

Appariement de schémas semi-structurés

Fabien Duchateau (prénom.nom@univ-lyon1.fr)

Contexte. L'appariement de schémas consiste à détecter les correspondances entre concepts équivalents de schémas hétérogènes. Ce processus d'appariement est crucial lorsque l'on souhaite intégrer ou fusionner des données de différentes sources, par exemple pour interroger ces sources de manière uniforme ou pour en extraire de nouvelles connaissances. L'appariement de schémas a été longuement étudié en relationnel et en XML [1], mais peu de travaux portent sur l'appariement de sources semi-structurées comme le BSON/JSON (utilisé par exemple dans MongoDB). Bien que ces modèles sont dits "flexibles" ou "sans schéma", il existe tout de même une organisation ou structuration *cachée* des données qu'il est possible d'exploiter pour l'appariement.

Objectifs. Ce projet cherche à atteindre les objectifs suivants :

- Dresser un état de l'art sur l'appariement de schémas JSON [2, 3, 4] et rédiger un rapport présentant les verrous scientifiques liés à ce type de schémas [5, 6].
- Définir ou développer un benchmark (métriques d'évaluation et jeux de données) pour l'appariement de schémas JSON. Des algorithmes d'appariement provenant du framework BigGorilla¹ ou JSONGlue [7, 4] peuvent être utilisés pour valider le benchmark.

Contraintes et compétences. Lecture d'articles en anglais, bonne qualité rédactionnelle.

Bibliographie

- [1] Philip A Bernstein, Jayant Madhavan, and Erhard Rahm. Generic schema matching, ten years later. *Proceedings of the VLDB Endowment*, 4(11):695–701, 2011. [Lien VLDB2011](#).
- [2] Mohamed-Amine Baazizi, Housseem Ben Lahmar, Dario Colazzo, Giorgio Ghelli, and Carlo Sartiani. Schema Inference for Massive JSON Datasets. In *Extending Database Technology (EDBT)*, Venice, Italy, 2017. [Lien EDBT2017](#).
- [3] Javier Luis Cánovas Izquierdo and Jordi Cabot. JSONDiscoverer: Visualizing the schema lurking behind JSON documents. *Knowledge-Based Systems*, 103:52–55, 2016. [Lien KBS2016](#).
- [4] Vitor Marini Blaselbauer and Joao Marcelo Borovina Josko. Jsonglue: A hybrid matcher for json schema matching. In *Companion Proceedings*, page 77, 2020. [Lien SBBD2020](#).
- [5] Javier Luis Cánovas Izquierdo and Jordi Cabot. Discovering implicit schemas in JSON data. In *International Conference on Web Engineering*, pages 68–83. Springer, 2013. [Lien ICWE2013](#).
- [6] Ekaterini Ioannou, Nataliya Rassadko, and Yannis Velegrakis. On Generating Benchmark Data for Entity Matching. *Journal on Data Semantics*, 2(1):37–56, 2013. [Lien JODS2013](#).
- [7] Chen Chen, Behzad Golshan, Alon Y Halevy, Wang-Chiew Tan, and AnHai Doan. BigGorilla: An Open-Source Ecosystem for Data Preparation and Integration. *IEEE Data Eng. Bull.*, 41(2):10–22, 2018. [Lien DE2018](#).

¹<http://www.biggorilla.org/>