DATA LINKING AND KNOWLEDGE DISCOVERY IN RDF DATA: METHODS AND SOME FEEDBACK FROM AGRONOMIC APPLICATIONS

FATIHA SAÏS

LAHDAK@LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAy UNIVERSITY
Joint work with: N. Pernelle, L. Papaleo, J. Raad and D. Symeonidou

1ST DATAIA DAYS « LIFE SCIENCES & AI», DEC. 4TH 2019
Linked Open Data (LOD)

FAIR Principles

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

Gene Ontology: 807473 triples
Lipid Ontology: 15406 triples

An RDF Graph is a set of triples.

- Its nodes are (labelled by) the subjects and objects appearing in the triples.
- Its edges are labelled by the properties.
NEED OF KNOWLEDGE
THE ROLE OF KNOWLEDGE IN AI

[Artificial Intelligence 47 (1991)]

ON THE THRESHOLDS OF KNOWLEDGE

Douglas B. Lenat
MCC
3500 W. Balcones Center
Austin, TX 78759

Edward A. Feigenbaum
Computer Science Department
Stanford University
Stanford, CA 94305

Abstract

We articulate the three major findings of AI to date: (1) The Knowledge Principle: if a program is to perform complex tasks well, it must know a great deal about the world in which it operates. (2) A plausible extension of that principle, called the Breadth Hypothesis: there are two additional abilities necessary for intelligent behavior in unexpected situations: falling back on increasingly general knowledge, and analogizing to specific but far-flung knowledge. (3) AI as Empirical Inquiry: we must test our ideas experimentally, on large problems. Each of these three hypotheses proposes a particular threshold to cross, which leads to a qualitative change in emergent intelligence. Together, they determine a direction for future AI research.

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."
"An ontology is an **explicit**, **formal** specification of a shared conceptualization.”
[Thomas R. Gruber, 1993]

**RDFS – Resource Description Framework Schema**
- Lightweight ontologies

**OWL – Web Ontology Language**
- Expressive ontologies
**OWL – WEB ONTOLOGY LANGUAGE**

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

- **Axioms**
  - `owl:subClassOf`
  - `owl:subPropertyOf`
  - `owl:inverseProperty`
  - `owl:FunctionalProperty`
  - `owl:minCardinality`
  - …

---

**Disjunction Constraints**

- FoodProduct $\perp$ Component
- Spaghetti $\perp$ Macaroni
- Fiber $\perp$ Vitamin
- Vitamin C $\perp$ Vitamin B
KNOWLEDGE

GRAPHS
WHO IS DEVELOPING KNOWLEDGE GRAPHS?

Academic side

2007
- DBpedia
- Wikidata

2012
- YAGO
- Wikidata

Commercial side

2012
- Google Knowledge Graph

2013
- Yahoo’s new SERP designs mobile and knowledge graph

2015
- LinkedIn Graph

2016
- Microsoft Graph
KNOWLEDGE GRAPH REFINEMENT

Completeness

Correctness
KNOWLEDGE GRAPH REFINEMENT

Completeness
Data Linking
Key discovery
Data Fusion

Correctness
Link Invalidation
Contextual identity
• Introduction
• **Key discovery for data linking**
• Link Invalidation
• Contextual identity
• Conclusion
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. person, article, protein).
Data linking or Identity link detection consists in detecting whether two descriptions of resources refer to the same real world entity (e.g. person, article, protein).
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. person, article, protein).

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

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

  versus

- **Global approaches:** consider data type properties (attributes) as well as object properties (relations) 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

- **Some surveys:**

Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for 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:

\[ \text{sameAs}(\text{museum11}, \text{museum21}) \]
\[ \text{sameAs}(\text{museum12}, \text{museum22}) \]
\[ \text{sameAs}(\text{museum13}, \text{museum23}) \]
Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for 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:

\[
\text{sameAs}(\text{museum11}, \text{museum21}) \\
\text{sameAs}(\text{museum12}, \text{museum22}) \\
\text{sameAs}(\text{museum13}, \text{museum23})
\]

A key: is a set of properties that uniquely identifies every instance of a class
Rule-based data linking approaches [Saïs et al. 2009, Al Bakri et al. 2015]: need for 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:

\[ \text{sameAs}(\text{museum}11, \text{museum}21) \]
\[ \text{sameAs}(\text{museum}12, \text{museum}22) \]
\[ \text{sameAs}(\text{museum}13, \text{museum}23) \]

A key: is a set of properties that uniquely identifies every instance of a class

How to automatically discover keys from KGs?
A key is a set of properties that **uniquely identifies** every instance in the data.

<table>
<thead>
<tr>
<th>FirstName</th>
<th>LastName</th>
<th>City</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>Paris</td>
</tr>
<tr>
<td></td>
<td>Marie</td>
<td>Tompson</td>
<td>Berlin</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>Rome</td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>Roche</td>
<td>Montpellier</td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>Ser</td>
<td>Paris</td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>Khan</td>
<td>London</td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>Martin</td>
<td>Lyon</td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>Blanc</td>
<td>Nantes</td>
</tr>
</tbody>
</table>

*Is [FirstName,LastName] a key? ✔*

*Is [City] a key? ✗*
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>City</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>Paris</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>Berlin</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>Toulouse</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>Rome</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>Roche</td>
<td>Montpellier</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>Ser</td>
<td>Paris</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>Khan</td>
<td>London</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>Martin</td>
<td>Lyon</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>Blanc</td>
<td>Nantes</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

*Is [FirstName, LastName] a key? ✔*

*Is [City] a key? ❌*

*Is [City] a key with 2 exceptions? ✔*
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ property combinations
  - need of efficient filtering and prunings
KEY DISCOVERY: A COMPLEX PROBLEM

- Find all the minimal keys requires at least $2^n$ 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 at least $2^n$ 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></th>
<th>FirstName</th>
<th>LastName</th>
<th>City</th>
<th>Profession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person1</td>
<td>Anne</td>
<td>Tompson</td>
<td>Paris</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person2</td>
<td>Marie</td>
<td>Tompson</td>
<td>Berlin</td>
<td>Actor</td>
</tr>
<tr>
<td>Person3</td>
<td>Marie</td>
<td>David</td>
<td>Toulouse</td>
<td>Actor</td>
</tr>
<tr>
<td>Person4</td>
<td>Vincent</td>
<td>Solgar</td>
<td>Rome</td>
<td>Actor, Director</td>
</tr>
<tr>
<td>Person4</td>
<td>Simon</td>
<td>Roche</td>
<td>Montpellier</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Jane</td>
<td>Ser</td>
<td>Paris</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Sara</td>
<td>Khan</td>
<td>London</td>
<td>Teacher</td>
</tr>
<tr>
<td>Person4</td>
<td>Theo</td>
<td>Martin</td>
<td>Lyon</td>
<td>Teacher, Researcher</td>
</tr>
<tr>
<td>Person4</td>
<td>Marc</td>
<td>Blanc</td>
<td>Nantes</td>
<td>Teacher</td>
</tr>
</tbody>
</table>

*Is [LastName] a non-key?* ➔ scan only a part of the data
SAKEY: N-ALMOST KEY DISCOVERY

- SAKey allows $n$ exceptions in the data

- n-almost key: a set of properties where $|E_P| \leq n$

- n-non key: a set of properties where $|E_P| \geq n+1$

$n=4$

All sets of properties that contain at least 5 exceptions

All sets of properties that contain at most 4 exceptions
APPLICATION TO SCIENTIFIC DATA

• Many scientific numerical data
  – Sensor data
  – Experimental data.

• Difficult to interpret numerical data
  – Different levels of precision
  – Different measure units…

• Better understand the numerical data

Danai Symeonidou, Isabelle Sanchez, Madalina Croitoru, Pascal Neveu, Nathalie Pernelle, Fatiha Saïs, Aurelie Roland-Vialaret, Patrice Buche, Aunur-Rofiq Muljarto, Remi Schneider: Key Discovery for Numerical Data: Application to Oenological Practices. ICCS 2016: 222-236
APPLICATION TO SCIENTIFIC DATA

Discover keys in numerical data

- **Keys**: combinations of properties that discriminate a resource

Evaluate their quality

- Experimental numerical data in 3 wine flavour datasets (2011-2014)

*How do we discriminate the wines?*
PROBLEM STATEMENT

Key discovery approaches consider all the values as symbolic

- Ex. $PH(\text{Wine1}, 3.455), PH(\text{Wine2}, 3.457)$

$\neq$

Key discovery in raw numerical data: Many not-significant keys can be found
PROPOSED METHOD STEPS

1. Data Pre-processing
2. Key Discovery
3. Key Quality Evaluation
4. n-non key discovery
5. n-almost key discovery

- Initial Data
- Symbolic Data
- Pre-processing

- q-quantiles
- n exceptions

- n-non keys
- n-almost keys
DATA PRE-PROCESSING

Objective: Interpret numerical data in a symbolic way

Solution: Use quantiles to group data values

- Quantiles: Cut points dividing a set of observations into equal-sized groups
  - Many quantiles ➔ Discovery of false keys
  - Few quantiles ➔ Lose of true keys

<table>
<thead>
<tr>
<th>Initial Data</th>
<th>PH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine1</td>
<td>3.15</td>
</tr>
<tr>
<td>Wine2</td>
<td>3.22</td>
</tr>
<tr>
<td>Wine3</td>
<td>3.23</td>
</tr>
<tr>
<td>Wine4</td>
<td>3.24</td>
</tr>
<tr>
<td>Wine5</td>
<td>3.56</td>
</tr>
<tr>
<td>Wine6</td>
<td>3.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>PH</td>
</tr>
<tr>
<td>Wine1</td>
</tr>
<tr>
<td>Wine2</td>
</tr>
<tr>
<td>Wine3</td>
</tr>
<tr>
<td>Wine4</td>
</tr>
<tr>
<td>Wine5</td>
</tr>
<tr>
<td>Wine6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Final Data</th>
<th>PH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wine1</td>
<td>1</td>
</tr>
<tr>
<td>Wine2</td>
<td>1</td>
</tr>
<tr>
<td>Wine3</td>
<td>2</td>
</tr>
<tr>
<td>Wine4</td>
<td>2</td>
</tr>
<tr>
<td>Wine5</td>
<td>3</td>
</tr>
<tr>
<td>Wine6</td>
<td>3</td>
</tr>
</tbody>
</table>
KEY QUALITY MEASURES

1) Key support
   - **Intuition:** The higher is the support the more sure is a key

2) Key exceptions
   - **Intuition:** A 0-almost key is considered more reliable than a 10-almost key

3) Key size
   - **Intuition:** Keys composed of few properties are preferred (easier to interpret)

4) **Property correlation:** The dependence of properties co-appearing in a key
   - **Intuition:** The less correlated are the properties participating in a key, the more informative the key is
## EXPERIMENTAL DATA

<table>
<thead>
<tr>
<th>Chemical families</th>
<th>Concentration levels in wine</th>
<th>Analyzed molecules</th>
<th>Analysis methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thiols</td>
<td>ppt</td>
<td>3MH = 3 mercaptohexanol&lt;br&gt;3MHA = 3 mercaptohexylacetate</td>
<td>LC-MS/MS</td>
</tr>
<tr>
<td>Esters</td>
<td>ppm</td>
<td>2PHEN = 2-phenylethanol&lt;br&gt;AH = hexyl acetate&lt;br&gt;AI = isoamyl acetate&lt;br&gt;ABPE = phenethyl acetate&lt;br&gt;DE = ethyl decanoate&lt;br&gt;HE = ethyl hexanoate&lt;br&gt;OE = ethyl octanoate&lt;br&gt;BE = ethyl butyrate&lt;br&gt;2HPE = Ethyl lactate&lt;br&gt;3HBE = Ethyl 3-hydroxybutyrate&lt;br&gt;2MBE = Ethyl 2-methylbutyrate&lt;br&gt;2MPE = Ethyl isobutyrate&lt;br&gt;2HICE = Ethyl leucate</td>
<td>GC-MS/MS</td>
</tr>
<tr>
<td>C13-noriprenoïds</td>
<td>Ppb</td>
<td>BDAM = beta-damascenone&lt;br&gt;BION = beta-ionone</td>
<td>GC-MS/MS</td>
</tr>
<tr>
<td>PDMS</td>
<td>Ppb</td>
<td>Dimethylsulfide potential = S-methylmethionine + others compounds</td>
<td>GC-MS/MS</td>
</tr>
<tr>
<td>GSH</td>
<td>ppm</td>
<td>Glutathione</td>
<td>LC-MS/MS</td>
</tr>
</tbody>
</table>
# Examples of Keys

<table>
<thead>
<tr>
<th>Key</th>
<th>Year</th>
<th>Quantiles</th>
<th>Support</th>
<th>Probability</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3MHA, BDAM, GSH, 2MPE, 3MH]</td>
<td>2014</td>
<td>5</td>
<td>73%</td>
<td>26%</td>
<td>5</td>
</tr>
<tr>
<td>[3MHA, GSH, OE, 2HICE, 3MH]</td>
<td>2014</td>
<td>5</td>
<td>73%</td>
<td>26%</td>
<td>5</td>
</tr>
<tr>
<td>[3MHA, AI, PDMS, 2MPE, 3MH]</td>
<td>2014</td>
<td>5</td>
<td>100%</td>
<td>26%</td>
<td>5</td>
</tr>
<tr>
<td>[3MHA, BE, PDMS, 2MPE, 3MH]</td>
<td>2014</td>
<td>5</td>
<td>100%</td>
<td>26%</td>
<td>5</td>
</tr>
<tr>
<td>[BDAM, OE, PDMS, 3MH]</td>
<td>2012</td>
<td>10</td>
<td>100%</td>
<td>17%</td>
<td>4</td>
</tr>
<tr>
<td>[GSH, OE, PDMS, 2PHEN]</td>
<td>2012</td>
<td>10</td>
<td>100%</td>
<td>17%</td>
<td>4</td>
</tr>
<tr>
<td>[AI, BDAM, 2HICE, 3MH]</td>
<td>2012</td>
<td>10</td>
<td>100%</td>
<td>17%</td>
<td>4</td>
</tr>
<tr>
<td>[3MHA, BDAM, GSH, 2MPE]</td>
<td>2013</td>
<td>5</td>
<td>63%</td>
<td>94%</td>
<td>4</td>
</tr>
<tr>
<td>[3MHA, BDAM, GSH]</td>
<td>2013</td>
<td>12</td>
<td>63%</td>
<td>64%</td>
<td>3</td>
</tr>
<tr>
<td>[AH, BDAM, GSH]</td>
<td>2013</td>
<td>12</td>
<td>63%</td>
<td>64%</td>
<td>3</td>
</tr>
<tr>
<td>[BE, 2HICE, 3MH]</td>
<td>2013</td>
<td>12</td>
<td>100%</td>
<td>64%</td>
<td>3</td>
</tr>
<tr>
<td>[BDAM, GSH, 3MH]</td>
<td>2013</td>
<td>12</td>
<td>63%</td>
<td>63%</td>
<td>3</td>
</tr>
<tr>
<td>[GSH, PDMS, 3HBE]</td>
<td>2013</td>
<td>10</td>
<td>63%</td>
<td>83%</td>
<td>3</td>
</tr>
<tr>
<td>[BDAM, GSH, 3MH]</td>
<td>2013</td>
<td>10</td>
<td>63%</td>
<td>83%</td>
<td>3</td>
</tr>
<tr>
<td>[GSH, PDMS, 3HBE]</td>
<td>2014</td>
<td>10</td>
<td>73%</td>
<td>63%</td>
<td>3</td>
</tr>
<tr>
<td>[PDMS, 3HBE, 3MH]</td>
<td>2014</td>
<td>12</td>
<td>100%</td>
<td>44%</td>
<td>3</td>
</tr>
<tr>
<td>[3MHA, GSH, PDMS]</td>
<td>2014</td>
<td>12</td>
<td>73%</td>
<td>44%</td>
<td>3</td>
</tr>
<tr>
<td>[BE, GSH, 3MH]</td>
<td>2014</td>
<td>12</td>
<td>73%</td>
<td>44%</td>
<td>3</td>
</tr>
</tbody>
</table>
VALIDATED KEYS

- 18 out of 104 keys (18%) were validated by the expert
- Support from 63% to 100%
  - Keys with low support can be as well significant

- Evaluated keys contain from 3 to 5 properties
  - Expert chose keys with big size (on contrary to the initial intuition)

- Example: Key \{AI, BDAM, 2HICE, 3MH\}
  - Correlations from 0.05 to 0.42
  - Properties not highly correlated ➔ interesting keys

- First step for predicting wine taste and wine component concentration
OUTLINE

- Introduction
- Key discovery for data linking
  - Link Invalidation
  - Contextual identity
- Conclusion
**OAEI*: RECENT RESULTS**

- Data Linking results for OAEI 2018 - SPIMBENCH Track

<table>
<thead>
<tr>
<th>SPIMBENCH Sandbox</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Time in ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML</td>
<td>0.8348</td>
<td>0.8963</td>
<td>0.8645</td>
<td>6220</td>
</tr>
<tr>
<td>Lily</td>
<td>0.8494</td>
<td>1.0</td>
<td>0.9185</td>
<td>1960</td>
</tr>
<tr>
<td>LogMap</td>
<td>0.9382</td>
<td>0.7625</td>
<td>0.8413</td>
<td>5887</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SPIMBENCH Mainbox</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
<th>Time in ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>AML</td>
<td>0.8385</td>
<td>0.8835</td>
<td>0.8604</td>
<td>37190</td>
</tr>
<tr>
<td>Lily</td>
<td>0.8546</td>
<td>1.0</td>
<td>0.9216</td>
<td>3103</td>
</tr>
<tr>
<td>LogMap</td>
<td>0.8925</td>
<td>0.7094</td>
<td>0.7905</td>
<td>23494</td>
</tr>
</tbody>
</table>

* OAEI: Ontology Alignment Evaluation Initiative
**OAEI*: RECENT RESULTS**

- Data Linking results for OAEI 2018 - SPIMBENCH Track

<table>
<thead>
<tr>
<th></th>
<th>SPIMBENCH Sandbox</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-measure</td>
<td>Time in ms</td>
</tr>
<tr>
<td>AML</td>
<td>0.8348</td>
<td>0.8963</td>
<td>0.8645</td>
<td>6220</td>
</tr>
<tr>
<td>Lily</td>
<td>0.8494</td>
<td>1.0</td>
<td>0.9185</td>
<td>1960</td>
</tr>
<tr>
<td>LogMap</td>
<td>0.9382</td>
<td>0.7625</td>
<td>0.8413</td>
<td>5887</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>SPIMBENCH Mainbox</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-measure</td>
<td>Time in ms</td>
</tr>
<tr>
<td>AML</td>
<td>0.8385</td>
<td>0.8835</td>
<td>0.8604</td>
<td>37190</td>
</tr>
<tr>
<td>Lily</td>
<td>0.8546</td>
<td>1.0</td>
<td>0.9216</td>
<td>3103</td>
</tr>
<tr>
<td>LogMap</td>
<td>0.8925</td>
<td>0.7094</td>
<td>0.7905</td>
<td>23494</td>
</tr>
</tbody>
</table>

* OAEI: Ontology Alignment Evaluation Initiative
In [Jaffri et al., 2008], the authors discuss how erroneous use of owl:sameAs in the interlinking of the DBpedia and DBLP datasets has resulted in publications becoming incorrectly assigned to different authors.

[Halpin et al. 2010] showed that 37% of owl:sameAs links randomly selected among 250 identity links between books were incorrect.

Automatic data linking tools do not guarantee 100% precision, because of:
- Errors, missing information, data freshness, etc.
IDENTITY LINK INVALIDATION

[Principle: use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies or error candidates in the linked resources descriptions.]

Laura Papaleo, Nathalie Pernelle, Fatiha Saïs, Cyril Dumont: Logical Detection of Invalid SameAs Statements in RDF Data. EKAW 2014: 373-384
**IDENTITY LINK INVALIDATION**

**Principle**: use of ontology axioms (functionality, local completeness, asymmetry, etc.) to detect inconsistencies or error candidates in the linked resources descriptions.

- Improvements in data linking **precision** up to **25%**

**Limits**:
- Scalability problems and need of uniform vocabulary in datasets

.nbPages is a Functional Property
IDENTITY PROBLEM AT LOD SCALE

www.bibsonomy.org

> Several domains
> 558 M. identity links
> 28 B. RDF triples
> 48 K. equiv. classes

[Beek et al., 2018]

Multilingual variations of DBpedia

http://sameas.cc/explicit/img
IDENTITY PROBLEM AT LOD SCALE

The largest identity set contains 177,794 terms:

Different countries
Different cities
Albert Einstein

→ quality problems
NETWORK BASED

- Considers the **identity network** build from the **explicit identity network** of sameAs links: removing of symmetric and reflexive links.

- Uses of Louvain **community detection** algorithm to detect subgraphs in the **identity network** that are highly connected.

- Defines a **ranking score** for each (intra-community and inter-community) identity link based on the **density of the community**.

Ranking of identity links

**intra-community erroneousness degree**

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

**inter-community erroneousness degree**

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

- LOD-a-lot dataset [Fernandez et al. 2017]: a compressed data file of 28B triples from LOD 2015 crawl
- An **explicit identity network** of 558.9M edges (links) and 179M nodes (resources)

Example: The *B. Obama* equality set that contain 440 nodes
Barack Obama’s Equality Set

DBpedia IRIs referring to the person Obama in different languages

DBpedia IRIs referring to the person Obama, his senator career

IRIs referring to the presidency and the Obama administration

DBpedia IRIs referring to the person Obama in different functions
Barack Obama’s Equality Set

Low err(e) for the links of this community

These two links have err(e) = 1

Most of the links have err(e) = 0.9
LINK INVALIDATION: NETWORK-BASED APPROACH EVALUATION [Raad et al. 2018]

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

No **benchmark** for qualitative evaluation

**Precision**: manual evaluation of 200 links

- The higher the error degree is the most likely the link will be 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]

**Recall**: **780 incorrect links** between **40 distinct** resources have been introduced in the explicit identity graph. **Recall = 93 %**
SOMETIMES, WE NEED WEAKER IDENTITY ...

- Identity is **context-dependent** [Geach, 1967]
  - allowing two medicines to be considered the same in terms of their chemical substance, but different in terms of their price
SOMETIMES, WE NEED WEEKER IDENTITY ...

- Identity is **context-dependent** [Geach, 1967]
  - allowing two medicines to be considered the same in terms of their chemical substance, but different in terms of their price

- **Identity over time** poses problems
  - a material could it be considered the same, even though some (or even all) of its original components have been replaced by new ones.

---

Bioreactor 1  

Bioreactor 2  

Same as **Bioreactor 1** ?
owl:sameAs predicate is too strict

- 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:

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

∀X∀Y owl:sameAs(X, Y) ∨ p(X, Z) ⇒ p(Y, Z)

Detection of weak-identity links ➔ Contextual Identity
New predicate :identi\text{ConTo} for expressing 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

CONTEXTUAL IDENTITY LINKS

[Raad et al., 2017]
**Contextual Identity Links**

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

Contextual Identity Link Example

\[ \Pi_{a}(Juice) = \{(Juice, \{rdf:Type, expiryDate\}, \{isComposedOf\}), (Banana, \{rdf:Type\}, \{isComposedOf^{-1}\}), (Strawberry, \{rdf:Type\}, \{hasAttribute, isComposedOf^{-1}\}), (Weight, \{rdf:Type, hasValue, hasUnit\}, \{hasAttribute^{-1}\}) \} \]

\[ \text{identiConTo}_{\Pi_{a}(Juice)}(juice1, juice2) \]
It automatically detects and adds these contextual identity links in the knowledge graph.

For each pair of instances \((i_1, i_2)\) of the source class, set of the most specific global contexts in which \((i_1, i_2)\) are identical.

[Raad et al., 2017]
Transformation of Micro-organisms

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

Digestion Process

- Classes: ≈ 5 000
- Individuals: ≈ 42 000
- Statements: ≈ 237 000

*A LIONES project funded by CDS Paris Saclay (2015-2018)*
Detect for each context $GC_i$, the measures $m_i$ where

$$identiConTo_{<GC_i>} (i_1, i_2) \cap observes(i_1, m_1) \rightarrow observes(i_2, m_2)$$

with $m_1 \simeq m_2$

$$identiConTo_{<GC_i>} (i1, i2) \rightarrow same(m_i)$$
### Detection of 38,844 rules

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

The domain experts has evaluated the plausibility of the best **20 rules** (in terms of error rate and support)
LIONES: CONTEXTUAL IDENTITY LINKS

[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>identiConTo_{GC_1} (x, y) → same(pH)</td>
<td>6.19 %</td>
<td>57</td>
</tr>
<tr>
<td>identiConTo_{GC_3} (x, y) → same(Dureté)</td>
<td>1.86 %</td>
<td>66</td>
</tr>
<tr>
<td>identiConTo_{GC_2} (x, y) → same(Friabilité)</td>
<td>4.52 %</td>
<td>647</td>
</tr>
</tbody>
</table>

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

3 Impossible
5 Not very probable
4 Can’t tell
5 Why not
3 Plausible

The error rate decreases of 12% when a global context is replaced by a more specific global context
WARM RULES – GRADUAL CAUSAL RULES DETECTION IN KNOWLEDGE GRAPHS

Application to plant development in climatic warming preoccupation

Phenology: the study of seasonal cycles of plants (timing and duration of flowering, fruiting, leaf out and leaf drop)
GRADUAL CAUSALITY RULE DISCOVERY IN RDF KGS

Pi1 \land Pi2 \land \ldots \land Pi_n \land Pvar_1 \land Pvar_2 \land \ldots \land Pvar_m \Rightarrow Peffect-var

Identity \quad Variability/cause \quad Effect

Example:

Same-leafSize(X,Y) \land SameGroundPh(X,Y) \land

Humidity(X,h1):t1 \land Humidity(Y,h2):t2 \land
Temp(X,temp1):t1 \land Temp(Y,temp2):t2 \land
(h1 > h2) \land (temp1 < temp2) \land (t1 < t2) \Rightarrow flowering-delay(Y)
CONCLUSION

- Semantic Web standards, agronomic data/knowledge and many applications are there

- Promising applications are emerging for which reasoning on data is central:
  - Information retrieval, decision-support, digital-assistants, …

- Many challenges remain to handle at large scale the incomplete, uncretain and evolving knowledge graphs
  - Combining numerical and symbolic AI is challenging but worthwhile to investigate more deeply.
DATA LINKING AND KNOWLEDGE DISCOVERY IN RDF DATA: METHODS AND SOME FEEDBACK FROM AGRONOMIC APPLICATIONS

FATIHA SAÏS

LAHDAK@LRI, PARIS SUD UNIVERSITY, CNRS, PARIS SACLAY UNIVERSITY

Joint work with: N. Pernelle, L. Papaleo, J. Raad and D. Symeonidou

1ST DATAIA DAYS « LIFE SCIENCES & AI», DEC. 4TH 2019
REFERENCES (1)

[Atencia et al. 2014] Manuel Atencia, Jérôme David, Jérôme Euzenat:
Data interlinking through robust linkkey extraction. ECAI 2014: 15-20

[Al-Balkri et al. 2015] Mustafa Al-Bakri, Manuel Atencia, Steffen Lalande, Marie-Christine Rousset:
Inferring Same-As Facts from Linked Data: An Iterative Import-by-Query Approach. AAAI 2015: 9-15


[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.

REFERENCES (2)

Assessing linked data mappings using network measures. In ESWC 2012

When owl:sameAs isn’t the same: An analysis of identity in Linked Data. In ISWC 2010.

Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. In JWS 2012.

[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.

An Automatic Key Discovery Approach for Data Linking. In Journal of Web Semantics
REFERENCES (3)

[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


[Saïs et al.09] Combining a Logical and a Numerical Method for Data Reconciliation.
Fatiha Saïs., Nathalie Pernelle and Marie-Christine Rousset.
In Journal of Data Semantics 2009.

[Soru et al. 2015] Tommaso Soru, Edgard Marx, Axel-Cyrille Ngonga Ngomo:
ROCKER: A Refinement Operator for Key Discovery. WWW 2015: 1025-1033


Non-key probability: The probability that a set of properties contains instances sharing the same values for this set

\[ Pr_k = 1 - e^{-\frac{n(n-1)}{2p}} \]

with

\[ p = \prod_{i=0}^{j} m_i \]

where

\( j \): # of properties
\( m_i \): # of distinct values
KEY QUALITY MEASURES

Non-key probability

- **Example:** 100 distinct wines

<table>
<thead>
<tr>
<th></th>
<th>Case1</th>
<th>Case2</th>
<th>Case3</th>
</tr>
</thead>
<tbody>
<tr>
<td>WineName</td>
<td>15 distinct values</td>
<td>50 distinct values</td>
<td>90 distinct values</td>
</tr>
<tr>
<td>YearProduction</td>
<td>10 distinct values</td>
<td>50 distinct values</td>
<td>80 distinct values</td>
</tr>
<tr>
<td>Non-key probability</td>
<td>1</td>
<td>0.87</td>
<td>0.49</td>
</tr>
</tbody>
</table>

- **Intuition:** Higher is the non-key probability of \{wineName, yearProduction\} more the discovered key is important
EXPERIMENTS

Experimental data: 3 wine aroma datasets

- Different chemical based flavourings of wine
  - Concentration of each flavour in a wine

<table>
<thead>
<tr>
<th></th>
<th># Instances</th>
<th># Flavours</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 (2011 – 2012)</td>
<td>63</td>
<td>19</td>
</tr>
<tr>
<td>D2 (2012 – 2013)</td>
<td>59</td>
<td>19</td>
</tr>
<tr>
<td>D3 (2013 – 2014)</td>
<td>44</td>
<td>19</td>
</tr>
</tbody>
</table>

Goal: Verify the interest of keys in numerical data

- Evaluate the **impact of quantiles** in the results
- Evaluate the **quality measures**
DATA
PREPROCESSING

Quantiles

- Use the non-key probability to define the number of quantiles

5, 10, 12 quantiles

- Setting it at less than 5 => no keys are obtained
- Using 5 to 12 quantiles ensured a significantly high probability