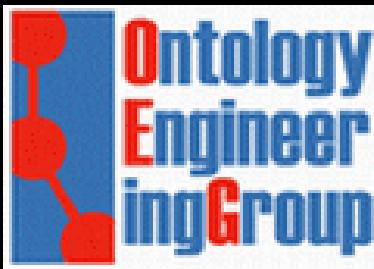


esDBpedia++

“An enhanced and
extended esDBpedia”



#esDBVienna2018

@marienorico



Thanks to TIN2016-78011-C4-4-R (AEI/FEDER, UE) and RTC-2016-4952-7

Detailed info

<https://goo.gl/vbS8GB>



In a ~~big, big~~ galaxy...

esDBpedia

*esDBpedia has
been without
modifications
for the last
years, but...*



*esDBpedia has
been without
modifications
for the last
years, but...*



*esDBpedia has
been without
modifications
for the last
years, but...*



*esDBpedia has
been without
modifications
for the last
years, but...*



We've got
reinforcements!



esDBpedia++

Company funded 2y project + Public Money

The screenshot shows the TAIGER website homepage. The header includes the logo, navigation links (SOLUTIONS, DISCOVER, ACADEMY, PARTNERS, CAREERS, COMPANY), and a 'REQUEST DEMO' button. The main headline reads 'Automate Manual yet Complex Tasks, Intelligently'. Below it, a quote says 'Awarded Gartner's "Cool Vendor of 2017"' with a 'REQUEST DEMO' button. The page features three main sections: 'REDUCE OPERATING COSTS' (with a bar chart icon), 'OPTIMISE PROCESSING TIME' (with a gear icon), and 'DRIVE REVENUE GROWTH' (with a bar chart icon). Each section has a brief description.



TAIGER's objective

**Enhance its Text Analytics
in-the-cloud service**

Semantic annotation

Semantic disambiguation



Our research

Enhance & extend esDBpedia

Enhance the quality

Extend to proprietary datasets



Enhancing quality

New techniques

ML + linked data

NLP + linked data



Extending the dataset

Alignment (ontologies & data)

Dataset 1

Dataset 2

It is a
secret
sauce!



ML + linked data

Supervised learning



ML + linked data

Supervised learning

Wrong mapping detection

Predicting Incorrect Mappings: A Data-Driven Approach Applied to DBpedia

Mariano Rico
Ontology Engineering Group,
Universidad Politécnica de Madrid
Madrid, Spain
mariano.rico@fi.upm.es

Heiko Paulheim
Data and Web Science Group,
University of Mannheim
Mannheim, Germany
heiko@informatik.uni-mannheim.de

Sebastian Hellmann
AKSW, Department of Computer
Science, Leipzig University
Leipzig, Germany
hellmann@informatik.uni-leipzig.d

Nandana Mihindukulasooriya
Ontology Engineering Group,
Universidad Politécnica de Madrid
Madrid, Spain
nmihindu@fi.upm.es

Dimitris Kontokostas
AKSW, Department of Computer
Science, Leipzig University
Leipzig, Germany
kontokostas@informatik.uni-leipzig.
de

ABSTRACT

DBpedia releases consist of more than 70 multilingual datasets that cover data extracted from different language-specific Wikipedia instances. The data extracted from those Wikipedia instances are transformed into RDF using mappings created by the DBpedia community. Nevertheless, not all the mappings are correct and consistent across all the distinct language-specific DBpedia datasets. As these incorrect mappings are spread in a large number of mappings, it is not feasible to inspect all such mappings manually to ensure their correctness. Thus, the goal of this work is to pro-



ML + linked data

Supervised learning

Wrong mapping detection

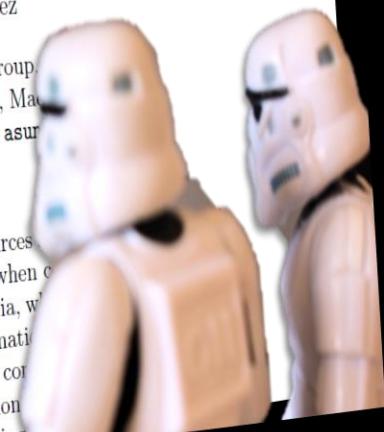
Inferring resource types

Inferring New Types on Large Datasets
Applying Ontology Class Hierarchy Classifiers:
The DBpedia Case

Mariano Rico*, Idafen Santana-Pérez, Pedro Pozo-Jiménez, and
Asunción Gómez-Pérez

Ontology Engineering Group
Universidad Politécnica de Madrid, Ma
{mariano.rico, isantana, ppozo, asur}@upm.es

Abstract. Adding type information to resources in knowledge graphs is a challenging task, specially when the graphs are generated collaboratively, such as DBpedia, where errors and noise produced during the transformation of data sources. It is important to assign the correct type to each resource in order to efficiently exploit the information it contains. In this work we explore how machine learning can be used to detect wrong mappings between resources and their types.



ML + linked data

Supervised learning

Wrong mapping detection

Inferring resource types

Predicting next query

Machine Learning-Based Query Augmentation for SPARQL
Endpoints

Mariano Rico¹, Rizkallah Touma², Anna Queralt² and María S. Pérez¹

¹ Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

² Barcelona Supercomputing Center (BSC), Barcelona, Spain

{mariano.rico, mperez}@fi.upm.es, {rizk.touma, anna.queralt}@bsc.es

Keywords:
Query Augmentation, Linked Data, Semantic Web, SPARQL E-
Triple Pattern

Abstract:
Linked Data repositories have become a popular source of publishing data through SPARQL endpoints. These endpoints usually launch several queries, either to find the information they need through trial-and-error or to modify the incoming queries to retrieve more data that is required by the user. In this paper, we propose a novel approach to query endpoints based on machine learning. Our approach separates the two main components of the system: the query parser and the query executor.



ML + linked data

Wrong mapping detection

Google Summer of Code 2018 Projects

Search for a project, organization, or student

VIEW BY ORGANIZATION >

STUDENT Search HELP SIGN IN

OpenML

Explore Data Task Flow Run Study Task type Measure People

The DBpedia mappings for each language are not aligned, causing

Predicting wrong DBpedia mappings

Created 05-05-2017 by Mariano Rico Visibility: public

DESCRIPTION 10 DATA SETS 4 TASKS 0 FLOWS 3 RUNS

This was an study started by Nandana and Mariano in 2016. We started with unsupervised methods, but we could not find good clusters. En 2017 we started with annotated data and here we are.

Summary

DBpedia releases consist of more than 70 multilingual datasets that cover data extracted from different language-specific Wikipedia instances. The data extracted from those Wikipedia instances are transformed into RDF using mappings created by the DBpedia community. Nevertheless, not all the mappings are correct and consistent across all the distinct language-specific DBpedia datasets. As these incorrect mappings are spread in a large number of mappings, it is not feasible to inspect all such mappings manually to

Predicting Incorrect Mappings: A Data-Driven Approach Applied to DBpedia

Mariano Rico
Ontology Engineering Group,
Universidad Politécnica de Madrid
Madrid, Spain
mariano.rico@fi.upm.es

Nandana Mihindukulasooriya
Ontology Engineering Group,
Universidad Politécnica de Madrid
Madrid, Spain
nmihindu@fi.upm.es

Dimitris Kontokostas
AKSW, Department of Computer
Science, Leipzig University
Leipzig, Germany
kontokostas@informatik.uni-leipzig.
de

Heiko Paulheim
Data and Web Science Group,
University of Mannheim
Mannheim, Germany
heiko@informatik.uni-mannheim.de

Sebastian Hellmann
AKSW, Department of Computer
Science, Leipzig University
Leipzig, Germany
hellmann@informatik.uni-leipzig.d
e

ACM Reference
Mariano Rico, N
Paulheim, Sebast
ing Incorrect Mapp
Proceedings of S
April 9–13, 201
<https://doi.org/>

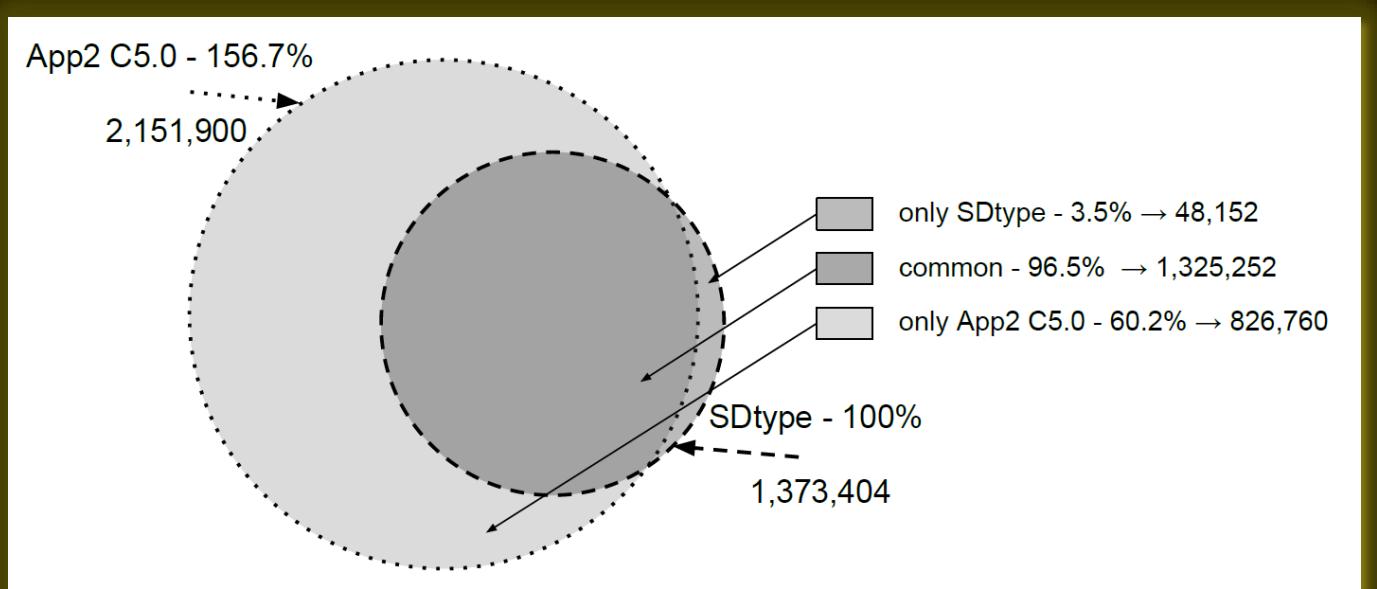
1 INT

ABSTRACT

DBpedia releases consist of more than 70 multilingual datasets that cover data extracted from different language-specific Wikipedia instances. The data extracted from those Wikipedia instances are transformed into RDF using mappings created by the DBpedia community. Nevertheless, not all the mappings are correct and consistent across all the distinct language-specific DBpedia datasets. As these incorrect mappings are spread in a large number of mappings, it is not feasible to inspect all such mappings manually to ensure their correctness. Thus, the goal of this work is to pro

ML + linked data

Inferring resource types

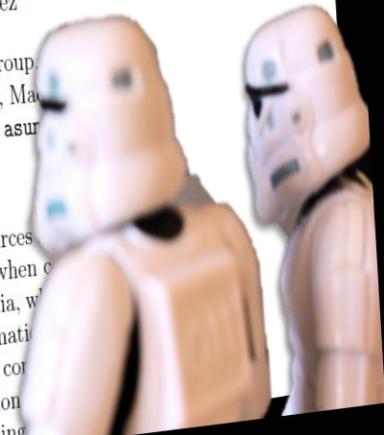


Inferring New Types on Large Datasets
Applying Ontology Class Hierarchy Classifiers:
The DBpedia Case

Mariano Rico*, Idafen Santana-Pérez, Pedro Pozo-Jiménez, and
Asunción Gómez-Pérez

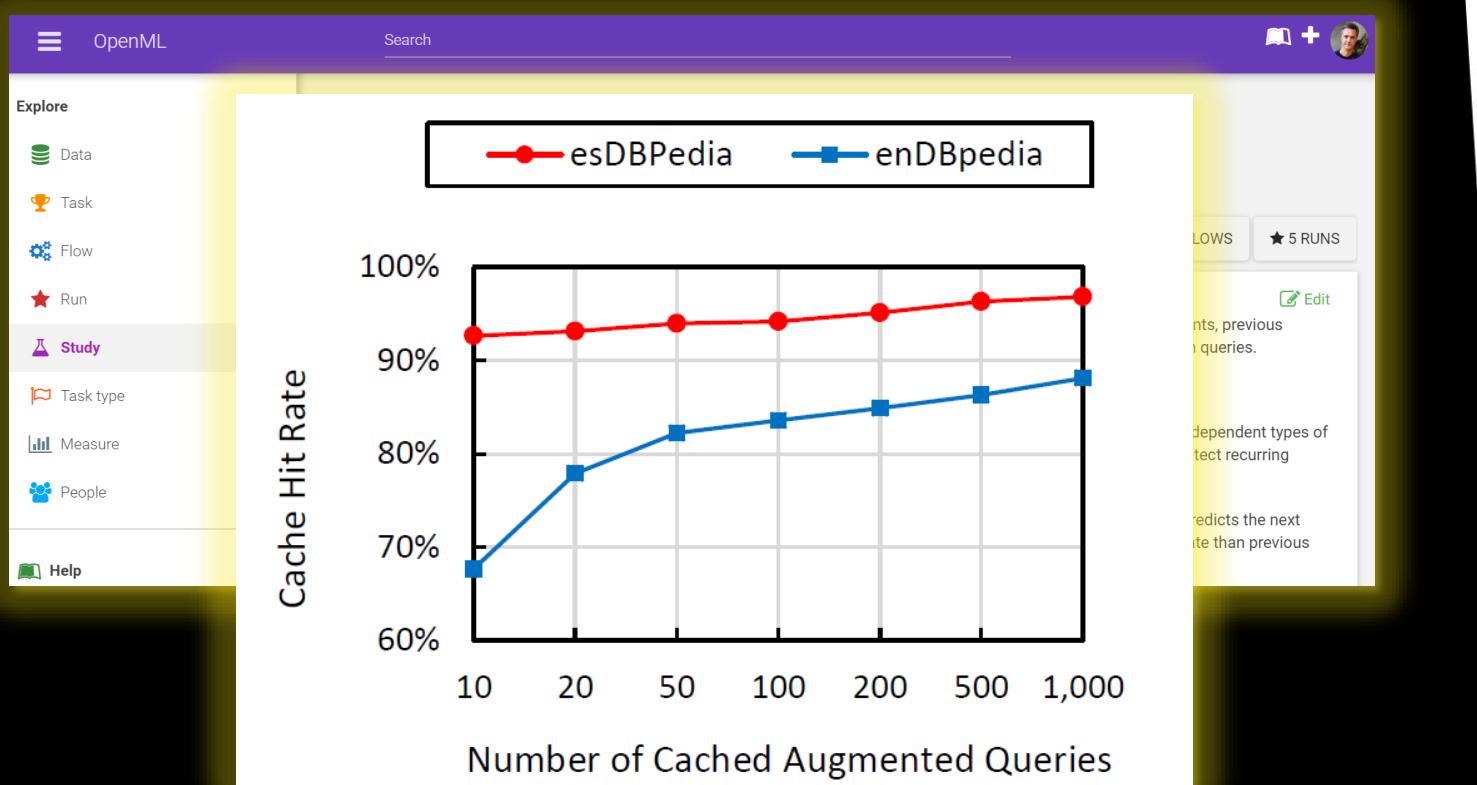
Ontology Engineering Group
Universidad Politécnica de Madrid, Ma
{mariano.rico, isantana, ppozo, asur}

Abstract. Adding type information to resources in knowledge graphs is a challenging task, specially when they are generated collaboratively, such as DBpedia, where there is a lack of type information and noise produced during the transformation from multiple data sources. It is important to assign the correct type to each resource in order to efficiently exploit the information contained in the graph. In this work we explore how machine learning can be used to infer new types for resources in DBpedia.



ML + linked data

Predicting next query



Machine Learning-Based Query Augmentation for SPARQL Endpoints

Mariano Rico¹, Rizkallah Touma², Anna Queralt² and María S. Pérez¹

¹Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

²Barcelona Supercomputing Center (BSC), Barcelona, Spain

{mariano.rico, mperez}@fi.upm.es, {rizk.touma, anna.queralt}@bsc.es

Keywords:
Query Augmentation, Linked Data, Semantic Web, SPARQL E
Triple Pattern

Abstract:
Linked Data repositories have become a popular source of public data through SPARQL endpoints. These endpoints usually launch several requests, either to find the information they need through trial-and-error or to execute each individual query sequentially. Our approach separates these requests by modifying the incoming queries to retrieve more data than is required. In this paper, we propose a novel approach to query augmentation based on machine learning. Our approach separates the incoming queries into two categories: those that require more data and those that do not. The first category is used to predict the next query, while the second is used to predict the previous query. This allows us to predict the next query before it is received, which can significantly reduce the response time of the endpoint.

NLP + linked data

NER & similarity



NLP + linked data

NER & similarity

“Resurrection” of URIs

Repairing Hidden Links in Linked Data

Enhancing the quality of RDF knowledge graphs

Nandana Mihindukulasooriya, Mariano Rico, Idafen Santana-Pérez, Raúl García-Castro and

Asunción Gómez-Pérez

{nmihindu,mariano.rico,isantana,rgarcia,asun}@f

Ontology Engineering Group, Universidad Politécnica

Madrid, Spain

ABSTRACT

Knowledge Graphs (KG) are becoming core components of most artificial intelligence applications. Linked Data, as a method of publishing KGs, allows applications to traverse within and even out of, the graph thanks to global dereferenceable identifiers denoting entities, in the form of IRIs. However, as we show in this work, after analyzing several popular datasets (namely DBpedia, LOD Cache, and Web Data Commons JSON-LD data) many entities are being represented using literal strings where IRIs should be used, diminishing the advantages of using Linked Data. To remedy this, we propose an approach for identifying such strings and replacing them with their corresponding entity IRIs. The proposed approach is based on identifying relations between entities based on both

1 INTRODU

Knowledge Graphs (KGs) are becoming core components of most artificial intelligence applications. Linked Data, as a method of publishing KGs, allows applications to traverse within and even out of, the graph thanks to global dereferenceable identifiers denoting entities, in the form of IRIs. However, as we show in this work, after analyzing several popular datasets (namely DBpedia, LOD Cache, and Web Data Commons JSON-LD data) many entities are being represented using literal strings where IRIs should be used, diminishing the advantages of using Linked Data. To remedy this, we propose an approach for identifying such strings and replacing them with their corresponding entity IRIs. The proposed approach is based on identifying relations between entities based on both

inside and external KGs.

NLP + linked data

NER & similarity

“Resurrection” of URIs

dbp properties to enhance
SPARQL queries

Data-Driven RDF Property
Semantic-Equivalence Detection Using NLP
Techniques

Mariano Rico^(✉), Nandana Mihindukulasooriya,
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[{mariano.rico,nmhiniudu,asun}](mailto:{mariano.rico,nmhiniudu,asun}@upm.es)

Abstract. DBpedia extracts most of its infoboxes. Manually-created “mappings” link DBpedia ontology properties (dbo properties) to DBpedia triples. However, infobox attributes produce triples with properties in a different namespace. In this position paper we point out that (a) the



NLP + ML + linked data

Not published yet

DBpedia resource type
from a textual description

The screenshot shows the NLP4Types demo tool interface. At the top right is the logo "NLP4Types" with a lightbulb icon. Below it is a sub-header: "Inferring Types using NLP analysis: The DBpedia use case - This demo tool showcases the type inference models we have developed for predicting types based on textual descriptions of resources." On the left, there's a text input field labeled "Description text" containing a description of Miguel de Cervantes Saavedra. To the right of the input field is a "Guess type" button. Below the input field, a "Writer" section shows the URL <http://dbpedia.org/ontology/Writer>. A large Stormtrooper figure from Star Wars is standing next to the computer screen, holding a large white card with JSON-like code on it. To the right of the Stormtrooper is a vertical list of options under "Is this type right?": "Yes, totally" (green), "Should be more precise" (yellow), "Should be more abstract" (cyan), "Somehow related" (light gray), and "No, it is totally wrong" (red). The JSON code on the card includes fields like "confidence": 0.3, "use_sw": true, "named_entity": "DBPER", "DBPEDIA_PERSON", "DBPEDIA_ARTIST", "DBPEDIA_LANGUAGE", and "DBPEDIA_LAN".

NLP + linked data

“Resurrection” of URIs

Table 7: String disambiguation performance

Class	# of strings	Disambiguated		Prec.	Recall
		#	Correct		
dbo:Athlete	100	51	50	98.04%	50%
dbo:SportsTeam	100	44	44	100%	44%
dbo:SportsEvent	100	58	55	94.83%	56%
Total	300	153	149	97.38%	49.67%

Table 8: Improvements in connectivity

Graph	edges	com- pon- ents	isolated	largest component	
				size	% total
Original	828,310	119,623	112,331	168,128	54.28
Repaired	1,035,912	99,137	93,507	192,805	64.16
Δ	+207,602	-20,486	-18,824	+24,677	+9.88
	+25.06%	-17.12%	-16.76%	+14.68%	

Repairing Hidden Links in Linked Data

Enhancing the quality of RDF knowledge graphs

Nandana Mihindukulasooriya, Mariano Rico, Idafen Santana-Pérez, Raúl García-Castro and

Asunción Gómez-Pérez

{nmihindu,mariano.rico,isantana,r.garcia,asun}@f

Ontology Engineering Group, Universidad Politécnica
Madrid, Spain

ABSTRACT

Knowledge Graphs (KG) are becoming core components of most artificial intelligence applications. Linked Data, as a method of publishing KGs, allows applications to traverse within and even outside of, the graph thanks to global dereferenceable identifiers denoting entities, in the form of IRIs. However, as we show in this work, after analyzing several popular datasets (namely DBpedia, LOD Cache, and Web Data Commons JSON-LD data) many entities are being represented using literal strings where IRIs should be used, diminishing the advantages of using Linked Data. To remedy this, we propose an approach for identifying such strings and replacing them with their corresponding entity IRIs. The proposed approach is based on identifying relations between entities based on both

1 INTRODU

Knowledge Graphs (KGs) are becoming core components of most artificial intelligence applications. Linked Data, as a method of publishing KGs, allows applications to traverse within and even outside of, the graph thanks to global dereferenceable identifiers denoting entities, in the form of IRIs. However, as we show in this work, after analyzing several popular datasets (namely DBpedia, LOD Cache, and Web Data Commons JSON-LD data) many entities are being represented using literal strings where IRIs should be used, diminishing the advantages of using Linked Data. To remedy this, we propose an approach for identifying such strings and replacing them with their corresponding entity IRIs. The proposed approach is based on identifying relations between entities based on both

NLP + linked data

dbp properties to enhance SPARQL queries

1	PF	se	{	Table 2. Example of dbp → dbo property mappings. Δ_1 is the enhancement for the example query in Listing 1.1.						}
				DBpedia dbo prop	dbp prop			Δ_1		
	Syntactic		Semantic							
1	PF	se	{	English birthPlace birthPlace	birthplace	placeofbirth	cityofbirth	cityofbirthPlace	350%	}
2	PF	se	?V	birthPlac	birthdplace	birthPalce	cityOfBirth	birthLocation		
3				birthPlace PlaceOfBirth	laceOfBirth					
4	se			oplaceOfBirth birthPlace.	birthPlacE					
5				birthPalce birthPlae birthPace	birthPlaxe					
6				birthPlace birthPlcace bithPlace	brithPlace					
7				nbirthPlace birthplace	birghPlace					
8				birthdplace biRthPlace birth	placebirth					
9				placeOfBirth placOfBirth	birthPlaceOf					
10				birthPlae						
11				Spanish birthPlace lugarDeNacimiento	lugarNacimiento	ciudaddenacimiento			221%	
12				lugarNacimiento	lugarnacimiento	ciudadDenacimiento				
13			}	lugardenacimiento	lugarNacimiento	paisdenacimiento	paisNacimiento			
				lugarNacierto		birthPlace	birthplace	placeOfBirth		
				German birthPlace geburtsort	birthplace	birthPlace	geburtsland	countryofbirth	134%	
				placeOfBirth placeofbirth						

Data-Driven RDF Property Semantic-Equivalence Detection Using NLP Techniques

Mariano Rico^(✉), Nandana Mihindukulasooriya,
Asunción Gómez-Pérez
Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
[{mariano.rico,nmhini,asun}](mailto:{mariano.rico,nmhini,asun}@upm.es)

Abstract. DBpedia extracts most of its infoboxes. Manually-created “mappings” link DBpedia ontology properties (dbo properties) to DBpedia triples. However, infobox attributes produce triples with properties in a different namespace. In this position paper we point out that (a) the



Thanks a LOD!



esDBpedia++

“An enhanced and
extended esDBpedia”



#esDBVienna2018

@marienorico



Thanks to TIN2016-78011-C4-4-R (AEI/FEDER, UE) and RTC-2016-4952-7

NLP + ML + linked data

Not published yet

Resource type from
textual description

The screenshot shows a web-based application for inferring resource types. At the top right is a logo with a lightbulb icon and the text "NLP4Types". Below it is a sub-header: "Inferring Types using NLP analysis: The DBpedia use case - This demo tool showcases the type inference models we have developed for predicting types based on textual descriptions of resources." On the left, there's a text input field containing a description of Miguel de Cervantes Saavedra. To the right of the input field is a blue button labeled "Guess type". Further down, there's a section titled "Writer" with a URL link: <http://dbpedia.org/ontology/Writer>. At the bottom right, there's a section titled "Is this type right?" with five colored buttons: green (Yes, totally), yellow (Should be more precise), cyan (Should be more abstract), grey (Somehow related), and red (No, it is totally wrong). A small Stormtrooper figurine is visible at the bottom right corner of the screen.

Description text ⓘ

Miguel de Cervantes Saavedra (/sər'ventəs/ or /sər'ven̩təs/; Spanish: [miˈɣel deθeɾˈbantes saˈβeðɾa]; sometimes Anglicized as Michael Cervantes; 29 September 1547 (assumed) – 22 April 1616), was a Spanish writer who is widely regarded as

If you don't feel inspired, you can copy & paste from a random Wikipedia article ↗

Debug ⓘ

```
{"confidence": 0.3, "use_sw": true, "named_entities": "DBPEDIA_AGENT DBPEDIA_PERSON DBPEDIA_ARTIST DBPEDIA_WRITER DBPEDIA_ETHNICGROUP DBPEDIA_LANGUAGE DBPEDIA_LANGUAGE DBPEDIA_AGENT DBPEDIA_PERSON"}
```

Writer
<http://dbpedia.org/ontology/Writer>

Is this type right? ⓘ

- Yes, totally
- Should be more precise
- Should be more abstract
- Somehow related
- No, it is totally wrong

NLP + ML + linked data

Not published yet

Resource type from
textual description



The screenshot shows the NLP4Types demo tool interface. At the top right is a logo with a lightbulb icon and the text "NLP4Types". Below it is a sub-headline: "Inferring Types using NLP analysis: The DBpedia use case - This demo tool showcases the type inference models we have developed for predicting types based on textual descriptions of resources." On the left, there's a "Description text" input field containing a paragraph about Miguel de Cervantes Saavedra. To the right of the input field is a "Guess type" button. Below the input field, a message says "If you don't feel inspired, you can copy & paste from a random Wikipedia article." On the right side, there's a "Writer" section with a URL and a "Is this type right?" section with five options: "Yes, totally", "Should be more precise", "Should be more abstract", "Somehow related", and "No, it is totally wrong". A small 3D rendering of a white robot is positioned between the "Writer" and "Is this type right?" sections.

Description text ⓘ

Miguel de Cervantes Saavedra (/sər'vontēz/ or /sər'vāntē z/; Spanish: [miˈxel deθerˈbantes saˈβeðɾa]; sometimes Anglicized as Michael Cervantes; 29 September 1547 (assumed) – 22 April 1616), was a Spanish writer who is widely regarded as

If you don't feel inspired, you can copy & paste from a random Wikipedia article ↗

Guess type

Writer

<<http://dbpedia.org/ontology/Writer>>

Is this type right? ⓘ

Yes, totally

Should be more precise

Should be more abstract

Somehow related

No, it is totally wrong

Debug ⓘ

```
{"confidence": 0.3, "use_sw": true, "named_entities": "DBPEDIA_PERSON DBPEDIA_ARTIST DBPEDIA_WRITER DBPEDIA_GROUP DBPEDIA_LANGUAGE DBPEDIA_LANGUAGE DBPEDIA_AWARD DBPEDIA_AWARD", "type": "DBPEDIA_PERSON", "type_label": "Person", "type_id": "http://dbpedia.org/ontology/Person", "type_url": "http://dbpedia.org/ontology/Person", "type_label_id": "http://dbpedia.org/resource/Person", "type_label_url": "http://dbpedia.org/resource/Person"}
```

A 3D rendering of a white robot is standing next to the interface.