IDENTITY MANAGEMENT IN THE WEB OF DATA

FATIHA SAÏS

LRI, PARIS SUD UNIVERSITY, CNRS, ORSAY, FR
OPENNESS AND PRIVACY BALANCE

[source1] https://cdn-images-1.medium.com/max/2600/1*xHzO_5cSSVetWnjpAbQABw.png
OPENNESS AND PRIVACY BALANCE

Privacy
Open data contains the most detailed information, granular data often includes *personally sensitive* information.

Openness
Open data enables varied and detailed analyses, granular data is the most *interesting and useful* for businesses, policymakers, researchers, and the public.

[source1] https://cdn-images-1.medium.com/max/2600/1*xHzO_5cSSVetWnjpAbQABw.png
OPEN DATA
LINKED OPEN DATA
FROM THE WWW TO THE LINKED OPEN DATA

- applying the principles of the WWW to data

data is links, not only properties
LINKED DATA PRINCIPLES

① Use HTTP URIs as identifiers for resources
   → so people can look up the data

② Provide data at the location of URIs
   → to provide data for interested parties

③ Include links to other resources
   → so people can discover more information
   → bridging disciplines and domains
   → unlock the potential of isolated repositories (islands)
RDF – RESOURCE DESCRIPTION FRAMEWORK

• Statements of < subject predicate object >

http://dbpedia.org/resource/Louvre

dbo:hasLocation

“Paris”

Subject

Predicate

Object

… is called a triple
**LINKED OPEN DATA**

Linked Data - Datasets under an open access
- 1,139 datasets
- over 100B triples
- about 500M links
- several domains

Ex. DBPedia : 1.5 B triples

NEED OF KNOWLEDGE
The knowledge principle: “if a program is to perform a complex task well, it must know a great deal about the world in which it operates.”
ONTOMETRY, A DEFINITION

“An ontology is an **explicit, formal specification** of a **shared conceptualization**.”

[Thomas R. Gruber, 1993]

**Conceptualization**: abstract model of domain related expressions

**Specification**: domain related

**Explicit**: semantics of all expressions is clear

**Formal**: machine-readable

**Shared**: consensus (different people have different perceptions)
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema
• Lightweight ontologies

OWL – Web Ontology Language
• Expressive ontologies

Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png
OWL ONTOLOGY

OWL – Web Ontology Language

- Represents rich and complex knowledge about things
- Based on Description Logic
- Can be used to verify the consistency of knowledge
- Can make implicit knowledge explicit

- **Classes:** concepts or collections of objects (individuals)
- **Properties:**
  - owl:DataTypeProperty (attribute)
  - owl:ObjectProperty (relation)
- **Hierarchy:**
  - owl:subClassOf
  - owl:subPropertyOf
- **Individuals:** ground-level of the ontology (instances)
ONTOMOLOGY LEVELS

Conceptual level:
- classes, properties (relations)

Instance level:
- facts (individuals)
OWL ONTOLOGY - AXIOMS

- **Axioms**: knowledge definitions in the ontology that were *explicitly defined* and have not been proven true.

- **Reasoning over an ontology**
  → Implicit knowledge can be made explicit by logical reasoning

- **Example**:
  
  Pompidou museum is an **Art Museum**
  
  `<Pompidou_museum rdf:type ArtMuseum>` .

  Pompidou museum contains **Musicircus**
  
  `<Pompidou_museum ao:contains Musicircus>` .

- **Infer that**:
  
  → Pompidou museum is a **CulturalPlace**
  
  `<Pompidou_museum rdf:type CulturalPlace>` .

  Because: **Museum** subsumes **ArtMuseum** and **CulturalPlace** subsumes **Museum**

  → **Musicircus** is a **Work**
  
  `<Musicircus rdf:type ao:Work>` .

  Because: the **range** of the object property **contains** is the class **Work**.
IDENTITY MANAGEMENT

- Detection of **identity links** between different descriptions of entities

- Discovery of **identification rules**, such as keys

- Detection of **erroneous identity links** and propose alternate links
OUTLINE

- Introduction
- Part 1: Data Linking
- Part 2: Key Discovery
- Part 3: Identity Link Invalidation
- Summary and Future Challenges
PART 1: DATA LINKING
Data linking or Identity link detection consists in detecting whether two descriptions of resources refer to the same real world entity (e.g. same person, same article, same gene).
DATA LINKING: DIFFICULTIES

- Data linking or Identity link detection consists in detecting whether two descriptions of resources refer to the same real world entity (e.g. same person, same article, same gene).

Incomplete Information:
- date and place of birth ?
- museum phone number ?
- .... ?

Misspelling errors

Different Vocabularies

NativeLabel

La Gioconda (it)

La Joconde

Mona Lisa

fb:Artist

fb:Painting

rdf:type

dbo:Author

wikidata:Q762

wikidata:Q12418

wikidata:Q90

http://fr.dbpedia.org/resource/Versailles

http://fr.dbpedia.org/resource/Musée_du_Louvre

http://fr.dbpedia.org/resource/Paris

http://fr.dbpedia.org/resource/La_Joconde

http://fr.dbpedia.org/resource/Musée_du_Louvre

http://fr.dbpedia.org/resource/Paris
DATA LINKING PROBLEM

- **Identity link detection:** detecting whether two descriptions of resources refer to the same real world entity (e.g. same person, same article, same gene).

**Definition (Link Discovery)**

- Given two sets $U_1$ and $U_2$ of resources
- Find a partition of $U_1 \times U_2$ such that:
  - $S = \{(s,t) \in U_1 \times U_2: \text{owl:sameAs}(s,t)\}$ and
  - $D = \{(s,t) \in U_1 \times U_2: \text{owl:differentFrom}(s,t)\}$

- A method is **total** when $(S \cup D) = (U_1 \times U_2)$
- A method is **partial** when $(S \cup D) \subset (U_1 \times U_2)$
- **Naïve complexity** $\in O(U_1 \times U_2)$, i.e. $O(n^2)$
Problem which exists since the data exists ... and under different terminologies: record linkage, entity resolution, data cleaning, object coreference, duplicate detection, data linkage ....

Automatic Linkage of Vital Records*

[SKAJ, Science 1959]

Computers can be used to extract “follow-up” statistics of families from files of routine records.


Record linkage: used to indicate the bringing together of two or more separately recorded pieces of information concerning a particular individual or family (1). Defined in this broad manner, it includes almost any use of a file of records to determine what has subsequently happened to people about whom one has some prior information.
DATA LINKING IS MORE COMPLEX FOR GRAPHS THAN TABLES (WHY?)

<table>
<thead>
<tr>
<th></th>
<th>Databases</th>
<th>Semantic Web</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schema/Ontologies</td>
<td>Same schema</td>
<td>Possibly different schema or ontologies</td>
</tr>
<tr>
<td>Multiple types</td>
<td>Single relation</td>
<td>Classes, hierarchically organized</td>
</tr>
<tr>
<td>Open World Assumption</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>UNA-Unique Name Assumption</td>
<td>Yes</td>
<td>May be no</td>
</tr>
<tr>
<td>Data volume</td>
<td>XX Thousands</td>
<td>XX Millions/Billions (e.g., DBpedia has 1.5 billion triples)</td>
</tr>
<tr>
<td>Multiple values for a property</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1 hasAuthor “Michel Chein”</td>
</tr>
<tr>
<td></td>
<td></td>
<td>P1 hasAuthor “Marie-Christine Rousset”</td>
</tr>
</tbody>
</table>

- Can propagate similarity decisions ➔ more expensive but better performance
- Can be generic and use domain knowledge, e.g. ontology axioms
DATA LINKING APPROACHES:
DIFFERENT CONTEXTS

• Datasets conforming to the same ontology

• Datasets conforming to different ontologies

• Datasets without ontologies
DATA LINKING APPROACHES

- **Local approaches**: consider data type properties to compare pairs of instances independently

  versus

- **Global approaches**: consider data type properties as well as object properties to propagate similarity scores/linking decisions (collective data linking)

- **Supervised approaches**: need samples of linked data to learn models, or need interactions with expert

  versus

- **Informed approaches**: need knowledge to be declared in the ontology or in other format
LOCAL APPROACHES

- Consider (path of) properties to compare pairs of instances independently
GLOBAL APPROACHES

- **Graph-based approaches**: (collective data linking): propagate similarity scores/linking decisions
SUPERVISED APPROACHES

- Need an expert to build samples of identity links to train models (or interactive approaches)

Examples of identity links

Learning of parameters, similarity functions, thresholds, ...

Identity link detection

Identity links
DATA LINKING APPROACHES: EVALUATION

- **Effectiveness**: evaluation of linking results in terms of recall and precision
  - **Recall** = (#correct-links-sys) / (#correct-links-groundtruth)
  - **Precision** = (#correct-links-sys) / (#links-sys)
  - **F-measure (F1)** = (2 x Recall x Precision) / (Recall + Precision)
DATA LINKING APPROACHES: EVALUATION

- **Effectiveness**: evaluation of linking results in terms of recall and precision
  - Recall = (#correct-links-sys) / (#correct-links-groundtruth)
  - Precision = (#correct-links-sys) / (#links-sys)
  - F-measure (F1) = (2 x Recall x Precision) / (Recall + Precision)

- **Efficiency**: in terms of time and space (i.e. minimize the linking search space and the interaction actions with an expert/user).

- **Robustness**: override errors in the data

- **Generality**: applicable to different datasets and different domains

- **Use of benchmarks**, like those of **OAEI** (Ontology Alignment Evaluation Initiative) or **Lance**
EXAMPLE:
KNOFUSS (LOCAL, UNSUPERVISED)

[Nikolov et al’12]

• Learns linking rules using genetic algorithms:

\[
\text{Sim}(i1, i2) = f_{ag}(w_{11}\text{sim}_{11}(V11,V21), \ldots w_{mn}\text{sim}_{mn}(V1m,V2n))
\]

• \(f_{ag}\): aggregation function for the similarity scores
• \(\text{sim}_{ij}\): similarity measure between values \(V1i\) and \(V2j\)
• \(w_{ij}\): weights in \([0..1]\)

• Assumptions:
  • Unique name assumption (UNA), i.e., two different URIs refer to two different entities.
  • Good coverage rate between the two datasets

See [Ferrara et al 2013] for a survey
Examples of linking rules learned on the OAEI’10 benchmark

<table>
<thead>
<tr>
<th>Test case</th>
<th>Similarity function</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>max(tokenized-jaro-winkler(soc_sec_id;soc_sec_id); monge-elkan(phone_number;phone_number))</td>
<td>≥0.87</td>
</tr>
<tr>
<td>Person2</td>
<td>max(jaro(phone_number;phone_number); jaro-winkler(soc_sec_id;soc_sec_id))</td>
<td>≥0.88</td>
</tr>
<tr>
<td>Restaurants (OAEI)</td>
<td>avg(0.22<em>tokenized-smith-waterman(phone_number;phone_number); 0.78</em>tokenized-smith-waterman(name;name))</td>
<td>≥0.91</td>
</tr>
<tr>
<td>Restaurants (fixed)</td>
<td>avg(0.35<em>tokenized-monge-elkan(phone_number;phone_number); 0.65</em>tokenized-smith-waterman(name;name))</td>
<td>≥0.88</td>
</tr>
</tbody>
</table>

Results in term of F-Measure on OAEI’10

<table>
<thead>
<tr>
<th>Dataset</th>
<th>KnoFuss+GA</th>
<th>ObjectCoref</th>
<th>ASMOV</th>
<th>CODI</th>
<th>LN2R</th>
<th>RiMOM</th>
<th>FBEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td><strong>1.00</strong></td>
<td><strong>1.00</strong></td>
<td><strong>1.00</strong></td>
<td>0.91</td>
<td><strong>1.00</strong></td>
<td><strong>1.00</strong></td>
<td>N/A</td>
</tr>
<tr>
<td>Person2</td>
<td><strong>0.99</strong></td>
<td>0.95</td>
<td>0.35</td>
<td>0.36</td>
<td>0.94</td>
<td>0.97</td>
<td>0.79</td>
</tr>
<tr>
<td>Restaurant (OAEI)</td>
<td>0.78</td>
<td>0.73</td>
<td>0.70</td>
<td>0.72</td>
<td>0.75</td>
<td><strong>0.81</strong></td>
<td>N/A</td>
</tr>
<tr>
<td>Restaurant (fixed)</td>
<td><strong>0.98</strong></td>
<td>0.89</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>0.96</td>
</tr>
</tbody>
</table>
RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

\[
\text{homepage}(X, Y) \land \text{homepage}(Z, Y) \Rightarrow \text{sameAs}(X, Z)
\]

<table>
<thead>
<tr>
<th>museum11</th>
<th>homepage</th>
<th>museum21</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="http://www.louvre.com">www.louvre.com</a></td>
<td></td>
</tr>
<tr>
<td>museum12</td>
<td><a href="http://www.musee-orsay.fr">www.musee-orsay.fr</a></td>
<td>museum22</td>
</tr>
<tr>
<td>museum13</td>
<td><a href="http://www.quai-branly.fr">www.quai-branly.fr</a></td>
<td>museum23</td>
</tr>
<tr>
<td>museum14</td>
<td>...</td>
<td>museum24</td>
</tr>
</tbody>
</table>
RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

\[ \text{homepage}(X, Y) \land \text{homepage}(Z, Y) \Rightarrow \text{sameAs}(X, Z) \]

Then we may infer:

\[ \text{sameAs}(\text{museum11}, \text{museum11}) \]
\[ \text{sameAs}(\text{museum12}, \text{museum22}) \]
\[ \text{sameAs}(\text{museum13}, \text{museum23}) \]
RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

\[ \text{homepage}(X, Y) \land \text{homepage}(Z, Y) \Rightarrow \text{sameAs}(X, Z) \]

Then we may infer:

\[ \text{sameAs}(\text{museum11}, \text{museum11}) \]
\[ \text{sameAs}(\text{museum12}, \text{museum22}) \]
\[ \text{sameAs}(\text{museum13}, \text{museum23}) \]

A key: is a set of properties that uniquely identifies every instance in the KG
RULE-BASED DATA LINKING APPROACHES

Informed approaches: need knowledge to be declared in an ontology language or other languages.

\[ \text{homepage}(X, Y) \land \text{homepage}(Z, Y) \Rightarrow \text{sameAs}(X, Z) \]

Then we may infer:

\[
\begin{align*}
\text{sameAs}(\text{museum11, museum11}) \\
\text{sameAs}(\text{museum12, museum22}) \\
\text{sameAs}(\text{museum13, museum23}) \\
\end{align*}
\]

A key: is a set of properties that uniquely identifies every instance in the KG

How to automatically discover keys from KGs?
OUTLINE

- Introduction
- Part 1: Data Linking
  - Part 2: Key Discovery
  - Part 3: Identity Link Invalidation
- Summary and Future Challenges
PART 2:
KEY DISCOVERY
**KEY SEMANTICS**

- **OWL2 Key for a class**: a combination of properties that uniquely identify each instance of a class

\[
\text{hasKey}( \text{CE ( OPE}_1 \ldots \text{OPE}_m ) ( \text{DPE}_1 \ldots \text{DPE}_n ) )
\]

\[
\forall X, \forall Y, \forall Z_1, \ldots, Z_n, \forall T_1, \ldots, T_m \land c(X) \land c(Y) \bigwedge_{i=1}^{n} (\text{ope}_i(X, Z_i) \land \text{ope}_i(Y, Z_i)) \\
\bigwedge_{i=1}^{m} (\text{dpe}_i(X, T_i) \land \text{dpe}_i(Y, T_i)) \Rightarrow X = Y
\]

\text{owl:hasKey(Book(Author) (Title)) means:}

\text{Book}(x_1) \land \text{Book}(x_2) \land \text{Author}(x_1, y) \land \text{Author}(x_2, y) \land \text{Title}(x_1, w) \land \text{Title}(x_2, w) \\
\Rightarrow \text{sameAs}(x_1, x_2)
A key is a set of properties that **uniquely identifies** every instance in the data.

<table>
<thead>
<tr>
<th></th>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>15/02/88</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>25/01/72</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person5</td>
<td>Simon</td>
<td>Roche</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person6</td>
<td>Jane</td>
<td>Ser</td>
<td>15/05/87</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person7</td>
<td>Sara</td>
<td>Khan</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person8</td>
<td>Theo</td>
<td>Martin</td>
<td>06/12/90</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person9</td>
<td>Marc</td>
<td>Blanc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

*Is [LastName] a key? ✗*

*Is [FirstName, LastName] a key? ✔*

**Exact keys**
### KEY VALIDITY

A key is a set of properties that **uniquely identifies** every instance in the data.

<table>
<thead>
<tr>
<th></th>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompston</td>
<td>15/02/88</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompston</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>25/01/72</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person5</td>
<td>Simon</td>
<td>Roche</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person6</td>
<td>Jane</td>
<td>Ser</td>
<td>15/05/87</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person7</td>
<td>Sara</td>
<td>Khan</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person8</td>
<td>Theo</td>
<td>Martin</td>
<td>06/12/90</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person9</td>
<td>Marc</td>
<td>Blanc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

**Is [FirstName, LastName] a key?**  ✔

**Is [Birthdate] a key with 2 exceptions?**  ✔

*Exact keys*

*Almost keys*
A key is a set of properties that uniquely identifies every instance in the data.

<table>
<thead>
<tr>
<th>Person</th>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>15/02/88</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>25/01/72</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person5</td>
<td>Simon</td>
<td>Roche</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person6</td>
<td>Jane</td>
<td>Ser</td>
<td>15/05/87</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person7</td>
<td>Sara</td>
<td>Khan</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person8</td>
<td>Theo</td>
<td>Martin</td>
<td>06/12/90</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person9</td>
<td>Marc</td>
<td>Blanc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

Is [FirstName,LastName] a key? ✔

Is [Birthdate] a key with 2 exceptions? ✔

Is [Birthdate and (Profession =“Actor”)] a key? ✔

**Exact keys**

**Almost keys**

**Conditional keys**
KEY DISCOVERY: A COMPLEX PROBLEM

• Find all the minimal keys requires $2^n$ property combinations

• For each combination scan all the instances
KEY DISCOVERY: A COMPLEX PROBLEM

• Find all the minimal keys requires $2^n$ property combinations need of efficient filtering and prunings

• For each combination scan all the instances
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires $2^n$ property combinations need of efficient filtering and prunings
- For each combination scan all the instances

maximal non-keys \(\xrightarrow{\text{derive}}\) minimal keys
KEY DISCOVERY: A COMPLEX PROBLEM

• Find all the minimal keys requires $2^n$ property combinations need of efficient filtering and prunings
• For each combination scan all the instances

maximal non-keys ➔ derive ➔ minimal keys

<table>
<thead>
<tr>
<th>FirstName</th>
<th>LastName</th>
<th>Birthdate</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>15/02/88</td>
<td>Actor</td>
</tr>
<tr>
<td>Person1</td>
<td>Tompson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>02/09/75</td>
<td>Actor</td>
</tr>
<tr>
<td>Person2</td>
<td>Tompson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>15/02/85</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>David</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>25/01/72</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Solgar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Roche</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>15/05/87</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Ser</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Khan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>06/12/90</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Martin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>27/10/84</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Blanc</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Is [LastName] a non-key?

➔ scan only a part of the data

Is [FirstName, LastName] a key?

➔ scan all the data
KEY DISCOVERY IN KNOWLEDGE GRAPHS

• Data characteristics:
  • Conforms to an ontology
  • Multi-valued properties
  • Incomplete data
  • Errors
  • Big datasets
  • No exact key can be discovered

• Assumptions:
  • UNA-Unique Name assumption
  • OWA-Open World assumption
KEY DISCOVERY IN KNOWLEDGE GRAPHS: CONTRIBUTIONS

• Data characteristics:
  • Conforms to an ontology
  • Multi-valued properties
  • Incomplete data
  • Errors
  • Big datasets
  • No exact key can be discovered

• Assumptions:
  • UNA-Unique Name assumption
  • OWA-Open World assumption

• KD2R: Exact keys discovery [Pernelle et al. ’13]
  • Multi-valued properties
  • Incomplete data

• SAKey: Almost-keys discovery [Symeonidou et al. 2014]
  • Few errors

• VICKey: Conditional keys discovery [Symeonidou et al. 2017]
  • No exact key is valid
KEY DISCOVERY IN KNOWLEDGE GRAPHS: CONTRIBUTIONS

- **KD2R**: Exact keys discovery [Pernelle et al. 2013]
  - Derives minimal keys from maximal non-keys
  - Key inheritance pruning
  - Key monotonicity and non-key anti-monotonicity prunings

- **SAKey**: Almost-keys discovery [Symeonidou et al. 2014]
  - Derives minimal n-almost keys from maximal (n+1) non-keys
  - Key monotonicity and non-key anti-monotonicity prunings
  - Singleton pruning and single key pruning
  - Potential (n+1) non-key computation
  - Semantic dependencies pruning

- **VICKey**: Conditional keys discovery [Symeonidou et al. 2017]
  - Derives minimal conditional keys from maximal non-keys
  - Key monotonicity and non-key anti-monotonicity prunings
  - Key support and coverage
(n+1) potential non-key construction: filtering of combinations of properties not needed be explored

- Incomplete data
- Properties referring to different classes

Potential n-non keys: Sets of properties that possibly refer to n-non keys
SAKEY: N-NON-KEY DISCOVERY (SAKEY)

(n+1)-non key discovery:

Intersections between sets of properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Set(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HasActor</td>
<td>{{f_1, f_2, f_3}, {f_2, f_3, f_4}}</td>
</tr>
<tr>
<td>HasDirector</td>
<td>{{f_1, f_2, f_3}, {f_2, f_3, f_6}}</td>
</tr>
<tr>
<td>ReleaseDate</td>
<td>{{f_2, f_6}}</td>
</tr>
<tr>
<td>HasName</td>
<td>{{f_2, f_6}}</td>
</tr>
<tr>
<td>HasLanguage</td>
<td>{{f_4, f_5}}</td>
</tr>
</tbody>
</table>

Final Map

\{hasActor, director\} \rightarrow 3-non key
KD2R, SAKEY AND VICKEY: EVALUATION

- **Scalability and runtime evaluation**
  - SAKey can handle size classes much larger than KD2R (DB: Natural Place more than 16 million triples and 243 properties (non-key discovery in 1min and key derivation 5min)
  - The use of prunings decreases the number of nodes to explore (e.g. a decrease of 50% for KD2R on DBpedia person) and the runtime (e.g. a decrease of 23% of runtime in SAKey)
### KD2R, SAKEY and VICKEY: Evaluation

- **Scalability and runtime evaluation**
  - **SAKey** can handle size classes much larger than **KD2R** (DB: Natural Place more than 16 million triples and 243 properties (non-key discovery in 1min and key derivation 5min))
  - The use of prunings decreases the number of nodes to explore (e.g. a decrease of 50% for KD2R on DBpedia person) and the runtime (e.g. a decrease of 23% of runtime in SAKey)

- **Relevance of keys for the data linking (with equality of literals)**
  - When keys (KD2R on OAEI2010:Person) are used F-Measure increases from 0.24 to 0.76
  - When conditional keys (VICKEY on Dbpedia and Yago) are used F-Measure increases from 0.08 to 0.55
  - When 3-almost keys (SAKey on OAEI 2013:Film) are used the F-measure is of 0.81
KEY DISCOVERY IN KGS: CHALLENGES

- Choose the good **key semantics** using the data characteristics (e.g. completeness)
- Define **holistic approaches** to discover different kinds of dependency constraints (e.g., denial constraints, key graphs and referring expressions)
- Define **incremental approaches** taking into account data evolution
OUTLINE

- Introduction
- Part 1: Data Linking
- Part 2: Key Discovery
- Part 3: Identity Link Invalidation
- Summary and Future Challenges
PART 3: IDENTITY LINK INVALIDATION
IDENTITY IS COMPLEX ...

“Lessons Learned: Managing Identity is Hard”
Jamie Taylor in ISWC 2017

“Biggest Problem: Identity”
Alan Patterson in ISWC 2018

Source: Aaron Bradley Twitter, October 26th, 2018
IDENTITY IS COMPLEX ...

From a Philosophical Point of View [Beek, 2018]

① Identity does not hold across modal contexts

- Allowing Lois Lane to believe that Superman saved her without requiring her to believe that Clark Kent saved her.
IDENTITY IS COMPLEX ...

From a Philosophical Point of View [Beek, 2018]

① Identity does not hold across modal contexts

② Identity is context-dependent [Geach, 1967]

- Allowing two medicines with the same chemical structure to be considered the same in a scientific context, but different in a commercial context (e.g., because they are produced by different companies).

\[\text{C}_9\text{H}_8\text{O}_4\text{ headache} \]

\[\text{ns3} \]

\[\text{ns1:Aspirin} \]

\[\text{chemical} \quad \text{cures} \]

\[\text{B Inc.} \quad €10,- \]

\[\text{ns2} \]

\[\text{ns1:Aspirin} \]

\[\text{chemical} \quad \text{cures} \]

\[\text{C}_9\text{H}_8\text{O}_4 \quad \text{headache} \]
Identity does not hold across modal contexts

Identity is context-dependent [Geach, 1967]

Identity over time poses problems

- since a car may be considered the same car, even though some (or even all) of its original components have been replaced by new ones.
IDENTITY IS COMPLEX ...

From an Operational Point of View

① Unless two resources are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

 comunicación: Only 3.6K owl:differentFrom triples compared to 558M owl:sameAs (LOD-a-lot dataset, 2015 crawl of the LOD Cloud)
IDENTITY IS COMPLEX ...

From an Operational Point of View

① Unless two resources are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

② Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource

◆ Barack Obama the person vs URL referring to his Web page
  (Problem of Sense and Reference [Halpin, 2010])
IDENTITY IS COMPLEX ...

From an Operational Point of View

① Unless two things are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical.

② Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource.

③ Modelers have different opinions about whether two objects are the same.

From a set of 250 owl:sameAs links, one Semantic Web expert judged that only 73 are correct identity links, whilst two other experts have judged 132 and 181 as true identity links, respectively [Halpin et al., 2010].
IDENTITY IS COMPLEX ...

From an Operational Point of View

1. Unless two things are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical

2. Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource

3. Modelers have different opinions about whether two objects are the same

4. Data linking approaches are rarely 100% precise
   - *Precision usually between 67% and 86% [OAEI 2017, OAEI 2018]*
IDENTITY IS COMPLEX ...

From an Operational Point of View

① Unless two things are explicitly said to be different, the absence of an identity statement between them does not mean that they are not identical.

② Hard to distinguish between the IRI referring to a non-information resource and its corresponding information resource.

③ Modelers have different opinions about whether two objects are the same.

④ Data linkage approaches are rarely 100% precise.

⑤ Lack of alternative well-defined and standardized identity predicates.
  - *rdfs:seeAlso, skos:exactMatch, etc.* → Lack of formal semantics.
THE ‘SAMEAS PROBLEM’

Web of Data contains a large* number of erroneous owl:sameAs

*~21% [Halpin et al., 2010]
Manual evaluation of 250 owl:sameAs from the Web

*~2.8% [Hogan et al., 2012]
Manual evaluation of 1K identical pairs from the Web

*~4% [Raad, 2018]
Manual evaluation of 300 owl:sameAs from the LOD Cloud + error degree distribution of 558M owl:sameAs
THE ‘SAMEAS PROBLEM’

The largest identity set contains 177,794 terms that 'should' refer to the same real world entity

However:

http://dbpedia.org/resource/Albert_Einstein
http://dbpedia.org/resource/Basketball
http://dbpedia.org/resource/Coca-Cola
http://dbpedia.org/resource/Deauville
http://dbpedia.org/resource/Italy
http://dbpedia.org/resource/Lists_of_christian_religions
... 

Full list at: https://sameas.cc/term?id=4073
HOW TO LIMIT THIS ‘SAMEAS PROBLEM’?

- Detect erroneous identity links / Validate correct ones
  - Inconsistency-based Approaches
  - Content-based Approaches
  - Network-based Approaches

- Propose alternative semantics for identity
  - Weak-Identity and Similarity predicates
  - Contextual Identity
owl:sameAs, indicates that two different descriptions refer to the same entity

a strict semantics,

1) Reflexive,
2) Symmetric,
3) Transitive and
4) Fulfils property sharing:

\[ \forall X \forall Y \text{owl:sameAs}(X, Y) \land p(X, Z) \Rightarrow p(Y, Z) \]
IDENTITY LINK INVALIDATION

- **Link invalidation** consists in determining whether an identity link is **erroneous**

- **Different kinds of information can be used:**
  - Resource descriptions
  - Consistency constraints
  - Source trustworthiness
  - Identity network metrics
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

owl:sameAs(b1, b2)?
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

\[ \text{owl:sameAs}(b1, b2)? \]
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

- Author: Grigoris Antoniou
- Title: A Semantic Web Primer
- Publication Year: 2007
- Number of Pages: 288
- Trustworthiness: UNA

Identity Network:

- owl:sameAs(b1, b2)?
- owl:sameAs(b2, b10)?
- owl:sameAs(b1, b6)?
- owl:sameAs(b6, b3)?
- owl:sameAs(b3, b5)?
- owl:sameAs(b5, b9)?
- owl:sameAs(b9, b8)?
- owl:sameAs(b8, b10)?

Additional information:
- nbPages: 208
- nbPages: 288
- pubYear: 2007
- aName: Paul Grauth
- aName: Grigoris Antoniou
- aName: P. Grauth
- aName: G. Antoniou
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Which kind of information to use for detecting erroneous Identity links?

Ontology Axioms:

- \( \text{Func(nbPages)} \)
- \( \text{LC(author)} \)
- \( \text{Func(title)} \)
- \( \text{Disj(Science-fiction, Memoir)} \), ...

Identity Network

UNA

Trustworthiness

Content
1. DETECTION OF ERRONEOUS IDENTITY LINKS

- Inconsistency-based
  - UNA
  - Trustworthiness
  - Ontology axioms
    - Cudré-Mauroux et al. 2009
    - Papaleo et al. 2014
    - Paulheim 2014

- Content-Based

- Network-Based
  - Community Detection
    - Raad et al. 2018
  - Network Metrics
    - Guéret et al. 2016

- [ de Melo, 2013 ]
- [ Valdestilhas et al., 2017 ]
INCONSISTENCY-BASED
1. DETECTION OF ERRONEOUS IDENTITY LINKS

- Inconsistency-based
  - UNA
  - Trustworthiness
  - Ontology axioms
    - Cudré-Mauroux et al. 2009
    - Papaleo et al. 2014
    - Paulheim 2014

- Content-Based

- Network-Based
  - Community Detection
  - Network Metrics
    - Raad et al. 2018
    - Guéret et al. 2016

[ de Melo, 2013 ]
[ Valdestilhas et al., 2017 ]
Principle: use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies and possible errors in the linked resources.

nbPages is a Functional Property

[Inconsistency-based and content-based onontology axiom violation]

[Papaleo et al., 2014]
[Hogan et al. 2012]
A logical **ontology-based method** to detect invalid sameAs statements

Builds a contextual graph «around» each one of the two resources involved in the sameAs by exploiting ontology axioms on:

- **functionality** and **inverse functionality** of properties and
- **local completeness** of some properties, e.g., the author list of a book.

Exploit the descriptions provided in these contextual graphs to eventually detect inconsistencies or high dissimilarities.
R the set of rules

(inverse) functional properties

- \( R_{1FDP} \) : sameAs\((x, y) \land p_i(x, w_1) \land p_i(y, w_2) \rightarrow \text{synVals}(w_1, w_2) \)
- \( R_{2FOP} \) : sameAs\((x, y) \land p_j(x, w_1) \land p_j(y, w_2) \rightarrow \text{sameAs}(w_1, w_2) \)
- \( R_{3IFP} \) : sameAs\((x, y) \land p_k(w_1, x) \land p_k(w_2, y) \rightarrow \text{sameAs}(w_1, w_2) \)

local complete properties

- \( R_{4LC} \) : sameAs\((x, y) \land p(x, w_1) \rightarrow p(y, w_1) \)

Apply Unit Resolution on \{F ∪ R\}.
[F set of facts, R set of rules]
ONTOMETRY AXIOM VIOLATION

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].


2-degree contextual graph
phone_number, hasAddress & city
(possible synvals computation)
ONTOLOGY AXIOM VIOLATION

- OAEI 2010 dataset on Restaurants
- Use of the output of different linking tools [1], [2] and [3].

<table>
<thead>
<tr>
<th>Linking Method</th>
<th>LM Precision</th>
<th>IM Recall</th>
<th>IM Precision</th>
<th>IM Accuracy</th>
<th>LM+IM precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>[120]</td>
<td>95.55%</td>
<td>75%</td>
<td>37%</td>
<td>93.34%</td>
<td>98.85%</td>
</tr>
<tr>
<td>[110]</td>
<td>69.71%</td>
<td>88.4%</td>
<td>88.4%</td>
<td>92.9%</td>
<td>95.19%</td>
</tr>
<tr>
<td>[138]</td>
<td>90.17%</td>
<td>100%</td>
<td>42.30%</td>
<td>86.60%</td>
<td>100%</td>
</tr>
</tbody>
</table>

IM: Invalidation method
LM: Linking method

Improvement in precision
1. DETECTION OF ERRONEOUS IDENTITY LINKS

- Inconsistency-based
  - UNA
    - Trustworthiness
    - Ontology axioms
    - Cudré-Mauroux et al. 2009
    - Papaleo et al. 2014
    - Paulheim 2014

- Content-Based

- Network-Based
  - Community Detection
    - Raad et al. 2018
  - Network Metrics
    - Guéret et al. 2012

[ de Melo, 2013 ]
[ Valdestilhas et al., 2017 ]
CONTENT BASED

[Paulheim, 2014]

**Principle**: links follow some patterns, links that violate those patterns are erroneous.

- A multi-dimensional and scalable **outlier detection** approach for finding **erroneous identity links**.
- Projection of links into **Vector Space**: each link is a point in an n-dimensional vector space.
CONTENT BASED

[Paulheim, 2014]

Principle: links follow some patterns, links that violate those patterns are erroneous.

- A multi-dimensional and scalable outlier detection approach for finding erroneous identity links.
- Projection of links into Vector Space: each link is a point in an n-dimensional vector space.
CONTENT BASED

- **Feature Vector**: resource types and ingoing/outgoing properties
  - e.g. LHS_foaf:based_near and RHS_foaf:based_near are distinct features.

- **Different strategies of creating vectors**: direct types only, all ingoing and outgoing properties, or a combination

- Several outlier detection methods were tested: LOF, CBLOF, LOP, 1-class SVM etc.

- Each method assign a score to each data point indicating the likeliness of being an outlier ➔ *incorrect link*.

[Paulheim, 2014]
CONTENT BASED

- **Dataset**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Peel Session</th>
<th>DBpedia</th>
<th>DBTropes</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Links</td>
<td>2,087</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Types</td>
<td>3</td>
<td>31</td>
<td>2</td>
<td>79</td>
</tr>
<tr>
<td># Properties</td>
<td>4</td>
<td>56</td>
<td>18</td>
<td>124</td>
</tr>
</tbody>
</table>

- **Gold Standard**: 100 randomly sampled links from D1 and D2

- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)

[Paulheim, 2014]
CONTENT BASED

[Paulheim, 2014]

- **Dataset**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Peel Session</th>
<th>DBpedia</th>
<th>DBTropes</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Links</td>
<td>2,087</td>
<td></td>
<td></td>
<td>4,229</td>
</tr>
<tr>
<td># Types</td>
<td>3</td>
<td>31</td>
<td>2</td>
<td>79</td>
</tr>
<tr>
<td># Properties</td>
<td>4</td>
<td>56</td>
<td>18</td>
<td>124</td>
</tr>
</tbody>
</table>

- **Gold Standard**: 100 randomly sampled links from D1 and D2

- Use of RapidMiner with anomaly detection and LOD extensions (6 methods)

- **Best performance on D1**:
  - 1-class SVM (AUC = **0.857**, F1= **0.471**)

- **Best performance on D2**:
  - LOF (AUC = **0.619**, F1= **0.5**)
CONTENT BASED

Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Peel Session</th>
<th>DBpedia</th>
<th>DBTropes</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td># Links</td>
<td>2,087</td>
<td>4,229</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Types</td>
<td>3</td>
<td>31</td>
<td>2</td>
<td>79</td>
</tr>
<tr>
<td># Properties</td>
<td>4</td>
<td>56</td>
<td>18</td>
<td>124</td>
</tr>
</tbody>
</table>

Gold Standard: 100 randomly sampled links from D1 and D2

Use of RapidMiner with anomaly detection and LOD extensions (6 methods)

Best performance on D1:
- 1-class SVM (AUC = 0.857, F1 = 0.471)

Best performance on D2:
- LOF (AUC = 0.619, F1 = 0.5)

Examples of typical source of errors for D1:
- Linking of songs to albums with the same name.
- Linking of different persons of the same name.
  e.g., a blues musician named Jimmy Carter to the U.S. president.

[Paulheim, 2014]
1. DETECTION OF ERRONEOUS IDENTITY LINKS

Inconsistency-based

Content-Based

Network-Based

UNA

Trustworthiness

Ontology axioms

Network Metrics

Community Detection

Raad et al. 2018

Guéret et al. 2012

Cudré-Mauroux et al. 2009

Papaleo et al. 2014

Paulheim 2014

[ de Melo, 2013 ]

[ Valdestilhas et al., 2017 ]
Principle

- The quality of a link can be evaluated based on **how connected a node** is within the **network** (data graph, sameAs network) in which it appears.
- How **network metrics** can help to detect erroneous links?
NETWORK BASED

Node in-degree and out-degree, Centrality, Clustering coefficient ….
NETWORK BASED

Density of detected community structures
Overall idea

Use the community structure of the network containing solely owl:sameAs links to assign error degree for each link.
4 main steps

Step 1: Extraction of the explicit identity statements

Step 2: Partition into equality sets (theoretically...the same entity)

Step 3: Detection of the community structure of each equality set using the Louvain algorithm [Blondel et al. 2008]

Step 4: Assignation of an error degree to each sameAs
Error degree based on the weight \((w=2\) when the same\(As\) is symmetric) and on the density of the involved community(ies)

**intra-community link**

\[
a) \quad err(e_C) = \frac{1}{w(e_C)} \times \left(1 - \frac{W_C}{|C| \times (|C| - 1)}\right)
\]

**inter-community link**

\[
b) \quad err(e_{C_{ij}}) = \frac{1}{w(e_{C_{ij}})} \times \left(1 - \frac{W_{C_{ij}}}{2 \times |C_i| \times |C_j|}\right)
\]
Experimentation - Dataset

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of 28B triples from LOD 2015 crawl
- Step 1: extraction of an explicit identity network of 558.9M sameAs links (179M nodes)
- Step 2: Partitionned in 48.9M of non singleton equality sets

Example: The B. Obama equality set which contains 440 nodes
NETWORK BASED

Step 3: Detection of the community structure for each equality set

Example: the community structure of the *Barack Obama’s* Equality Set

- **$C_0$**: DBpedia IRIs referring to the person Obama in different languages
- **$C_1$**: IRIs referring to the Obama administration, government
- **$C_2$**: IRIs referring to the person Obama in different functions
- **$C_3$**: IRIs referring to the person Obama, his senator career

[Raad et al., 2018]
Step 4: Computation of the error degrees

Low error degrees for the links of this community

\[ \text{err}(e) = 1 \]

For these 2 links

[Raad et al., 2018]
**NETWORK BASED**

- **Scales up** to a graph of **28 billion** triples: **11 hours for the 4 steps**

- **Finding the threshold**: manual evaluation of 200 randomly chosen links

<table>
<thead>
<tr>
<th></th>
<th>0-0.2</th>
<th>0.2-0.4</th>
<th>0.4-0.6</th>
<th>0.6-0.8</th>
<th>0.8-1</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>same</td>
<td>35(100%)</td>
<td>22(100%)</td>
<td>18(85.7%)</td>
<td>7(77.7%)</td>
<td>15(68.1%)</td>
<td>97(88.9%)</td>
</tr>
<tr>
<td>related</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>unrelated</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>related + unrelated</td>
<td>0(0%)</td>
<td>0(0%)</td>
<td>3(14.2%)</td>
<td>2(22.2%)</td>
<td>7(31.8%)</td>
<td>12(11%)</td>
</tr>
<tr>
<td>can’t tell</td>
<td>5</td>
<td>18</td>
<td>19</td>
<td>31</td>
<td>18</td>
<td>91</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>200</td>
</tr>
</tbody>
</table>

The higher an error degree is the most likely a link is erroneous
- 100% of owl:sameAs with an **error degree <0.4** are correct

- Can theoretically **invalidate a large set of owl:sameAs links** on the LOD:
  - **1.26M** owl:sameAs have an **error degree** in [0.99, 1]
Different approaches: consistency-based, content-based or network-based relaying on different kinds of information (UNA, axioms, mappings, textual values/types/properties or network metrics).

Some approaches are global (collective), some are instance-based (pairs of resources are considered independently).
ERRONEOUS LINK DETECTION: SUMMARY

- **Different approaches**: consistency-based, content-based or network-based relaying on different kinds of information (UNA, axioms, mappings, textual values/types/properties or network metrics)

- Some approaches are global (collective), some are instance-based (pairs of resources are considered independently).

**Limitations**

- **Evaluation** are often conducted on few links, on specific datasets

- Some **assumptions** cannot be made on the LOD:
  - UNA is not always fulfilled
  - Ontology and Ontology axioms are not always available
  - Differences are rarely available: useful for inconsistency checking
  - Network-based approaches do not need such assumptions + scalable but they cannot decide for small equality sets, and higher precision is needed.

- Need of alternate links
CONTEXTUAL IDENTITY LINKS
\begin{itemize}
  
  \item **Weaker** kinds of \textit{identity} can be expressed by considering a \textit{subset of properties} with respect to which two resources can be considered to be the same.
  
  \item Identity is \textbf{context-dependent} [Geach, 1967]
    
    \begin{itemize}
      
      \item allowing two medicines to be considered \textit{the same in terms of their chemical substance, but different in terms of their price} (e.g., because they are produced by different companies).
    
    \end{itemize}
  
\end{itemize}
CONTEXTUAL IDENTITY LINKS

- New **contextual identity** relation

- An **algorithm** for automatic detection of the **most specific contexts** in which two instances (resources) are identical
  - the detection process can further be guided by a set of **semantic constraints** that are provided by domain experts.

- Contexts are defined as a sub-ontology of the domain ontology

- All the possible contexts are organized in a lattice using an order relation.

[Raad et al., 2017]
\[ \text{GC}_1 = (C = \{ \text{Drug, Paracetamol, Lactose, Weight} \}, \text{OP} = \{ \text{isComposedOf, hasWeight} \}, \text{DP} = \emptyset, A = \{ \text{domain(isComposedOf)} = \text{Drug}, \text{range(isComposedOf)} = \text{Lactose} \cup \text{Paracetamol}, \text{domain(hasWeight)} = \text{Lactose} \cup \text{Paracetamol}, \text{range(hasWeight)} = \text{Weight} \}) \]
## CONTEXTUAL IDENTITY LINKS

\[
\text{GC}_1= (C = \{\text{Drug, Paracetamol, Lactose, Weight}\}, \\
OP = \{\text{isComposedOf, hasWeight}\}, \ DP = \emptyset, \\
A = \{\text{domain(isComposedOf) = Drug,} \\
\text{range(isComposedOf) = Lactose } \sqcup \text{ Paracetamol,} \\
\text{domain(hasWeight) = Lactose } \sqcup \text{ Paracetamol,} \\
\text{range(hasWeight) = Weight}\})
\]

\[
\leq
\]

\textbf{(more specific than)}

\[
\text{GC}_2= (C = \{\text{Drug, Lactose}\}, \\
OP = \{\text{isComposedOf, DP = \{\emptyset\}}\}, \\
A = \{\text{domain(isComposedOf) = Drug,} \\
\text{range(isComposedOf) = Lactose}\})
\]
For each pair of individuals \((i_1, i_2)\) of the target class set of the most specific global contexts in which \((i_1, i_2)\) are identical
CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]

- Micro-organisms stabilization
- Dairy Gels Transformation

(Ibanescu et al., 2016)

A Process and Observation Ontology

- Classes: ≈ 4 700
- Instances: ≈ 415 000
- Statements: ≈ 1 700 000

- Classes: ≈ 5 000
- Instances: ≈ 42 000
- Statements: ≈ 237 000
### CONTEXTUAL IDENTITY LINKS

**Experiment 1**

- **PO³ (A Process and Observation Ontology)**
  - CellExtraDry
  - Classes: \(\approx 4700\)
  - Instances: \(\approx 415,000\)
  - Statements: \(\approx 1,700,000\)

**DECIDE**

DEtection of Contextual IDEntity

**Experiment 2**

- **PO³ (A Process and Observation Ontology)**
  - Carredas
  - Classes: \(\approx 5,000\)
  - Instances: \(\approx 42,000\)
  - Statements: \(\approx 237,000\)

<table>
<thead>
<tr>
<th></th>
<th>CellExtraDry</th>
<th>Carredas</th>
</tr>
</thead>
<tbody>
<tr>
<td># Instances (type:Mixture)</td>
<td>210</td>
<td>619</td>
</tr>
<tr>
<td># Possible Pairs</td>
<td>21,945</td>
<td>191,271</td>
</tr>
<tr>
<td># Dependant Classes (Total Classes)</td>
<td>191 (208)</td>
<td>488 (555)</td>
</tr>
<tr>
<td># Graph Nodes per pair</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td># Identity Links</td>
<td>33,092</td>
<td>239,410</td>
</tr>
<tr>
<td># Identity Links per pair</td>
<td>1.41</td>
<td>1.25</td>
</tr>
<tr>
<td># Different Global Contexts</td>
<td>28</td>
<td>233</td>
</tr>
<tr>
<td>Execution Time (approx. minutes)</td>
<td>2</td>
<td>26</td>
</tr>
</tbody>
</table>

[Raad et al., 2017]
Detect for each context $\text{GC}_i$, the measures $m_i$ where

$$\text{identiConTo}_{<\text{GC}_i>}(i_1, i_2) \cap \text{observes}(i_1, m_1) \rightarrow \text{observes}(i_2, m_2)$$

with $m_1 \approx m_2$

$$\text{identiConTo}_{<\text{GC}_i>}(i_1, i_2) \rightarrow \text{same}(m_i)$$
Detection of 38 844 rules

<table>
<thead>
<tr>
<th>Règle</th>
<th>Taux d’erreur</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{identiConTo}_{&lt;GC_1&gt;}(x, y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow$ same(pH)</td>
<td>6.19 %</td>
<td>57</td>
</tr>
<tr>
<td>$\text{identiConTo}_{&lt;GC_3&gt;}(x, y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow$ same(Dureté)</td>
<td>1.86 %</td>
<td>66</td>
</tr>
<tr>
<td>$\text{identiConTo}_{&lt;GC_2&gt;}(x, y)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rightarrow$ same(Friabilité)</td>
<td>4.52 %</td>
<td>647</td>
</tr>
</tbody>
</table>

The domain experts have evaluated the plausibility of the best 20 rules (in terms of error rate and support)

[Raad et al., 2017]
Detection of 38 844 rules

<table>
<thead>
<tr>
<th>Règle</th>
<th>Taux d’erreur</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{identiConTo}_{&lt;GC_1&gt;}(x, y) \rightarrow \text{same}(pH)$</td>
<td>6.19 %</td>
<td>57</td>
</tr>
<tr>
<td>$\text{identiConTo}_{&lt;GC_3&gt;}(x, y) \rightarrow \text{same}(\text{Dureté})$</td>
<td>1.86 %</td>
<td>66</td>
</tr>
<tr>
<td>$\text{identiConTo}_{&lt;GC_2&gt;}(x, y) \rightarrow \text{same}(\text{Friabilité})$</td>
<td>4.52 %</td>
<td>647</td>
</tr>
</tbody>
</table>

The domain experts have evaluated the plausibility of the best 20 rules (in terms of error rate and support)

<table>
<thead>
<tr>
<th>plausibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>3</td>
</tr>
</tbody>
</table>

Impossible  Not very probable  Can’t tell  Why not  Plausible

The error rate decreases of 12% when a global context is replaced by a more specific global context
CONCLUSION AND FUTURE CHALLENGES
IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship
IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship

owl:sameAs  ----> Genuine identity

Ivont:somewhatSameAs  ----> Near/weak identity
IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship

owl:sameAs: Genuine identity

lvont:somewhatSameAs: Near/weak identity

:sameBook
:sameExpression: Subjective identity
:sameArtWork
:dentiConTo<na>: Contextual identity
IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship
- Need of hybrid methods

Network Topology

Source Reliability

Link Content

Ontology Axioms

owl:sameAs

Ivont:somewhatSameAs

b1

b2

:sameBook

:sameExpression

:sameArtWork

:depeviConTo<πa>
IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship
- Need of hybrid methods
- Link quality assessment is not a matter of one unique dimension

Network Topology

Source Reliability

Link Content

Ontology Axioms

owl:sameAs
lvont:somewhatSameAs
:sameBook
:sameExpression
:sameArtWork
:dentiConTo<πα>

Link Validity:
Inconsistent equivalent classes, Invalid links, Contextual links

Link Properties:
Transitivity, symmetry, ...

Link added-value:
Information gain, reachability, ...

Link meta-data:
availability, evolution
IDENTITY MANAGEMENT: SUMMARY AND CHALLENGES

- Different kinds of identity relationship
- Need of hybrid methods
- Link quality assessment is not a matter of one unique dimension

What is about the distinctness relation?

Network Topology
Source Reliability
Link Content
Ontology Axioms

owl:sameAs
lvont:somewhatSameAs
:sameBook
:sameExpression
:sameArtWork
:dentiConTo<πa>

Link Validity:
Inconsistent equivalent classes, Invalid links, Contextual links

Link Properties:
Transitivity, symmetry, …

Link added-value:
Information gain, reachability, …

Link meta-data:
availability, evolution
Mastering open data and entity identification technologies will lead to master data access control and data de-identification.
IDENTITY MANAGEMENT IN THE WEB OF DATA

FATIHA SAÏS
LRI, PARIS SUD UNIVERSITY, CNRS, ORSAY, FR

SAIS@LRI.FR
REFERENCES (1)

[Beek et al., 2016] A contextualised semantics for owl: sameas.
W. Beek, S. Schlobach, and F. van Harmelen. In ESWC 2016

[CudreMauroux et al., 2009] idmesh: graph-based disambiguation of linked data.

[de Melo, 2013] Not quite the same: Identity constraints for the web of linked data.


[Halpin et al., 2010] When owl:sameAs isn’t the same: An analysis of identity in Linked Data.

[Hogan et al., 2012] Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora.
REFERENCES (2)

[Jaffri et al., 2008 ] URI disambiguation in the context of linked data.
   A. Jaffri, H. Glaser, and I. Millard. In LDOW@WWW 2008.


[Papaleo et al., 2014] Logical detection of invalid sameas statements in rdf data.

[Raad et al., 2017] Detection of contextual identity links in a knowledge base.

[Raad et al., 2018 ] Detecting Erroneous Identity Links on the Web using Network Metrics. J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs. ISWC 2018

REFERENCES (3)


[Symeonidou et al. 2014] SAKey: Scalable almost key discovery in RDF data.

Danai Symeonidou, Luis Galarraga, Nathalie Pernelle, Fatiha Saïs and Fabian Suchanek.
In ISWC 2017.