Métadonnées des lacs de données et principes FAIR

Ahlame Diouan[*,**], Eric Ferey[**], Sabine Loudcher[*], Jérôme Darmont[*], Camille Noûs [***]

[*]Univ Lyon, Univ Lyon 2, UR ERIC – [**]BIAL-X – [***]Laboratoire Cogitamus

Depuis le début du 21e siècle, les usages des organisations dans les processus de prise de décision sont bouleversés par la disponibilité de grands volumes de données hétérogènes appelées mégadonnées (*big data*). Ces mégadonnées constituent une véritable opportunité pour les organisations, mais elles s'accompagnent entre autres de problématiques de volume, de vélocité et de variété qui surpassent les capacités des systèmes traditionnels de stockage et de traitement des données (Miloslavskaya et Tolstoy, 2016). C'est dans ce contexte que Dixon (2010) introduit le concept de lac de données (*data lake*), en guise de solution aux problèmes induits par l'hétérogénéité des mégadonnées. Un lac de données a besoin de métadonnées qui permettent de décrire les données stockées dans le lac, ainsi qu'un système efficace de gestion de ces métadonnées, qui nécessite à son tour un modèle de métadonnées pour sa mise en œuvre (Sawadogo et al., 2019). L'étude des modèles de métadonnées des lacs de données est un sujet de recherche très actif et fait l'objet de plusieurs propositions, dont les modèles MEDAL (Sawadogo et al., 2019), DAMMS (Ravat et Zhao, 2019), HANDLE (Eichler et al., 2020), ainsi que le métamodèle goldMEDAL (Scholly et al., 2021).

Pour évaluer les modèles et les systèmes de métadonnées, Scholly et al. (2021) se réfèrent à plusieurs critères, notamment des critères techniques ou fonctionnels : enrichissement sémantique, polymorphisme des données, versionnement, suivi d'utilisation, catégorisations, liaisons de similarité, propriétés des métadonnées et niveaux de granularité multiples. De plus, il existe dans la littérature d'autres critères qui pourraient être complémentaires, comme les principes FAIR, qui permettent de qualifier des données et des métadonnées (Wilkinson et al., 2016). Ces principes ne constituent pas une norme, mais agissent comme un guide pour les éditeurs et les responsables de la gestion de données. Ceci les aide à évaluer leurs implémentations pour rendre leurs ressources numériques trouvables, accessibles et faciles à utiliser. Comme un lac de données est susceptible de prendre en charge des données hétérogènes et de rendre les données interopérables grâce à son système de gestion de métadonnées, cela peut être pertinent de confronter le concept de lac de données aux principes FAIR, qui suivent.

— Faciles à trouver (*Findable*) : les données sont décrites par des métadonnées riches et précises, les (méta)données possèdent un identifiant unique et pérenne, et elles sont enregistrées ou indexées dans un dispositif permettant de les rechercher.
— Accessibles (*Accessible*) : les (méta)données doivent être récupérables par des protocoles de communication standardisés, libres et ouverts, avec authentification éventuelle.
— Interopérables (*Interoperable*) : les (méta)données sont décrites avec un vocabulaire contrôlé permettant l'interopérabilité et la gestion des références entre elles.
— Réutilisables (*Reusable*) : les (méta)données doivent être publiées avec une licence d'utilisation claire et accessible, toutes en étant décrites par une pluralité d'attributs indiquant leur provenance.

Zhao (2021) a évalué les approches de gestion des métadonnées dans les lacs selon les principes FAIR, sans pourtant utiliser tous les sous-principes. Nous complétons cette évaluation car, parmi les quinze sous-principes FAIR, nous estimons que certains se réfèrent aux modèles et d'autres aux systèmes de gestion des métadonnées.

L'objectif du poster associé à cet article est de présenter l'évaluation des modèles, d'une part, et des systèmes de métadonnées, d'autre part, au prisme des principes FAIR. Pour chacun des quinze sous-principes, nous précisons leur degré de prise en charge (ne s'applique pas, non pris en charge ou impossible, possible mais pas mentionné, explicitement traité).

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*, Volume 12393 of *LNCS*, pp. 73–88.

Miloslavskaya, N. et A. Tolstoy (2016). Big data, fast data and data lake concepts. *Procedia Computer Science 88*, 300–305.

Ravat, F. et Y. Zhao (2019). Metadata management for data lakes. In *European Conference on Advances in Databases and Information Systems*, Volume 1064 of *CCIS*, pp. 37–44.

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*, Volume 1064 of *CCIS*, pp. 440–451.

Scholly, E., P.-N. 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 *International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data*, Volume 2840 of *CEUR*, pp. 31–40.

Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data 3*(1), 1–9.

Zhao, Y. (2021). *Metadata Management for Data Lake Governance*. Ph. D. thesis, Université Toulouse 1 Capitole, France.

Summary

Several criteria in the literature have been proposed to evaluate metadata models and metadata management systems for data lakes. Among such criteria, the FAIR principles serve to guide data managers to make digital resources Findable, Accessible, Interoperable and Reusable. In this poster paper, we evaluate data lake metadata models and systems with respect to FAIR principles. Moreover, we provide degrees of achievement for each FAIR subprinciple (does not apply, not supported or impossible, possible but not documented, explicitly supported).