Title: Defining a sustainable workflow for semantically enriching territorial Open Data

Context:
Like many other territorial collectivities, the *Métropole de Lyon*[1] is publishing as Open Data several datasets that are produced either by the *Métropole de Lyon* itself or by its partners[2]. As is often the case, these datasets are available in relatively standard formats (*e.g.* CSV, JSON, GeoJSON, ...), providing a good level of syntactic interoperability. Yet, semantics is quite shallow. The *meaning* of data is often conveyed out of band, through textual documentation in the metadata as well as private communications between data consumers and the data.grandlyon.com help desk. Moreover, sometimes consumers rely on their own interpretations, which certainly may be erroneous. As such, much of the potential value of the published Open Data is hard to exploit:

- the semantics of each dataset has to be hard-coded into each client application,
- making these applications prone to break each time the underlying data evolve,
- preventing data consumers from easily building value-added applications mixing data from multiple datasets.

Linked Data[3] is a set of technologies and standards recommended by the W3C, aiming to solve the problems above. Such technologies rely on a graph-oriented data model (RDF), which facilitates data integration and fosters the use of shared vocabularies equipped with explicit semantics. The extra effort required from data publishers to comply with Linked Data principles benefits all the consumers, as it renders each dataset easier to use and to link with other datasets, the latter being even more important.

The Métropole de Lyon plans to have its Open Data progressively comply with Linked Data principles. Such a long term goal provides the general context of this PhD proposal.

Topic:
A number of languages and tools have already been proposed, allowing to migrate legacy data to Linked Data, *e.g.*: R2RML[4], RML[5] and "Plate-forme Territoire[6]", the latter being explicitly dedicated to territorial Open Data. As for vocabularies, the growing interest for Linked Data has lead to the publication of many vocabularies and ontologies (cf. the LOV[7] directory), usable in different domains including those relevant to territorial Open Data.

The first task of the PhD student will consist in reviewing the existing tools and vocabularies, in order to single out those ones being the most suitable for the Open Data published by the *Métropole de Lyon*. Potentially, such a task will call for an extension of the existing solutions.

---

[1] https://www.grandlyon.com/
[2] https://data.grandlyon.com/
[3] https://5stardata.info/en/
[4] https://www.w3.org/TR/r2rml/
[5] http://rml.io/
[6] http://territoire.emse.fr/
[7] https://lov.linkeddata.es/

The second step will consist in making sure that the migration workflow is viable and maintainable in the long run. Indeed, semantic integration is a continuous, never-ending process:

- during the definition of the mapping between legacy data and some target linked vocabularies, misinterpretations can easily occur. When discovered, they obviously require an update of the mapping.
- Datasets are updated with different frequencies, ranging from yearly to real-time. Of course, semantics should follow any evolution in the data production/update workflow, which may require an update of the mapping.
- The way data are used in practice may reveal some subtle semantic aspects, not initially taken into account, which may also necessitate some update of the mapping.

Therefore, the proposed migration process must be robust with respect to changes in the source data. It must also enable the staff of the *Métropole de Lyon* to identify the evolutions to be applied to the entire workflow and the best way to implement them.

The Phd student:
will work in Lyon (France), part-time at the *Métropole de Lyon* and part-time at the LIRIS laboratory (UMR 5205 CNRS). S·he should have a master degree or equivalent in computer science, with a strong background in Web technologies and Knowledge representation. Ideally, s·he would also have some experience in Semantic Web and Linked Data. Good communication skills, especially in English (written and oral) are also required. Some knowledge of the French language would be a plus, but is not a strict requirement.

Contacts:
Alessandro Cerioni <acerioni@grandlyon.com>
Pierre-Antoine Champin <pierre-antoine.champin@liris.cnrs.fr>