Understanding heterogeneous data sources with ConnectionLens

Nelly Barret, 1st year PhD student at Inria and LIX (UMR 7161)
Supervised by Ioana Manolescu

Context

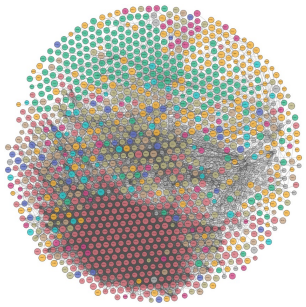
- Open data initiative has led to a set of big heterogeneous data sources
- Heterogeneous data sources are difficult to integrate and understand/exploit

How to help a human user grasp the content of a data source?

1. Analyse and exploit the structure of the data
2. Compute a form of semantics on the content of the data

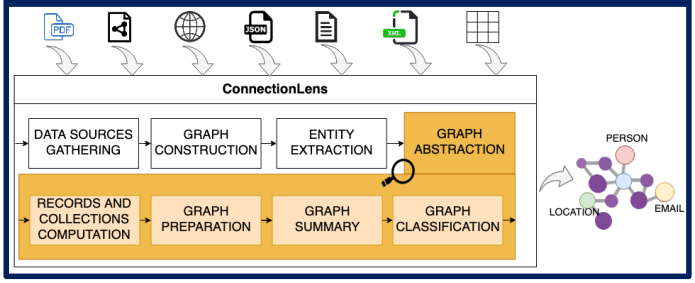
Existing works

- Summarizing semi-structured data (statistical, structured, logical approaches)
  → Based only on structure, not on content
- Schema inference
  → Specific to one data model



Goasdoué F, Guzewicz P, Manolescu I. RDF graph summarization for first-sight structure discovery. VLDBJ 2020.
Baazizi MA, Colazzo D, Ghelli G. *et al.* Parametric schema inference for massive JSON datasets. VLDB 2019.

Overview of the ConnectionLens Approach



Assumptions

We are given a directed graph containing:

- Data nodes
- Extracted entities nodes
- Data, similarity and equivalence edges
- A set of data item categories: *Person, Event, Creative work, …*
- For each category, a set of attributes names

Understanding what a data source is about (in an automatic way)

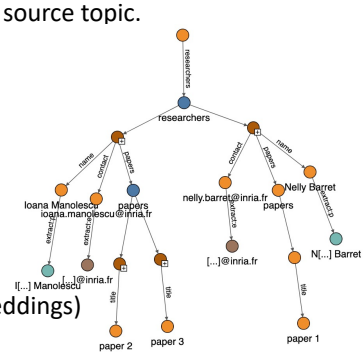
Intuition: the structure and the content of a ConnectionLens graph are useful to compute the data source topic.

Goal 1: interpret the structure of the data

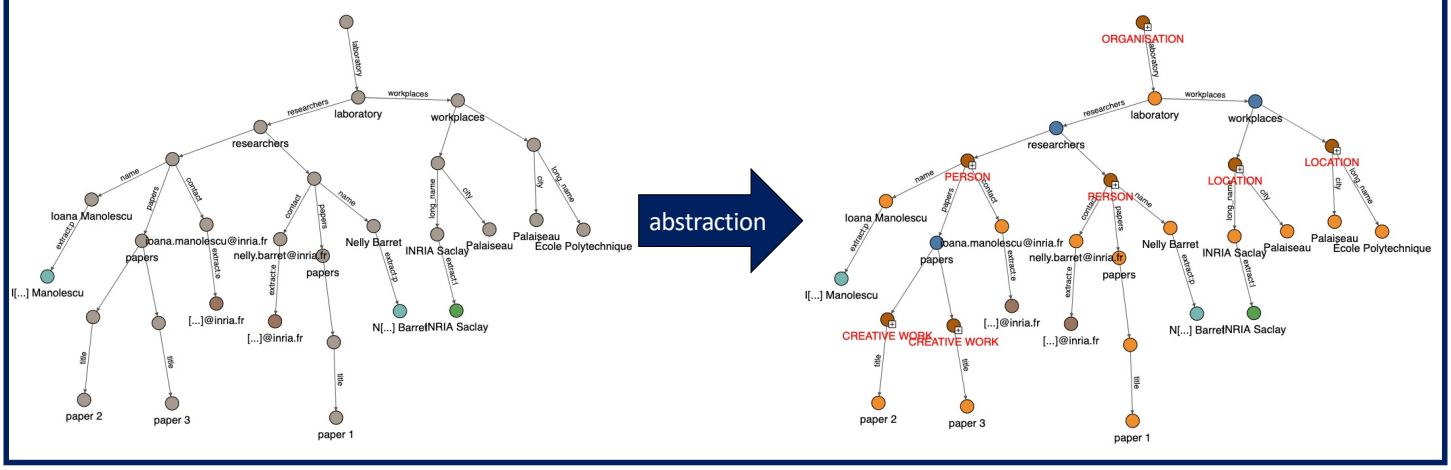
1. Detect *Records* (things), possibly organised in *Collections* (set of similar things)

Goal 2: exploiting the content of the data

1. Build a DataGuide summary of each record of each collection
2. Classify each record using its properties and the data items categories:
   a. Compute the *signature* of each property
   b. Find the best category for each attribute (compare it with category attributes using embeddings)
   c. Classify the record using a majority vote



A scenario in ConnectionLens



Future work

- Enrich the data source to improve the understanding, e.g. using knowledge bases or web information
- Generate semi-automatically the attributes of a category using external resources
- Create expressive summaries of what a data source is about