

JDEV 2017

Marseille / 4-7 juillet



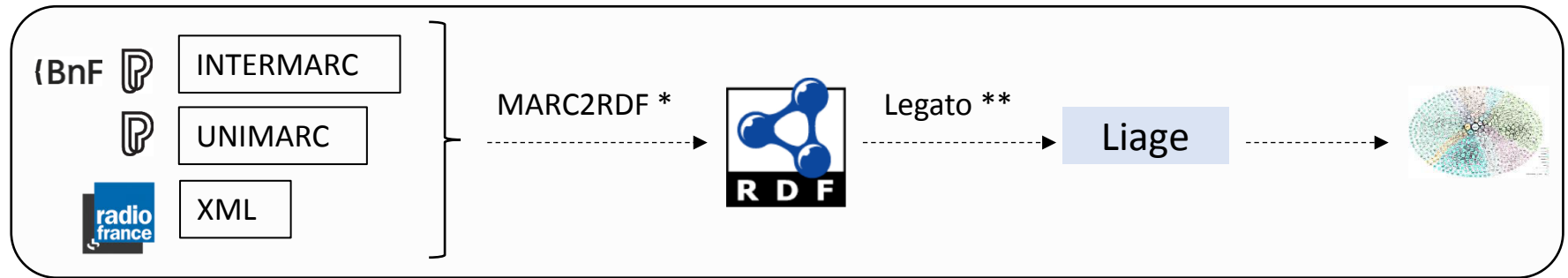
Interconnexion de Données du Web avec SILK

Manel Achichi & Konstantin Todorov

LIRMM / Université de Montpellier



Projet DOREMUS

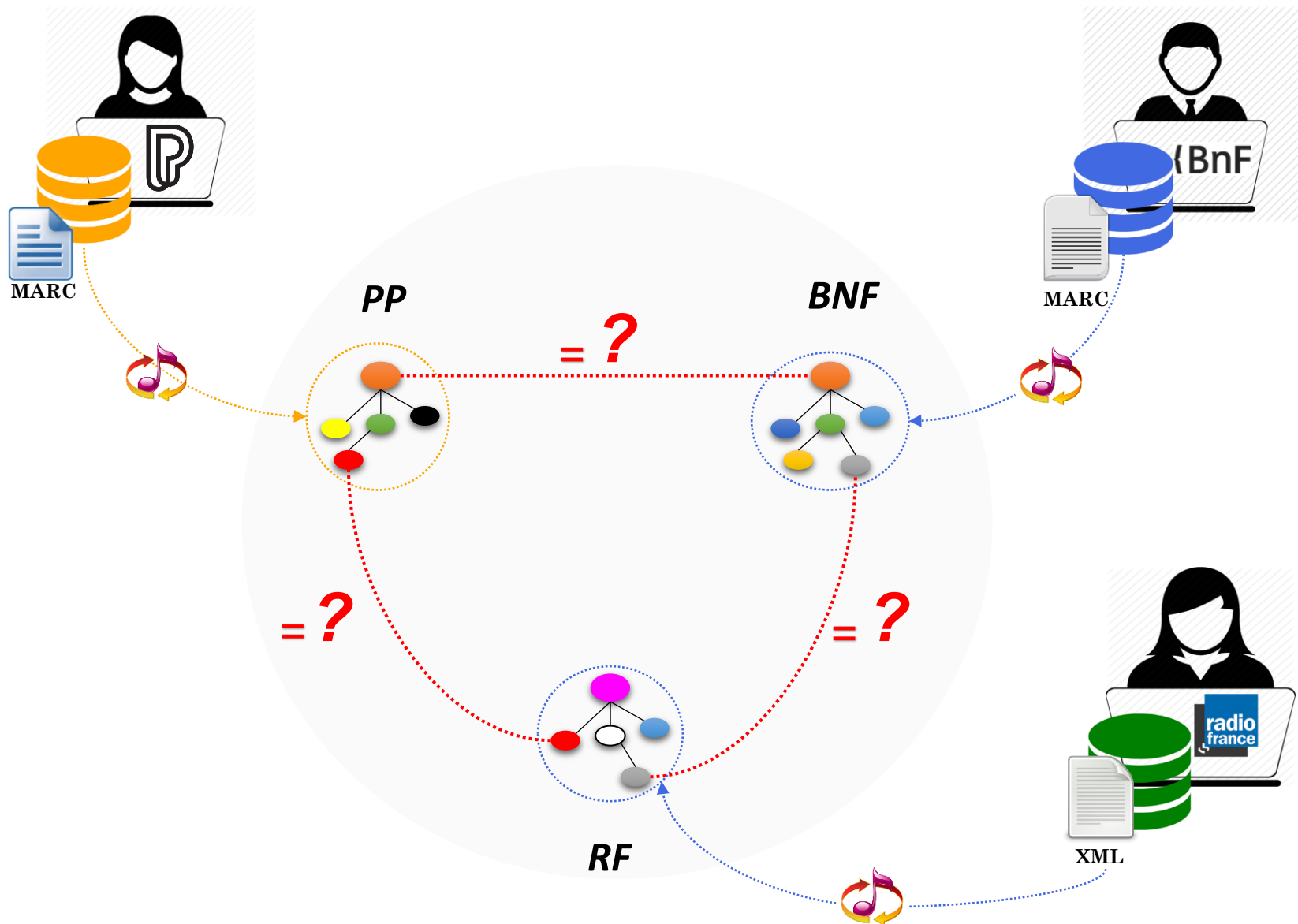


MEANING ENGINES

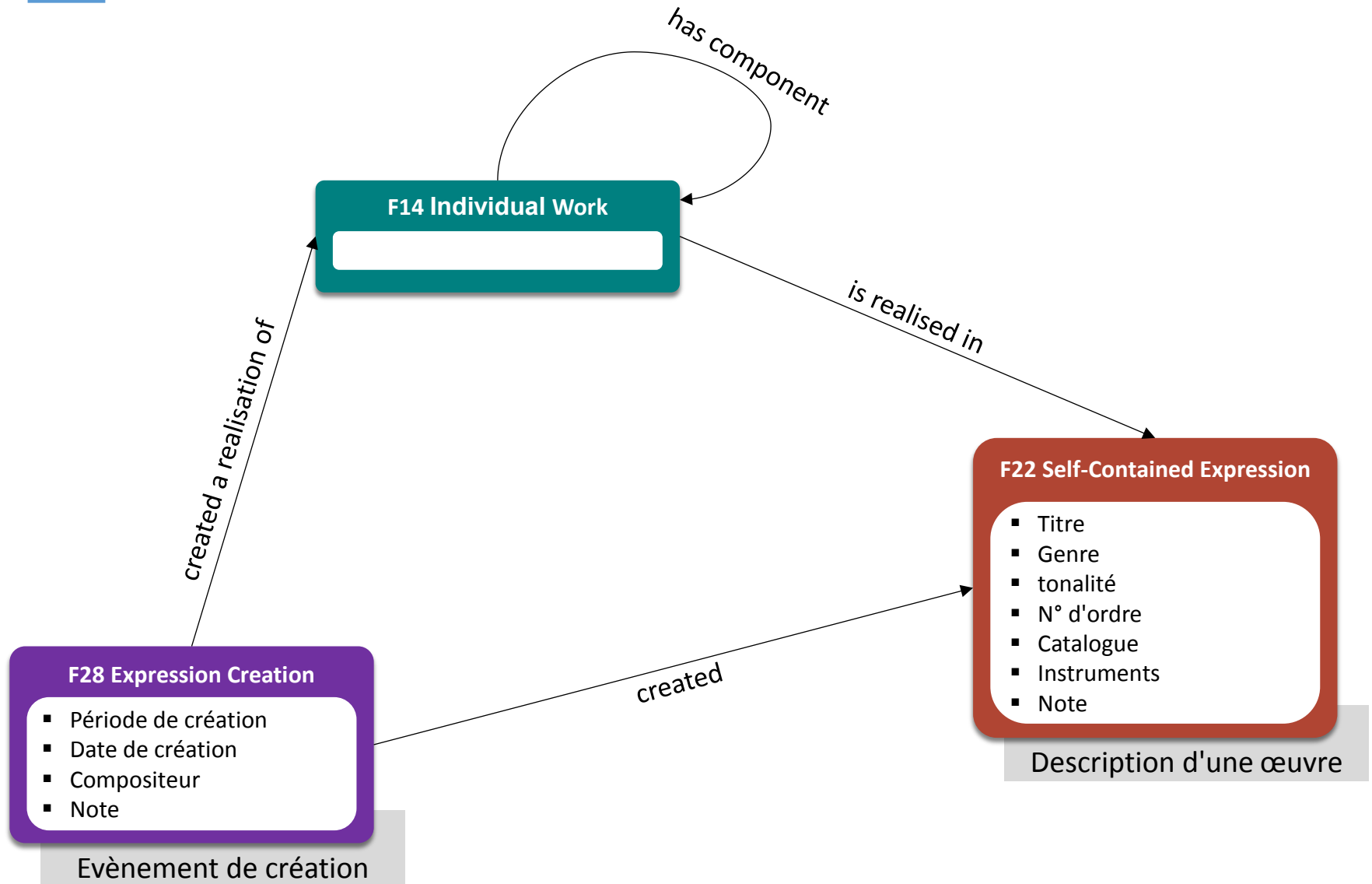
(*) : <https://github.com/DOREMUS-ANR/marc2rdf>

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

Projet DOREMUS



Modèle de DOREMUS



Données de DOREMUS

created

<http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>

F22

a efrbroo:F22_Self-Contained_Expression ;
mus:U70_has_title "Clair de lune"@fr, "Sonate Clair de lune"@fr, "Quasi una fantasia"@it,
"Mondschein-Sonate"@de, "Sonates"@fr, "Quasi una fantasia"@it, "Sonata quasi una fantasia"@it,
"Moonlight sonata"@en ;
mus:U10_has_order_number "14"^^xsd:int ;
mus:U11_has_key <http://data.doremus.org/vocabulary/key/cxm> ;
mus:U12_has_genre <http://data.doremus.org/vocabulary/iaml/genre/sn> ;
mus:U13_has_casting <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea/casting/1> ;
mus:U17_has_opus_statement <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea/opus/27-2> ;
dcterms:identifier "13908188" ;

<http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892>

F28

a efrbroo:F28_Expression_Creation ;
efrbroo:R17_created <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>;
efrbroo:R19_created_a_realisation_of <http://data.doremus.org/work/30256b51-d277-3688-ad62-560ae982ff2f> ;
ecrm:P9_consists_of <http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892/activity/1> ;
ecrm:P4_has_time-span <http://data.doremus.org/event/3f9d2fae-da75-3c66-902d-fa3a0755d892/time> ;

<http://data.doremus.org/work/30256b51-d277-3688-ad62-560ae982ff2f>

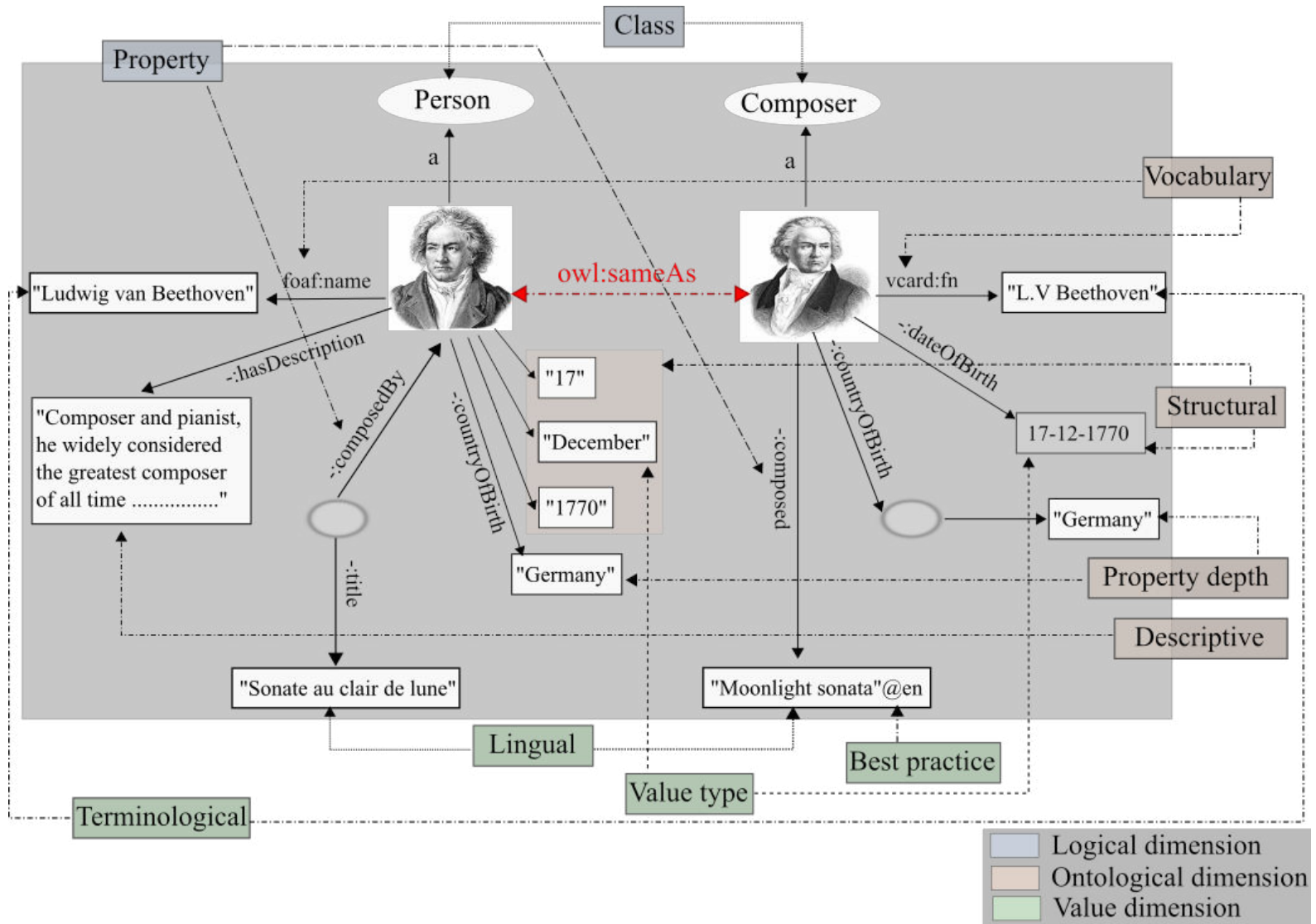
F14

a efrbroo:F14_Individual_Work ;
efrbroo:R9_is_realised_in <http://data.doremus.org/expression/d72301f0-0aba-3ba6-93e5-c4efbee9c6ea>;

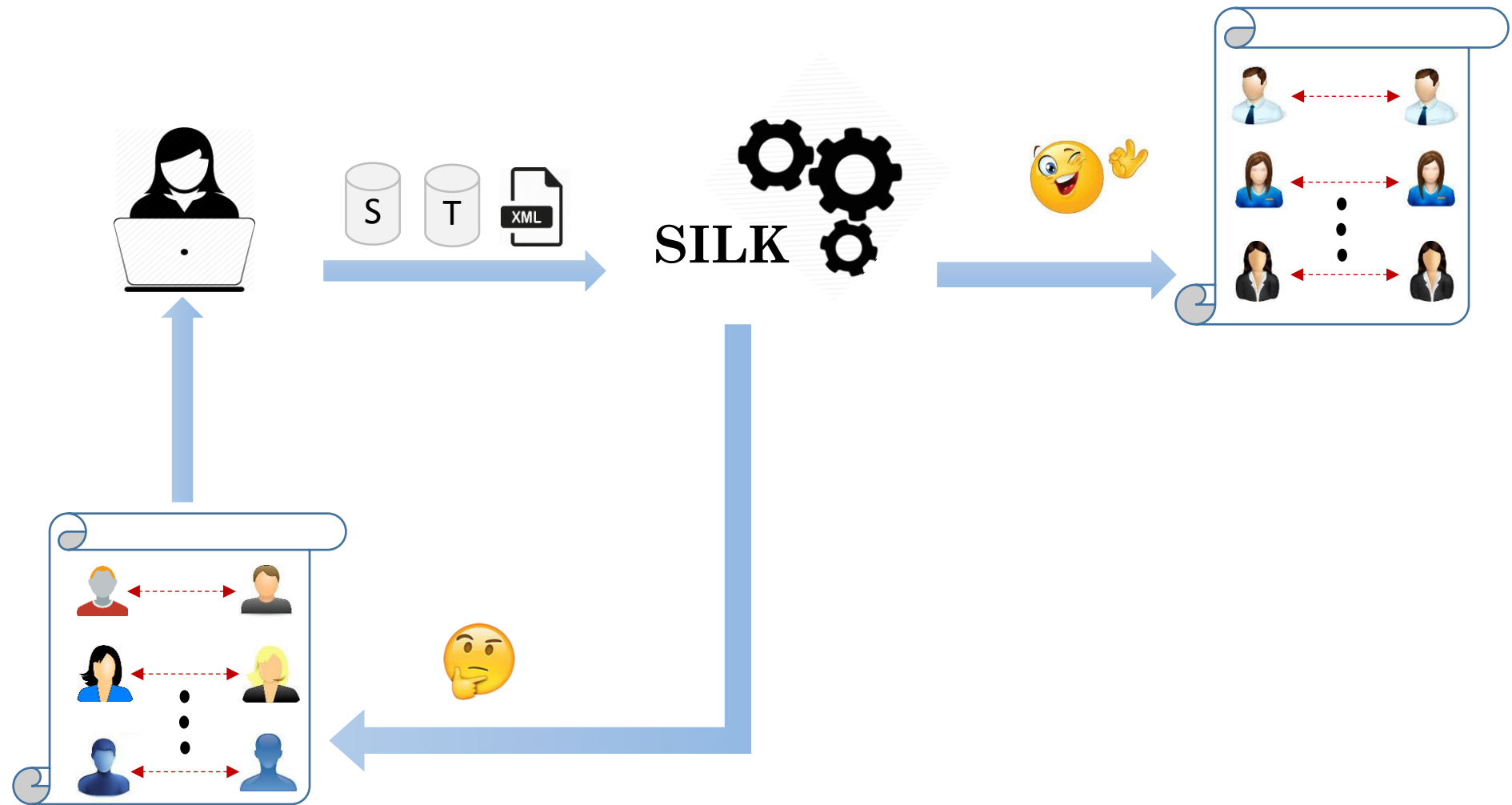
created a realisation of

is realised in

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

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <Silk>
3 <Prefixes>
4 <Prefix id="rdf" namespace="http://www.w3.org/1999/02/22-rdf-syntax-ns#" />
5 <Prefix id="property" namespace="http://example.property.org/ontology#" />
6 </Prefixes>
7 <DataSources>
8 <DataSource type="file" id="file1">
9 <Param name="file" value="C:/filePath/source.ttl" />
10 <Param name="format" value="TURTLE"/>
11 </DataSource>
12 <DataSource type="file" id="file2">
13 <Param name="file" value="C:/filePath/target.ttl" />
14 <Param name="format" value="TURTLE"/>
15 </DataSource>
16 </DataSources>
17 <Interlinks>
18 <Interlink id="resources">
19 <LinkType>owl:sameAs</LinkType>
20 <SourceDataset dataSource="file1" var="a">
21 <RestrictTo ?a a property:resourceType . </RestrictTo>
22 </SourceDataset>
23 <TargetDataset dataSource="file2" var="b">
24 <RestrictTo ?b a property:resourceType . </RestrictTo>
25 </TargetDataset>
26 <LinkageRule>
27 <Aggregate type="average">
28 <Compare metric="levenshtein" threshold="1" required="true">
29 <Input path="?a/property:p1" />
30 <Input path="?b/property:p2" />
31 </Compare>
32 </Aggregate>
33 <Filter limit="1" />
34 </LinkageRule>
35 <Outputs>
36 <Output type="file" minConfidence="0.8">
37 <Param name="file" value="results.rdf" />
38 <Param name="format" value="alignment" />
39 </Output>
40 <Output type="file" maxConfidence="1.0">
41 <Param name="file" value="verify.rdf" />
42 <Param name="format" value="alignment" />
43 </Output>
44 </Outputs>
45 </Interlink>
46 </Interlinks>
47 </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

```
7   <DataSources>
8   <DataSource type="file" id="file1">
9     <Param name="file" value="C:/filePath/source.ttl" />
10    <Param name="format" value="TURTLE"/>
11  </DataSource>
12  <DataSource type="file" id="file2">
13    <Param name="file" value="C:/filePath/target.ttl" />
14    <Param name="format" value="TURTLE"/>
15  </DataSource>
16 </DataSources>
```

2. Lien vers le SPARQL endpoint

```
7   <DataSource id="dbpedia">
8     <EndpointURI> http://dbpedia.org/sparql </EndpointURI>
9     <Graph> http://dbpedia.org </Graph>
10    <PageSize>10000</PageSize>
11  </DataSource>
12  <DataSource id="geonames">
13    <EndpointURI> http://localhost:8890/sparql </EndpointURI>
14  </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 chaînes de caractères basée sur la mesure de Jaro.
2. jaroWinklerSimilarity: Similarité entre chaînes de caractères basée sur la mesure de Jaro-Winkler.
3. qGramSimilarity: Similarité entre chaînes de caractères basée sur Q-Grams.
4. stringEquality: Retourne 1 si les chaînes de caractères 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.