ers 2 Internship in Software Engineering / Knowledge Engineering
Building a Real-World Evidence Hypergraph

Dates:	April 2026 – October 2026 (flexible)
Duration:	6 months (desired start anytime between February and April 2026)
Host labs:	CESP (Université Paris-Saclay, UVSQ, Inserm)
Address:	16 Av. Paul Vaillant Couturier, 94800 Villejuif, France
Supervisor:	Pr Lamiae Grimaldi (PU-PH, Inserm CESP) — lamiae.grimaldi@aphp.fr
Co-supervisors:	Clémence REDA (CNRS research associate) — reda@bio.ens.psl.eu — https://clreda.github.io Albert Buchard (MD, MSc., University of Geneva) — albert.buchard@hug.ch
Allowance:	Gratification

Internship title

Knowledge Engineering for a Real-World Evidence hypergraph: metamodel, tooling, and large-scale construction of an epidemiological evidence knowledge base

Keywords

Knowledge engineering; hypergraph data model; real-world evidence; knowledge base; metamodel; ontologies; provenance; data quality; evidence ingestion; validation; querying; causal inference; epidemiology; reproducibility; auditability; LLM-based agents; automated curation

Internship description

Real-world evidence studies leverage routinely collected healthcare data and modern epidemiological designs to estimate associations and causal effects between exposures (e.g., drugs), outcomes (e.g., disease onset, progression, adverse events), and patient-level factors across a broad spectrum of conditions. These results are increasingly produced with causal inference workflows and are critical to accelerate evidence synthesis for applications such as drug repurposing and clinical decision support.

A large Real-World Evidence Knowledge Base (RWE-KB) has recently been developed to structure and connect this evidence at scale. It compiles findings from epidemiological studies (results, populations, exposures, comparators, outcomes, covariates, bias/limitations, metadata), links treatments to targets, and is enriched with ontological and mechanistic knowledge. The resource is naturally represented as a hypergraph, enabling n-ary relations, hierarchical node/edge types, contextualized assertions, explicit evidence levels, and end-to-end provenance. However, the current hypergraph is still sparse and heterogeneous, and scaling it to a level that supports downstream tasks, such as AI development and clinician-facing products, requires stronger validation, data quality, provenance, and robust ingestion/curation workflows.

The core objective of this internship is to grow the existing RWE-KB into a large-scale, high-trust evidence hypergraph, with explicit provenance, quality signals, and conflict-aware aggregation. The intern will drive the expansion of the hypergraph by integrating new epidemiological evidence end-to-end, from normalization to representation, while strengthening the metamodel and validation rules that keep the KB consistent. Building on the current tooling, they will harden ingestion and curation workflows to improve key performance indicators and optimize LLM-based curation agents that reconcile inconsistent sources, handle deduplication, and reduce manual burden while keeping an auditable review loop. The outcome is a substantially larger, cleaner, and more reliable knowledge base designed to power downstream AI pipelines and clinician-facing applications.

Main responsibilities:

- Curate and extend the hypergraph metamodel for robust provenance and auditability.
- Harden the ingestion/curation stack (ETL, validation, tests, data quality, dedup, entity resolution, data contracts) and improve observability/monitoring.
- Improve existing LLM-based curation agents to resolve conflicts and produce structured, auditable updates (no predictive model development).
- Grow the RWE-KB evidence hypergraph by integrating new epidemiological studies into the knowledge base, with measurable targets (coverage growth, validation pass rate, conflict-resolution traceability).

Profile

- M2 student in Software Engineering / Knowledge Engineering / Computer Science
- Strong Python proficiency; able to deliver reliable, well-tested, maintainable code.
- Experience with MVP web tooling (React/Node.js) is a plus, but not required (opportunity to learn best practices)
- Sound foundations in data modeling, software design, testing, and version control.
- Interest in knowledge bases and graph/hypergraph representations, with attention to provenance and data quality.
- Interest in LLM-based agents/workflows for assisted ingestion, conflict resolution, and curation (experience welcome, not required).
- Comfortable in an interdisciplinary environment (computer science, epidemiology, healthcare); autonomous and well organized.

How to apply
Interested candidates should apply either in English or French to reda@bio.ens.psl.eu and lamiae.grimaldi@aphp.fr with a detailed CV and a motivation letter.
Desired starting date: anytime between February and April 2026. **Duration:** six months.