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-