Enrichment of RDF Knowledge Graphs with Contextual Identity Links and Fuzzy Temporal Data

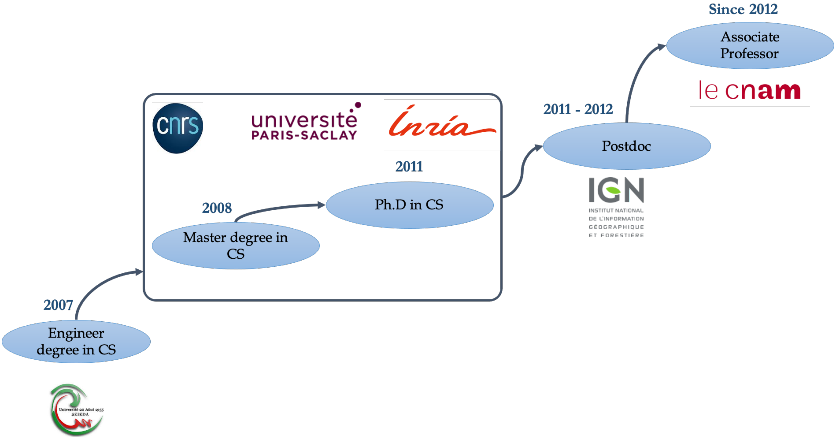
Fayçal Hamdi

Laboratoire CEDRIC, Equipe ISID
Conservatoire National des Arts et Métiers, Paris, France

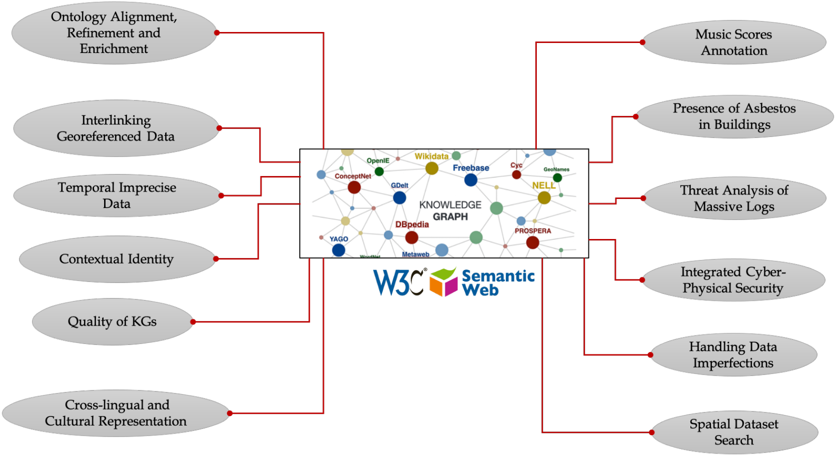
Soutenance d'Habilitation à Diriger des Recherches - 5 novembre 2020

Outline

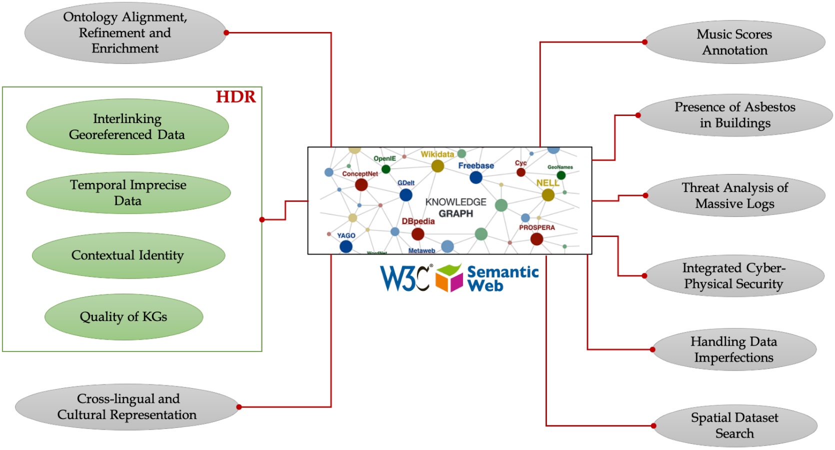
Curriculum Vitae



Research Statement

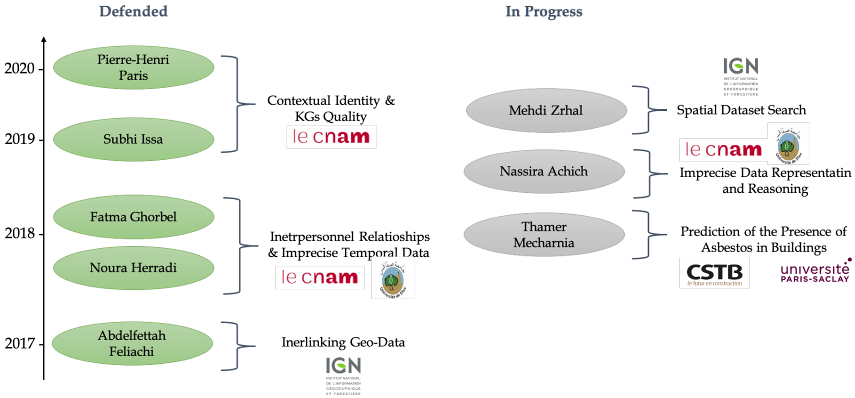


Research Statement



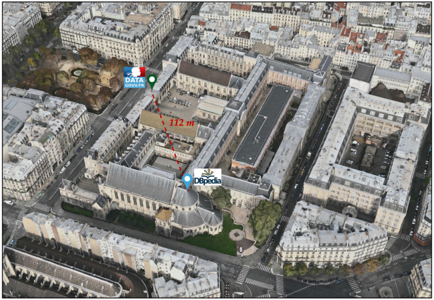
Research Statement

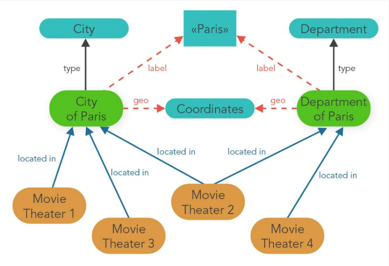
Thesis Supervision



Outline

Context





James defended his thesis at the Sorbonne University at the end of the 90s

from 1996 to 2000?

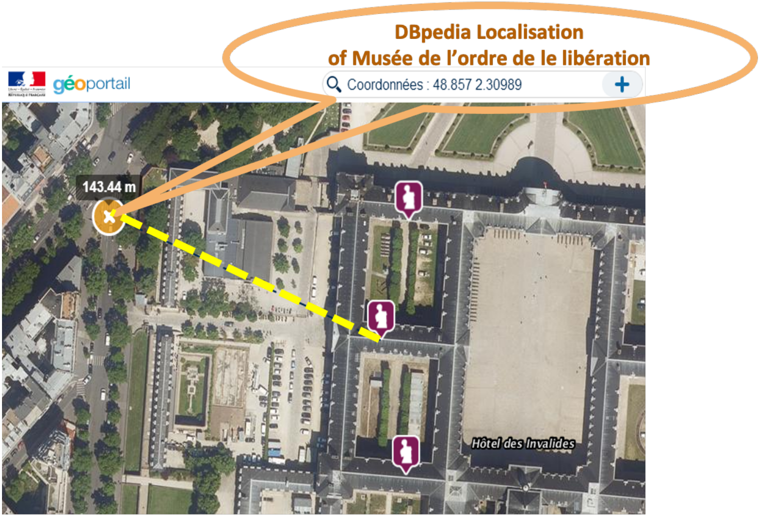
Outline

1. Curriculum Vitae
2. Context
3. Enriching KGs
   - Geo-Domain Identity Links
   - Contextual Identity Links
   - Fuzzy Temporal Data
4. Quality of KGs: Completeness and Conciseness
   - Completeness
   - Conciseness
5. Conclusion

Outline

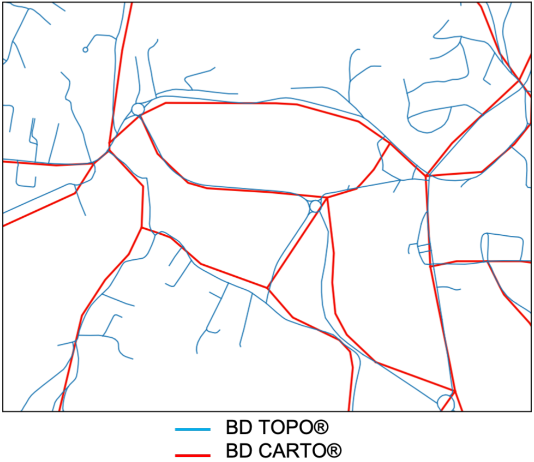
Heterogeneity of geometries on the Web of data

- Difference in planimetric accuracies



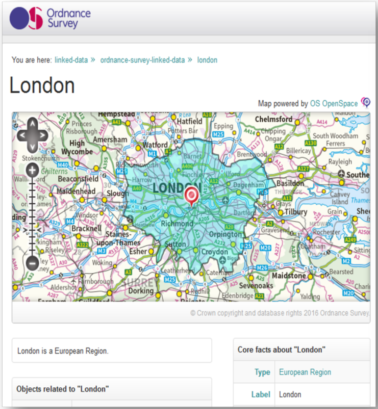
Heterogeneity of geometries on the Web of data

- Difference in geometric resolutions



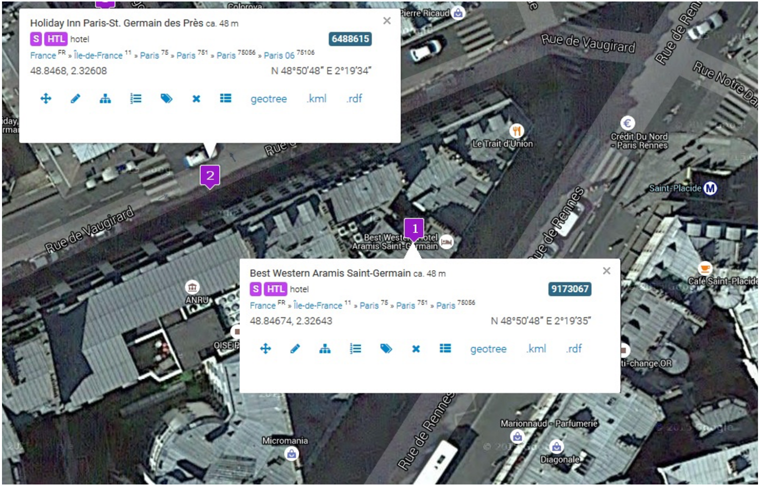
Heterogeneity of geometries on the Web of data

- Difference in geometric modeling



Heterogeneity of geometries on the Web of data

- Internal heterogeneity



The XY Semantics Ontology

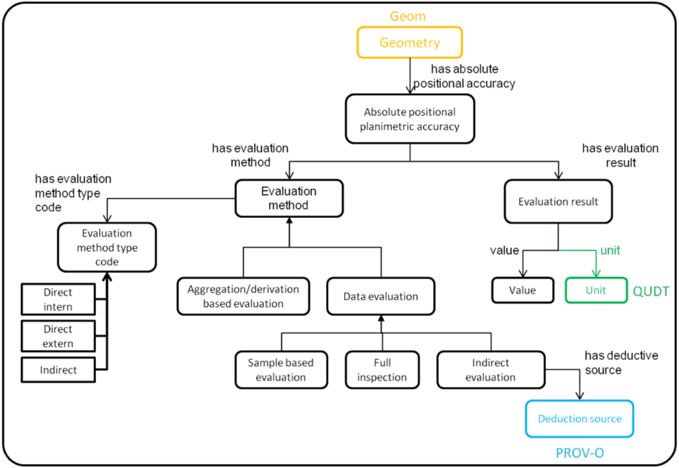
Characteristics that are more likely to affect the setting of a spatial data matching process:

- The absolute positional accuracy of geometries
- The geometry capture rules (geometric modeling)
- The vagueness of the spatial characteristics of the geographic entities represented by the geometries
- The level of detail of the data sources

The XY Semantics Ontology

http://data.ign.fr/def/xysemantics

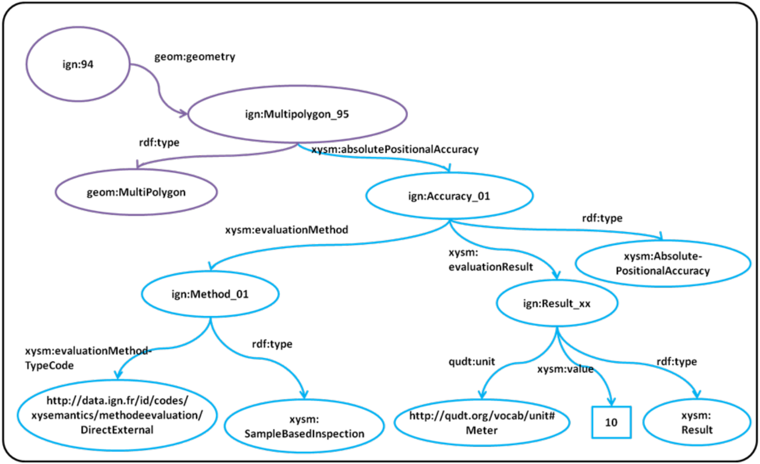
- An excerpt describing the planimetric accuracy of geometries



The XY Semantics Ontology

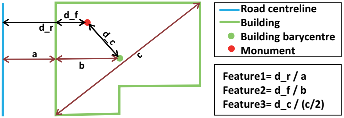
http://data.ign.fr/def/xysemantics

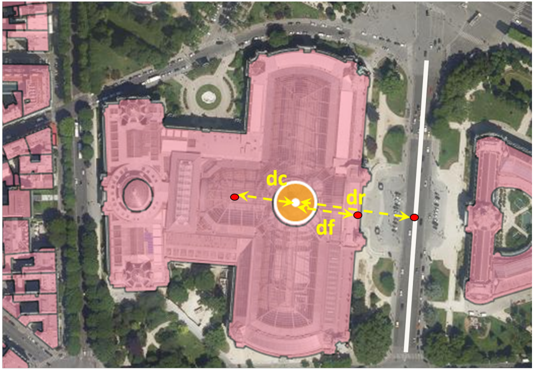
- An excerpt describing the planimetric accuracy of geometries



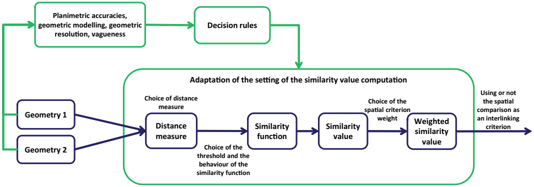
Populating the XY Ontology

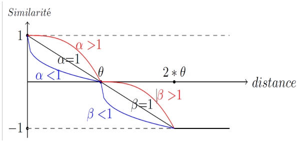
- When geometric metadata are not provided

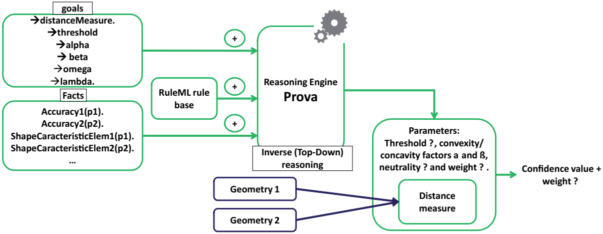




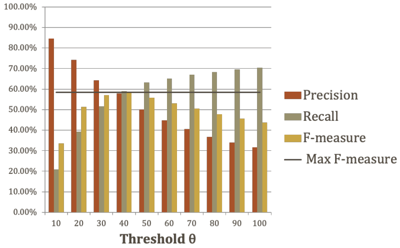
Our Adaptive Interconnection Approach



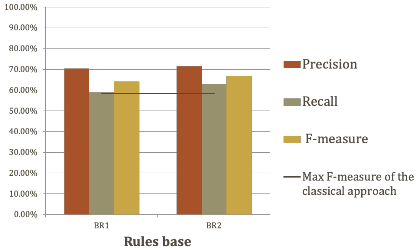




Results

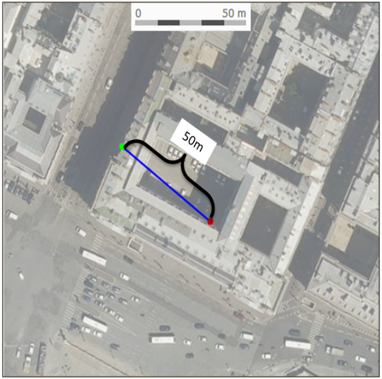


Classical Approach

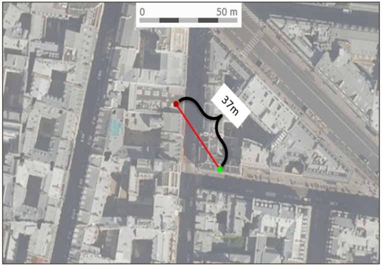


Adaptative Approach

Results



Generated Link



Avoided Link

Outline

1. Curriculum Vitae
2. Context
3. Enriching KGs
   - Geo-Domain Identity Links
   - Contextual Identity Links
   - Fuzzy Temporal Data
4. Quality of KGs: Completeness and Conciseness
   - Completeness
   - Conciseness
5. Conclusion

Leibniz's law of Identity

owl:sameAs semantics is based on:

Identity of indiscernibles:

$$\forall x, \forall y (\forall p, \forall o, (\langle x, p, o \rangle \text{ and } \langle y, p, o \rangle) \rightarrow x = y)$$

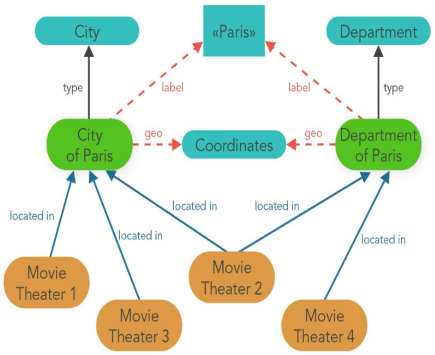
Indiscernibility of identicals:

$$\forall x, \forall y (x = y \rightarrow \forall p, \forall o, (\langle x, p, o \rangle \rightarrow \langle y, p, o \rangle))$$

$\Longrightarrow$ property-value couples can be propagated from one entity to another identical entity and thus, increase completeness

Context

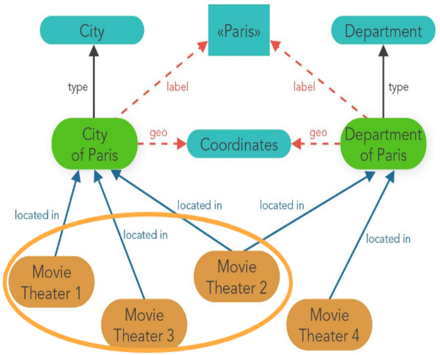
- Both the city and the department of Paris are different in a legal context
- But, they are identical in a geographical context
- What if a user want to retrieve movie theaters in Paris?
  - Only 3 are connected to the city
  - Only 2 are connected to the department
  - GeoNames
- **Contextual identity** is a possible answer



$\Longrightarrow$ Contextual identity must allow the propagation of properties in certain cases

Context

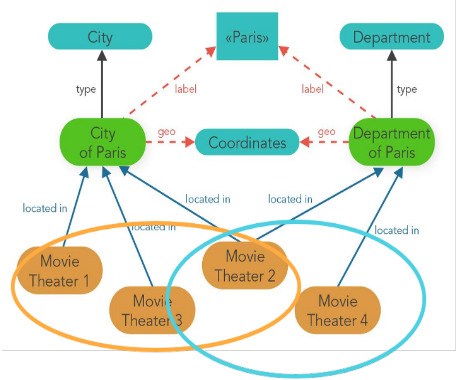
- Both the city and the department of Paris are different in a legal context
- But, they are identical in a geographical context
- What if a user want to retrieve movie theaters in Paris?
  - Only 3 are connected to the city
  - Only 2 are connected to the department
  - GeoNames
- **Contextual identity** is a possible answer



$\Longrightarrow$ Contextual identity must allow the propagation of properties in certain cases

Context

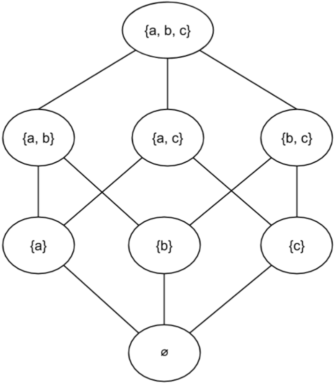
- Both the city and the department of Paris are different in a legal context
- But, they are identical in a geographical context
- What if a user want to retrieve movie theaters in Paris?
  - Only 3 are connected to the city
  - Only 2 are connected to the department
  - GeoNames
- **Contextual identity** is a possible answer



$\Longrightarrow$ Contextual identity must allow the propagation of properties in certain cases

Related Work

- Identity context = set of properties (indiscernibility set)
  - Entities must share the same value for each property
- Contexts can be represented with a lattice



But there is no clue on what to do with other properties
$\Longrightarrow$ No propagation

Beek W, Schlobach S, van Harmelen F. **A contextualised semantics for owl: sameAs.** In European Semantic Web Conference. Springer, Cham, 2016.

Related Work

Identity context = indiscernibility set (Π) + propagation set (Ψ) + alignment procedure ($\approx$)

$$x =_{(\Pi,\Psi,\approx)} y \leftrightarrow \forall (p_1, p_2) \in \Pi^2 \text{ with } p_1 \approx p_2$$
$$\text{and } \forall v_1, v_2 \text{ with } v_1 \approx v_2 : \langle x, p_1, v_1 \rangle \leftrightarrow \langle y, p_2, v_2 \rangle$$

$$x =_{(\Pi,\Psi,\approx)} y \rightarrow \forall (p_1, p_2) \in \Psi^2 \text{ with } p_1 \approx p_2$$
$$\text{and } \forall v_1, v_2 \text{ with } v_1 \approx v_2 : \langle x, p_1, v_1 \rangle \leftrightarrow \langle y, p_2, v_2 \rangle$$

$\Longrightarrow$ Users must **provide everything**

Idrissou, Al Koudous, et al. **Is my: sameAs the same as your: sameAs? Lenticular lenses for context-specific identity.** In Proceedings of the Knowledge Capture Conference. 2017.

How to find a propagation set of properties?

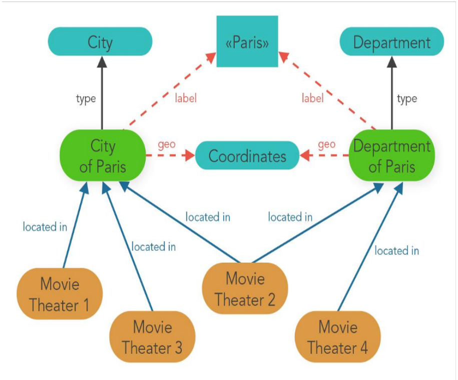
- Identity context based on Idrissou et al.'s definition
- Tobler's first law: "Everything is related to everything else, but near things are more related than distant things."

  $\Longrightarrow$ Propagable properties could be semantically related to indiscernible properties

- Sentences describing properties could be transformed into numerical vectors
- Vectors representing propagable properties must be close to vectors representing indiscernible properties

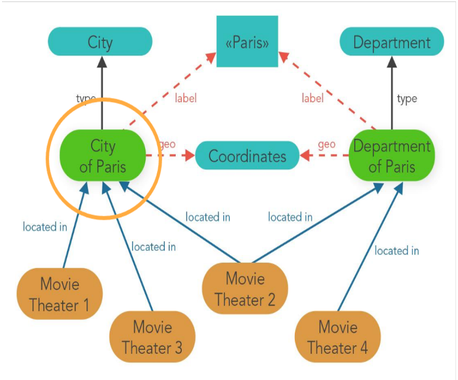
Property Propagation

- Sample knowledge graph about Paris and its movie theaters
- We consider the City of Paris as the seed of the identity lattice.



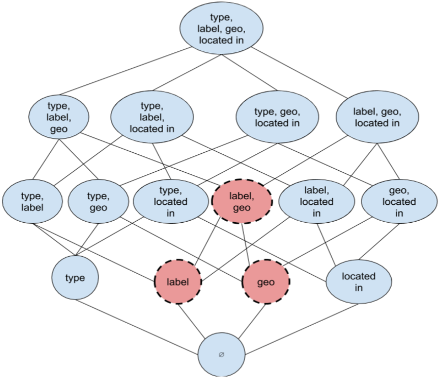
Property Propagation

- Sample knowledge graph about Paris and its movie theaters
- We consider the City of Paris as the seed of the identity lattice.



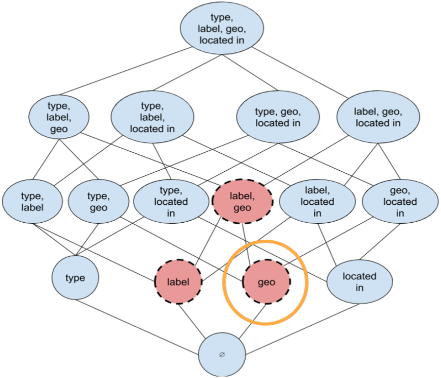
Property Propagation

- Simplified identity lattice
- Each node correspond to the an indiscernibility set
- Only red nodes have contextually identical entities



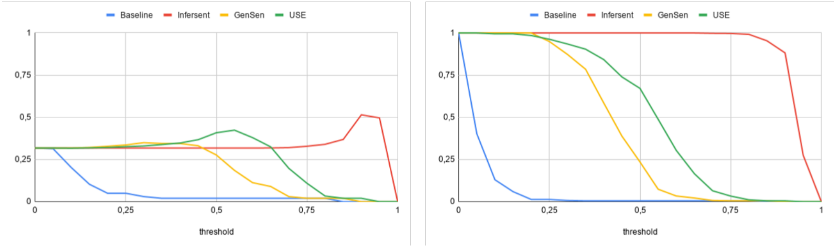
Property Propagation

- Candidate properties for propagation = "type", "label" and "located in"
- We compute the embeddings of the descriptions of the four properties
- The vector representing "located in" is close to the vector representing "geo"



$\Longrightarrow$ "located in" can be propagated for the indiscernibility set geo

Results

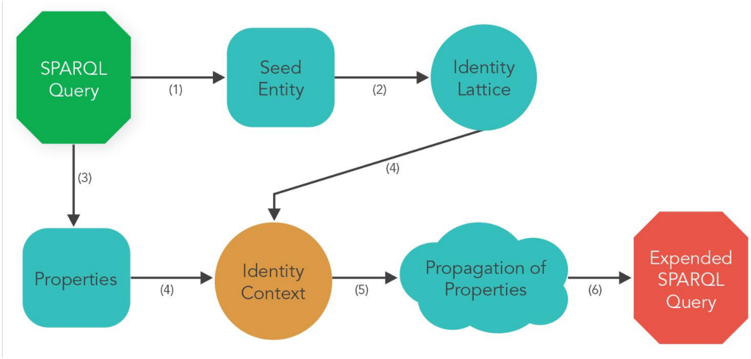


- Gold standard = 100 entities (5 classes)
- Baseline vs Infersent vs GenSen vs USE $\Longrightarrow$ The winner is Infersent

<u>Conclusions:</u>

- Textual descriptions are useful to discover properties that are propagable
- Highly dependent on the encoder

Framework for Propagation of Properties

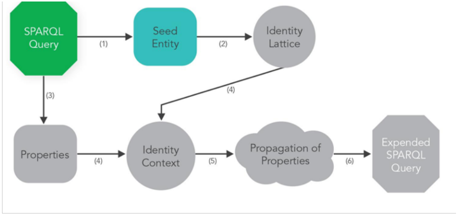


Example

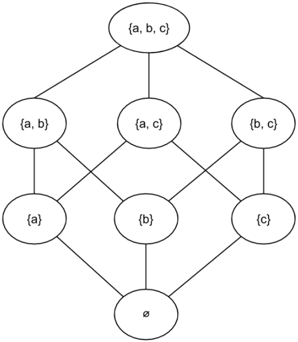
Who are the convicted members of Les Républicains?

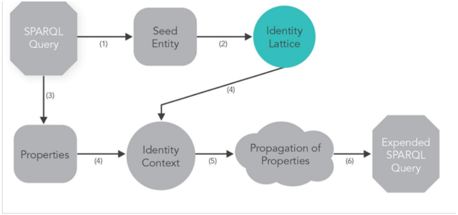
```
SELECT DISTINCT ?politician ?crime
WHERE {
  ?politician :memberOf :TheRepublicans ;
       :convictedOf ?crime .
}
```

# of results w/o context	2



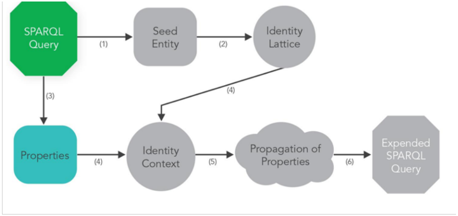
Example


{a, b, c}
{a, b}
{a, c}
{b, c}
{a}
{b}
{c}
∅


SPARQL Query
(1)
Seed Entity
(2)
Identity Lattice
(3)
(4)
Properties
(4)
Identity Context
(5)
Propagation of Properties
(6)
Expended SPARQL Query

Example

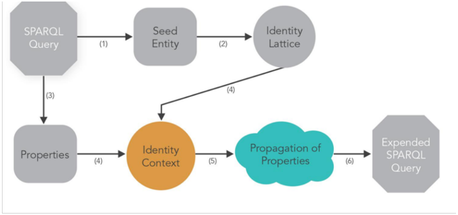
```
SELECT DISTINCT ?politician ?crime
WHERE {
  ?politician :memberOf :TheRepublicans ;
       :convictedOf ?crime .
}
```



Example

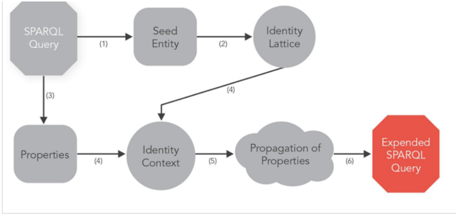
The user must choose the most appropriate identity context among those proposed.

Seed	The Republicans
Ψ	member of, political party
Π	country, political, ideology
Contextually identical entities	UMP, RPR, UDR, UNR



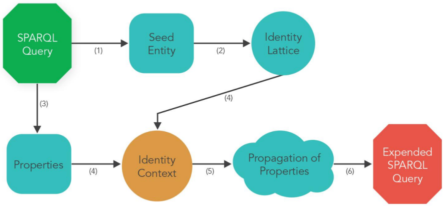
Example

```
SELECT DISTINCT ?politician ?crime
WHERE {
  VALUES (?party) {
    (:TheRepublicans) (:UMP) (:RPR) (:UDR)
     (:UNR)
  }
  ?politician :memberOf ?party ;
       :convictedOf ?crime .
}
```



Example

Seed	The Republicans
Ψ	member of, political party
Π	country, political, ideology
Contextually identical entities	UMP, RPR, UDR, UNR
# of results w/o context	2
# of results w/ context	**13**



Outline

Imprecise time interval

- How to represent and reason about:

Alexandre was married to Nicole around 1981 until the end of the 90s

Imprecise time interval

- How to represent and reason about:

Alexandre was married to Nicole around 1981 until the end of the 90s

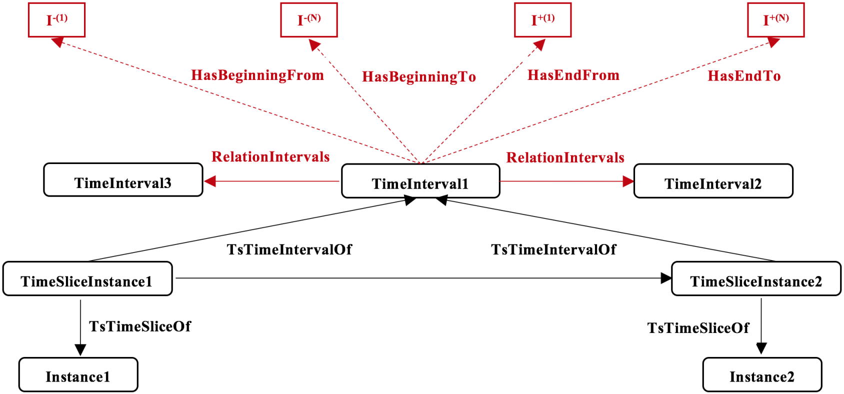
from 1980 to 1982? **from 1995 to 2000?**

Our Approaches

- A Crisp-Based Approach
  - Extend the 4D-fluents model to represent imprecise time intervals and their crisp relationships in OWL 2
  - Reason on imprecise time intervals by extending the Allen's interval algebra in a crisp way
  - Infer interval relations via a set of SWRL rules
- A Fuzzy-Based Approach
  - Extend the 4D-fluents model to represent imprecise time intervals and their relationships in Fuzzy-OWL 2
  - Reason on imprecise time intervals by extending the Allen's interval algebra in a fuzzy gradual personalized way
  - Infer fuzzy interval relations using a set of Mamdani IF-THEN rules

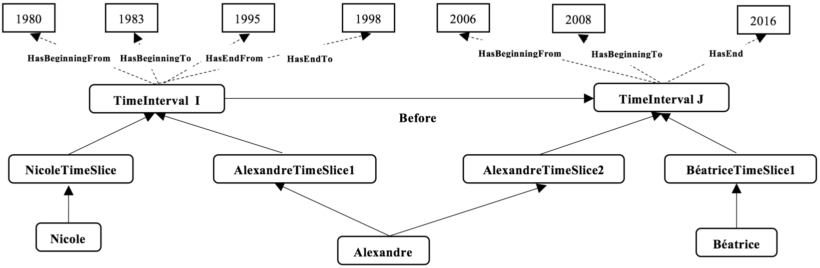
A Crisp-Based Approach

4D-Fluents Extension



A Crisp-Based Approach

4D-Fluents Extension



A Crisp-Based Approach

Crisp temporal interval relations

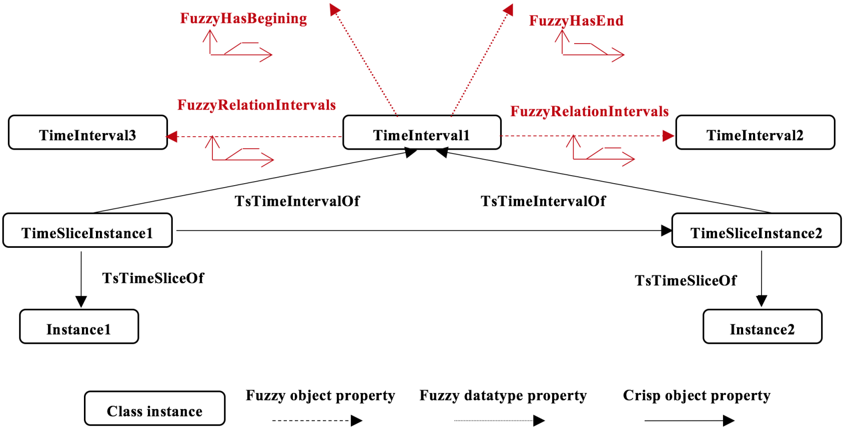
Relation	Inverse	Interpretation	Relations between interval bounds
$Before(I,J)$	$After(I,J)$	$\forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-} : (I^{+(i)} < J^{-(j)})$	$I^{+(N)} < J^{-(1)}$
$Meets(I,J)$	$MetBy(I,J)$	$\forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-} : (I^{+(i)} = J^{-(j)})$	$(I^{+(1)} = J^{-(1)}) \wedge (I^{+(N)} = J^{-(N)})$
$Overlaps(I,J)$	$OverlappedBy(I,J)$	$\forall I^{-(i)} \in I^{-}, \forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-}, \forall J^{+(j)} \in J^{+} : (I^{-(i)} < J^{-(j)}) \wedge (J^{-(j)} < I^{+(i)}) \wedge (I^{+(i)} < J^{+(j)})$	$(I^{-(N)} < J^{-(1)}) \wedge (J^{-(N)} < I^{+(1)}) \wedge (I^{+(N)} < J^{+(1)})$
$Starts(I,J)$	$StartedBy(I,J)$	$\forall I^{-(i)} \in I^{-}, \forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-}, \forall J^{+(j)} \in J^{+} : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} < J^{+(j)})$	$(I^{-(1)} = J^{-(1)}) \wedge (I^{-(N)} = J^{-(N)}) \wedge (I^{+(N)} < J^{+(1)})$
$During(I,J)$	$Contains(I,J)$	$\forall I^{-(i)} \in I^{-}, \forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-}, \forall J^{+(i)} \in J^{+} : (J^{-(j)} < I^{-(i)}) \wedge (I^{+(i)} < J^{+(i)})$	$(J^{-(N)} < I^{-(1)}) \wedge (I^{+(N)} < J^{+(1)})$
$Ends(I,J)$	$EndedBy(I,J)$	$\forall I^{-(i)} \in I^{-}, \forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-}, \forall J^{+(j)} \in J^{+} : (I^{-(i)} < J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$(J^{-(N)} < I^{-(1)}) \wedge (I^{+(1)} = J^{+(1)}) \wedge (I^{+(N)} = J^{+(N)})$
$Equal(I,J)$	$Equal(I,J)$	$\forall I^{-(i)} \in I^{-}, \forall I^{+(i)} \in I^{+}, \forall J^{-(j)} \in J^{-}, \forall J^{+(j)} \in J^{+} : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$(I^{-(1)} = J^{-(1)}) \wedge (I^{-(N)} = J^{-(N)}) \wedge (I^{+(1)} = J^{+(1)}) \wedge (I^{+(N)} = J^{+(N)})$

- A SWRL Rule:

$$TimeInterval(I) \wedge TimeInterval(J) \wedge HasEndFrom(I,a) \wedge HasBeginningFrom(J,b) \wedge Equals(a,b) \wedge HasEndTo(I,c) \wedge HasBeginningTo(J,d) \wedge Equals(c,d) \rightarrow Meet(I,J)$$

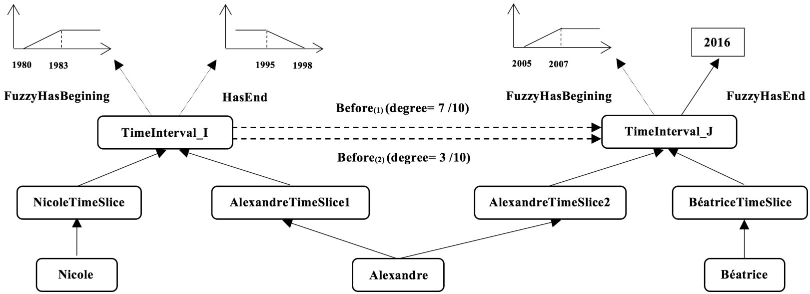
A Fuzzy-Based Approach

4D-Fluents Extension



A Fuzzy-Based Approach

4D-Fluents Extension



A Fuzzy-Based Approach

Fuzzy gradual personalized temporal interval relations

Relation	Inverse	Relations between bounds	Definition
$Before^{(\alpha,\beta)}_{(K)}(I,J)$	$After^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : (I^{+(i)} < J^{-(j)})$	$Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{-(1)})$
$Meets^{(\alpha,\beta)}(I,J)$	$MetBy^{(\alpha,\beta)}(I,J)$	$\forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^- : (I^{+(i)} = J^{-(j)})$	$Min(Same^{(\alpha,\beta)}(I^{+(1)}, J^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(N)}, J^{-(N)}))$
$Overlaps^{(\alpha,\beta)}_{(K)}(I,J)$	$OverlappedBy^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} < J^{-(j)}) \wedge (J^{-(j)} < I^{+(i)}) \wedge (I^{+(i)} < J^{+(j)})$	$Min(Precede^{(\alpha,\beta)}_{(K)}(I^{-(N)}, J^{-(1)}) \wedge Precede_(K)^{(\alpha,\beta)}(J^{-(N)}, I^{+(1)}) \wedge Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{+(1)}))$
$Starts^{(\alpha,\beta)}_{(K)}(I,J)$	$StartedBy^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} < J^{+(j)})$	$Min(Same^{(\alpha,\beta)}(I^{-(1)}, J^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{-(N)}, J^{-(N)}) \wedge Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{+(1)}))$
$During^{(\alpha,\beta)}_{(K)}(I,J)$	$Contains^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(i)} \in J^+ : (J^{-(j)} < I^{-(i)}) \wedge (I^{+(i)} < J^{+(i)})$	$Min(Precede^{(\alpha,\beta)}_{(K)}(J^{-(N)}, I^{-(1)}) \wedge Precede^{(\alpha,\beta)}_{(K)}(I^{+(N)}, J^{+(1)}))$
$Ends^{(\alpha,\beta)}_{(K)}(I,J)$	$EndedBy^{(\alpha,\beta)}_{(K)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} < J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$Min(Precede^{(\alpha,\beta)}_{(K)}(J^{-(N)}, I^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(1)}, J^{+(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(N)}, J^{+(N)}))$
$Equal^{(\alpha,\beta)}(I,J)$	$Equal^{(\alpha,\beta)}(I,J)$	$\forall I^{-(i)} \in I^-, \forall I^{+(i)} \in I^+, \forall J^{-(j)} \in J^-, \forall J^{+(j)} \in J^+ : (I^{-(i)} = J^{-(j)}) \wedge (I^{+(i)} = J^{+(j)})$	$Min(Same^{(\alpha,\beta)}(I^{-(1)}, J^{-(1)}) \wedge Same^{(\alpha,\beta)}(I^{-(N)}, J^{-(N)}) \wedge Same^{(\alpha,\beta)}(I^{+(1)}, J^{+(1)}) \wedge Same^{(\alpha,\beta)}(I^{+(N)}, J^{+(N)}))$

- A Mamdani IF-THEN rule:

(define-concept Rule0 (g-and (some Precede$_{(1/1)}$ Fulfilled) (some Precede$_{(1/2)}$ Fulfilled) Fulfilled) (some Precede$_{(1/3)}$ Fulfilled) (some Overlaps$_{(1)}$ True)))

Outline

Linked Data Quality

What is the meaning of Quality?

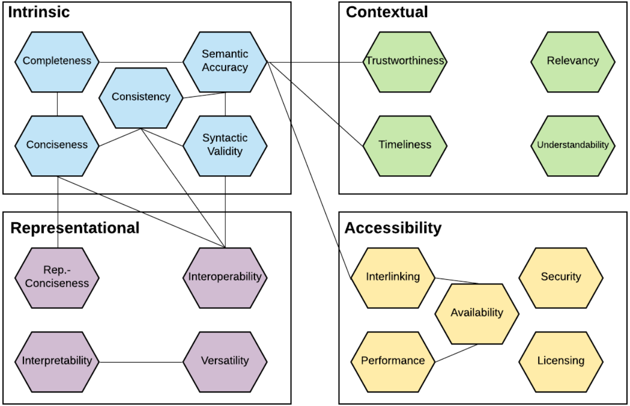
- Quality is defined as *fitness for use*
- The degree to which data suits requirements

Dimensions: accuracy, completeness, consistency, timeliness,...

Consequence of Poor Quality

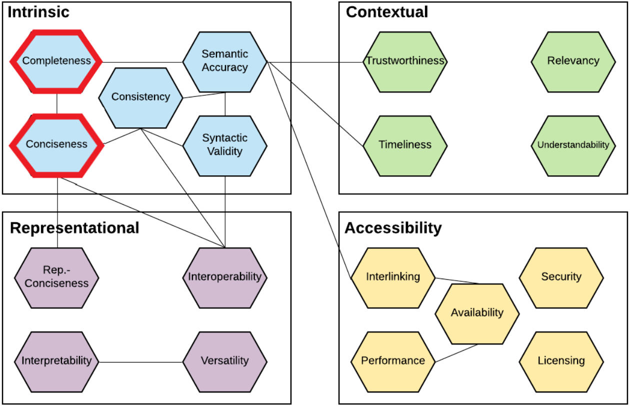
- Poor data model (relevancy, level of detail, granularity, etc.)
- Inconsistency of data values (accuracy, completeness, trustworthiness, etc.)
- Integration issues (interlinking with other data sources, applicability for federated query)
- Loss in output leading to extra charges (time, cost, etc.)

Linked Data Quality Dimensions



Adapted from "Quality Assessment for Linked Data: A Survey", Zaveri et al. 2014

Linked Data Quality Dimensions



Adapted from "Quality Assessment for Linked Data: A Survey", Zaveri et al. 2014

Outline

Motivating Example

```
SELECT ?subject WHERE {
      ?subject rdf:type dbo:Scientist .
}

For each subject do
SELECT ?property ?value WHERE {
      subject ?property ?value .
}
return Scientist_schema
```

Is every scientist described by all the **properties**?

First name, last name, birth date, birth place, etc.

Motivating Example

We need a **reference schema** to calculate completeness

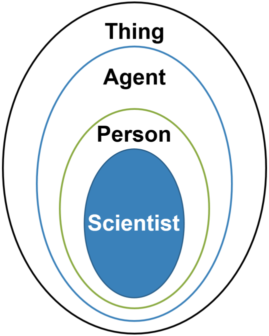
- A reference scientist schema (ontology) could be:

$$Scientist\_Schema = \{Properties\ on\ Scientist\} \cup \{Properties\ on\ Person\} \cup \{Properties\ on\ Agent\} \cup \{Properties\ on\ Thing\}$$

such that: $Scientist \sqsubseteq Person \sqsubseteq Agent \sqsubseteq Thing$

Motivating Example

$$Comp(Albert\_Einstein) = \frac{|Properties\ on\ Albert\_Einstein|}{|Scientist\_Schema|}$$
$$= \frac{21}{664} = 3.16\%$$



Is this schema relevant?

Motivating Example

$$Comp(Albert\_Einstein) = \frac{|Properties\ on\ Albert\_Einstein|}{|Scientist\_Schema|} = \frac{21}{664} = 3,16\%$$

The property **weapon** is in *Scientist_Schema*, but it is not relevant to the instance *Albert_Einstein* Data completeness can be achieved

with a suitable schema containing **mandatory properties**

The approach overview

Goal

Elaborate a solution for RDF data completeness assessment in the absence of a reference/gold schema

- Explore instances to get an idea how they are actually describing
- Property frequently used by several instances of a class is **more important** than less often used one

$\Longrightarrow$ Extracting **properties used more frequently** than others to describe instances of a given class and calculating a completeness in respect to these properties

The Mining-based Approach

The Mining-based Approach includes two steps:

1. **Properties mining**: Applying the well known FP-growth algorithm for mining maximal frequent itemsets $\mathcal{MFP}$
2. **Completeness calculation**: Using the apparition frequency of items (properties) in $\mathcal{MFP}$, to give each of them a weight and calculate the completeness of each transaction (regarding the presence or absence of properties)

Properties mining

Example

Instance	Transaction
The_Godfather	{director, musicComposer}
Goodfellas	{director, editing}
True_Lies	{director, editing, musicComposer}

Let $\xi = 60\%$ and the set of frequent patterns

$$\mathcal{FP} = \{\{director\}, \{musicComposer\}, \{editing\}, \{director, musicComposer\} \\ \{director, musicComposer\}\}$$

The $\mathcal{MFP}$ set would be:

$$\mathcal{MFP} = \{\{director, musicComposer\}, \{director, editing\}\}$$

Completeness calculation

- Carry out for each transaction, a comparison between its corresponding properties and the $\mathcal{MFP}$ set

Definition (*Completeness $\mathcal{CP}$*)

Let $\mathcal{I}'$ a subset of instances, $\mathcal{T}$ the set of transactions constructed from $\mathcal{I}'$, and $\mathcal{MFP}$ a set of maximal frequent pattern. The completeness of $\mathcal{I}'$ corresponds to the completeness of its transaction vector $\mathcal{T}$ obtained by calculating the average of the completeness of $\mathcal{T}$ regarding each pattern in $\mathcal{MFP}$. Therefore, we define the completeness $\mathcal{CP}$ of a subset of instance $\mathcal{I}'$ as follows:

$$\mathcal{CP}(\mathcal{I}') = \frac{1}{|\mathcal{T}|} \sum_{k=1}^{|\mathcal{T}|} \sum_{j=1}^{|\mathcal{MFP}|} \frac{\delta(E(t_k), \hat{P}_j)}{|\mathcal{MFP}|} \tag{1}$$

such that: $\hat{P}_j \in \mathcal{MFP}$, and $\delta(E(t_k), \hat{P}_j) = \begin{cases} 1 & \text{if } \hat{P}_j \subset E(t_k) \\ 0 & \text{otherwise} \end{cases}$

Completeness calculation

Example

Instance	Transaction
The_Godfather	{director, musicComposer}
Goodfellas	{director, editing}
True_Lies	{director, editing, musicComposer}

The completeness of this subset of instances regarding $\mathcal{MFP} = \{\{director, musicComposer\}, \{director, editing\}\}$, would be:

$$\mathcal{CP}(\mathcal{I}') = (2 * (1/2) + (2/2))/3 = 0.67$$

Prototype: LOD-CM

Welcome

A tool designed to help users of RDF knowledge graphs.

What is LOD-CM?

LOD-CM is a tool that produces a Conceptual Model (CM) through a UML class diagram. It mines maximal frequent patterns (also known as maximal frequent itemset) upon properties used by instances of a given OWL class to build the most appropriate CMs.

For a given dataset, you can **choose a class** among its classes, then **choose a threshold** corresponding to the minimum percentage of instances having a set of properties, and we compute CMs. For each group of properties simultaneously present above the threshold, we create a class diagram.

But why would I use that?

- UML class diagrams are *easy to read and understand*.
- CMs allow a user to *explore* dataset *without prior knowledge*.
- A user can easily *compare* two CMs *to choose* the better suited dataset.

Let's try it!

Select a dataset | Select a class | Select a threshold | Let's go!

Prototype: LOD-CM

Conceptual model for ***Film*** class in ***DBpedia***

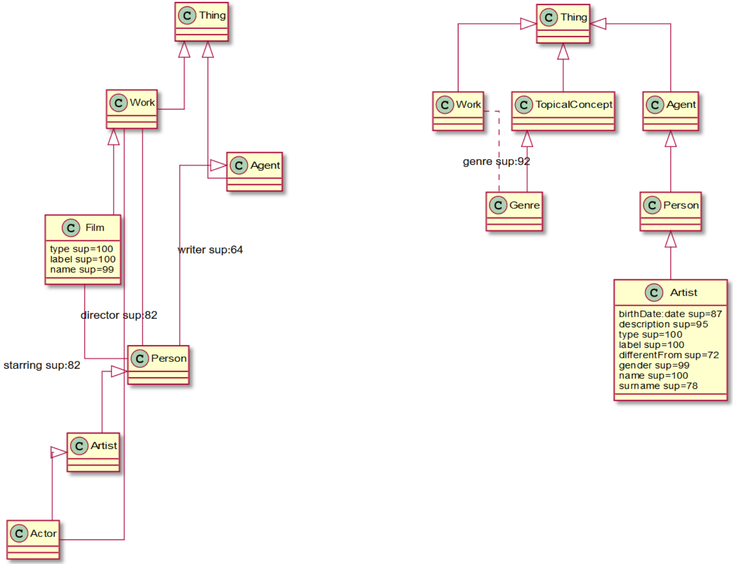
Current threshold is set to 50%, i.e. all properties of a group are present together in at least 50% of Film instances.

Select a group of maximal frequent itemset:

Each property group is present simultaneously in 50% of instances.

- (●) director, label, name, runtime, starring, type
- ( ) director, label, name, starring, type, writer
- ( ) label, name, runtime, type, writer

Prototype: LOD-CM



Evaluation

- Experiments were performed on the well-known real-world datasets, DBpedia, publicly available on the LOD cloud
- We chose two relatively distant versions; **v3.6** generated in **March/April 2013**, and **v2015-04** generated in **February/March 2015**
- For each dataset we have chosen a couple of categories. $C = \{Film, Organisation, Scientist, PopulatedPlace\}$

Evaluation

- For the properties used in the resources descriptions, we have chosen the English datasets ***mapping-based properties***, ***instance types***, and ***labels***

Table 1: Number of resources/category

	Film	Organisation	PopulatedPlace	Scientist
v3.6(2013)	53,619	147,889	340,443	9,726
v2015-04	90,060	187,731	455,398	20,301

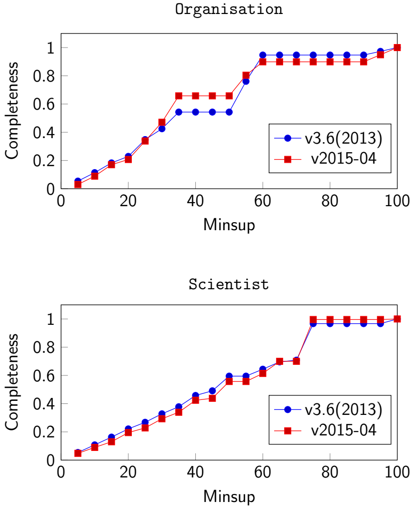


Figure 1: Completeness of DBpedia v3.6 and v2015-04 when varying the minimum support ξ

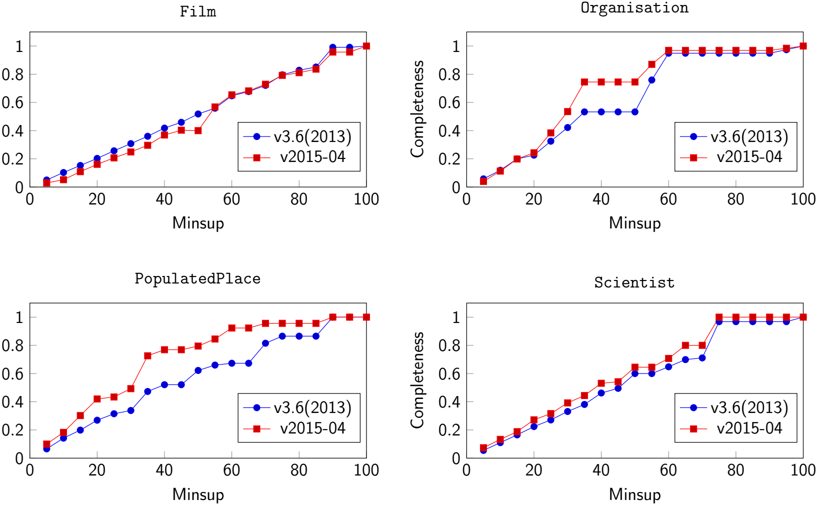


Figure 2: Completeness of equivalent resources from DBpedia v3.6 and v2015-04

Outline

Conciseness Dimension

Conciseness aims to avoid repetition through elements having the same meaning with different identifiers or names

Dataset is concise if does not contain:

- two equivalent **classes/predicates** with different names (Schema level)
- two equivalent **objects** with different names (Instance level)

Conciseness Dimension

Conciseness aims to avoid repetition through elements having the same meaning with different identifiers or names

Dataset is concise if does not contain:

- two equivalent **classes/predicates** with different names (Schema level)
- two equivalent objects with different names (Instance level)

Our objective: Discovering **synonymously used** predicates (conciseness at schema level)

Motivating Example

```
SELECT ?s WHERE { ?s birthPlace France }
```

Subject	Predicate	Object
Emma Watson	nationality	British
Emma Watson	bornIn	France
Emma Watson	bornOn	15-04-1990
Antoine Griezmann	birthPlace	France
Antoine Griezmann	height	1,74
Antoine Griezmann	type	Footballer

Motivating Example

```
SELECT ?s WHERE { ?s birthPlace France }
```

Subject	Predicate	Object
Emma Watson	nationality	British
Emma Watson	bornIn	France
Emma Watson	bornOn	15-04-1990
Antoine Griezmann	birthPlace	France
Antoine Griezmann	height	1,74
Antoine Griezmann	type	Footballer

Motivating Example

```
SELECT ?s WHERE { ?s birthPlace France }
```

Subject	Predicate	Object
Emma Watson	nationality	British
Emma Watson	bornIn	France
Emma Watson	bornOn	15-04-1990
Antoine Griezmann	birthPlace	France
Antoine Griezmann	height	1,74
Antoine Griezmann	type	Footballer

Motivating Example

```
SELECT ?s WHERE { ?s birthPlace France }
```

Subject	Predicate	Object
Emma Watson	nationality	British
Emma Watson	bornIn	France
Emma Watson	bornOn	15-04-1990
Antoine Griezmann	birthPlace	France
Antoine Griezmann	height	1,74
Antoine Griezmann	type	Footballer

```
SELECT * WHERE {
  {?s1 birthPlace France}
Union
  {?s2 bornIn France}
}
```

Data publisher ignores the ontology (schema)

Related Work

An approach for generating and evaluating synonym candidate pairs

1. Range content filtering
   - Mining **predicates** of each distinct **object**
   - Retrieving **frequent** candidate pairs
2. Schema analysis
   - Mining **predicates** of distinct **subject**
   - Keeping pairs with high **negative correlation**

The algorithm produces too many **false positives**

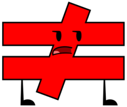
Abedjan Z, Naumann F. **Synonym analysis for predicate expansion.** In Extended semantic web conference. Springer, Berlin, Heidelberg, 2013.

Conciseness Dimension

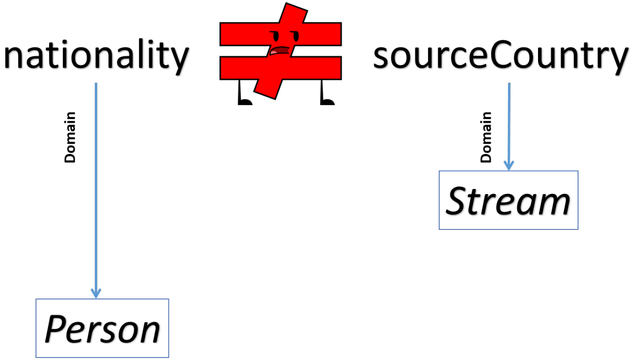
Our objective is to **decrease false positive** results through:

- Semantic analysis
- NLP-based analysis

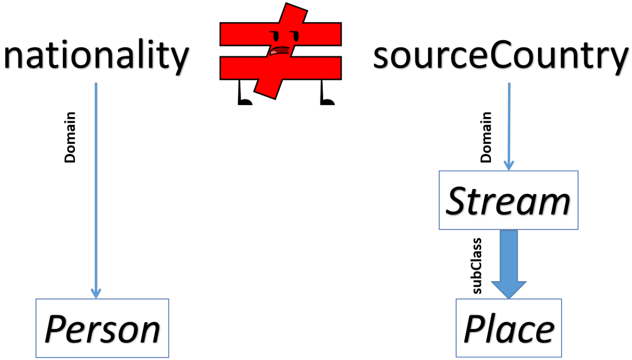
Semantic Analysis

nationality  sourceCountry

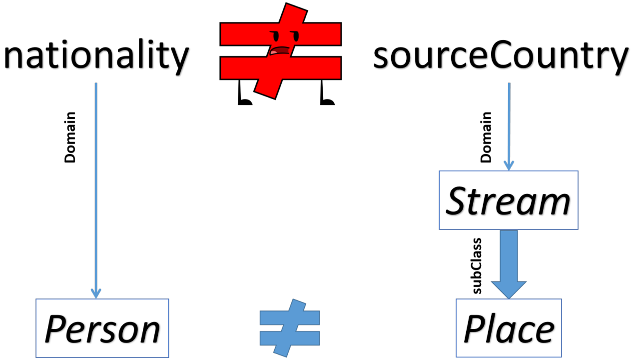
Semantic Analysis



Semantic Analysis



Semantic Analysis



Semantic Analysis

- Excluding candidates having incompatible semantic features.
- Semantic features:
  Domain restriction, Range restriction ,Functional properties, Transitive properties, Symmetric properties, Max cardinality

Semantics features	**Description**
Domain restriction	p_1 & p_2 cannot be synonyms if: $\exists p_1.\top \sqsubseteq C_1 \quad \wedge \quad \exists p_2.\top \sqsubseteq C_2 \quad \wedge \quad C_1 \sqcap C_2 \sqsubseteq \bot$
Range restriction	p_1 & p_2 cannot be synonyms if: $\top \sqsubseteq \forall p_1.C_1 \quad \wedge \quad \top \sqsubseteq \forall p_2.C_2 \quad \wedge \quad C_1 \sqcap C_2 \sqsubseteq \bot$
Functional properties	p_1 & p_2 cannot be synonyms if: p_1 *is a FunctionalProperty* $\wedge$ p_2 *is a Non FunctionalProperty*

NLP-based Analysis

Excluding predicates that are semantically similar but non-equivalent (e.g. *composer* and *artist*)

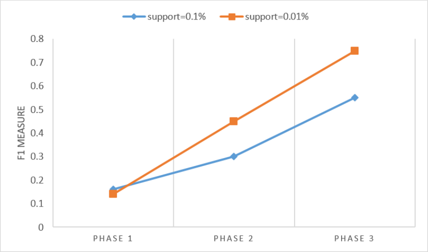
- Considers the meaning of predicates in a specific context using learning algorithms
  - *Word embedding*: using algorithms, such as *Word2vec*, to map predicates to vectors of numbers; two predicates sharing common contexts are located close to each other in the space vector
  - Applying a cosine similarity to compare pairs of vectors

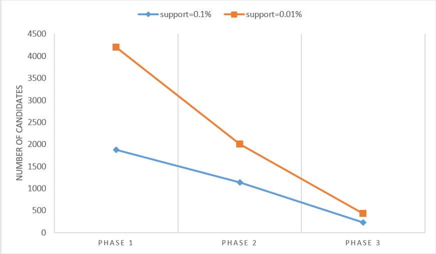
$$\cos(\mathbf{t},\mathbf{e}) = \frac{\mathbf{te}}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}}$$

First Experiment

Support threshold=0.01%

- Statistical analysis: NC=4197, F1=0.14
- Semantic analysis: NC=2006, F1=0.45
- NLP-based analysis: NC=429, F1=0.76





- Semantic analysis eliminates 52.2**%** of false positives and NLP-based analysis eliminates 78.6**%** of false positives
- Filters the predicates that share the same semantic features but are non-equivalents (e.g. *author* and *composer*)

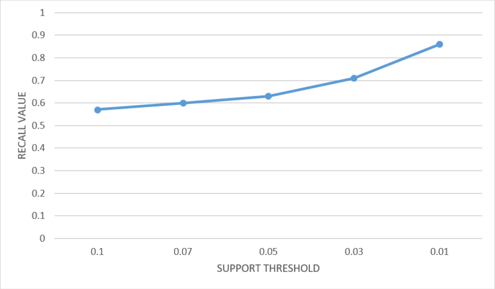
Second Experiment

- Performs tests between the predicates of different datasets (i.e. DBpedia & YAGO datasets)
- Compares with a gold standard containing mappings between the predicates of these two datasets

Predicate 1 (YAGO)	Predicate 2 (DBpedia)
diedIn	deathPlace
diedOnDate	deathDate
isCitizenOf	nationality
livesIn	residence
hasPopulation	populationTotal

Second Experiment

- Support threshold=0.01%, Recall value=0.86



- Our approach:
  - Finds a good number of equivalent predicates (recall at roughly 86%)
  - Fails to find all the equivalent predicates (e.g. *isbn* and *hasISBN*) relies on the fact that some predicate pairs share insufficient number of objects

Outline

Conclusion and Research Perspectives

Geo-Domain Identity Links

- An ontology to represent knowledge about geometry positional accuracy and capture rules
- An approach to extract XY semantics by using automatic supervised learning
- A data matching approach that relies on XY semantics to adapt the comparison of geometries

Perspectives:

- Time complexity should be improved by adding a cache system for the reasoning results
- Further tests with bigger and more heterogeneous datasets
- Consider the geometry resolution and its vagueness in both populating and interlinking approaches

Conclusion and Research Perspectives

Contextual Identity Links

- An approach **to compute a set of propagable** properties given a set of indiscernible properties:
  - Based on Tobler's first law and sentence embedding
  - A full framework to increase completeness of SPARQL queries

Perspectives:

- Not rely only on description of properties
- Try to use values of properties or semantic features of the property
- Challenge our work with a combination of distinct KGs

Conclusion and Research Perspectives

Fuzzy Temporal Data

- A Crisp-Based Approach
  - Extend the 4D-fluents model to represent imprecise time intervals and their crisp relationships in OWL 2
  - Extend the Allen's interval algebra in a crisp way and infer interval relations via a set of SWRL rules
- A Fuzzy-Based Approach
  - Extend the 4D-fluents model to represent imprecise time intervals and their relationships in Fuzzy-OWL 2
  - Extend the Allen's interval algebra in a fuzzy gradual personalized way and Infer fuzzy interval relations using a set of Mamdani IF-THEN rules

Perspectives:

- Define a composition table between the resulting relationships of precise and imprecise time intervals
- Extend our approach to represent and reason over time intervals that are both imprecise and uncertain

Conclusion and Research Perspectives

Linked Data Quality

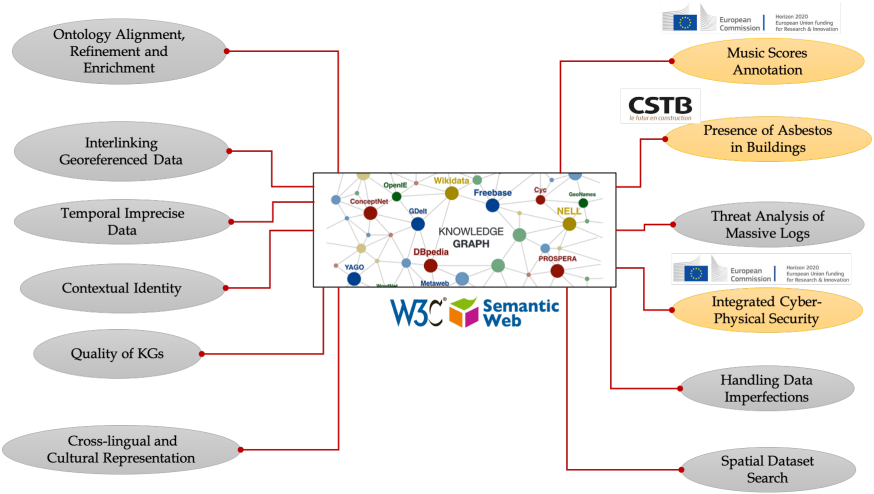
- Developing an approach for Linked Data completeness assessment
- Implementing "LOD-CM" prototype to reveal conceptual schema from linked datasets
- Providing an approach for assessing the conciseness of a dataset by discovering equivalent predicates

Perspectives:

- Investigating the effectiveness of the approaches against additional Linked Open Data datasets such as Wikidata
- Allowing the user to compare conceptual schemas from different datasets
- Dealing with uncommon predicates to discover equivalent predicates

Conclusion and Research Perspectives

Future Research Projects



Conclusion and Research Perspectives

Long Term Plan:

- Studying the fully automatic adaptation of the Knowledge Graph interlinking, enrichment, refinement, and reasoning to the context of use
- Exploring deep learning algorithms towards the automation of the consideration of contexts in the various processes

Thank You!
Questions?