JDEV 2017

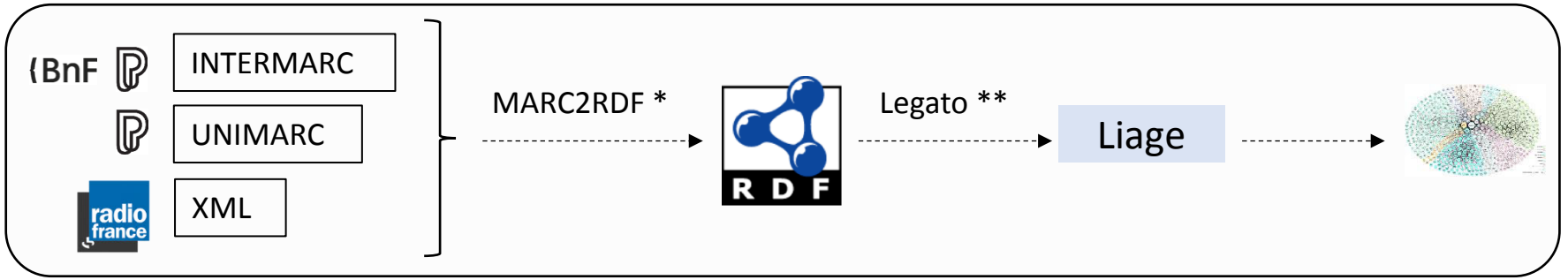
Marseille / 4-7 juillet

Interconnexion de Données du Web avec SILK

Manel Achichi & Konstantin Todorov

LIRMM / Université de Montpellier

Projet DOREMUS

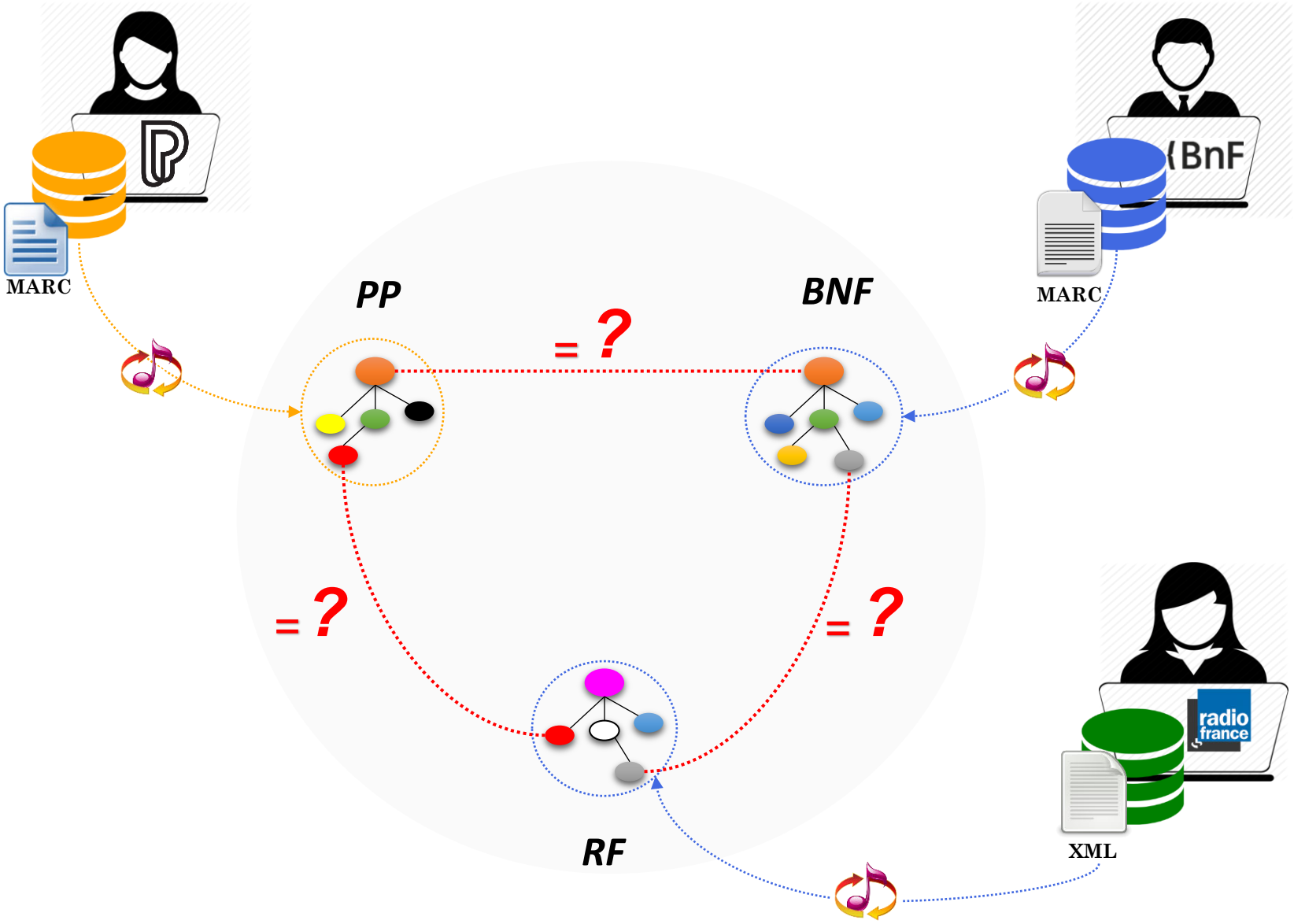


LIRMM EURECOM ROUEN Gérico MEANING ENGINES

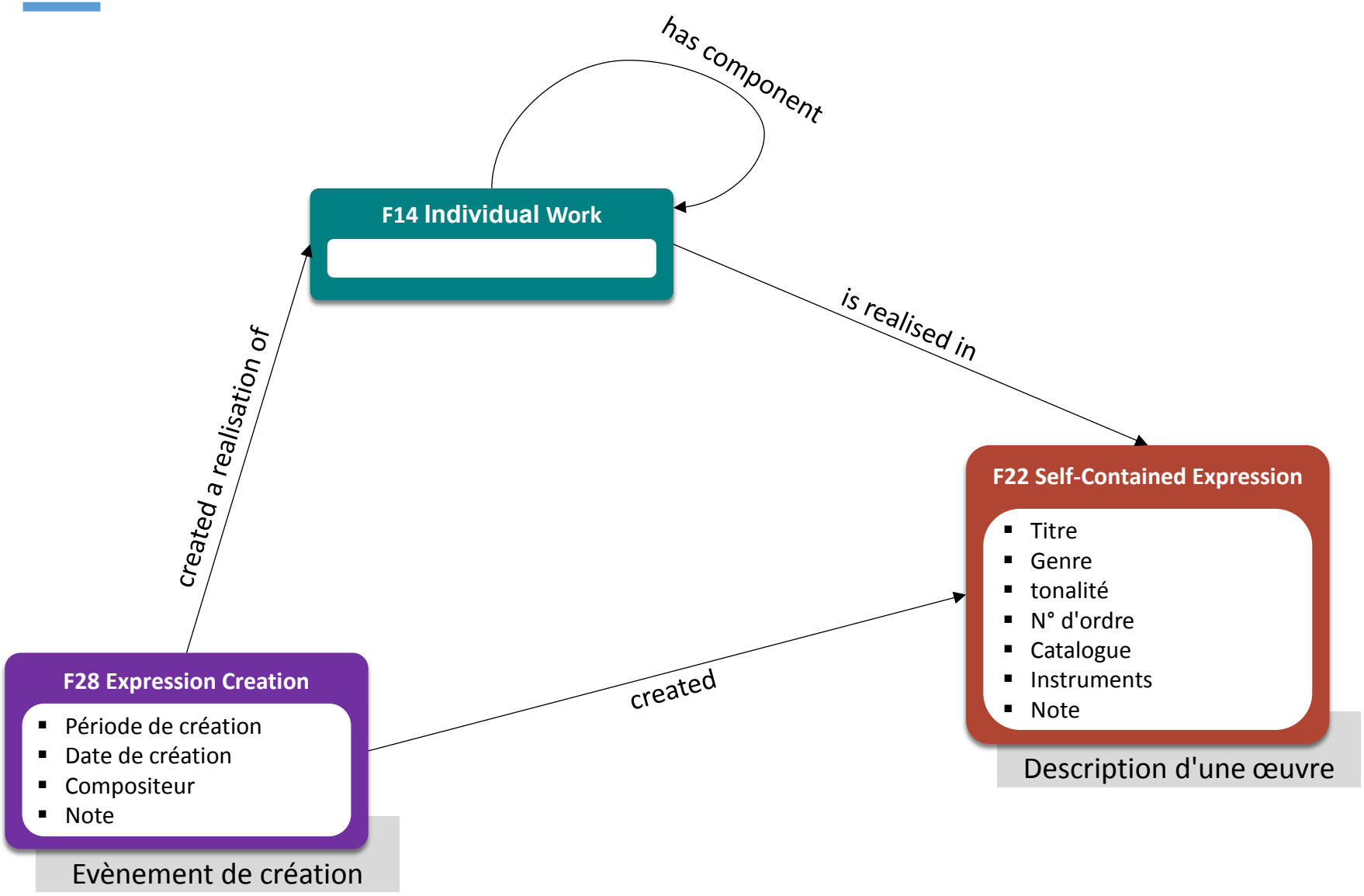
(*) : https://github.com/DOREMUS-ANR/marc2rdf

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

Projet DOREMUS

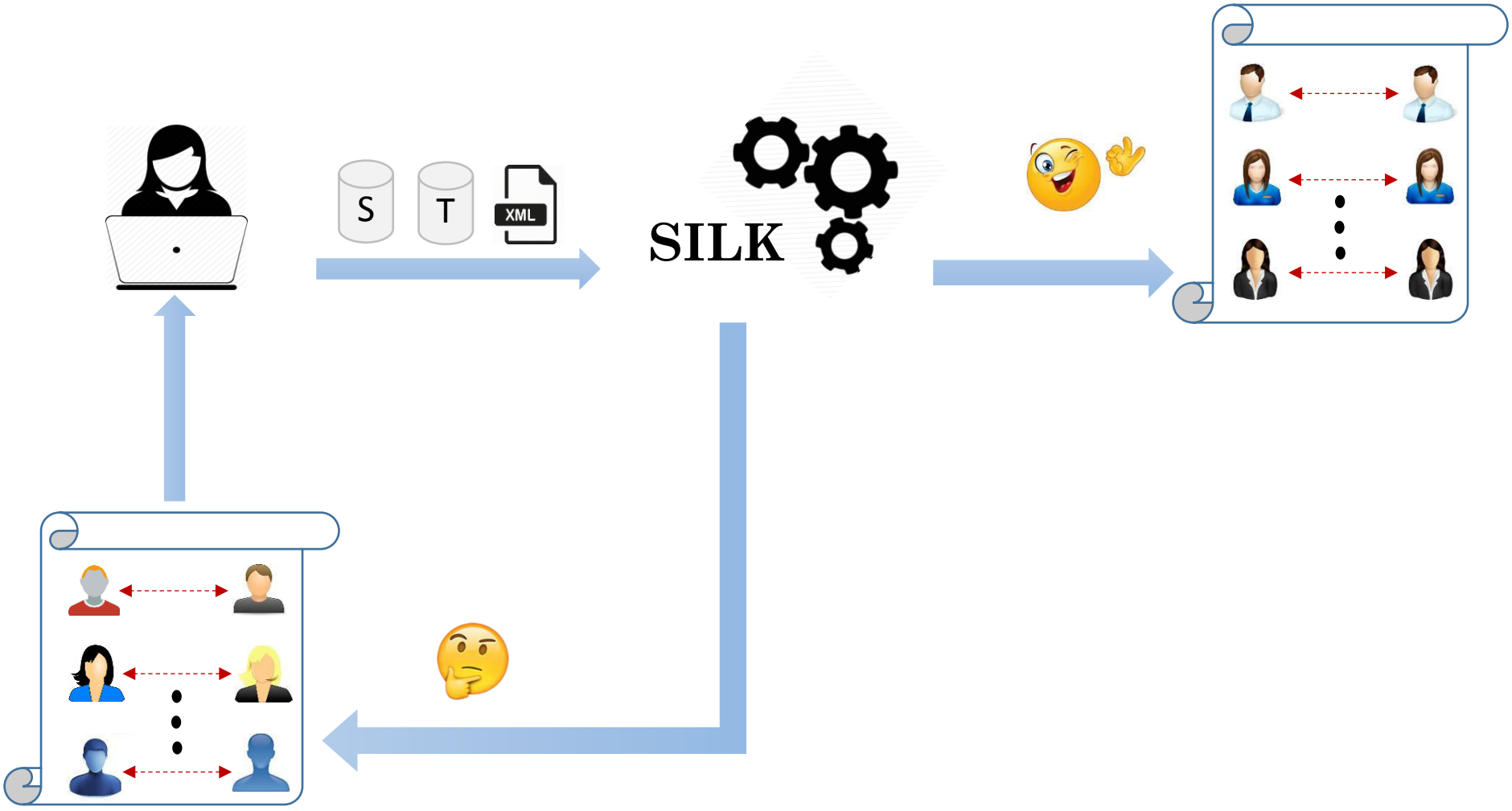


Modèle de DOREMUS

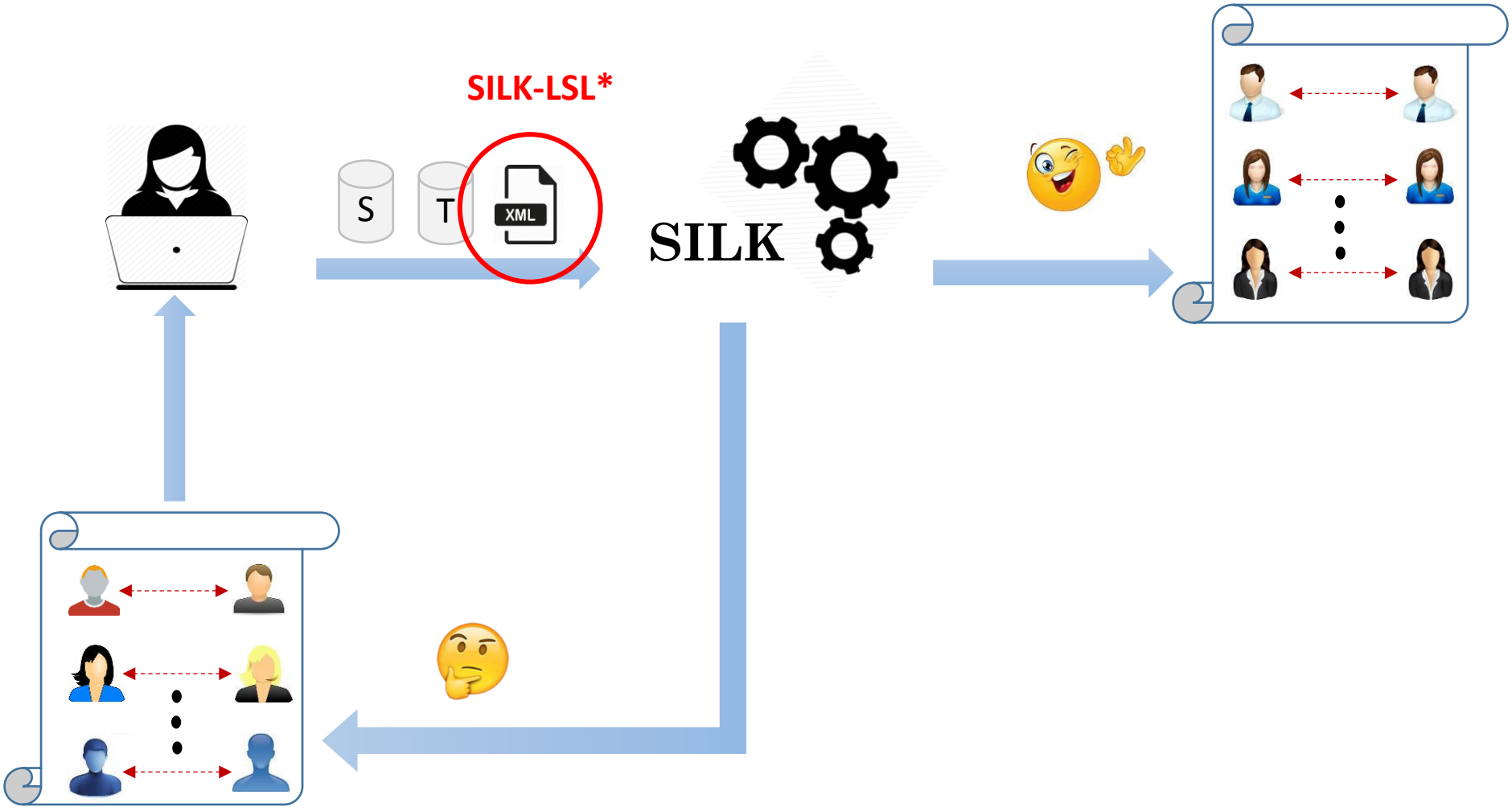


Données de DOREMUS

SILK: Framework de Découverte de Liens



SILK: Framework de Découverte de Liens



(*) LSL: Link Specification Language

SILK-LSL

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-LSL

```
<?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

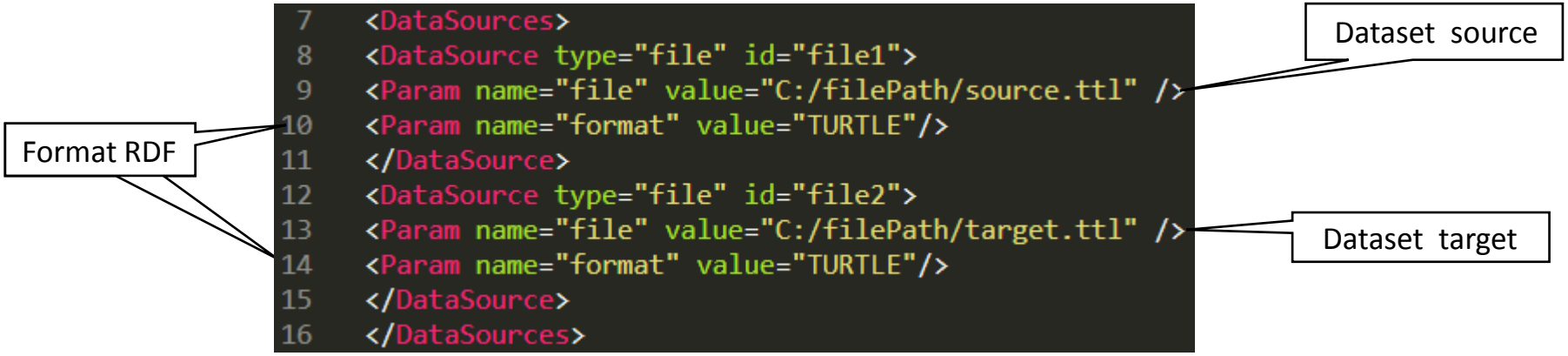
Format d'alignement

Fichier de sortie "liens sûrs"

Fichier de sortie "liens à vérifier"

SILK-LSL : Les sources de données

1. Chemins vers des dumps en RDF



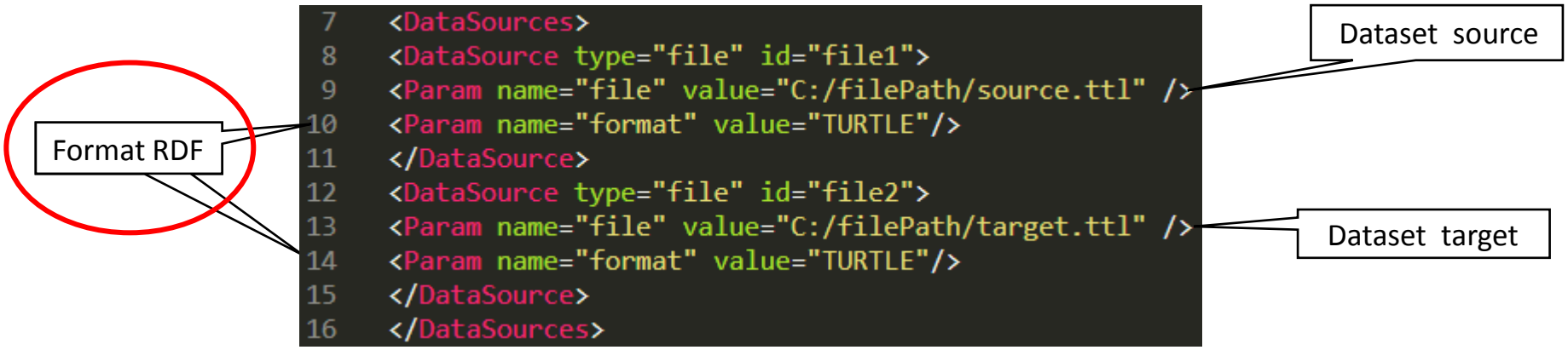

```
<DataSources>
<DataSource type="file" id="file1">
<Param name="file" value="C:/filePath/source.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
<DataSource type="file" id="file2">
<Param name="file" value="C:/filePath/target.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
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

1. Chemins vers des dumps en RDF

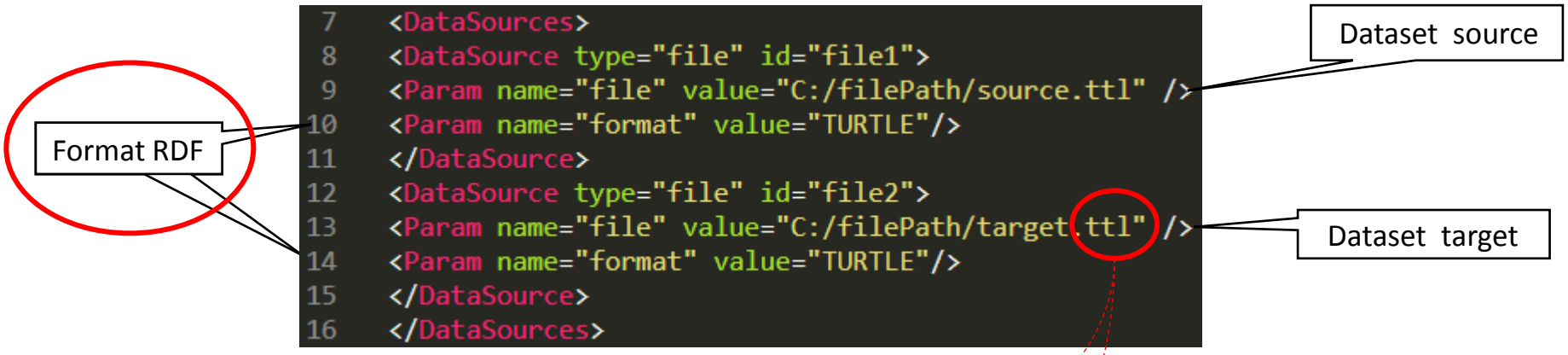



```
<DataSources>
<DataSource type="file" id="file1">
<Param name="file" value="C:/filePath/source.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
<DataSource type="file" id="file2">
<Param name="file" value="C:/filePath/target.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
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

Format ("RDF/XML", "N-TRIPLE", "TURTLE", "TTL", "N3").

SILK-LSL : Les sources de données

1. Chemins vers des dumps en RDF

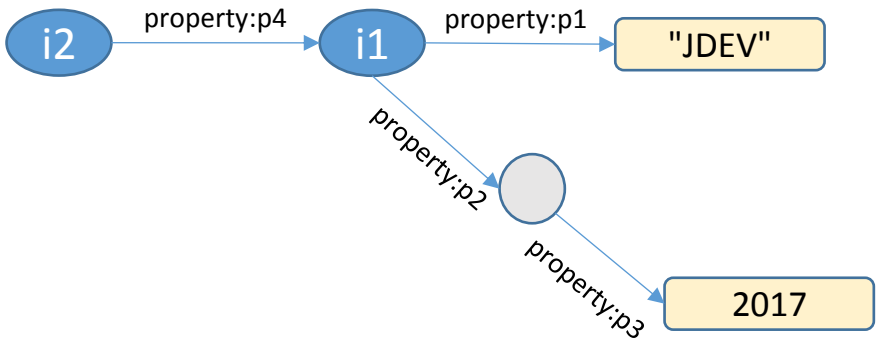



```
<DataSources>
<DataSource type="file" id="file1">
<Param name="file" value="C:/filePath/source.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
<DataSource type="file" id="file2">
<Param name="file" value="C:/filePath/target.ttl" />
<Param name="format" value="TURTLE"/>
</DataSource>
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

Format ("RDF/XML", "N-TRIPLE", "TURTLE", "TTL", "N3").

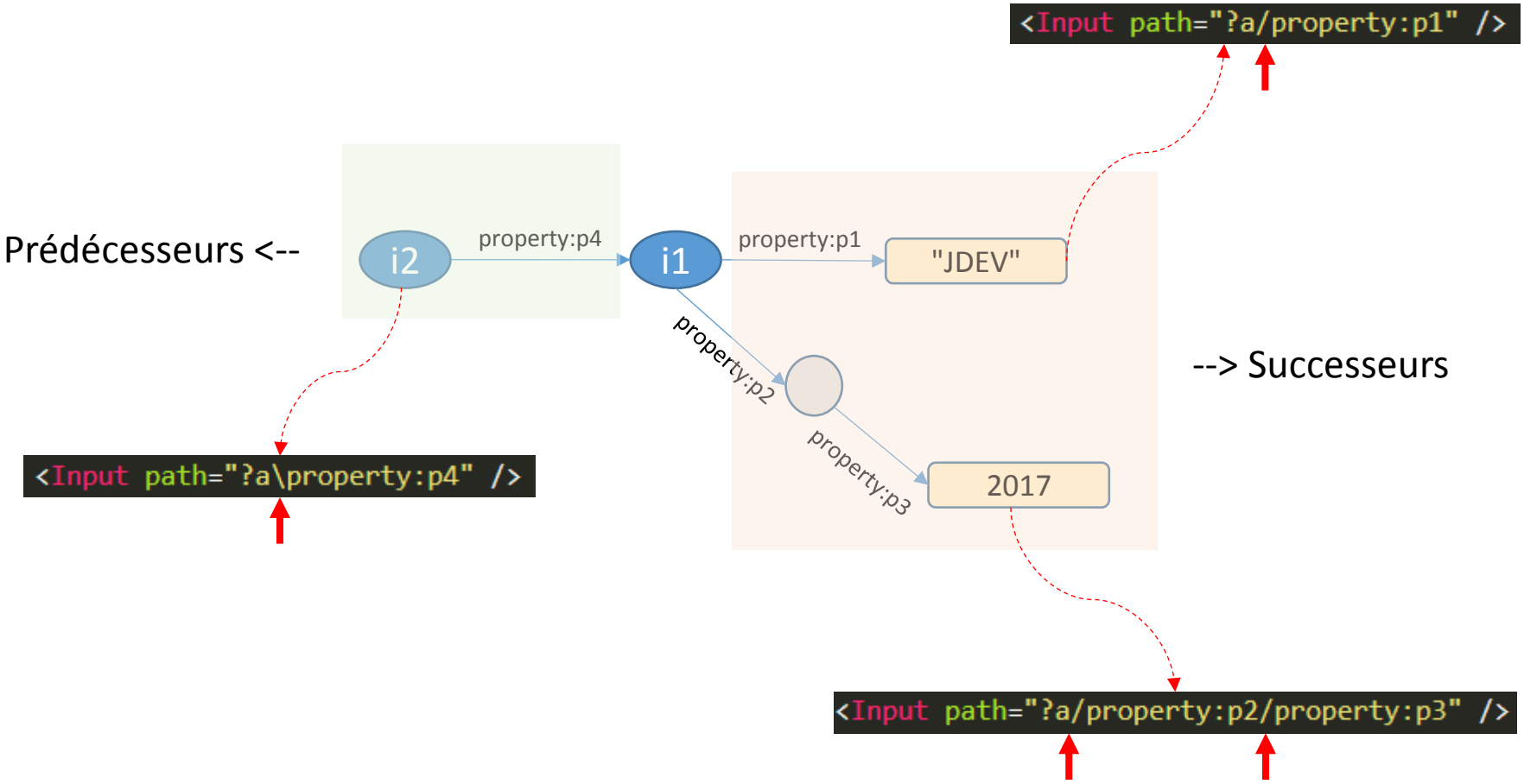
SILK-LSL : Régles de Liage

1. Chemin vers les propriétés à comparer

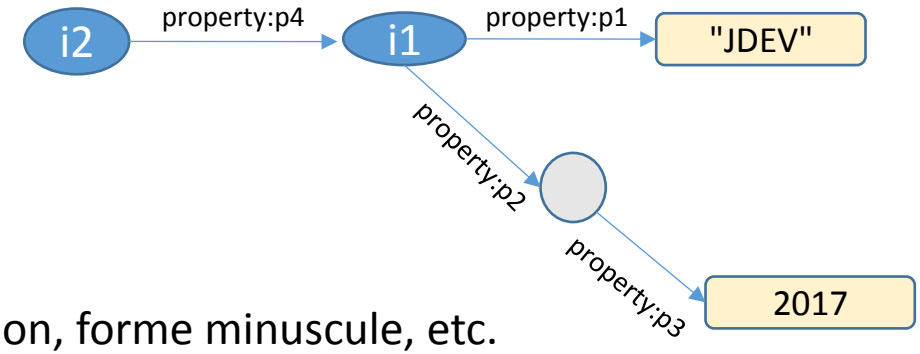


SILK-LSL : Régles de Liage

1. Chemin vers les propriétés à comparer



SILK-LSL : Régles de Liage



1. Chemin vers les propriétés à comparer

2. Fonctions de transformation: tokenisation, forme minuscule, etc.

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

3. Comparaison: mesures de similarité, seuil, etc.

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

SILK-LSL : Aggrégations

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

Type	Description
average	La moyenne des valeurs de confiance
min	La valeur de confiance la moins élevée
max	La valeur de confiance la plus élevée

- Valeur de confiance: SILK convertit une distance calculée (par exemple la valeur 2 de Levenshtein) en une valeur de confiance entre -1 et +1.

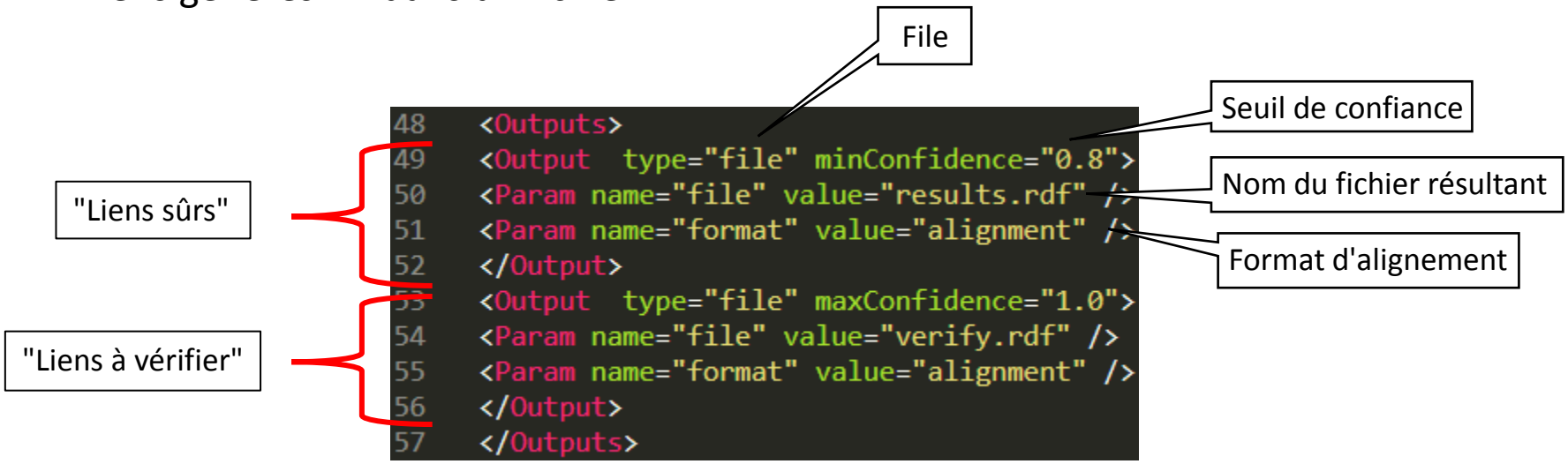
  Plus la valeur augmente et plus la similarité entre les instances augmente

SILK-LSL : Mesures de similarité

Mesure de similarité	Description
levenshteinDistance	Nombre minimum d'éditions (insertion/suppression/substitution) pour transformer un string à un autre
levenshtein	Distance levenshtein normalisée dans un intervalle de [0,1]
jaro	Similarité entre chaines de caratcères basée sur la mesure de Jaro
jaroWinkler	Similarité entre chaines de caratcères basée sur la mesure de Jaro-Winkler
equality	Retourne 0 si les chaînes de caratères sont égales, sinon 1
date	similarité entre 2 dates au format AAAA-MM-JJ
jaccard	Similarité entre chaines de caratcères basée sur la distance de Jaccard

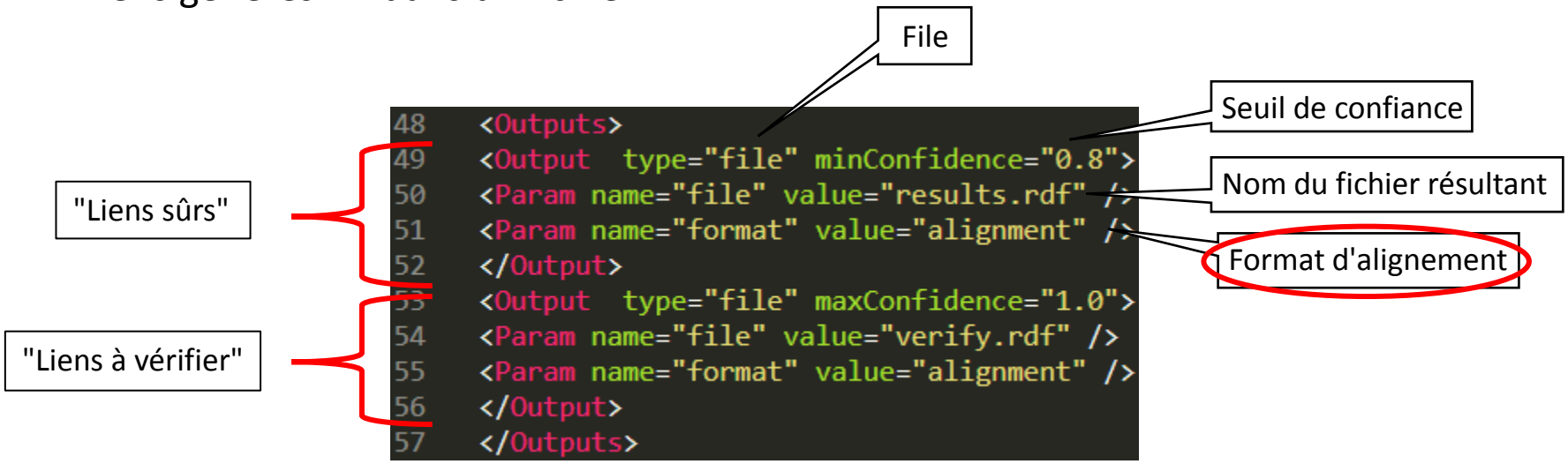
SILK-LSL : Les fichiers de sortie

- Liens générés --> dans un fichier RDF



SILK-LSL : Les fichiers de sortie

- Liens générés --> dans un fichier RDF



Format d'alignement ("N-TRIPLE", "alignement").

Possibilité de définir de 1 à n (n>1) blocs de <output> … </output>

SILK-LSL : Définir plusieurs fichiers de sortie

```
59  <Outputs>
60      <Output  type="file" minConfidence="lower threshold" maxConfidence="upper threshold">
61          <Param name="file" value="file name" />
62          <Param name="format" value="alignment format" />
63      </Output>
64      ...
65      <Output  type="file" minConfidence="lower threshold" maxConfidence="upper threshold">
66          <Param name="file" value="file name" />
67          <Param name="format" value="alignment format" />
68      </Output>
69  </Outputs>
```

SILK: Mode Ligne de Commande

- Entrées:

  - 2 jeux de données:

  - 1 fichier de configuration (LSL):
  Un exemple détaillé --> https://app.assembla.com/wiki/show/silk/Link_Specification_Language

  - "SILK.jar":
  https://github.com/silk-framework/silk/releases (la version de SILK utilisée dans ce tutoriel est 2.6.1)

- Commande:

  java -DconfigFile= configFile.xml -jar silk.jar

SILK: Exo-1

- Entrées:

  - 2 jeux de données "source" et "target": https://github.com/manoach/JDEV2017-SILK-/tree/master/DFP

  - 1 fichier de configuration (LSL):
    Un exemple détaillé --> https://app.assembla.com/wiki/show/silk/Link_Specification_Language

  - "SILK.jar": https://github.com/manoach/JDEV2017-SILK-/tree/master/DFP

- A faire:

  - Lier les instances de type <http://erlangen-crm.org/efrbroo/F22_Self-Contained_Expression> en comparant leur(s):
    + Titres <http://erlangen-crm.org/current/P102_has_title> (en tokenisant les valeurs)
    + Numéro d'ordre <http://data.doremus.org/ontology#U10_has_order_number>
    + La note de leur opus <http://data.doremus.org/ontology#U17_has_opus_statement>

  - Mesure : Levenshtein
  - Seuil Levenshtein = 2
  - Seuil d'aggrégation = 0.8
  - Mapping attendu --> 1:1 (une instance "source" possède au plus une "instance" target).

SILK: Exo-2

- Entrées:

  - 2 jeux de données "source" et "target": https://github.com/manoach/JDEV2017-SILK-/tree/master/DHT

  - 1 fichier de configuration (LSL):
    Un exemple détaillé --> https://app.assembla.com/wiki/show/silk/Link_Specification_Language

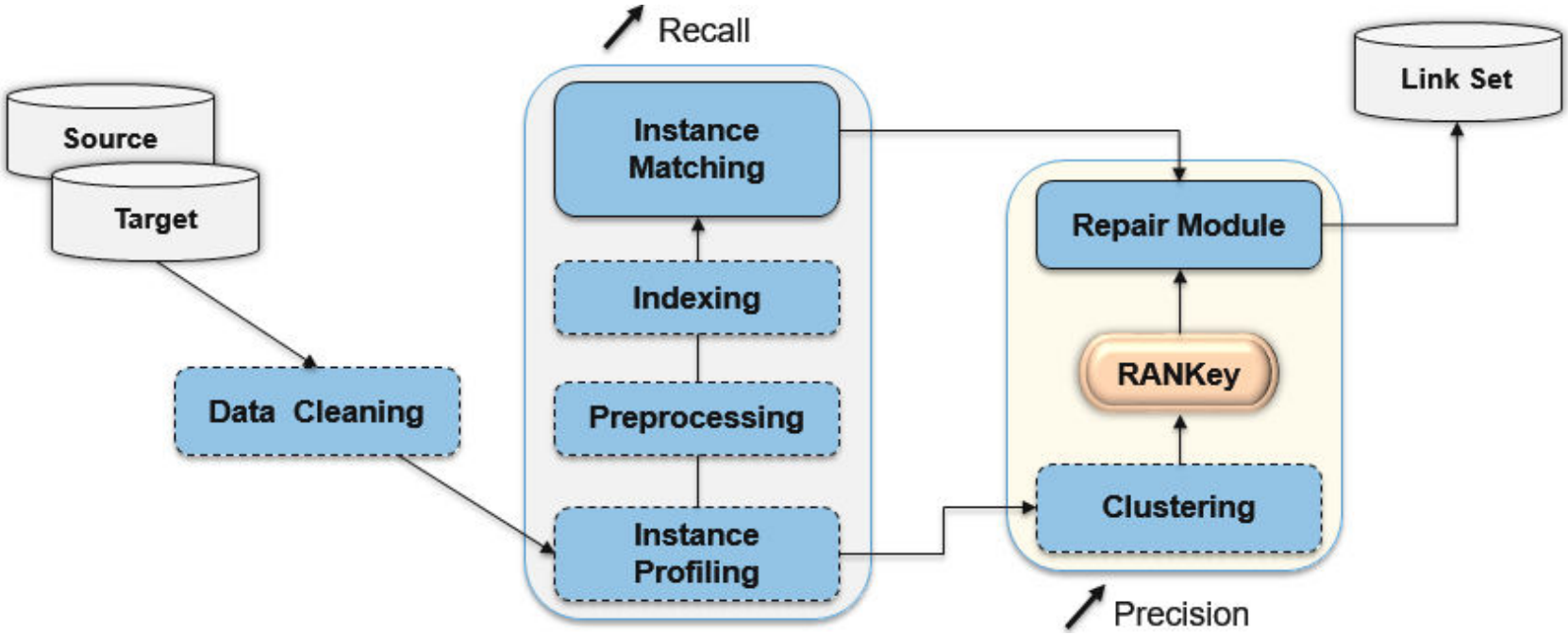
  - "SILK.jar": https://github.com/manoach/JDEV2017-SILK-/tree/master/DHT

- A faire:

  - Lier les instances de type <http://erlangen-crm.org/efrbroo/F22_Self-Contained_Expression> en comparant leur(s):

    + Notes <http://erlangen-crm.org/current/P3_has_note> (avec les paramètres de votre choix)

Legato : Workflow



Legato: Instance Profiling