JDEV 2017

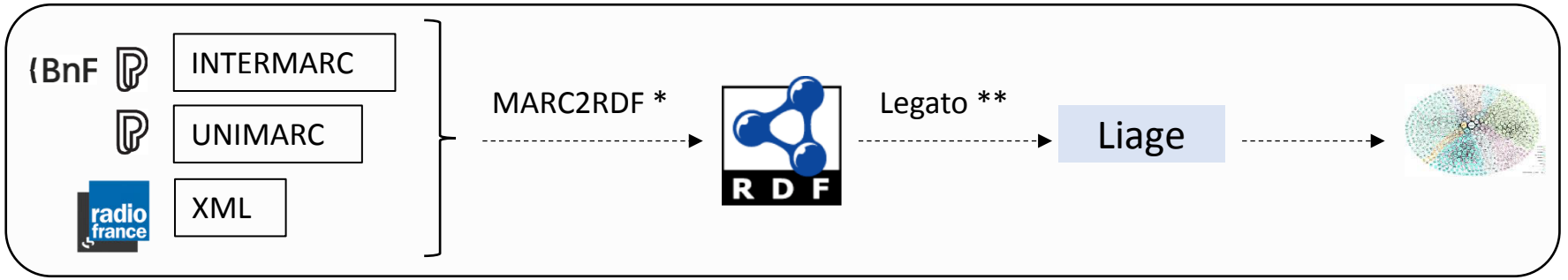
Marseille / 4-7 juillet

Interconnexion de Données du Web avec SILK

Manel Achichi & Konstantin Todorov

LIRMM / Université de Montpellier

Projet DOREMUS

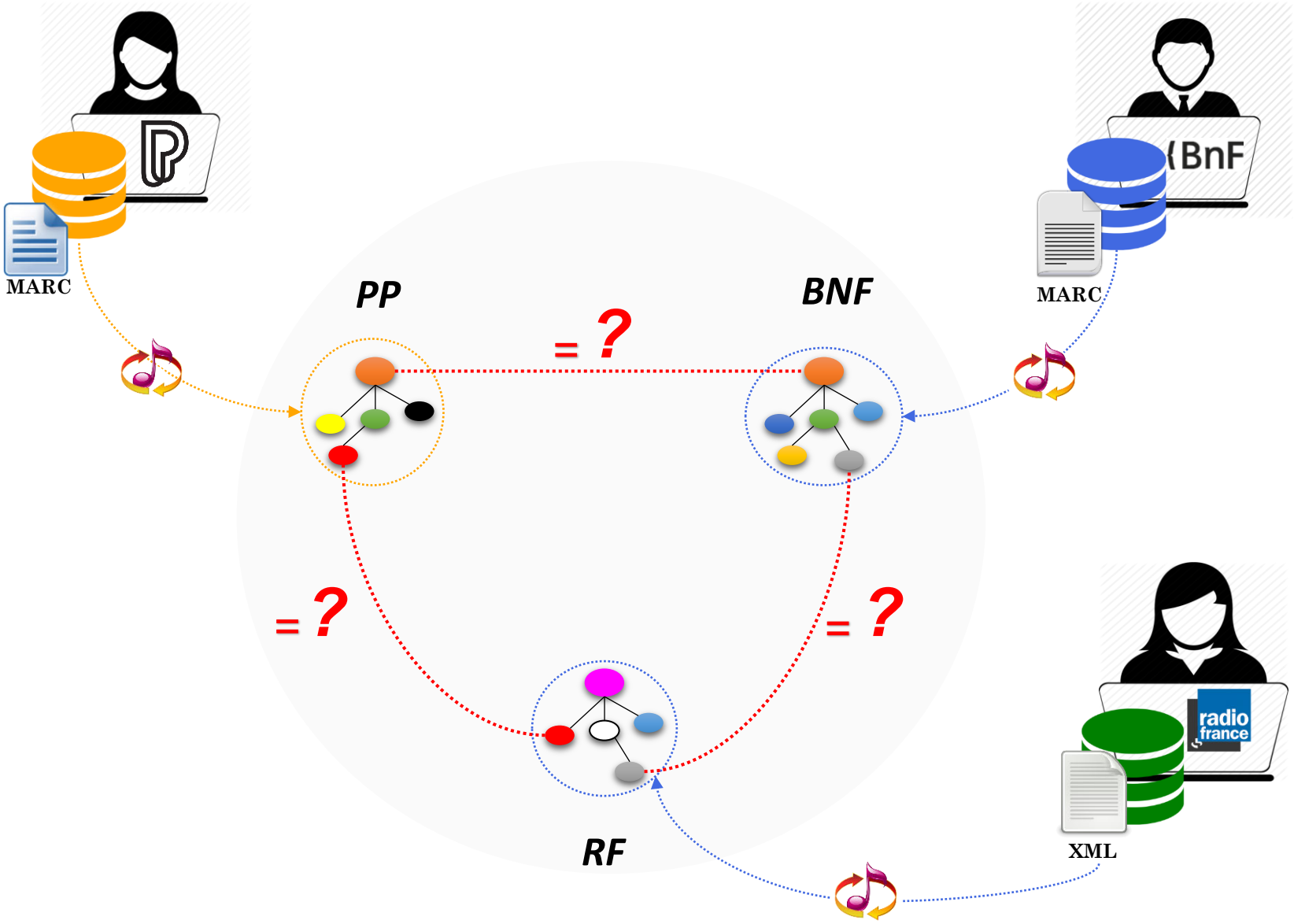


LIRMM EURECOM ROUEN Gérico MEANING ENGINES

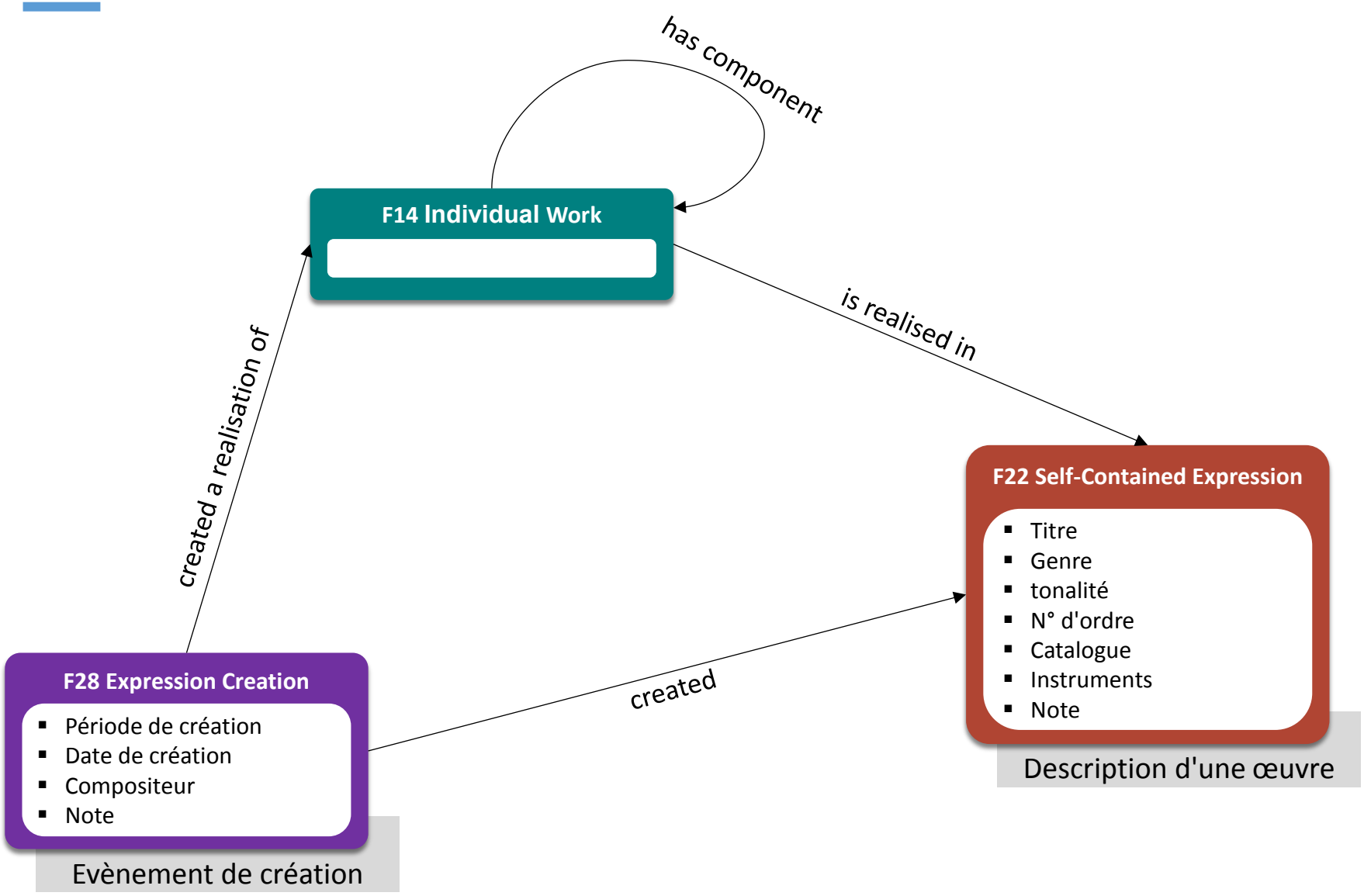
(*) : https://github.com/DOREMUS-ANR/marc2rdf

(**) : https://github.com/DOREMUS-ANR/legato

Projet DOREMUS

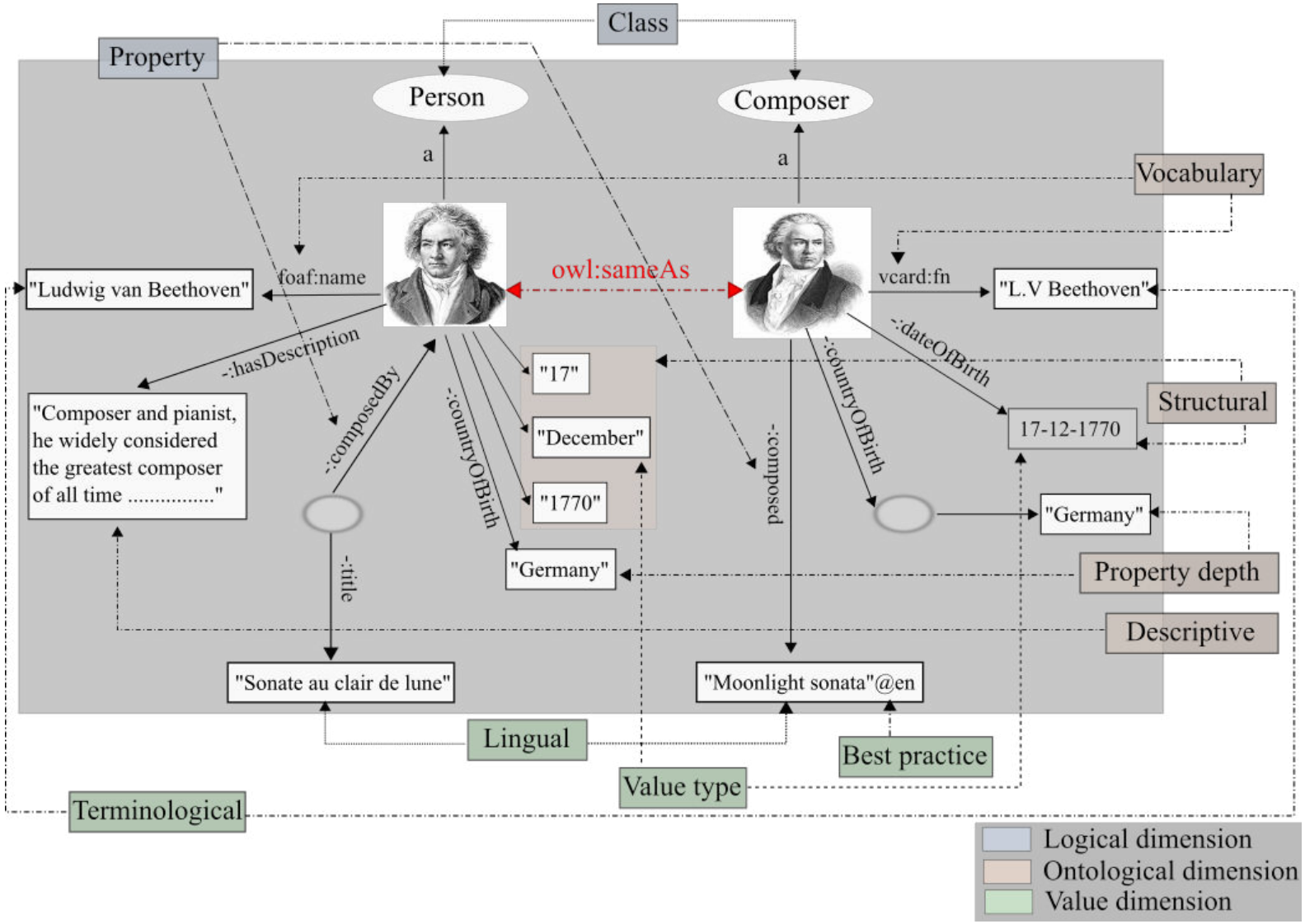


Modèle de DOREMUS

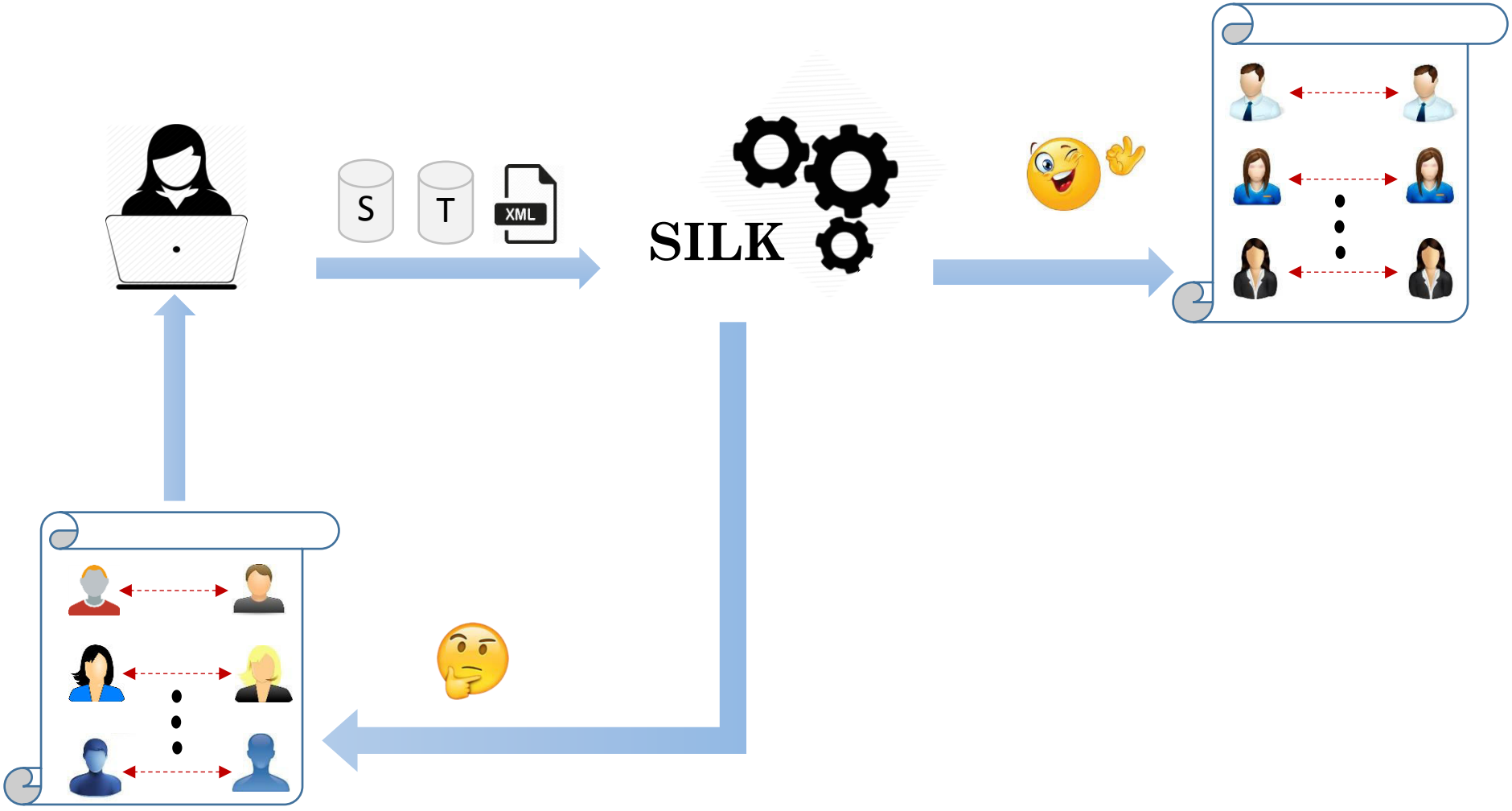


Données de DOREMUS

Liage des Données: Challenges



SILK



SILK : Fichier de Configuration

1. Les préfixes
2. Les sources de données
    a. Jeu de données "source"
    b. Jeu de données "target"
3. Les types
    a. Lien à générer
    b. Ressources à comparer
4. Les règles de liage
    a. Mesures de similarité
    b. Propriétés à comparer
5. Les Paramètres de sortie
    a. Liens sûrs
    b. Liens à vérifier

SILK : Fichier de Configuration

```
<?xml version="1.0" encoding="UTF-8"?>
<Silk>
<Prefixes>
<Prefix id="rdf" namespace="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>
<Prefix id="property" namespace="http://example.property.org/ontology#" />
</Prefixes>
<DataSources>
<DataSource type="file" id="file1">
<Param name="file" value="C:/filePath/source.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
<DataSource type="file" id="file2">
<Param name="file" value="C:/filePath/target.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
</DataSources>
<Interlinks>
<Interlink id="resources">
<LinkType>owl:sameAs</LinkType>
<SourceDataset dataSource="file1" var="a">
<RestrictTo> ?a a property:resourceType . </RestrictTo>
</SourceDataset>
<TargetDataset dataSource="file2" var="b">
<RestrictTo> ?b a property:resourceType . </RestrictTo>
</TargetDataset>
<LinkageRule>
<Aggregate type="average">
<Compare metric="levenshtein" threshold="1" required="true">
<Input path="?a/property:p1" />
<Input path="?b/property:p2" />
</Compare>
</Aggregate>
<Filter limit="1" />
</LinkageRule>
<Outputs>
<Output  type="file" minConfidence="0.8">
<Param name="file" value="results.rdf" />
<Param name="format" value="alignment" />
</Output>
<Output  type="file" maxConfidence="1.0">
<Param name="file" value="verify.rdf" />
<Param name="format" value="alignment" />
</Output>
</Outputs>
</Interlink>
</Interlinks>
</Silk>
```

Préfixes

Dataset source

Dataset target

Type de lien

Types de ressources à comparer

Propriétés à comparer

Seuil de similarité

Format d'alignement

Fichier de sortie "liens sûrs"

Fichier de sortie "liens à vérifier"

SILK : Les sources de données

1. Chemins vers des dumps en RDF

```
<DataSources>
<DataSource type="file" id="file1">
<Param name="file" value="C:/filePath/source.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
<DataSource type="file" id="file2">
<Param name="file" value="C:/filePath/target.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
</DataSources>
```

2. Lien vers le SPARQL endpoint

```
<DataSource id="dbpedia">
    <EndpointURI> http://dbpedia.org/sparql </EndpointURI>
    <Graph> http://dbpedia.org </Graph>
    <PageSize>10000</PageSize>
</DataSource>
<DataSource id="geonames">
    <EndpointURI> http://localhost:8890/sparql </EndpointURI>
</DataSource>
```

SILK : Propriétés à comparer

1. Possibilité de comparer plusieurs paires de propriétés --> plusieurs blocs de <compare> … </compare>

2. Plusieurs mesures de similarité

3. Plusieurs fonctions de transformation: tokenisation, normalisation, concaténation, etc.

```
36   <Aggregate type="average">
37   <Compare metric="levenshtein" threshold="1" required="true">
38       <TransformInput function="tokenize">
39           <Input path="?a/property:p1" />
40       </TransformInput>
41       <TransformInput function="tokenize">
42           <Input path="?b/property:p2" />
43       </TransformInput>
44   </Compare>
45   </Aggregate>
```

SILK : Mesures de similarité

1. jaroSimilarity: Similarité entre chaines de caratcères basée sur la mesure de Jaro.

2. jaroWinklerSimilarity: Similarité entre chaines de caratcères basée sur la mesure de Jaro-Winkler.

3. qGramSimilarity: Similarité entre chaines de caratcères basée sur Q-Grams.

4. stringEquality: Retourne 1 si les chaînes de caratères sont égales, sinon 0.

5. numSimilarity: similarité numérique

6. dateSimilarity: similarité entre 2 dates

7. uriEquality: Retourne 1 si deux URIs sont égales, sinon 0.

8. taxonomicSimilarity: Mesure basée sur la distance entre deux concepts.

9. maxSimilarityInSet: Retourne la similarité la plus élevée entre un seul élément à tous les éléments d'un ensemble.

10. setSimilarity: similarité entre 2 ensemble d'éléments.