Scalable integration and interpretation of heterogeneous data sources

Ioana Manolescu, Inria and Institut Polytechnique de Paris ioana.manolescu@inria.fr
Karen Bastien, WeDoData karen@wedodata.fr

July 21, 2020

1 Overview

Digital data, whether text (news articles), semi-structured (tweets, other social media content) or structured (RDF or CSV files) is produced and shared at very large speed today. Events such as elections, public manifestations, commercial or cultural events etc. unfold under the form of new data items joining a growing global corpus of available information. Integrating and making sense of this wealth of data is a highly prized goal.

We consider the particular setting of a **data lake**, which is a set of **heterogeneous data sources**, in which some common elements (e.g., people, places, organization names, or semantic concepts) may be found across data sources. For instance, a media article may mention a company that sets up a new office in a small city, a PDF document may list the recipients of public funding, or addresses of suppliers of a merchandise or service (e.g., child clothing or baby-sitting), a spreadsheet may list cultural or leisure equipments (e.g., fitness equipments, parks or museums etc.) in the various cities of a geographic area etc. Such resources may be created at small scale by individual user, or large(r) scale by companies, public administrations, or non-governmental organizations (NGO).

A natural question in such a setting is: how to take advantage of **a large, dynamic set of heterogeneous sources of information** in an **integrated** fashion, with **little effort as possible spent on integrating the data**? Here are some sample scenarios:

- A young mother would like, given a PDF of local businesses and a spreadsheet with leisure equipments, to identify quickly the cities where baby-sitting services and sport equipments can be found.
- A company looking for local suppliers could use results not only from the Yellow Pages (*Registre des Entreprises*) but also information about the public funding they may have received, as well as snippets from recent Web and social media articles about these suppliers.
- Valuable digital cultural assets, such as the French press archive digitized by Retronews or those of the Institut National de l'Histoire de l'Art (INHA), could be used in innovative ways by interconnecting their databases. This would allow building applications where users navigate from a painter like Corot, to the places where he worked or that he painted, e.g. Igny (91), to media coverage of that place along the years etc.[1]

The support currently available to answer such questions is either missing or has significant limitations:

1. **Manual stitching:** this is the case where users have to open or query each data source using a separate tool (e.g., a spreadsheet editor and a Web browser) and "stitch" (combine) results manually. This is tedious for users, and time-consuming. Further, some popular data formats (e.g., JSON or RDF) are not accessible to non-technical users.
2. **Integrated portals:** some Web sites or applications have been built by IT specialists for a public or private client organization, with the goal of exposing a certain number of datasets to the users. Moreover, their maintenance is difficult, as adding a new dataset requires IT effort to integrate it in the portal, and their extensibility is limited, as it cannot be expected from the application owners to continuously extend it to accomodate one novel category of goods or services.

Vision The PhD project aims at building a **generic, efficient integration platform**, which will ingest heterogeneous, structured or unstructured data sources and provide **user-friendly search and interactions with the data**, based on the paradigm of **integrated graphs**.
The project would benefit *from* and *to* an AI Chair in Artificial Intelligence, awarded by the ANR to I. Manolescu (Inria), with WeDoData as a supporting (non-funded) partner, as we explain below.

[1] Retronews and INHA are among the clients that WeDoData has worked for; this inspired our examples.

2 Scientific background and relation with AI Chair SourcesSay

Data integration systems provide a single access point to a set of heterogeneous data sources, which users can exploit through queries [7, 14]; they are frequently used in large organizations to integrate relational databases. Among data integration systems, data lake tools, e.g., by IBM or Oracle, open-source projects such as Apache Drill or PrestoDB, or *polystore* systems [1, 8, 12] assume an *explicit logical connection* (or *mapping*) between the schema of the data sources, and the *integrated schema* accessible to the users; writing mappings requires an advanced technical expertise. In contrast, we do not aim at an integrated schema, since this is unfeasible for large sets of highly heterogeneous data, and requires effort to design and to maintain when new sources are added. Instead, we provide users with a *graph view* of the data sources, on which we enable intuitive querying and exploration, along the lines of the "Data Spaces" vision [9]. The **Linked Open Data (LOD)** stack of technologies advocated by the W3C, based on the RDF data model and on semantic (ontology) languages, also exposes data as graphs. However, this approach supposes an *initial ontology (or schema)* design, to which each data source is then mapped; such mappings must be updated to integrate other data. In contrast, we aim to integrate heterogeneous data into graphs, based on the presence of co-occurring entities (or, more generally, *linkable elements* - see below).

Data cleaning (or **entity resolution**) aims to identify and eliminate duplicates in structured databases [10, 3, 13]; many modern methods rely on machine learning [11]. These techniques need to be revisited in the presence of unstructured *text content*, frequent in digital arenas, which hold many text documents and text fields in other formats such as JSON, RDF, CSV, etc. Further, our applications may feature little-known entities, for whom we do not have background information (e.g., a reference knowledge base). This is the case, for instance, of names of places that are only well known to inhabitants of a certain area, historical characters of lesser notoriety etc. etc. Yet, we still want to use entity mentions, when possible, as *linkable elements*, to interconnect the data around them. For instance, we may identify the name of a person or a place in a painting's title, to interconnect the painting with information about the place, other famous artists from that place etc.

A **graph view** of the data lake is a data graph (*i*) modeling each data source and its internal structure (when present); (*ii*) including structure discovered (extracted) from text, e.g., mentions of entities or linkable elements, which lead to new nodes; (*iii*) interconnecting data sources based on the presence of common linkable elements they contain. Starting in 2018, Inria has developed a first prototype, called **ConnectionLens**, which builds such a graph view out of heterogeneous data sources [6]. Currently, ConnectionLens is capable of answering *interconnection queries* between nodes that match some given keywords [2]. For instance, it can find all the connections *"What are all the connections between "Corot" and "Palaiseau"?* (Answer: he painted in Igny, which is a neighbor of Palaiseau.)

Research on ConnectionLens has received support from the ANR through the **AI Chair in Artificial Intelligence SourcesSay** (2020-2024). In SourcesSay, learning-based methods will be developed to (*i*) improve the extraction of linkable elements, (*ii*) learn what graph connections are interesting, and (*iii*) devise novel visualization methods for graph views of heterogeneous data. The present proposal aims to strengthen the collaboration between Inria and WeDoData, by carrying our research together oriented by their concrete applications.

3 PhD project: scaling up graph construction and analysis

The goal of the PhD is to *scale up by orders of magnitude the construction and storage of the graph*, while at the same time *devising novel graph analysis algorithms specific to ConnectionLens graphs*. Overall, **Inria will guide the literature review, algorithm design, implementation, and experiments**. WeDoData will help t**arget datasets characteristic of those they encountered in their past projects** (e.g., for OCDE, Paris School of Economics, France Télévisions, Île-de-France's Open data portal etc.) and devise **concrete applications of ConnectionLens data integration**. Further, they will also **provide input on the choices of tools and libraries** for the platform to be robust, extensible, easy to maintain and to deploy in a variety of practical settings.

3.1 Scaling up graph construction

Building a ConnectionLens graph requires: a pass over the data sources to transform them into nodes and edges, this is also where we extract linkable elements; then, a pass over the *pairs of nodes* thus obtained, to determine *when we should unify two nodes* for instance, two mentions of "Camille Corot" in two distinct datasets), and *when we should connect them through a "similar" edge*. As an example for the latter, a novel character mentioned as "Carlota" in one source may or may not be the same as "Carlotta" mentioned in another source. In some cases, a knowledge base (KB) may help identify the entity these mentions refer to, and chose the correct spelling. However, as explained above, we may encounter mentions with no support in the KB. In this case, adding a "similar" edge records a possible connection. Based on it, if and when sufficient evidence accumulates later (e.g., we find the

"Carlotta" spelling in a large number of sources), we may decide to align the "Carlota" name to it, also; or, a KB we acquire in the future may help us solve this entity reference problem at that time.

As known from the data cleaning literature, *the node comparison stage is the most expensive*, as it involves many pairwise comparisons. This remains challenging despite recent advances in entity matching and resolution, mostly because *assumptions made in prior work* (either all sources are relational, or all are textual; KBs are available and they cover all entities; a full crawl of the Web can be used as reference etc.) *do not hold*. Currently, ConnectionLens relies on storing nodes and edges in PostgreSQL, an open-source reliable relational database management system; it retrieves pairs of nodes to be compared through a set of SQL queries we crafted, and applies different similarity functions according to the best duplicate detection/data cleaning practices [5].

The PhD will study alternative algorithms for scaling up the comparison process. Avenues to consider include: switching to a distinct back-end based on a faster store (possibly a graph store, existing or to be prototyped specifically for our needs); parallelizing the effort through multi-threading and/or running in a cluster (Inria owns one); splitting the work into a first "rough" step to be done when loading the data and a long-running "daemon-style" stage to be run subsequently to incrementally improve/enhance the graph.

3.2 Making sense of ConnectionLens graphs

The second part of the PhD project will investigate novel graph analysis (or refinement) methods to produce, from the "rough" graph obtained through ConnectionLens' integration, a different *meaningful* graph, possibly smaller. Three graph transformation principles are envisioned here: (*i*) preserve the *entities* and the *relationships* connecting them, not the surrounding data fields, attributes etc. For instance, a tweet in JSON format has about 120 nodes, while conceptually, it only connects a fews entities and hashtags; a similar "entity focus" could be applied to general ConnectionLens graphs, based on the observation that entities are the most meaningful for end users; (*i*) identify *the most significant paths* to compactly represent the possibly numerous paths which may connect two entities in a graph; (*ii*) *customize* the extracted graph to focus only on entities matching some specification of user interest, e.g., through search keywords or focusing on a geographic area. These problems are challenging and may have very high complexity, as they involve analyzing paths and connecting trees in the ConnectionLens graph; efficient heuristics will be sought where appropriate. First steps toward such heuristics have been made in [4].

Practical information

The PhD student will be employed by Inria and will split his time between our office in Palaiseau (Ecole Polytechnique) and Paris (WeDoData). The thesis is financed through a PRPhD project funded by DIM RFSI, an organization fostering Computer Science and Technology projects in the larger Parisian area. The PhD will start in the fall of 2020 or in early 2021.

A successful candidate should have a strong background in data management and algorithms, and solid programming skills. Knowledge of data visualization, Web standards such as XML, RDF and JSON, and good French speaking skills would be a plus.

To apply, send to Ioana Manolescu (`ioana.manolescu@inria.fr`) and Karen Bastien (`karen@wedodata.fr`) a recent CV, M2 (or equivalent) academic transcripts, and any material referring to your development experience.

References

[1] R. Alotaibi, D. Bursztyn, A. Deutsch, I. Manolescu, and S. Zampetakis. Towards Scalable Hybrid Stores: Constraint-Based Rewriting to the Rescue. In *SIGMOD*, 2019. 2

[2] A. Anadiotis, M.-Y. Haddad, and I. Manolescu. Graph-based keyword search in heterogeneous data sources. In 36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA), 2020. 2

[3] A. Arasu, S. Chaudhuri, Z. Chen, K. Ganjam, R. Kaushik, and V. R. Narasayya. Towards a domain independent platform for data cleaning. *IEEE Data Eng. Bull.*, 34(3), 2011. 2

[4] I. Burger, I. Manolescu, E. Pietriga, and F. M. Suchanek. Toward Visual Interactive Exploration of Heterogeneous Graphs. In *SEAdata 2020 Workshop, in conjunction with EDBT/ICDT*, Mar. 2020. 3

[5] O. Bălălău, C. Conceiçao, H. Galhardas, I. Manolescu, T. Merabti, J. You, and Y. Youssef. Graph integration of structured, semistructured and unstructured data for data journalism. In 36ème Conférence sur la Gestion de Données – Principes, Technologies et Applications (BDA), 2020. 3

[6] C. Chanial, R. Dziri, H. Galhardas, J. Leblay, M.-H. Le Nguyen, and I. Manolescu. ConnectionLens: Finding Connections Across Heterogeneous Data Sources (demonstration). *PVLDB*, 11, 2018. 2

[7] A. Doan, A. Y. Halevy, and Z. G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012. 2

[8] A. Elmore, J. Duggan, M. Stonebraker, and al. A Demonstration of the BigDAWG Polystore System. *PVLDB*, 2015. 2

[9] M. J. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record*, 34(4), 2005. 2

[10] H. Galhardas, D. Florescu, D. E. Shasha, E. Simon, and C. Saita. Declarative data cleaning: Language, model, and algorithms. In *VLDB*, 2001. 2

[11] I. F. Ilyas and X. Chu. *Data Cleaning*. Morgan & Claypool, 2019. 2

[12] B. Kolev, P. Valduriez, C. Bondiombouy, R. Jiménez-Peris, R. Pau, and J. Pereira. CloudMdSQL: querying heterogeneous cloud data stores with a common language. *Distributed and parallel databases*, 2016. 2

[13] G. Simonini, G. Papadakis, T. Palpanas, and S. Bergamaschi. Schema-agnostic progressive entity resolution. In *ICDE*, 2018. 2

[14] M. Stonebraker and I. F. Ilyas. Data integration: The current status and the way forward. *IEEE Data Eng. Bull.*, 41(2), 2018. 2