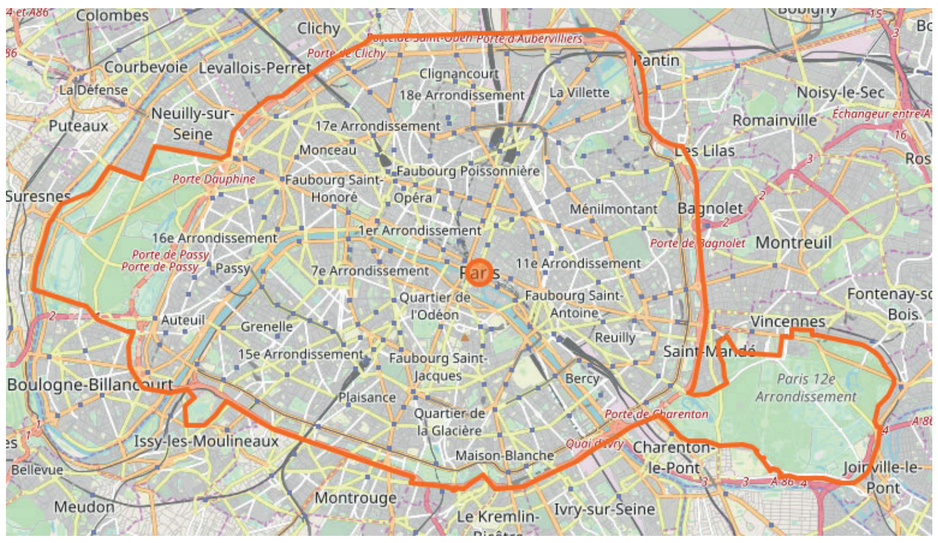
ASSESSING THE QUALITY OF owl:sameAs LINKS

le cnam
Cédric iSiD
Pierre-Henri PARIS
pierre-henri.paris@upmc.fr

1 MOTIVATION



The city of Paris and the department of Paris have the same boundaries: they are **the same in a geographical context**. But in a **legal or administrative context, they are different**. In France, a department is a much higher subdivision of the territory than a city: they do not serve the same purpose.

▸ IDENTITY IS CONTEXTUAL

2 owl:sameAs SEMANTICS

According to Leibniz:
The identity of indiscernibles (e.g.):

$$\forall P:(P(x,o)\leftrightarrow P(y,o))\rightarrow x=y$$

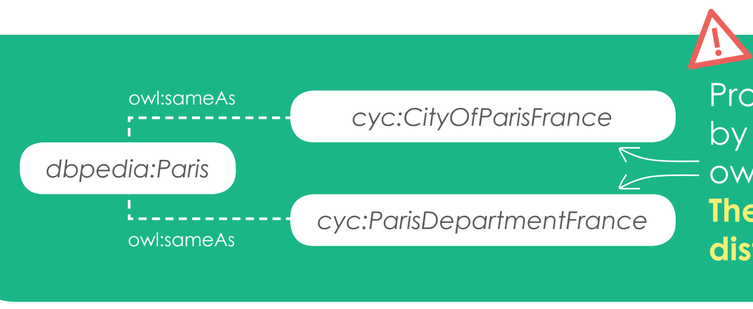
And its converse, the indiscernibility of identicals:

$$x=y\rightarrow \forall P,(P(x,o)\leftrightarrow P(y,o))$$

▸ A VERY STRICT SEMANTICS

3 PROBLEM STATEMENT

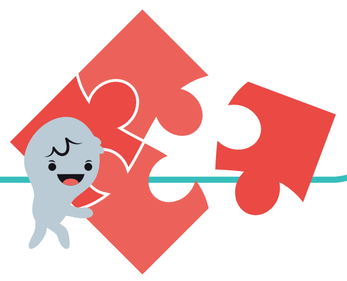
- Continuous and independent **evolution** of Linked Data
- High **dependence on the context** of use
- **Suspicion** about the accuracy and nature of owl:sameAs links:
  - error
  - lack of precision
  - incompleteness



Problem introduced by the two owl:sameAs links: **They are two distinct entities**

RESEARCH METHODOLOGY AND APPROACH 4

1. Gather **useful and precise statistics** about owl:sameAs links and semantics in general
2. **Reproduce** independently some results from previous work
3. **Identify** data interlinking quality defects (depending on context)
   a. For each type of defect, associate a **set of assessment methods** and algorithms
   b. Propose methods to **handle** those links
4. **Implement** the proposed solutions through a Web application and/or a Web service to support Data Interlinking evaluation and improvement
5. Develop solutions that **scale well**



EVALUATION PLAN 5

1. Dataset **selection**
2. For each selected datasets, defined several **contexts**
   a. For each dataset **evaluate several dimensions** of data quality (e.g. completeness) against predefined contexts
   b. Apply our approach to **assess existing links**
   c. Then **modify** datasets according to our approach's results
   d. **Compute again** the same data quality dimensions against selected datasets