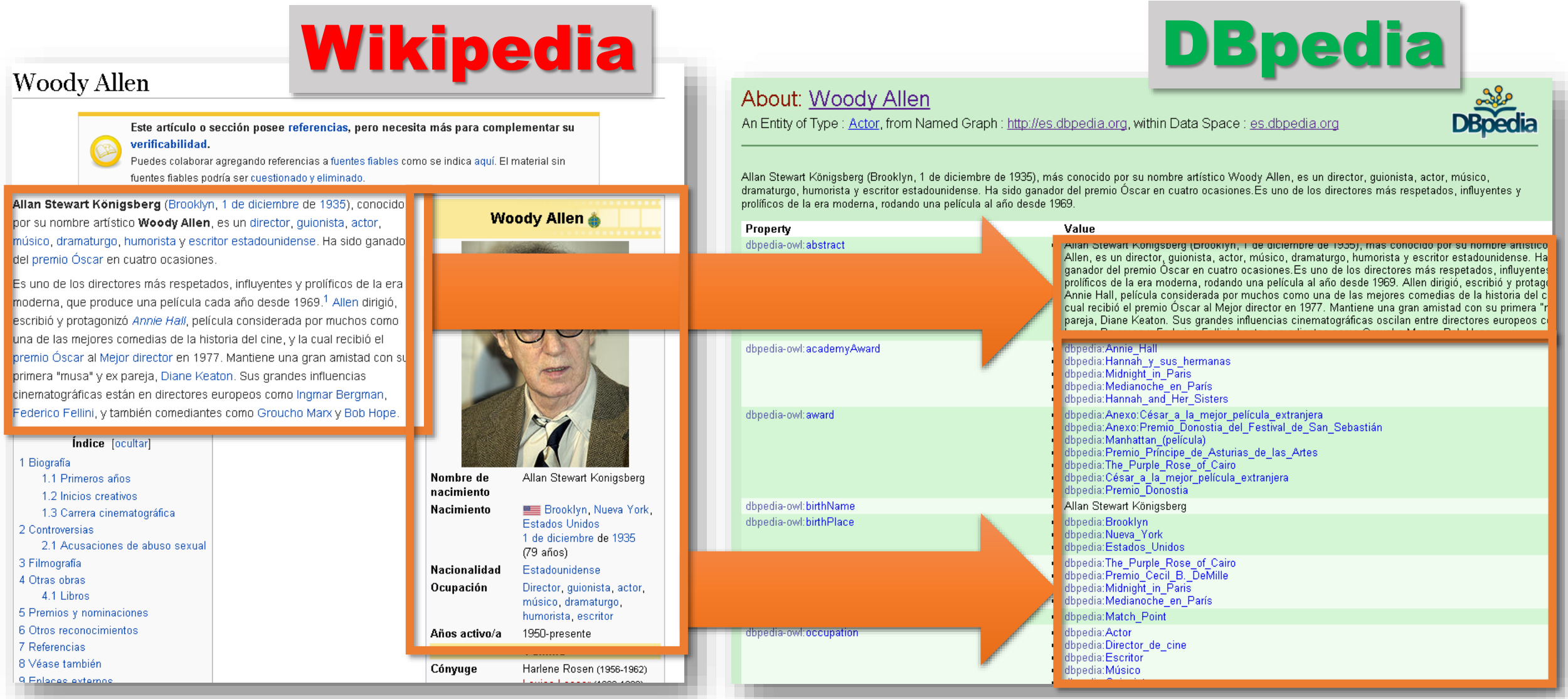
Predicting Incorrect Mappings: A Data-Driven Approach Applied to DBpedia

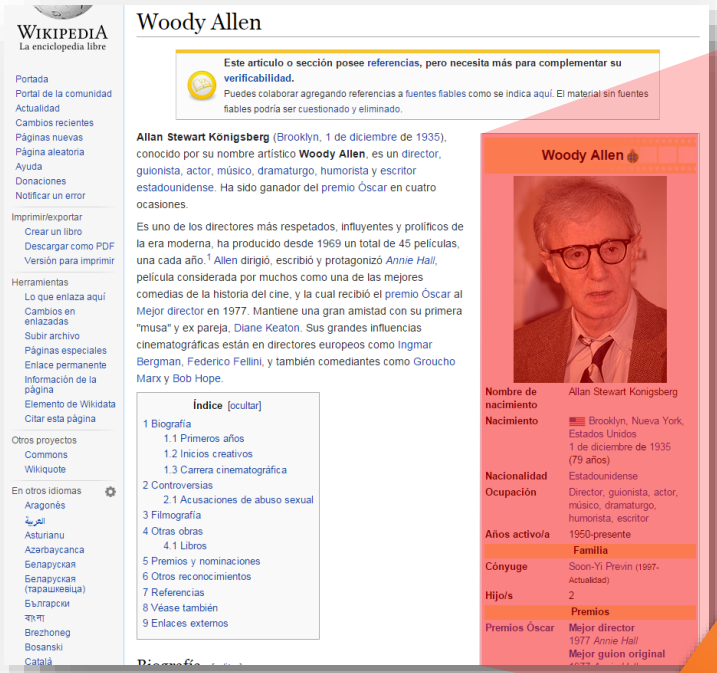
Mariano Rico, Nandana Mihindukulasooriya, Dimitris Kontokostas, Heiko Paulheim, Sebastian Hellmann, Asunción Gómez-Pérez

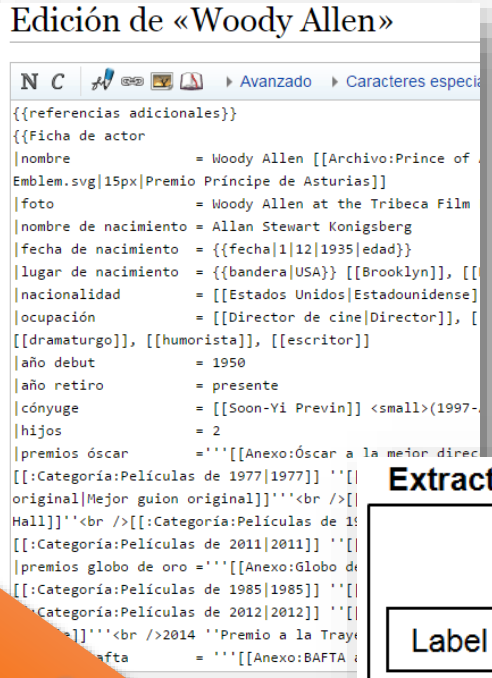
European Union
European Regional Development Fund
"A way to build Europe"

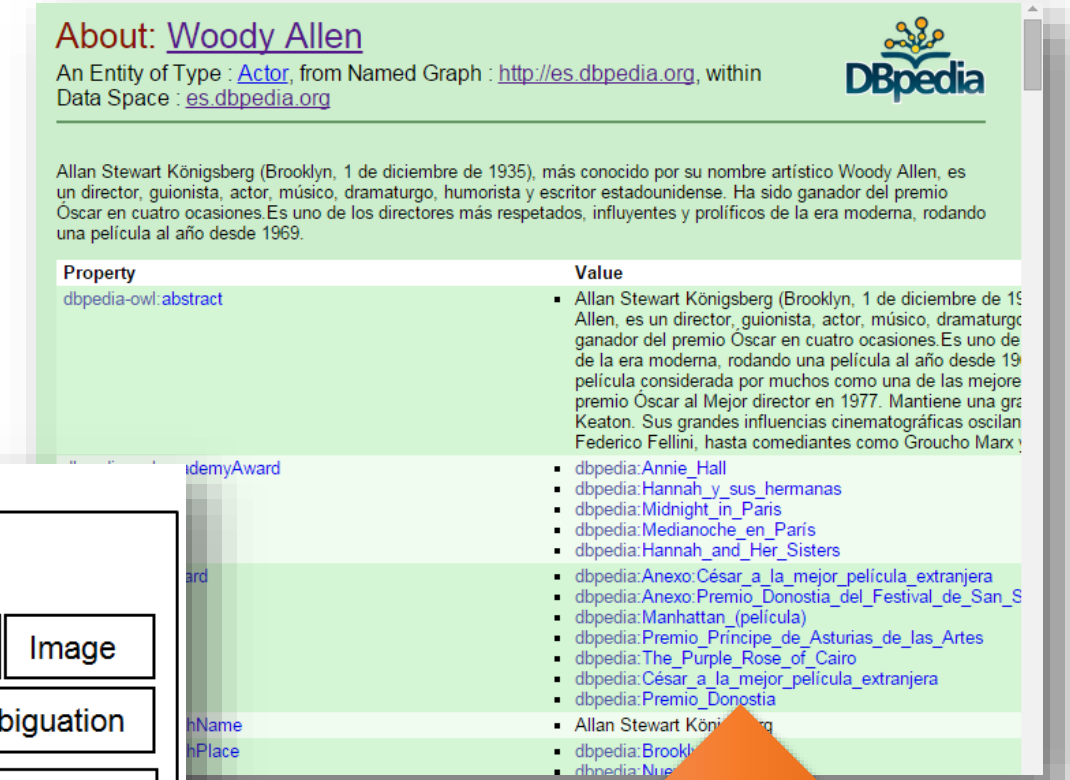
This work was partially funded by the Spanish MINECO Ministry (projects RTC-2016-4952-7 and TIN2013-46238-C4-2-R), the BES-2014-068449 grant, and grants from the EU's H2020 Programmes for the ALIGNED project (GA 644055). Also, this work has been partially funded by the project Data 4.0 (TIN2016-78011-C4-4-R), from the Spanish State Investigation Agency of the MINECO and FEDER Funds.

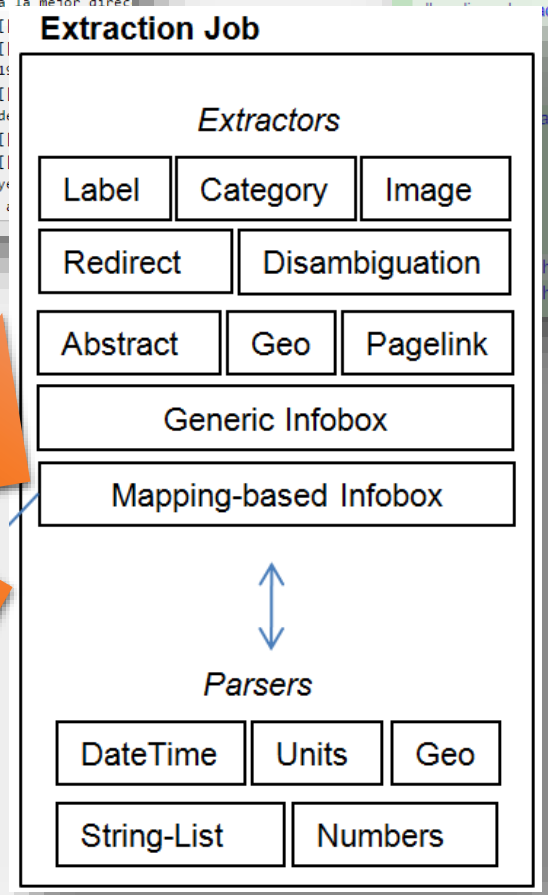
How DBpedia works. From text to data











infobox

mappings

RDF

Rendering

Triple store

How DBpedia works. Mappings

Wikipedia

```
{{referencias adicionales}}
{{Ficha de actor
|nombre                = Woody Allen [[Archivo:Prince of Asturias Foundation
Emblem.svg|15px|Premio Príncipe de Asturias]]
|foto                  = Woody Allen at the Tribeca Film Festival.jpg
|nombre de nacimiento = Allan Stewart Konigsberg
|fecha de nacimiento  = {{fecha|1|12|1935|edad}}
|lugar de nacimiento  = {{bandera|USA}} [[Brooklyn]], [[Nueva York]], [[Estados
Unidos]]
|nacionalidad          = [[Estados Unidos|Estadounidense]]
|ocupación             = [[Director de cine|Director]], [[guionista]], [[actor]],
[[músico]], [[dramaturgo]], [[humorista]], [[escritor]]
|año debut             = 1950
|año retiro            = presente
|cónyuge               = Harlene Rosen <small>(1956-1962)</small><br />[[Louise
Lasser]] <small>(1966-1969)</small><br />[[Soon-Yi Previn]] <small>(1997-
Actualidad)</small>
|pareja                = [[Diane Keaton]] <small>(1973-1979)</small><br />[[Mia
Farrow]] <small>(1980-1992)</small>
|hijos                 = '''Adoptivos'''<br />Malone Farrow<br />Moshe Farrow<br
/>Bechet Dumaine <small>(1999)</small><br />Manzie Tio <small>(2000)</small><br
/>'''Biológicos'''<br />Ronan Farrow <small>(1987)</small>
|premios óscar         ='''[[Anexo:Óscar a la mejor dirección|Mejor
director]]'''<br />[[:Categoría:Películas de 1977|1977]] ''[[Annie Hall]]''<br
/>'''[[Anexo:Óscar al mejor guion original|Mejor guion original]]'''<br />
```

DBpedia

About: Woody Allen

An Entity of Type : Actor, from Named Graph : http://es.dbpedia.org, within Data Space : es.dbpedia.org

Ontology Class

Allan Stewart Königsberg (Brooklyn, 1 de diciembre de 1935), más conocido por su nombre artístico Woody Allen, es un director, guionista, actor, músico, dramaturgo, humorista y escritor estadounidense. Ha sido ganador del premio Óscar en cuatro ocasiones.Es uno de los directores más respetados, influyentes y prolíficos de la era moderna, rodando una película al año desde 1969.

Data

Property	Value
dbpedia-owl:abstract	Allan Stewart Königsberg (Brooklyn, 1 de diciembre de 1935), más conocido por su nombre artístico Allen, es un director, guionista, actor, músico, dramaturgo, humorista y escritor estadounidense. Ha ganador del premio Óscar en cuatro ocasiones.Es uno de los directores más respetados, influyentes prolíficos de la era moderna, rodando una película al año desde 1969. Allen dirigió, escribió y protago Annie Hall, película considerada por muchos como una de las mejores comedias de la historia del cir cual recibió el premio Óscar al Mejor director en 1977. Mantiene una gran amistad con su primera "m pareja, Diane Keaton. Sus grandes influencias cinematográficas oscilan entre directores europeos co Ingmar Bergman y Federico Fellini, hasta comediantes como Groucho Marx y Bob Hope.
dbpedia-owl:academyAward	dbpedia:Annie_Hall dbpedia:Hannah_y_sus_hermanas dbpedia:Midnight_in_Paris dbpedia:Medianoche_en_París dbpedia:Hannah_and_Her_Sisters
dbpedia-owl:award	dbpedia:Anexo:César_a_la_mejor_película_extranjera dbpedia:Anexo:Premio_Donostia_del_Festival_de_San_Sebastián dbpedia:Manhattan_(película) dbpedia:Premio_Príncipe_de_Asturias_de_las_Artes dbpedia:The_Purple_Rose_of_Cairo dbpedia:César_a_la_mejor_película_extranjera dbpedia:Premio_Donostia
dbpedia-owl:birthName	Allan Stewart Königsberg
dbpedia-owl:birthPlace	dbpedia:Brooklyn dbpedia:Nueva_York dbpedia:Estados_Unidos
dbpedia-owl:goldenGlobeAward	dbpedia:The_Purple_Rose_of_Cairo dbpedia:Premio_Cecil_B._DeMille dbpedia:Midnight_in_Paris dbpedia:Medianoche_en_París
dbpedia-owl:goyaAward	dbpedia:Match_Point
dbpedia-owl:occupation	dbpedia:Actor dbpedia:Director_de_cine dbpedia:Escritor dbpedia:Músico dbpedia:Guionista

How DBpedia works. Multilingual mappings

- DBpedia mapping process
  - Only 19 DBpedias
    - Aka chapters (EPs)
  - Much more language mappings (49 languages)
    - 30 of them do not have chapter. Any volunteer? ☺
  - Wikipedia has 288 active languages

Mappings: typical errors

- Language details
  - E.g. dbo:elevation and dbo:height
    - In English: elevation is for mountains, height is for persons
    - In many non-English languages:height can be also for mountains
- Duplicated properties. Same semantics
  - E.g. dbo:formationYear and dbo:foundingYear
    - formationYear is used for music groups
    - foundingYear is used for companies
- Level of specificity
  - E.g. dbo:code and dbo:postalCode

Mappings with errors: effects

- Queries more complex
  - Example of queries for properties semantically equivalent

```
1  PREFIX dbo: <http://dbpedia.org/ontology/>
2  select ?s ?bp {
3     ?s dbo:birthPlace ?bp .
4  }
```

Listing 1.1. Original SPARQL query

```
1  PREFIX dbo: <http://dbpedia.org/ontology/>
2  PREFIX dbp: <http://dbpedia.org/property/>
3
4  select ?s ?bp  where {
5   ?s ?p ?bp .
6   VALUES ?p {
7    dbo:birthPlace #typical dbo property
8    #Alternative dbp properties
9    dbp:birthPlcace dbp:birthplace
10   dbp:birhPlace   dbp:bithPlace
11   dbp:birtPlace   dbp:biRthPlace
12  }
13 }
```

Listing 1.2. Enhanced SPARQL query

Source:
Data-Driven RDF Property Semantic-Equivalence Detection Using NLP Techniques, M Rico, N Mihindukulasooriya et al., EKAW 2016

Intuition

- Correct mapping

Table 1: Data from correct and consistent mappings

DBpedia	Subject	Predicate	Object
English	dbr:Mount_Everest	geo:long	86.925278
		dbo:mountainRange	dbr:Himalayas
Spanish	dbr-es:Monte_Everest (owl:sameAs dbr:Mount_Everest)	geo:long	86.925278
		dbo:mountainRange	dbr-es:Himalayas (owl:sameAs dbr:Himalayas)
Greek	dbr-el:Έβερεστ (owl:sameAs dbr:Mount_Everest)	geo:long	86.925278
		dbo:mountainRange	dbr-el:Ιμαλάια (owl:sameAs dbr:Himalayas)
German	dbr-de:Mount_Everest (owl:sameAs dbr:Mount_Everest)	geo:long	86.925278
		dbo:mountainRange	dbr-de: Mahalangur_Himal (owl:sameAs dbr:Himalayas)

Intuition

- Incorrect mapping

DBpedia Dataset	Subject	Predicate	Object
English	dbr:Mount_Everest	dbo:elevation	8848
Spanish	dbr-es:Monte_Everest	**dbo:height**	8848
Greek	dbr-el:Έβερεστ	dbo:elevation	8848
German	dbr-de:Mount_Everest	dbo:elevation	8848

  - False negatives
    - E.g. Juanita Reina (Spanish singer) was born in Seville. She also died in Seville

```
dbr:Juanita_Reina  dbo:birthPlace  dbr:Seville
dbr:Juanita_Reina  dbo:deathPlace  dbr:Seville
```

Supervised learning

- Training set for English-Spanish (EN-ES)

Template(en)	Attribute(en)	Template(es)	Attribute(es)	Prop(en)	Prop(es)	Annotation
Infobox_French_commune	elevation_m	Ficha_de_entidad_subnacional	elevación_media	dbo:elevation	dbo:height	**Wrong** mapping
Infobox_company	foundation	Ficha_de_organización	fundación	dbo:foundingYear	dbo:formationYear	**Wrong** mapping
Infobox_mountain	elevation_m	Ficha_de_montaña	Elevación	dbo:elevation	dbo:prominence	**Wrong** mapping
Infobox_album	Artist	Ficha_de_álbum	productor	dbo:artist	dbo:producer	**Correct**

Sometimes the ontology has similar properties for the same thing but this is still a wrong mapping

The producer is the same person as the artist in some albums

- We also used ES-DE, EN-NL and EN-GR
- We trained for **objects** as IRIs (e.g. EN-ES-IRI) and literals (e.g.EN-ES-lit).

Features

- Instance-based features
  - Direct: M1, M2, M3, M4. Values: integer
  - Indirect: C1, C2, C3, C4. Values: float in [0-1]
- Schema-based features
  - TB1 to TB11. Values 0/1

Features

- Instance-based features
  - **Direct**: M1, M2, M3, M4. Values: integer
    - M1: Number of resources with dbo:elevation in EN DBpedia that have dbo:height in ES DBpedia
    - M2: Number of resources with dbo:elevation in EN that have dbo:height in ES and the object values are identical (equivalent).
    - M3 (false positives in M2): Number of resources in EN that are also in ES, having same object, but different property. E.g. people can have same dbo:birthPlace and dbo:deathPlace but this doesn't mean that dbo:birthPlace and dbo:deathPlace are the same property or are wrongly mapped. Another example: people who is actor and producer (dbo:directedBy and dbo:actor).
    - M4 (EN): Number of resources in EN with dbo:elevation and dbo:height (wrong mapping) vs. dbo:birthPlace and dbo:deathPlace (different relations)
    - M4 (ES): analogous to M4(EN) but for ES

Features

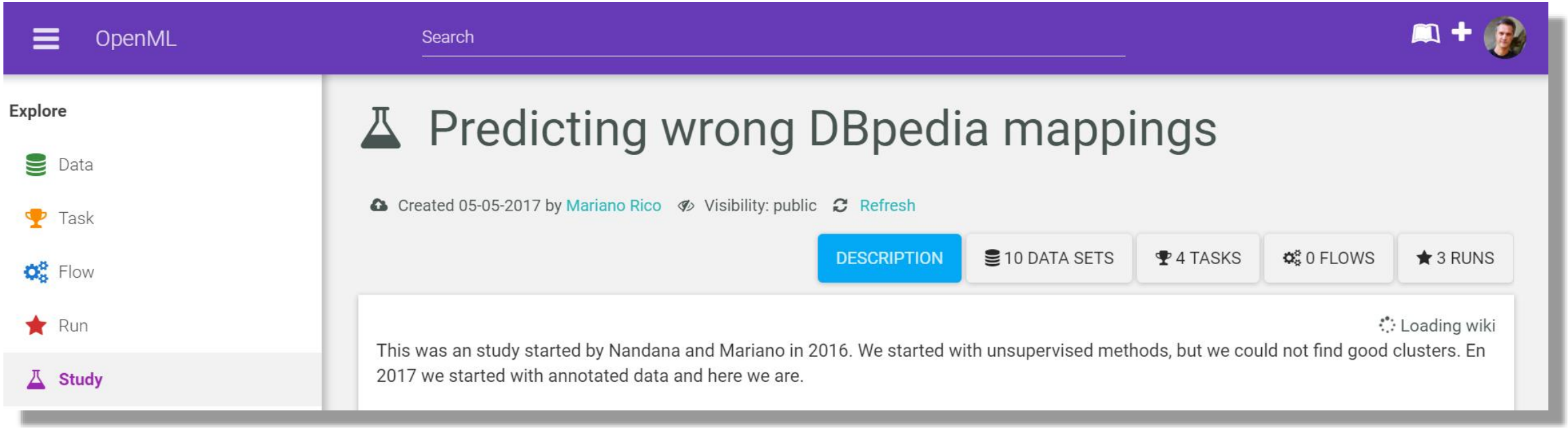
- Instance-based features
  - Direct: M1, M2, M3, M4. Values: integer
  - **Indirect**: C1, C2, C3, C4. Values: float in [0-1]
    - C1= M2/M1
    - C2 = M3/M1
    - C3 = C3(EN) = M4(EN)/M1
    - C4 = C3(ES) = M4(ES)/M1
- Schema-based features
  - TB1 to TB11. Values 0/1

Features

- **Schema-based features**
  - TB1 to TB11. Values 0/1
    - TB1: is dbo:elevation subproperty of dbo:height?
    - TB2 (reverse TB1): is dbo:height subproperty of dbo:elevation?
    - TB3: class in EN and ES is the same?
    - TB4: class in EN is subclass of class in ES?
    - TB5: class in ES is subclass of class in EN?
    - TB6: domain in ES and ES is the same?
    - TB7: domain in EN is subclass of domain in ES?
    - TB8: domain in ES is subclass of domain in EN?
    - TB9: range in ES and ES is the same?
    - TB10: range in EN is subclass of range in ES?
    - TB11: range in ES is subclass of range in EN?

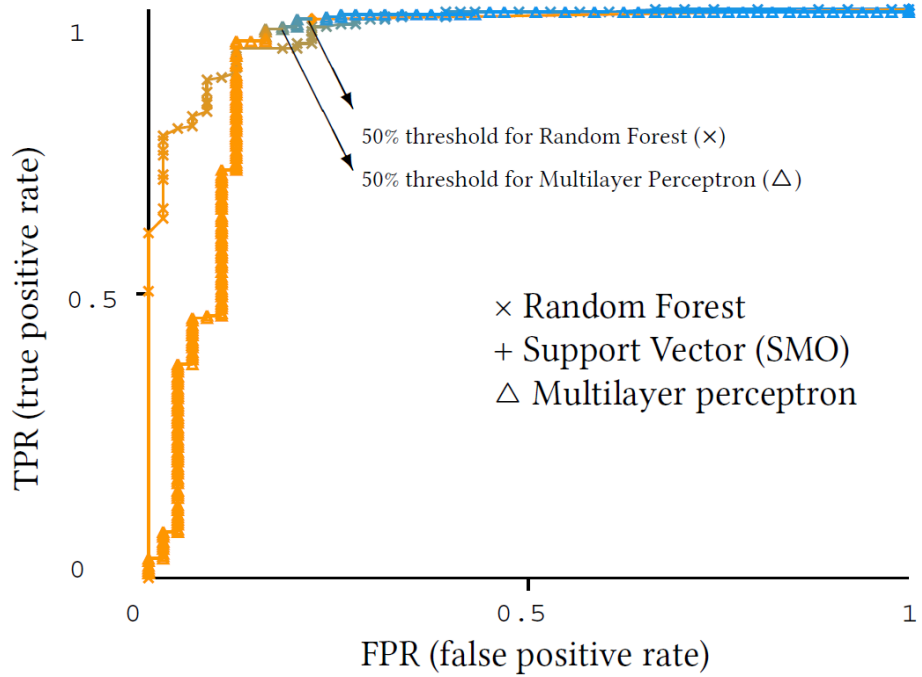
Reproducibility

- All available at OpenML
  - OpenML project https://www.openml.org/s/53
    - Training datasets
    - Parameters and results of executions



First tests. Predicting for EN-ES-lit

- Accuracy for different ML methods
  - Random Forest (RF)
  - Support Vector Machine (SMO)
  - Multilayer Perceptron (MP)
- Best results (higher Area Under the Curve) for RF
- Method: 10-fold cross validation (with fold stratified, i.e., keeping class proportions in each fold)



	Random Forest	Multilayer Perceptron	SMO
Correctly Classified Instances	211 (93.36%)	213 (94.25%)	211 (93.36%)
Incorrectly Classified Instances	15 (6.64%)	13 (5.75%)	15 (6.64%)
Kappa statistic	0.7865	0.8117	0.7748
Mean absolute error	0.1101	0.0641	0.0664
Root mean squared error	0.2288	0.2276	0.2576
Relative absolute error	34.8987%	20.3324%	21.0402%
Root relative squared error	57.7554%	57.4747%	65.0442%
Total Number of Instances	226	226	226

Other languages

- Accuracy for Random Forest (RF)
- Can we get a unique model for all languages? NO (tested for EN-ES-lit)
  - It is better an Ad hoc model for each language

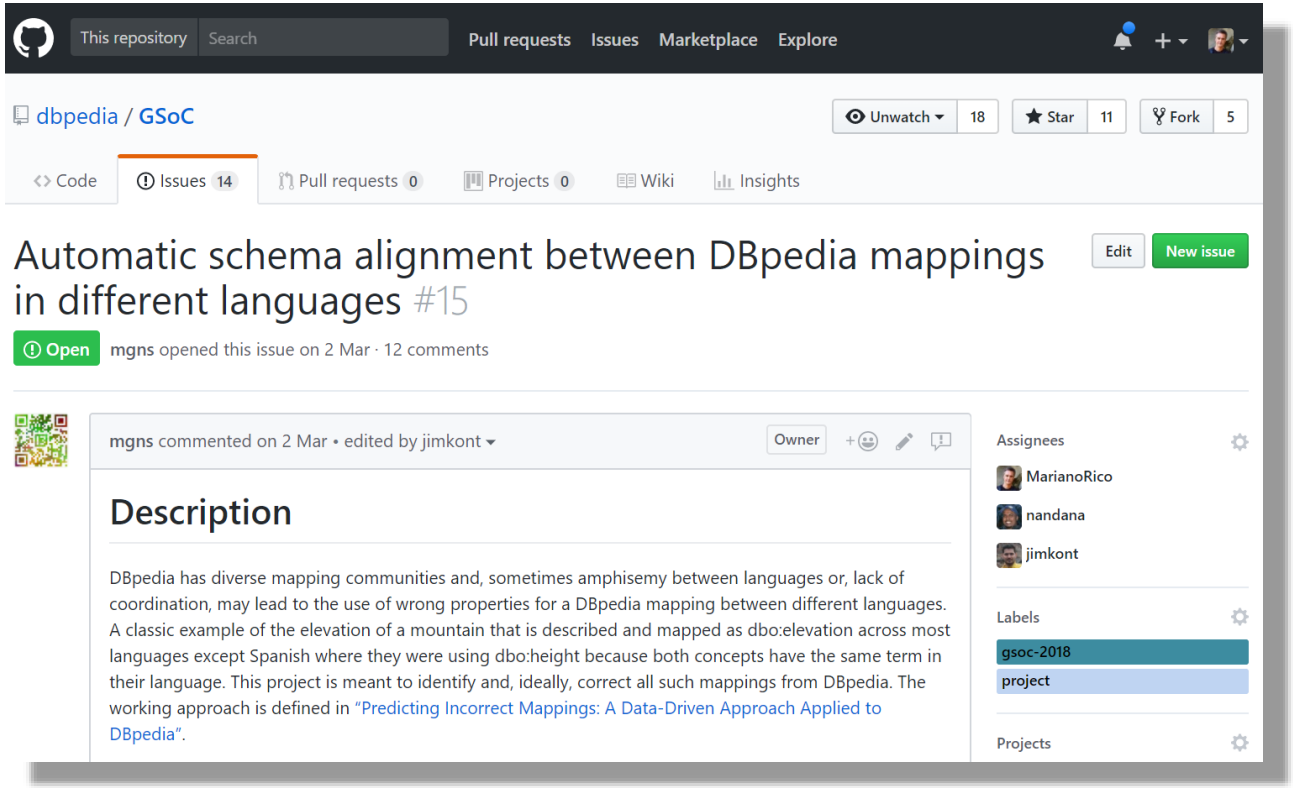
		EN-ES		ES-DE		EN-NL		EN-GR	
		lit	IRI	lit	IRI	lit	IRI	lit	IRI
Accuracy	Ad hoc model	93.36%	**95.00%**	N.A.	**96.36%**	**71.08%**	**~94%**	73.02%	**89.71%**
	En-ES-lit model		65.00%	N.A.	87.28%	61.45%	67.86%	**77.78%**	88.24%
Annotations	Total instances	211	80		110	83	28	63	68
	'Correct' instances	175	71		102	35	9	33	44
	'Incorrect' instances	36	9		8	48	19	30	24
Number of mappings		799	4979		4999	1329	4971	328	2785

Limitations

- The training set is difficult to get
  - You require people fluent in pairs of languages
    - Easy for Englis-*
    - Difficult for the rest
  - Annotation is hard --> Few annotations
    - The good news is that with 240 annotations we get good accuracy (up to 93%)

Future work

- To create a user interface to assist annotators
  - We have got a Google Summer of Code project
    - Web app
      - Balance high probably wrong mappings versus low number of mapping instances
      - Several experts should evaluate the same mappings to achieve a common agreement about correct/incorrect mappings
      - Deal with highly unbalanced mappings (e.g. few incorrect).



Thanks a LOD!!

Mariano.Rico@upm.es
@marianorico

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

Mariano Rico, Nandana Mihindukulasooriya, Dimitris Kontokostas, Heiko Paulheim, Sebastian Hellmann, Asunción Gómez-Pérez

European Union
European Regional Development Fund
"A way to build Europe"

This work was partially funded by the Spanish MINECO Ministry (projects RTC-2016-4952-7 and TIN2013-46238-C4-2-R), the BES-2014-068449 grant, and grants from the EU's H2020 Programmes for the ALIGNED project (GA 644055). Also, this work has been partially funded by the project Data 4.0 (TIN2016-78011-C4-4-R), from the Spanish State Investigation Agency of the MINECO and FEDER Funds.