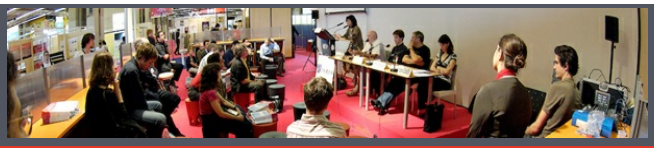
INTERNATIONAL RELATIONS



INTERNSHIP SUBJECT

2902 - Efficient parallel abstraction of heterogeneous data for AI-based ex

Modern data lakes support heterogeneous formats, from the classic relational format, which is widely used, to more flexible and less regular formats like XML, graphs, etc. The ConnectionLens project (https://team.inria.fr/cedar/connectionlens/) integrates data of any format (relational, CSV, JSON, XML, RDF) into a directed data graph model, enriched, with the help of AI (Information Extraction), with entities they contain, such as: People, Organizations, Locations, emails, URIs, etc. This system can be seen as a “data lake” platform. Based on ConnectionLens, the Abstra (https://team.inria.fr/cedar/projects/abstra/) tool automatically calculates, from the directed data graph, an abstraction, very close to an entity-relationship (ER) model. Thus, Abstra automatically identifies:

- Entities, i.e. data objects with an internal structure (attributes). Unlike known entities in relational database design, the attributes of Abstra entities can themselves be nested (have an internal structure);
- Relationships, which connect entities, and may in turn have attributes (including internally structured attributes).

Currently Abstra is capable of only identifying binary relationships (between two entities). Further, it can abstract only one dataset at a time.

The CEDAR team developed ConnectionLens as well as Abstra mainly targeting users who are not experts in data-related technologies, and in particular journalists from Le Monde, ICIJ, etc. A major challenge with data lakes is their size, sometimes going upto terabytes of data. Thus, the goal of this internship will be to scale Abstra to make it usable for such data lakes by:

1. Developing a parallel implementation for Abstra to abstract fragments of the same dataset in parallel.
2. Extending Abstra to abstract multiple datasets at a time, while exploiting the regular structure of some datasets to quickly approximate the schema.

Required Skills

The selected student should have a strong background in algorithms, databases, systems and good programming skills. Development can take place in Java (preferred) and SQL, in a collaborative team, synchronizing and reconciling code versions using Git.

General Information

- **Research Theme :** Data and Knowledge Representation and Processing
- **Locality :** Palaiseau
- **Level :** Master
- **Period :** 1st January 2026 -> 31st March 2026 (3 months)

These are approximative dates. Please contact the training supervisor to know the precise period.

- **Deadline to apply :** 1st July 2025 (midnight)

Contacts

- **Training Supervisor :** Ioana Manolescu / Ioana.Manolescu@inria.fr
- **Second Training Supervisor :** MOHANTY Madhulika / madhulika.mohanty@inria.fr
- **Team Manager :** Ioana Manolescu / Ioana.Manolescu@inria.fr

More information

- **Inria Team :** CEDAR
- **Inria Center :** Centre Inria de Saclay