Approche Guidée par la Valeur et la Variété pour Concevoir des Entrepôts de Données Etendus

Value and Variety Driven Approach for Extended Data Warehouses Design

Nabila Berkani [1], Selma Khouri [1], Ladjel Bellatreche [2]

[1] Ecole nationale Supérieure d'Informatique, BP 68M, 16309, Oued-Smar, Alger, Algérie
n_berkani@esi.dz,s_khouri@esi.dz

[2] Laboratoire d'Informatique et d'Automatique pour les Systèmes. LIAS/ISAE-ENSMA et Université de Poitiers 86961 Futuroscope cedex, France
bellatreche@ensma.fr

RÉSUMÉ. En un laps de temps assez court (1990 à nos jours), la technologie des entrepôts de données est passée par toutes les phases de la vie d'un produit technologique : une introduction sur le marché, une croissance, une maturité et une baisse de régime sentie avec l'apparition des données massives (Big Data). Dans le paysage des Big Data, l'arrivée des Linked Open Data ($\mathcal{LOD}$) transforme la menace Big Data en une opportunité pour les $\mathcal{ED}$, car elles sont porteuses de valeurs ajoutées et de connaissances que nous ne trouvons pas dans les sources internes alimentant un $\mathcal{ED}$. Cependant, l'introduction des $\mathcal{LOD}$ augmente la variété des sources, qui doit être gérée efficacement. Dans cet article, nous présentons une nouvelle approche de conception d'$\mathcal{ED}$ guidée par la valeur et la variété que nous appliquons à une étude de cas du domaine des SHS.

ABSTRACT. In a very short time (1999-present), the data warehouse ($\mathcal{DW}$) technology has gone through all the phases of a technological product's life : introduction on the market, growth, maturity and decline, signaled by the appearance of Big Data. In the big data landscape, the arrival of Linked Open Data (LOD) transforms the Big Data threat into an opportunity for $\mathcal{DW}$s, because they bring added value and knowledge that we do not find in the internal sources feeding a $\mathcal{DW}$. However, the consideration of $\mathcal{LOD}$s increases the variety of sources, which must be managed effectively. In this paper, we present a new value and variety driven approach for $\mathcal{DW}$ design that we apply to a case study of the SHS domain.

MOTS-CLÉS. Entrepôt de Données Etendu, Données Ouvertes et Liées, Variété, Valeur.

KEYWORDS. Data warehouse, linked open data, Variety, Value.

1. Introduction

L'industrie des systèmes de stockage des données représente actuellement un marché très fructueux générant plusieurs milliards de dollars par an. Les entrepôts de données ($\mathcal{ED}$) sont un exemple de ces systèmes qui contribuent à gérer un des capitaux les plus importants de toute organisation représentant ses données. L'arrivée des données massives et les besoins d'analyser en continu des données volumineuses a contribué à la baisse du régime de la technologie des $\mathcal{ED}$, ce qui a également impacté son écosystème. Cette situation a poussé certains à anticiper la disparition des $\mathcal{ED}$ au profit des systèmes Big Data. En conséquence, la communauté "Data Analytics" est divisée entre trois mouvements : (1) les fournisseurs de solutions Big Data qui écartent catégoriquement les solutions d'entreposage. (2) les pros de la technologie des $\mathcal{ED}$ qui militent pour l'extension de la technologie d'$\mathcal{ED}$ afin de prendre en compte certaines dimensions de Big Data et ses différents V. (3) Un mouvement combinant les deux solutions. Notons que dans la mouvance Big Data, une autre ère autour des Linked Open Data ($\mathcal{LOD}$) est aussi bien présente. Elle a montré un intérêt important auprès des gouvernements, des usagers, des chercheurs, etc. A titre d'exemple, les sciences sociales et humaines ($\mathcal{SHS}$) est un domaine qui joue un rôle primordial dans la compréhension et l'interprétation du contexte économique, culturel et social dans lequel vivent et agissent les populations. L'évolution de la recherche dans ce domaine passe inévitablement par l'échange et le partage des connaissances entre les chercheurs. La mise à disposition des chercheurs d'un contenu de données automatiquement exploitable, partageable et facilement utilisable pour la prise de décision a été rapidement perçue comme un enjeu cruciale, et a laissé émerger de nombreuses sources d'informations $\mathcal{SHS}$ au format $\mathcal{LOD}$. Ces données représentent une valeur ajoutée pour les consommateurs des solutions d'$\mathcal{ED}$. En d'autres termes, au lieu d'exploiter uniquement les sources de données internes dans le processus de construction d'un $\mathcal{ED}$ dédiée au données $\mathcal{SHS}$, l'inclusion

d'autres ressources comme les $\mathcal{LOD}$ représente une valeur ajoutée pour les acteurs et analystes du domaine. Par ailleurs, la prise en compte de la valeur ajoutée lors de la conception des entrepôts de données suscite beaucoup d'intérêt de la communauté. Un numéro spécial autour de cette thématique vient d'être lancé [1]. Cependant, cette quête de la valeur augmente la variété des sources d'informations induites par le formalisme (orienté graphe) et le vocabulaire de ces sources. Dans une vision classique, ces deux V (variété et valeur) sont traités mais leurs corrélations et leur mise à niveau n'est pas considérée, car ils sont gérés par différents acteurs et durant des phases différentes du projet d'entreposage. L'objectif de cette mise à niveau est d'évaluer le gain en termes de valeur (connaissances issues des sources LOD) tout en traitant la variété qui augmente indéniablement par la considération de ces nouvelles sources. Dans cet article, nous présentons une nouvelle approche de conception d'$\mathcal{ED}$ guidée par la valeur et la variété, où différents scénarii d'intégration des $\mathcal{LOD}$ impactant les processus ETL sont étudiés. Ces scénarios sont déduits des politiques organisationnelles réalistes pour intégrer les $\mathcal{LOD}$ dans la construction d'un $\mathcal{ED}$: (a) une première politique où l'$\mathcal{ED}$ est construit à partir de zéro par une intégration simultanée de sources de données internes et externes, et (b) une seconde où l'$\mathcal{ED}$ est déjà opérationnel lorsque l'entreprise décide d'intégrer les $\mathcal{LOD}$.

Cet article comporte les sections suivantes : la section 2 présente un état de l'art des principaux travaux du domaine. La section 3 présente notre méthode de construction d'un $\mathcal{ED}$ intégrant les $\mathcal{LOD}$ comme une ressource externe. La section 4 présente une validation de la méthode sur une étude de cas exploitant des données $\mathcal{SHS}$. La section 5 conclut l'article.

2. État de l'art : $\mathcal{LOD}$ une Valeur Ajoutée pour les $\mathcal{ED}$

Les $\mathcal{LOD}$ ont intégré l'environnement $\mathcal{ED}$ où certains travaux ont consolidé les efforts faits dans les $\mathcal{ED}$ issus de sources internes pour l'unification des formalismes (Ravat *et al.*, 2017; Baldacci *et al.*, 2017; Deb Nath *et al.*, 2015; Etcheverry *et al.*, 2014), et l'unification des vocabulaires en utilisant des structures ad-hoc telles que des tables de correspondance (utilisant des mesures de similarité) (Ravat *et al.*, 2017; Deb Nath *et al.*, 2015) ou en utilisant une ontologie partagée (Alberto *et al.*, 2016). Pour l'ensemble de ces travaux, la politique organisationnelle de l'entreprise est considérée comme figée où la plupart des travaux considèrent un seul scénario d'intégration des $\mathcal{LOD}$ dans un $\mathcal{ED}$ opérationnel. Ce scénario contraint les concepteurs à gérer la variété des $\mathcal{LOD}$ selon l'$\mathcal{ED}$ cible (suivant ses contraintes techniques). Ceci se fait soit : (a) *à priori* au niveau du formalisme d'unification par une approche médiateur (Ravat *et al.*, 2017) ou une approche d'entreposage où des fragments du $\mathcal{LOD}$ son dupliqués dans l'$\mathcal{ED}$ pour différentes raisons comme la réparation des informations manquantes (Alberto *et al.*, 2016) ou l'unification des cubes internes et externes (Deb Nath *et al.*, 2015). Peu de travaux traitent en particulier le processus ETL (Baldacci *et al.*, 2017; Deb Nath *et al.*, 2015; Kämpgen *et al.*, 2012). (b) *A postériori* lors de l'interrogation de l'$\mathcal{ED}$ (Saad *et al.*, 2013; Matei *et al.*, 2014). En examinant ces travaux, nous remarquons qu'il n'existe pas une étude d'intégration des $\mathcal{LOD}$ dans le processus de construction d'un $\mathcal{ED}$ couvrant l'ensemble de phases de cycle de vie. De plus, contrairement aux études existantes, notre approche propose trois contributions principales : (i) une conceptualisation de la variété en présence de sources de internes et externes, (ii) elle propose différents scénarios, définis au niveau organisationnel, d'intégration de $\mathcal{LOD}$ dans un $\mathcal{ED}$, (iii) notre approche étudie les corrélations entre la variété (gérée via un processus ETL) et la valeur dans un projet d'entreposage.

3. Approche proposée

Notre approche s'articule autour de trois principales étapes : la conceptualisation de la variété, les scénarios d'intégration, et la quantification de la valeur ajoutée.

3.1. Conceptualisation de la variété

L'objectif de cette première étape est de fournir un méta-modèle ETL suffisamment générique pour couvrir la variété des des sources internes et $\mathcal{LOD}$ (leurs différents formalismes et vocabulaires) ainsi que les différents scénarios d'intégration possibles (abordés à l'étape suivante). La formalisation de la variété des sources de données est effectuée sur la base de la logique de description (DL). Dans la terminologie DL, une base de connaissances est composée de deux composants : <T Box, ABox>. Le TBox (boîte terminologique) indique la connaissance intentionnelle (schéma) et le ABox (boîte d'assertion) indique la connaissance d'extension (instances). Grâce au formalisme DL, le modèle conceptuel des sources a été formalisé. L'environnement ETL doit ainsi prendre en compte la variété à différents niveaux : (a) les concepts et (b) les formalismes des sources pour gérer la variété

1. https ://www.elsevier.com/journals/international-journal-of-information-management/0268-4012/guide-for-authors

des sources en entrée. (c) Les opérateurs, (b) les activités et (c) les workflow pour gérer la variété des scénarios d'intégration. Basé sur la WfMC [2], nous proposons le méta-modèle illustré dans la figure 3.1). La définition d'un méta-modèle permet de rendre générique la représentation ETL et de gérer la variété des sources. Le méta-modèle prend en compte un ensemble de workflows considérés comme une collection globale d'activités ETL et de transitions entre eux. Une transition détermine la séquence d'exécution des activités pour générer un workflow à partir des sources vers l'$\mathcal{ED}$ cible. Les transformations ETL sont réalisées grâce à des opérateurs ETL définis avec une signature d'éléments (Exp. *Extract*, *Context*, *Filter*, etc). La figure 3.1 présente les opérateurs ETL conventionnels ainsi que les opérateurs requis par l'introduction du $\mathcal{ED}$, qui sont illustrés par des points rouges dans la Figure. Sur le plan technique, la traduction de ces opérateurs peut se faire soit selon un modèle générique défini, soit selon un modèle pivot choisi parmi les sources candidates. Pour nos expérimentations, nous avons opté pour ce second scénario, considérant les $\mathcal{LOD}$ comme source 'élue', car elle approte une représentation générique souvent utilisée comme formalisme pivot (orienté graphe), ainsi que des ontologies explicitant la sémantique du vocabulaire utilisé. L'unification des formalismes des sources internes se fait donc en considérant un formalisme orienté graphe, et l'unification de leur vocabulaire se fait selon la sémantique des $\mathcal{LOD}$.

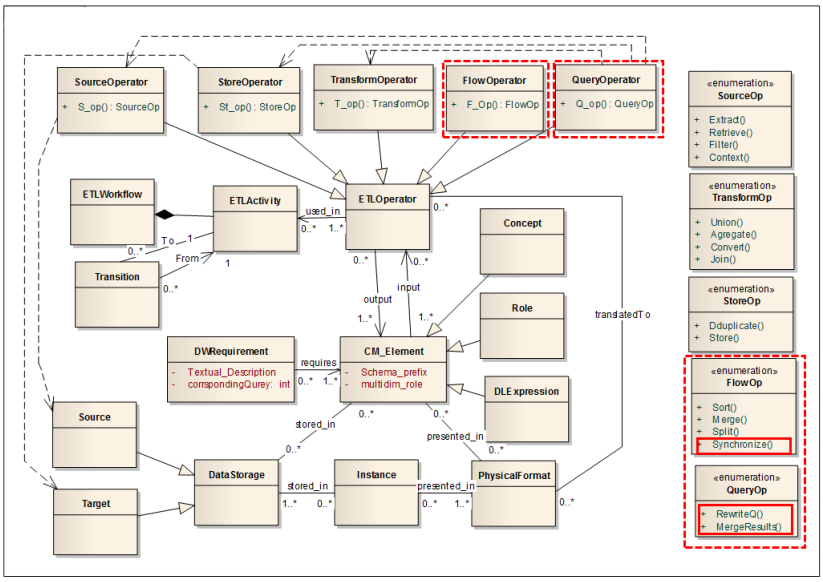

FIGURE 3.1. *Le méta-modèle de workflow ETL.*

3.2. Les scénarios d'intégration Variété-Valeur

Dans ce qui suit, nous proposons trois scénarios permettant d'intégrer les $\mathcal{LOD}$ dans la construction d'un $\mathcal{ED}$.

Intégration des $\mathcal{LOD}$ en série. Ce scénario correspond à une conception *conventionnelle* d'$\mathcal{ED}$. Les $\mathcal{LOD}$ sont considérés comme étant une nouvelle source sémantique à gérer en plus des sources internes.

Intégration des $\mathcal{LOD}$ en parallèle. Ce scénario procède à l'intégration des sources internes et des $\mathcal{LOD}$, tout en supposant que l'$\mathcal{ED}$ est opérationnel. Le processus ETL du $\mathcal{LOD}$ (ETL-$\mathcal{LOD}$) est généré puis *synchronisé* avec le processus ETL initial (ETL-$\mathcal{ED}$) défini à partir des sources internes. Ce scénario nécessite la consolidation de deux workflow (internes et externes) pour maintenir l'entrepôt cible à jour. Nous formalisons le problème de la consolidation en introduisant l'opérateur *synchronize* dans le méta modèle ETL proposé. Cette opération correspond à une synchronisation entre deux workflows : (i) le *Workflow actuel* : ETL workflow qui satisfait les n exigences en cours à l'instant t, et (ii) le *Nouveau Workflow* : ETL workflow qui satisfait les exigences à venir à l'instant $t+1$. L'opérateur *Synchronize* correspond à une série d'opérations fréquemment rencontrées dans la gestion de workflow : (i) AND-Join : identifier les deux flux ETL, appliquer les opérations de verrouillage potentielles et effectuer l'opération de jointure entre les concepts, (ii) OR-Join : correspond à une opération de fusion de concepts et propriétés effectuée à l'aide de l'opérateur Merge et (iii) Clean : effectue un nettoyage des données, vérifie les valeurs nulles et supprime les données en double avant de les charger dans l'$\mathcal{ED}$.

Intégration des $\mathcal{LOD}$ à la demande. Ce scénario correspond à un ETL à la demande ("orienté requêtes") où les données des $\mathcal{LOD}$ sont extraites puis chargées dans l'$\mathcal{ED}$ uniquement lorsque ces données sont nécessaires à la satisfaction des exigences exprimées en requêtes OLAP.. Ce scénario nécessite d'abord la réécriture des requêtes OLAP sur les $\mathcal{LOD}$ dans le but d'extraire le fragment répondant aux exigences. Il requiert ensuite l'application des transformations nécessaires sur le processus ETL dédié aux $\mathcal{LOD}$ en utilisant la classe *Transform-Operator*

2. http ://www.wfmc.org/

du méta-modèle proposé (Figure 3.1). Le résultat de ce processus ETL est matérialisé en utilisant classe *Store-Operator*. Les résultats de cette opération sont d'abord intégrés dans un cube de données dédié à l'analyse des données $\mathcal{LOD}$ puis *fusionnés* avec les résultats des requêtes OLAP exécutées sur l'$\mathcal{ED}$, pour enfin être affichés à l'utilisateur final. Cette opération de fusion est représentée par la classe *Query-Operator*, qui est liée aux classes : *Source-Operator* pour extraire les fragments de $\mathcal{LOD}$, *Store-Operator* pour déployer le résultat de l'ETL $\mathcal{LOD}$ sur l'$\mathcal{ED}$ et *Transform-Operator* afin de gérer les transformations requises par le processus ETL dédié aux $\mathcal{LOD}$. Nous avons également enrichi la classe *Query-Operator* par les méthodes *Rewrite_Query* et *MergeResult_Queries* permettant l'unification des résultats obtenus et la gestion des différentes opérations d'interrogation mentionnées.

3.3. Quantification de la valeur ajoutée

Nous avons identifié deux principales métriques de valeur : (a) le taux de concepts multidimensionnels (MD) de l'$\mathcal{ED}$, représentant son expressivité pour les analystes du domaine ; (b) la capacité de l'$\mathcal{ED}$ à satisfaire les besoins des utilisateurs. Ces métriques ne sont pas explicitement représentées dans le modèle précédent, mais les éléments permettant de les calculer le sont. La première métrique (a) est calculée comme suit :

$$Valeur(S_{iMD}) = \frac{Nbre\_Concepts(S_i)}{NbreTotal\_Concepts(ED)} \quad [1]$$

où $Nbre\_Concepts(S_i)$ et $NbreTotal\_Concepts_ED$ représentent respectivement le nombre de concepts MD obtenus de la source $_i$ et le nombre total de concepts MD de $\mathcal{ED}$. Concernant la satisfaction des besoins, nous évaluons la valeur ajoutée en utilisant la valeur de chaque source comme suit :

$$Valeur(S_{iB}) = \frac{NbreReponsesBes(S_i)}{NbreReponsesReq(ED)} \quad [2]$$

où $NbreReponsesBes(S_i)$ et $NbreReponsesReq(ED)$ décrivent respectivement le nombre de réponses aux requêtes (correspondant à un besoin) et le nombre de réponses aux requêtes sur l'entrepôt.

4. Expérimentations

Nous présentons un ensemble d'évaluations pour montrer l'efficacité de notre proposition. Considérons le scénario où une organisation gouvernementale souhaite construire un $\mathcal{ED}$ pour analyser les événements politiques de différents pays. Pour ce faire, les organisations des différents pays doivent fusionner leurs bases de données pour construire un $\mathcal{ED}$ qui doit répondre à certaines exigences clés. Supposons que quatre organisations de différents pays participent à cette opération où chaque organisation est considérée comme étant une source de données : la France (S_{FR}), l'espagne (S_{ESP}), l'Allemagne (S_{AL}) et le royaume uni (S_{RU}). La particularité de ces sources est qu'elles sont dérivées du référentiel relatif aux évènements politiques européens issus de PoliMedia (`http://polimedia.nl/`). Le choix du déploiement de l'entrepôt a été défini pour Oracle relationnel et graphe. Dans ce contexte, le processus ETL doit faire face à la variété des données sources tout en satisfaisant les besoins. En considérant uniquement les sources ci-dessus, les besoins exprimés ne sont pas entièrement satisfaits, par conséquent, l'appel à des ressources externes devient pertinent pour ajouter de la valeur à l'entrepôt cible. Notre ressource externe correspond à un fragment du portail LOD décrivant les discours du parlement européens `http://www.talkofeurope.eu/data/`. Le fragment obtenu contient environ 7,9 x 10^6 Quads. Dans un premier temps, nous évaluons l'impact des efforts de conceptualisation sur la gestion de la variété dans l'entrepôt obtenu.Nous comparons le choix du format 'élu' qui est le format graphe du $\mathcal{LOD}$ utilisé avec le format générique proposé dans (Berkani & Bellatreche, 2017). La comparaison est faite en fonction des entités, des attributs, des relations et des instances, tout en considérant les trois scénarios. Les figures 4.2 décrivent les résultats obtenus. Ces derniers montrent clairement que notre modèle $\mathcal{LOD}$ orienté graphe capture plus d'éléments que le méta-modèle pivot. En effet, tous les éléments satisfaisant les exigences (fragment $\mathcal{LOD}$) sont matérialisés dans l'entrepôt.

L'évaluation a été menée également pour mesurer la valeur capturée par l'intégration des sources internes et des sources externes $\mathcal{LOD}$. Pour le premier critère à savoir taux de concepts multidimensionnels et d'instances intégrées, la figure 4.3 illustre les résultats obtenus, où le nombre de concepts et d'instances intégrés à partir des $\mathcal{LOD}$ est assez important. Il apparait clairement que la prise en compte du fragment $\mathcal{LOD}$ augmente le nombre de concepts multidimensionnels pour les trois scénarios. En comparant les trois scénarios, nous constatons qu'ils sont presque équivalents. Pour évaluer le second critère se rapportant à la satisfaction des besoins, nous avons formulé nos besoins utilisateurs sous la forme de requêtes OLAP exécutées sur l'$\mathcal{ED}$. L'exécution des requêtes OLAP a été

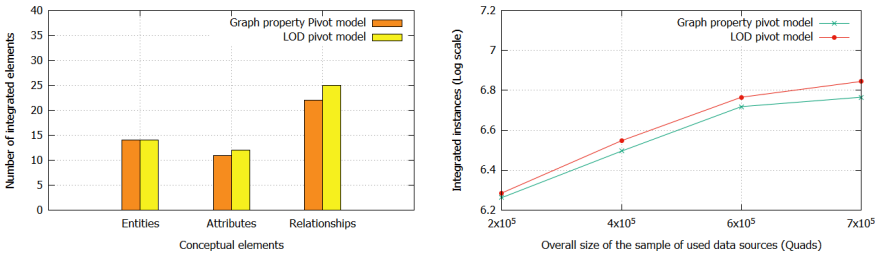

FIGURE 4.2. *Comparison entre les modèles pivot $\mathcal{LOD}$ Vs générique $\mathcal{LOD}$ et Graph property*

effectuée en quatre étapes. Les besoins des utilisateurs satisfaits par les sources de données internes représentent $\sim$ 65%. Une fois les $\mathcal{LOD}$ intégrées, ce taux augmente considérablement jusqu'à atteindre un taux maximum de 96%. Nous avons également remarqué que le troisième scénario (ETL à la demande) engendre le meilleur résultat et répond aux besoins des utilisateurs plus rapidement que les autres scénarios. Cela peut s'expliquer par le fait que ce scénario se concentre sur l'intégration des données relatives aux besoins des utilisateurs exprimés par des requêtes OLAP.

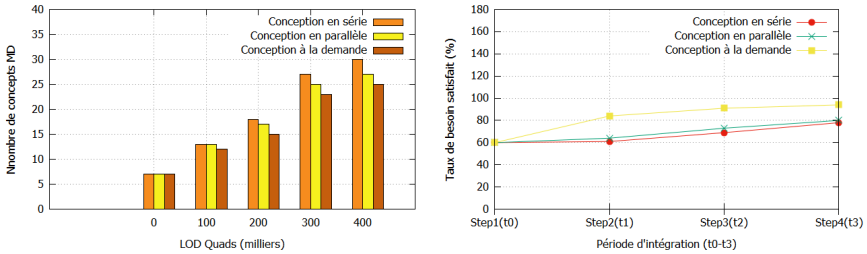

FIGURE 4.3. *La valeur ajoutée par l'intégration des $\mathcal{LOD}$*

5. Conclusion

Notre travail dans cet article montre l'intérêt d'exploiter les données $\mathcal{LOD}$ dans la construction d'$\mathcal{ED}$. Nous avons proposé un méta-modèle pivot ETL pour la gestion de la variété. Nous avons également étudié trois scénarios d'intégration tout en considèrant l'interaction entre la valeur et la variété des sources. Nos expérimentations basées sur des données réelles validant nos propositions. Ce travail ouvre plusieurs perspectives notamment : (1) la considération d'autres V de Big Data, par exemple le volume des données sources, (2) le déploiement et l'optimisation physique de l'entrepôt de données résultant.

Références

ALBERTO A., ENRICO G., MATTEO G., STEFANO R. & OSCAR R. (2016). Towards exploratory olap on linked data. In *SEBD*, p. 86–93.

BALDACCI L., GOLFARELLI M., GRAZIANI S. & RIZZI S. (2017). Qetl : An approach to on-demand etl from non-owned data sources. *DKE*, 112, 17–37.

BERKANI N. & BELLATRECHE L. (2017). A variety-sensitive ETL processes. In *DEXA*, p. 201–216.

DEB NATH R. P., HOSE K. & PEDERSEN T. B. (2015). Towards a programmable semantic extract-transform-load framework for semantic data warehouses. In *DOLAP*, p. 15–24.

ETCHEVERRY L., VAISMAN A. & ZIMÁNYI E. (2014). Modeling and querying data warehouses on the semantic web using qb4olap. In *DaWAK*, p. 45–56.

KÄMPGEN B., O'RIAIN S. & HARTH A. (2012). Interacting with statistical linked data via OLAP operations. In *ESWC (Satellite Events)*, p. 87–101.

MATEI A., CHAO K. & GODWIN N. (2014). OLAP for multidimensional semantic web databases. In *BIRTE*, p. 81–96.

RAVAT F., SONG J. & TESTE O. (2017). Vers un modèle unifié de données entreposées et de données ouvertes liées. concepts et expérimentations. *ISI*, 22(2), 35–67.

SAAD R., TESTE O. & TROJAHN C. (2013). Olap manipulations on rdf data following a constellation model. In *1st International Workshop on Semantic Statistics*.