

Leopard

ISWC Semantic Web Challenge 2017

René Speck^{1,2} and Axel-Cyrille Ngonga Ngomo³
speck@infai.org axel.ngonga@upb.de

¹Data Science Group, Institute for Applied Informatics, Germany

²Data Science Group, University of Leipzig, Germany

³Data Science Group, University of Paderborn, Germany

October 24th, 2017





■ Task one: attribute prediction

- Given:
 - organization-name
 - hasURL
- Prediction:
 - isDomiciledIn
 - hasLatestOrganizationFoundedDate
 - hasHeadquartersPhoneNumber

■ Task two: attribute validation

- Given:
 - organization-name
 - isDomiciledIn
- Validation:
 - hasURL
 - hasLatestOrganizationFoundedDate
 - hasHeadquartersPhoneNumber



knowledge graph by PermIDs (<http://permid.org>)



THOMSON REUTERS

■ Dataset one

- PermIDs: 14425
- Unique organization names: 14392
- Unique URLs: 13953

■ Dataset two

- PermIDs: 14351
- Unique organization names: 14309
- Statements: 41734

Duplicate examples

“Mcdonald’s” 17 times in dataset one, 30 times in dataset two

“<http://www.mcdonalds.com>” 79 times in dataset one, 75 times in dataset two



Leopard Pipeline

A BaseLine Approach to Attribute Prediction and Validation for Knowledge Graph Population.

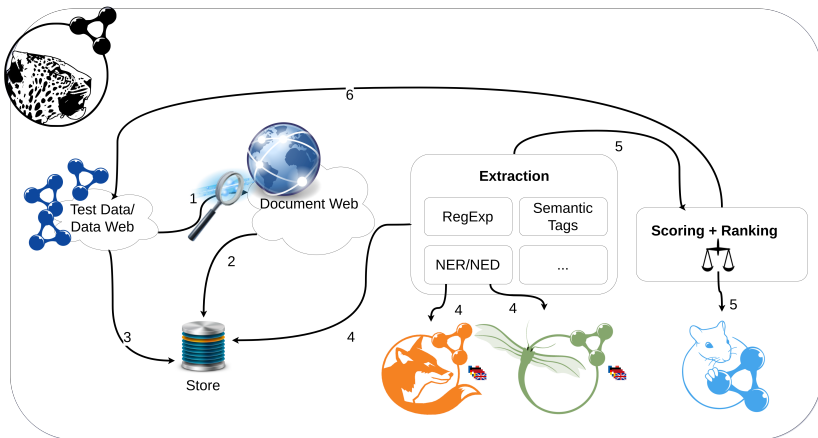


Figure : Overview of Leopards workflow



Leopard Extraction Modules

Phone number extraction to

`hasHeadquartersPhoneNumber` (0.5231 P, 0.0995 R),

`isDomiciledIn` (0.9754 P, 0.0094 R)

The University for the Information Society

Search (Google)

Warburger Straße 100 | 33098 Paderborn
Phone: +49 5251 600

Hotline student: +49 5251 60 50 40

Phone: +49 5251 600

PROSPECTIVE STUDENTS
INTERNATIONAL STUDENTS
BUSINESS
PRESS
ALUMNI

PEOPLE DIRECTORY
CAFETERIA
ONLINE APPLICATION
LIBRARY
PAUL

SITE NOTICE
LEGAL NOTICE
WEBRELAUNCH
LOGIN

NETWORKED WITH BUSINESS AND INDUSTRY
INNOVATION

COMPACT CAMPUS
67 COURSES OF STUDY

CUTTING-EDGE RESEARCH
INTERNATIONAL

DYNAMIC GROWTH
INTERDISCIPLINARY RESEARCH PROJECTS

<http://googlei18n/libphonenumber>



NER/NED to `isDomiciledIn`






- Website text to language detection
- NE of type PLACE with the multilingual version of Fox and Agdistis
- Find the country of the NE in DBpedia in case the NE is not a country
- Choose the country with the highest frequency
- 0.6837 P, 0.0355 R



Figure : Multilingual Fox and Agdistis (NER/NED)



Top Level Domain to `isDomiciledIn`

*.de		0.9678 P, 0.0321 R
*.fr		
*.uk		
...	...	
*.com		0.9005 P, 0.275 R
*.net		
...	...	



- Score each extraction module with Gerbil (precision)
- Leopard chooses the result of the module with the highest precision



Figure : Gerbil SWC is the evaluation platform for the Semantic Web Challenge at ISWC 2017



Annotator	F1 measure
Socrates	0.5539711491
Leopard	0.5343789728
Disco	0.5331521739

Figure : Task one attribute prediction results

Annotator	Area Under Curve
Socrates-KI	0.6801440802
MatchSoup	0.6518086946
Leopard	0.5308753416

Figure : Task two attribute validation results



Acknowledgement

The work presented in this talk has been funded by the H2020 project HOBBIT under the grant agreement number 688227.



InfAI®
Institute for Applied Informatics

<https://project-hobbit.eu>



Thank you! Questions?

René Speck
Data Science Group
speck@infai.org

<https://github.com/dice-group/Leopard>