goldMEDAL : une nouvelle contribution à la modélisation générique des métadonnées des lacs de données

Etienne Scholly[*,**], Pegdwendé N. Sawadogo[*], Pengfei Liu[*],
Javier A. Espinosa-Oviedo[*,***] Cécile Favre[*], Sabine Loudcher[*], Jérôme Darmont[*],
Camille Noûs[****]

[*]Université de Lyon, Lyon 2, UR ERIC
{etienne.scholly, pegdwende.sawadogo, pengfei.liu, javier.espinosa-oviedo,
cecile.favre, sabine.loudcher, jerome.darmont}@univ-lyon2.fr
[**]BIAL-X
[***]LAFMIA lab
[****]Laboratoire Cogitamus
camille.nous@cogitamus.fr

Résumé. Nous résumons ici un article publié en 2021 dans l'atelier international DOLAP adossé aux conférences jointes EDBT et ICDT. Nous y proposons goldMEDAL, un modèle générique de métadonnées pour les lacs de données basé sur quatre concepts et une modélisation en trois niveaux : conceptuel, logique et physique (Scholly et al., 2021).

L'essor des mégadonnées a révolutionné les pratiques d'exploitation des données et a conduit à l'émergence de nouveaux concepts. Parmi eux, les lacs de données sont de vastes dépôts de données hétérogènes qui peuvent être analysés par diverses méthodes (Dixon, 2010). Un lac de données efficace nécessite un système de métadonnées qui répond aux nombreux problèmes posés par le traitement de données volumineuses et hétérogènes. L'étude des modèles de métadonnées des lacs de données est un sujet de recherche très actif et fait l'objet de plusieurs propositions.

Parmi celles-ci, le modèle MEDAL propose de représenter les données à travers trois concepts principaux : les *objets* qui correspondent à un ensemble de données homogènes, les *représentations* qui résultent de transformations de l'objet associé et les *versions* qui représentent les mises à jour d'un objet (Sawadogo et al., 2019). Cependant, le modèle MEDAL ne peut pas représenter simultanément différents niveaux de granularité des données.

Ravat et Zhao (2019) proposent un modèle dont la principale contribution est la notion de métadonnées de *zone*, qui spécifie la zone où se trouvent les données (par exemple, zone de données brutes, zone de données traitées). Toutefois, ce modèle ne prend pas non plus en charge les niveaux de granularité multiples des données.

Eichler et al. (2020) introduisent le modèle HANDLE, qui utilise le concept générique d'*entité de données* pour représenter à la fois des fichiers de données et des parties de fichiers de données. Ceci lui permet de prendre en charge n'importe quel niveau de granularité. Chaque entité de données est associée à des étiquettes qui représentent des zones, des niveaux de granu-

larité ou des catégorisations. Cependant, HANDLE ne prend pas en compte le versionnement des données.

Nous constatons ainsi que les modèles de métadonnées existants (même les plus recents) sont soit adaptés à un cas d'utilisation spécifique, soit insuffisamment génériques pour gérer différents types de lacs de données, y compris notre premier modèle, MEDAL. Pour traiter ce problème de généricité, nous présentons le modèle goldMEDAL, une évolution de MEDAL. Ce nouveau modèle comprend trois niveaux de modélisation : conceptuel, logique et physique.

En nous inspirant des différentes notions introduites dans MEDAL, nous basons le modèle conceptuel de goldMEDAL sur quatre concepts principaux : entité de données, groupement, lien et processus (Figure 1).

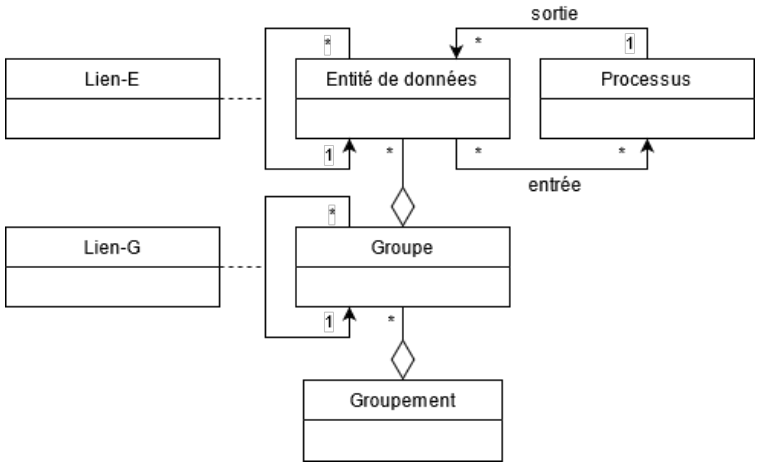

FIG. 1 – *Diagramme de classes UML des concepts de goldMEDAL*

— **Entité de données.** Les entités de données sont les unités de base du modèle de métadonnées. Elles sont flexibles en termes de granularité des données. Par exemple, une entité de données peut représenter un fichier de tableur, un document textuel ou semi-structuré, une image, une table de base de données, un tuple ou une base de données entière. L'introduction de tout nouvel élément dans le lac de données entraîne la création d'une nouvelle entité de données.

— **Groupement.** Un groupement est un ensemble de groupes ; un **groupe** rassemble des entités de données sur la base de propriétés communes. Par exemple, des zones de données brutes et prétraitées d'un lac forment les groupes d'un groupement de zones. Un autre exemple est un regroupement de documents textuels en fonction de la langue d'écriture.

— **Lien.** Les liens sont utilisés pour associer soit des entités de données entre elles, soit des groupes d'entités de données entre eux. Ils peuvent être orientés ou non. Ils permettent d'exprimer, par exemple, de simples liens de similarité entre entités de données ou des hiérarchies entre groupes. Par exemple, une hiérarchie temporelle mois → trimestre aurait les mois de janvier, février et mars liés au premier trimestre d'une année donnée.

— **Processus.** Un processus désigne toute transformation appliquée à un ensemble d'entités de données qui produit un nouvel ensemble d'entités de données.

Au niveau logique, les concepts de goldMEDAL sont représentés à travers un graphe. Ainsi, les entités de données sont représentées par des nœuds. Les liens deviennent des arêtes. Enfin, les groupes et processus sont traduits en hyper-arêtes.

Au niveau physique, nous avons implémenté goldMEDAL dans trois cas d'usage différents. Une première implémentation dénommée *HOUDAL* et dédié à l'habitat social utilise la base de données Neo4J [1] pour fournir un service de stockage et d'analyse de données principalement structurées. Le lac de données *AUDAL* supporte quant à lui des données tabulaires et textuelles. Il exploite les bases de données Neo4j, MongoDB [2] et ElasticSearch [3] pour analyser l'avancée de la servicisation et la digitalisation dans les PMI de la Région Rhône-Alpes-Auvergne. Enfin, la troisième implémentation, mise en œuvre au sein du projet HyperThesau et intitulée *Achaeo-DAL*, est dédiée à l'exploitation de données archéologiques constituées de données structurées, semi-structurées et non structurées (images et textes). Elle est basée sur Apache Atlas [4].

À travers les trois modèles physiques implementés avec goldMEDAL, nous démontrons la faisabilité ainsi que la flexibilité de notre modèle de métadonnées. De plus, les concepts de goldMEDAL généralisent ceux des modèles les plus récents : MEDAL, Ravat et Zhao (2019) et HANDLE. Cela fait de goldMEDAL le modèle le plus générique pour la modélisation des métadonnées de lacs de données à ce jour.

Remerciements

Le doctorat d'E. Scholly est financé par la société BIAL-X [5]. Le doctorat de P.N. Sawadogo est financé par la Région Auvergne-Rhône-Alpes à travers le projet AURA-PMI. Le projet HyperThesau est financé par le Laboratoire d'Excellence "Intelligences des Mondes Urbains" [6].

Références

Dixon, J. (2010). Pentaho, Hadoop, and Data Lakes. https ://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

Eichler, R., C. Giebler, C. Gröger, H. Schwarz, et B. Mitschang (2020). HANDLE-A Generic Metadata Model for Data Lakes. In *International Conference on Big Data Analytics and Knowledge Discovery (DaWak 2020), Bratislava, Slovakia*, pp. 73–88.

Ravat, F. et Y. Zhao (2019). Metadata management for data lakes. In *European Conference on Advances in Databases and Information Systems (ADBIS 2019), Bled, Slovenia*, pp. 37–44. Springer.

1. https://neo4j.com
2. https://www.mongodb.com
3. https://www.elastic.co
4. https://atlas.apache.org
5. https://www.bial-x.com
6. https://imu.universite-lyon.fr

Sawadogo, P. N., E. Scholly, C. Favre, E. Ferey, S. Loudcher, et J. Darmont (2019). Metadata systems for data lakes : models and features. In *International Workshop on BI and Big Data Applications (BBIGAP@ADBIS 2019), Bled, Slovenia*, pp. 440–451. Springer.

Scholly, E., P. Sawadogo, P. Liu, J. A. Espinosa-Oviedo, C. Favre, S. Loudcher, J. Darmont, et C. Noûs (2021). Coining goldmedal : A new contribution to data lake generic metadata modeling. In *23rd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP@EDBT/ICDT 2021)*, Volume 2840, pp. 31–40. CEUR.

Summary

We summarize here a paper published in 2021 in the DOLAP international workshop DOLAP associated with the EDBT and ICDT conferences. We propose goldMEDAL, a generic metadata model for data lakes based on four concepts and a three-level modeling: conceptual, logical and physical (Scholly et al., 2021).