CENTRALE LYON

Project: Big Data, Linked Open Data and SPARQL

Lamia Derrode

27 January 2026

Educational Objectives

1. Understand the challenges of **Big Data** in an application or scientific domain
2. Discover and exploit **Linked Open Data (LoD)**
3. Query heterogeneous datasets using **SPARQL**
4. Design a simple **data processing, analysis and visualization workflow**
5. Develop **scientific writing** and **oral presentation** skills

Project Organization

1. Work in **groups of three (3) students**

2. The **project topic** and **datasets** must be **validated by the supervisor in advance**

3. The project includes **two mandatory components**:
   - A **Summary Report**
   - A **data processing task based on SPARQL and Linked Open Data**

1. Summary Report (maximum 10 pages)

CENTRALE LYON

The report must address a topic linking **Big Data** with an application or scientific domain, such as:

- Big Data and medicine
- Big Data and environment
- Big Data and art and culture
- Big Data and ecology
- Big Data and a scientific project (astronomy, genomics, climate science, etc.)

1. Summary Report

Expected Content of the Report

1. **Introduction to the domain**
   a) Problem statement
   b) Big Data challenges
2. **Data description**
   a) Nature, volume and variety of the data
   b) Open Data / Linked Open Data sources
3. **Big Data and LoD technologies**
   RDF, SPARQL, endpoints, vocabularies
4. **Link with the practical part**
   a) Questions addressed through the data
   b) Analysis of results
5. **Conclusion**

2. Practical Part: Data Processing with SPARQL

CENTRALE LYON

1. Use **at least one public SPARQL endpoint**

2. Design several **advanced SPARQL queries**, such as:
   - FILTER, OPTIONAL, UNION
   - Aggregation functions (COUNT, AVG, GROUP BY)

3. Use **standard ontologies/vocabularies**

2. Practical Part: Data Processing with SPARQL

Example Data Sources

- **DBpedia** (culture, geography, people)
- **Bio2RDF** (biomedical data) : https://bio2rdf.org/
- **European Open Data Portal : https://data.europa.eu/en**
- **La Bibliothèque nationale de Franc**e : https://data.bnf.fr/
- **Yago** (people, cities, countries, movies, and organizations) : https://yago-knowledge.org/

2. Practical Part: Data Processing with SPARQL

CENTRALE LYON

Example SPARQL-Based problem

- Analysis of the geographical distribution of diseases using biomedical data
- Correlation between pollution, climate and health indicators
- Analysis of artworks by period, country and artistic movement
- Cross-analysis of environmental and socio-economic data

2. Practical Part: Data Processing with SPARQL

CENTRALE LYON

Expected Outcomes

- A set of commented SPARQL queries
- Analysis of results (tables, charts, maps, etc.)
- Critical discussion on data quality and limitations

Final Presentation

CENTRALE LYON

- 20 minutes presentation + 5 minutes for questions
- Collective presentation during the last class session
- Content should include:
  - ✓ Big Data context
  - ✓ Data sources
  - ✓ SPARQL demonstration
  - ✓ Results and analysis

Assessment

- Oral presentation: 20%
- Written report: 20%
- Evaluation criteria include:
  - ✓ Scientific and technical quality
  - ✓ Relevance and complexity of SPARQL queries
  - ✓ Effective use of Linked Open Data
  - ✓ Clarity of the report and presentation
  - ✓ Critical analysis

Submission

- The written report and the presentation slides must be submitted via the Moodle platform.
- All files must be compressed into a single ZIP archive, named with the last names of the three students.
- **The submission must be completed by march, 15.**
- Late submissions will be penalized.