaction de celui-ci. Toutes ces informations (performances des buts et des indicateurs ainsi que les conflits identifiés) fournissent la valeur organisationnelle des buts considérés qui est principalement représentée par la performance du but ultime en haut de la hiérarchie reflétant la stratégie de l'organisation pour le cube de données. Nous avons également développé un outil case destiné au concepteur pour illustrer notre approche. Une démonstration de l'outil est disponible dans le lien `https://drive.google.com/file/d/1b6xPuctvKhGePx9Ts61vsgdVZPlQlGF5/view?usp=sharing`.

7 Conclusion

Dans cet article, nous avons abordé la problématique d'introduction des LOD dans un cube de données selon la vision des décideurs de la politique organisationnelle, alors que les approches existantes fournissent une vision orientée sources qui correspond au point de vue des ingénieurs de données. Nous avons identifié l'élément clé liant ces deux visions : les KPI. Ces derniers permettent de mesurer la performance des buts reflétant cette politique organisationnelle adoptée, et sont évalués à partir des sources de données internes et externes (LOD). Nous avons ainsi proposé une approche orientée buts qui se base sur un méta-modèle liant les buts, leurs KPI aux sources de données internes et externes. L'approche propose un processus permettant la prise en compte incrémentale des buts et identifie la valeur organisationnelle du cube après l'introduction des ressources internes et externes requis par ces buts. L'étape clé de cette approche est le passage d'une spécification de KPI exprimée par les décideurs à une spécification Sparql utilisable sur le cube sémantique. Des expérimentations ont été menées pour valider notre approche. Ce travail doit être complété par la validation de l'algorithme de génération des requêtes Sparql des KPI sur des cas complexes et le développement d'algorithmes automatisant une stratégie de matérialisation selon les indicateurs fournis et selon la provenance des sources LOD.

Références

Abelló Gamazo, A., E. Gallinucci, M. Golfarelli, S. Rizzi Bach, et O. Romero Moral (2016). Towards exploratory olap on linked data. In *SEBD*, pp. 86–93.

Baldacci, L., M. Golfarelli, S. Graziani, et S. Rizzi (2017). QETL : An approach to on-demand etl from non-owned data sources. *DKE 112*, 17–37.

Barone, D., L. Jiang, D. Amyot, et J. Mylopoulos (2011). *Reasoning with Key Performance Indicators*, Volume 92, pp. 82–96. Springer Berlin Heidelberg.

Berkani, N., S. Khouri, et L. Bellatreche (2019). Value-driven approach for designing extended data warehouses. In *DOLAP*.

Bonifati, A., W. Martens, et T. Timm (2017). An analytical study of large sparql query logs. *Proceedings of the VLDB Endowment 11*(2), 149–161.

Chen, P. (2002). Entity-relationship modeling : historical events, future trends, and lessons learned. In *Software pioneers*, pp. 296–310. Springer.

Ciferri, C., R. Ciferri, L. Gómez, M. Schneider, A. Vaisman, et E. Zimányi (2013). Cube algebra : A generic user-centric model and query language for olap cubes. *International Journal of Data Warehousing and Mining (IJDWM) 9*(2), 39–65.

Deb Nath, R. P., K. Hose, et T. B. Pedersen (2015). Towards a programmable semantic extract-transform-load framework for semantic data warehouses. In *DOLAP*, pp. 15–24.

Djilani, Z., S. Khouri, L. Bellatreche, et A. Khiat (2017). Les besoins fonctionnels candidats à l'entreposage et l'analyse en ligne. In *EDA*, pp. 139–149.

Etcheverry, L. et A. Vaisman (2016). Querying semantic web data cubes. pp. 11–23.

Etcheverry, L., A. Vaisman, et E. Zimanyi (2014). Modeling and querying data warehouses on the semantic web using qb4olap. In *DaWAK*, pp. 45–56.

Gallinucci, E., M. Golfarelli, S. Rizzi, A. Abellï¿½, et O. Romero (2018). Interactive multidimensional modeling of linked data for exploratory olap. *Information Systems 77*, 86–104.

Horkoff, J., D. Barone, L. Jiang, E. Yu, D. Amyot, A. Borgida, et J. Mylopoulos (2014). Strategic business modeling : representation and reasoning. *SSM 13*(3), 1015–1041.

Kämpgen, B. et A. Harth (2011). Transforming statistical linked data for use in OLAP systems. In *I-SEMANTICS*, pp. 33–40.

Kämpgen, B., S. O'Riain, et A. Harth (2012). Interacting with statistical linked data via OLAP operations. In *ESWC (Satellite Events)*, pp. 87–101.

Khouri, S., A. R. Ghomari, et Y. Aouimer (2019). Thinking the incorporation of lod in semantic cubes as a strategic decision. In *To appear in International Conference on Model and Data Engineering*. Springer.

Khouri, S., K. Semassel, et L. Bellatreche (2015). Managing data warehouse traceability : A life-cycle driven approach. In *CAiSE*, pp. 199–213.

Maté, A., J. Trujillo, et J. Mylopoulos (2017a). Conceptual modeling for indicator selection. In *Conceptual Modeling Perspectives*, pp. 55–68. Springer.

Maté, A., J. Trujillo, et J. Mylopoulos (2017b). Specification and derivation of key performance indicators for business analytics : A semantic approach. *Data & Knowledge Engi-*

neering 108, 30–49.

Matei, A., K. Chao, et N. Godwin (2014). OLAP for multidimensional semantic web databases. In *BIRTE*, pp. 81–96.

Nasiri, A., W. Ahmed, R. Wrembel, et E. Zimányi (2017). Requirements engineering for data warehouses (re4dw) : From strategic goals to multidimensional model. In *International Conference on Conceptual Modeling*, pp. 133–143. Springer.

Princea, S. T., N. Guarino, G. Guizzardi, et J. Mylopoulos. (2017). An ontological analysis of value propositions. In *21st International Enterprise Distributed Object Computing Conference (EDOC)*, pp. 184–193. IEEE.

Ravat, F., J. Song, et O. Teste (2016). Designing multidimensional cubes from warehoused data and linked open data. In *RCIS*, pp. 1–12.

Saad, R., O. Teste, et C. Trojahn (2013). OLAP manipulations on RDF data following a constellation model. In *1st International Workshop on Semantic Statistics*.

Silva Souza, V. E., J.-N. Mazï¿½n, I. Garrigï¿½s, J. Trujillo, et J. Mylopoulos (2012). Monitoring strategic goals in data warehouses with awareness requirements. In *ACM - SAC'12*, pp. 10–75. ACM Press.

Summary

With the proliferation of Linked Open Data (LOD) sources, organizations have quickly seen the opportunity to extend their internal data cubes with such external data. Although LOD sources have a real added value, they have brought their own technical difficulties related to the management of this new source. However, the existing literature focus on managing these technical difficulties, whereas this decision is primarily 'organizational'. This omission is explained by the fact that the integration of LODs is studied according to a source-driven vision, whereas a requirements-driven vision would be more adequate to study effectively the relevance of the integration of LOD sources. In this paper, we propose an organizational (goal-oriented) vision in which Key Performance Indicators (KPIs) are the central element linking data sources to organizational objectives. We propose a goal-oriented approach to estimate the organizational value (via KPI values) of the LOD resources in the cube. The approach is based on a metamodel aligning the three models: the data model, the goal model and the KPI model. Experiments are conducted to validate our approach.