

Semantic Integration of Heterogeneous Data Sources

LVMT | École nationale des ponts et chaussées

Institut Polytechnique de Paris

Context

Supply-chain data integration involves recurring tasks, such as linking an order to a shipment or identifying the operational event of a delivery.

Differences in schemas, identifiers and the terminology used to represent the main components of a supply-chain process complicate these tasks [1].

This internship will draw on case-based reasoning [2] and the reuse of semantic models [3] to investigate whether using context information extracted from past, resolved integration tasks can provide guidance to define and solve new integration cases.

The internship contributes to ongoing research efforts in data integration for logistics within [ACME](#) (*Approvisionnement Collaboratif Multimodal Écologique*), a project funded through France 2030 under PEPR MOBIDEC.

Objectives and main tasks

The goal of the internship project is to develop, implement and evaluate the capability of a Python-based software that can be used to 1) extract context from an existing set of data integration tasks, including the problem being addressed, the data source characteristics, the operational constraints, *etc.*; and 2) rank existing tasks based on their relevance to the new data integration problem to resolve.

The project will involve four main tasks:

- To review selected literature on schema matching, semantic modelling and case-based reasoning.
- To implement and compare interpretable similarity measures based on lexical, structural and semantic information, and assess whether contextual information improves the retrieval of relevant past cases.
- To evaluate the capability of the proposed software on a set of representative integration problems and assess the relevance of the retrieved past cases.
- To analyse which elements of the retrieved cases, such as candidate correspondences, transformation patterns or known failure patterns, can inform a new integration problem and what adaptation they require.

Candidate Profile

- To be enrolled in a Master 2 or final-year engineering programme in computer science, artificial intelligence, data science or a related field.
- To have a working proficiency in the Python programming language and demonstrated experience in database and algorithm design.
- A familiarity with SQL, Semantic Web technologies (RDF(S), OWL, SPARQL), graph data models or ontologies is not required but would be a strong asset.
- An ability to read scientific literature and to communicate research findings in English.

Host and Practical details

Host: Laboratoire Ville Mobilité Transport (LVMT), École Nationale des Ponts et Chaussées.

Location: 77420 Champs-sur-Marne, France.

Duration: Six months within March–September 2027; exact dates to be agreed.

Gratification: Statutory hourly rate applicable in France (€4.50/hour in 2026).

Supervision and Application

Prospective candidates are required to send their CV, academic transcripts covering the last 3 years and a short statement of interest to both supervisors:

M. Mohammed Younes — mohammed.younes@enpc.fr

Dr Daphné Tuncer — daphne.tuncer@enpc.fr

Applications will be reviewed on a rolling basis until the position is filled.

Only complete applications sent to both supervisors will be considered.

Selected References

- [1] E. Rahm and P. A. Bernstein, “A survey of approaches to automatic schema matching,” *The VLDB Journal*, vol. 10, no. 4, pp. 334–350, 2001.
- [2] A. Aamodt and E. Plaza, “Case-based reasoning: Foundational issues, methodological variations, and system approaches,” *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994.
- [3] M. Taheriyani, C. A. Knoblock, P. Szekely, and J. L. Ambite, “Learning the semantics of structured data sources,” *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 37–38, pp. 152–169, 2016.