

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 21^e 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.