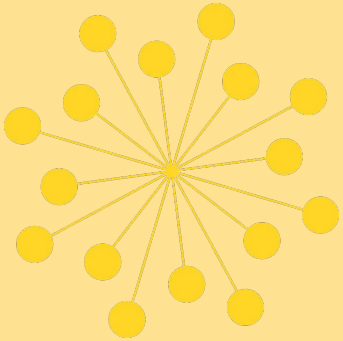


Part 1: Latest developments in DBpedia
Part 2: Entityclassfier.eu NER

Milan Dojčinovski

KEG Seminar
5 December, 2019

DBpedia Background

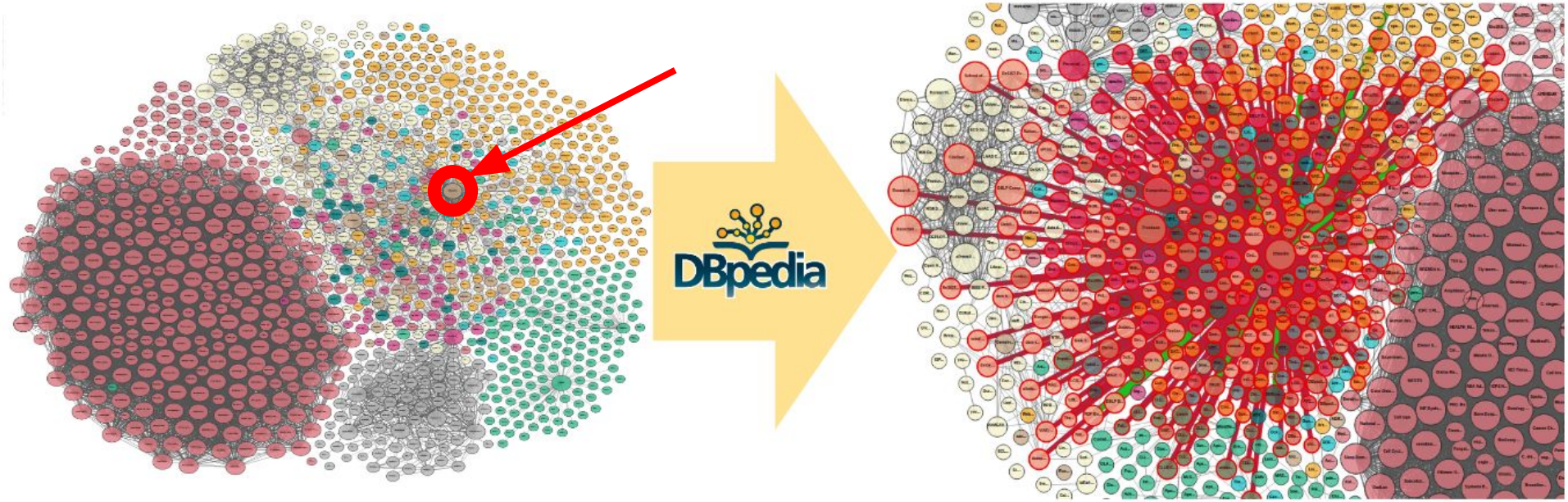
Initially started in 2007 in Leipzig, Germany

2008 - DBpedia is a crowd-sourced **community effort** to extract structured **information from Wikipedia** and make this information **available on the Web**.

2018 - New mission: Global and unified access to knowledge graphs

- Original definition still holds true
- Platform (matchmaking) to integrate your data with all other data
- DataBus https://databus.dbpedia.org

DBpedia allows to access data



Connecting data is about connecting people and organisations

Czech DBpedia

Po směru hodinových ručiček počínaje obrázkem nahoře: Pražský hrad, výškové budovy na Pankráci, Malá Strana, Staroměstské náměstí, Karlův most, Národní divadlo

Vlajka Prahy

Znak Prahy

heslo:	**Praga Caput Rei publicae**[p. 1] (dříve *Praha matka měst*)
status:	hlavní město, zároveň kraj a statutární město
historická země:	Čechy
LAU 2:	CZ0100 554782
kraj (NUTS 3):	**Hlavní město Praha** (CZ010)
okres (LAU 1):	**Hlavní město Praha** (CZ0100)
ISO 3166-2:CZ:	CZ-PR
Státní poznávací značka:	A
poštovní směrovací číslo:	100 00–199 00
katastrální výměra:	496 km²
obyvatel:	1 294 513[1]
rozpočtové výdaje:	60 991 mil. Kč (2010)[2]
hustota zalidnění:	2581,7 obyvatel/km²
zeměpisná šířka:	50° 05′ s. š.
zeměpisná délka:	14° 25′ v. d.
nadmořská výška:	177–399 m n. m.
nejvyšší bod:	vrch Teleček mezi Sobínem a Chrášťany *(399 m n. m.)*
nejnižší bod:	hladina Vltavy u Suchdola *(177 m n. m.)*
počet městských obvodů:	10
počet městských *(správních)* obvodů:	22
počet městských částí:	57
počet místních částí:	146

DBpedia

DBpedia

Search DBpedia... | @ http://cs.dbpedia.org | česky, čeština

CATEGORIES

TYPES

GALLERY

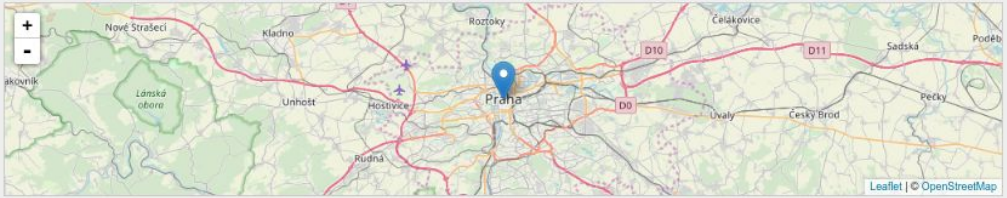
External Links

Born Here

Praha

Praha je hlavní a současně největší město Česka a 15. největší město Evropské unie. Leží mírně na sever od středu Čech na řece Vltavě, uvnitř Středočeského kraje, jehož je správním centrem, ale jako samostatný kraj není jeho součástí. Je sídlem velké části státních institucí a množství dalších organizací a firem. Sídlí zde prezident republiky, parlament, vláda, ústřední státní orgány a jeden ze dvou vrchních soudů.

cs.wikipedia.org/wiki/Praha



Property:	Value:
prop-cs:aprHi°c :	13.4 (xsd:double)
prop-cs:aprLo°c :	3.5 (xsd:double)
prop-cs:aprPrecipMm :	38.2 (xsd:double)
prop-cs:augHi°c :	23.5 (xsd:double)
prop-cs:augLo°c :	13 (xsd:integer)
prop-cs:augPrecipMm :	69.6 (xsd:double)
prop-cs:další :	nafotografovali, sestavili, úvodem a rejstříkem opatřili Barbora a Marek Lašťovkovi @cs
prop-cs:decHi°c :	2.1 (xsd:double)

Core Dataset Groups

Available extractions 13 billion facts total (200 GB)

- Generic (automatic)
- Mappings-based (rule-based)
- Text
- Wikidata

Based on the Wikimedia XML dumps

Generic extraction

- 132 languages, 30 datasets
- http://dbpedia.org/property/ properties
- https://en.wikipedia.org/w/index.php?title=Prague&action=edit
- http://dbpedia.org/page/Prague
- https://github.com/dbpedia/extraction-framework/tree/master/core/src/main/scala/org/dbpedia/extraction/mappings

Mappings based extraction

- 40 languages, 6 datasets
- http://dbpedia.org/ontology properties

```
| coordinates                = {{coord|50|05|N|14|25|E|region:CZ|display=inline,title}}
| subdivision_type           = Country
| subdivision_name           = [[Czech Republic]]
| established_title          = Founded
| area_urban_km2             = 298
```

Property Mapping (help)	
template property	area_urban_km2
ontology property	areaUrban
unit	squareKilometre

Geocoordinates Mapping (help)	
coordinates template property	coordinates

Geocoordinates Mapping (help)	
longitude degrees template property	longd
longitude minutes template property	longm
longitude seconds template property	longs
longitude direction template property	longEW
latitude degrees template property	latd
latitude minutes template property	latm
latitude seconds template property	lats
latitude direction template property	latNS

- http://dbpedia.org/resource/Prague

Text extraction

- 132 languages, 8 datasets
- Short and long abstracts
- Training data for text mining
- Fact extraction

Currently offline due to maintenance (refactoring)

Wikidata extraction

- Same approach as for Wikipedia:
  - Generic and Mappings-based
- Mappings in JSON

+ Allows unified access over Wikipedia and Wikidata
+ Wikidata has no ontology, DBpedia has 8 (DBO, Yago, Umbel,..)
+ Generic still extracts 584 million facts

```
"P279": [
    {
        "rdfs:subClassOf": "$getDBpediaClass"
    }
],
"P625": [
    {
        "rdf:type": "http://www.w3.org/2003/01/geo/wgs84_pos#SpatialT
    },
    {
        "geo:lat": "$getLatitude"
    },
    {
        "geo:long": "$getLongitude"
    },
    {
        "georss:point": "$getGeoRss"
    }
],
```

What's New in DBpedia

- 2+ years innovation phase (ongoing)
- Reengineered the released process -> DataBus
- More frequent releases -> monthly/bi-weekly
- Provenance support -> at fact level
  - same facts but different sources, basis for fusion
- Support for browsing data
- ID Management -> same entity in different Wikipedia editions
  - http://global.dbpedia.org
- Dataset dependency declaration
  - my service uses: dbpedia/Mapped-Geocoordinates/2017.05.10

Digital Factory Platform

https://databus.dbpedia.org/

Inspired by Maven Central

Databus - Digital Factory Platform

https://databus.dbpedia.org

- Hosts a public metadata repository
- Not exclusive to DBpedia data
- Free to use, published metadata must be CC-0
- Published files stay on publisher's server
  - Full control over access (HTTP-Auth) and license

Automated monthly releases

ABOUT

Classification of instances with the DBpedia Ontology. Contains triples of the form `<$resource> rdf:type <$dbpedia_ontology_class>` generated by the mappings extraction.

ARTIFACT VERSIONS

Version	Release Date	License
2019.10.01 (Documentation)	Oct 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.09.01 (Documentation)	Sep 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.08.30 (Documentation)	Aug 30th 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.08.01 (Documentation)	Aug 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.07.01 (Documentation)	Jul 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.06.01 (Documentation)	Jun 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.05.01 (Documentation)	May 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.04.20 (Documentation)	Apr 20th 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.04.01 (Documentation)	Apr 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.03.01 (Documentation)	Mar 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0
2019.01.01 (Documentation)	Jan 1st 2019	http://purl.oclc.org/NET/rdflicense/cc-by3.0

Data Provenance and Versioning: DataID

```
######
#     #   ##   #####   ##   #####  #    #  ####
#     #  #  #    #    #  #  #    # #    # #
#     # #    #   #   #    # #####  #    #  ####
#     # ######   #   ###### #    # #    #      #
#     # #    #   #   #    # #    # #    # #    #
######  #    #   #   #    # #####   ####   ####

# Plugin version 1.3-SNAPSHOT - https://github.com/dbpedia/databus-maven-plugin


@prefix databus: <https://databus.dbpedia.org/> .
@prefix dataid-mt: <http://dataid.dbpedia.org/ns/mt#> .
@prefix dct:   <http://purl.org/dc/terms/> .
@prefix dataid: <http://dataid.dbpedia.org/ns/core#> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dataid-cv: <http://dataid.dbpedia.org/ns/cv#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dcat:  <http://www.w3.org/ns/dcat#> .
@prefix prov:  <http://www.w3.org/ns/prov#> .
@prefix dataid-debug: <http://dataid.dbpedia.org/ns/debug.ttl#> .

<http://dbpedia-mappings.tib.eu/release/mappings/instance-types/2019.10.01/dataid.ttl#instance-types_lang=el.ttl.bz2>
        a                              dataid:SingleFile ;
        dataid:associatedAgent         <https://vehnem.github.io/webid.ttl#this> ;
        dataid:compression             "bzip2" ;
        dataid:contentVariant          "el" ;
        dataid:duplicates              "0"^^xsd:decimal ;
        dataid:file                    <https://databus.dbpedia.org/marvin/mappings/instance-types/2019.10.01/instance-types_lang=el.ttl.bz2> ;
        dataid:formatExtension         "ttl" ;
        dataid:isDistributionOf        <http://dbpedia-mappings.tib.eu/release/mappings/instance-types/2019.10.01/dataid.ttl#Dataset> ;
        dataid:nonEmptyLines           "102249"^^xsd:decimal ;
        dataid:preview                 "# started 2019-10-08T13:21:28Z\n<http://el.dbpedia.org/resource/ÎŸÎ¾Î¹Ï„Î±Î½Î¹ÎºÎ®_Î³Î»ÏŽÏƒÏƒÎ±> <http://www.w3.org/1999/02/22-rdf-syntax-
ns#type> <http://dbpedia.org/ontology/Language> .\n<http://el.dbpedia.org/resource/ÎŒÎ¼Î·ÏÎ¿Ï‚> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Person>
.\n<http://el.dbpedia.org/resource/Î›ÎµÏ…ÎºÏ‰ÏƒÎ¯Î±> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/City> .\n<http://el.dbpedia.org/resource/ÎœÏ…
Î¸Î¿Î»Î¿Î³Î¯Î±> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Work> .\n<http://el.dbpedia.org/resource/Î•Î»Î»Î·Î½Î¹ÎºÎ®_Î³Î»ÏŽÏƒÏƒÎ±>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Language> .\n<http://el.dbpedia.org/resource/Î•Î»ÎµÏ…Î¸ÎÏÎ¹Î¿Ï‚_Î'ÎµÎ½Î¹Î¶ÎÎ»Î¿Ï‚>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Politician> .\n<http://el.dbpedia.org/resource/Mega_Channel> <http://www.w3.org/1999/02/22-rdf-syntax-
ns#type> <http://dbpedia.org/ontology/TelevisionStation> .\n<http://el.dbpedia.org/resource/ÎšÏÏ€ÏÎ¿Ï‚> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://dbpedia.org/ontology/Country> .\n<http://el.dbpedia.org/resource/Î'Î»Î²Î±Î½Î¯Î±> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://dbpedia.org/ontology/Country> ." ;
        dataid:sha256sum               "3ef05a493808099ffe9d90563aa712249eab325f424e6f485dc5893b117f00f2" ;
        dataid:signature
"OxmICj7sQ8jIhzbGc9aohiBLoSa55poX4iUTid5bdjq/plk3519CdspZbyJD7+rvbTvs09JqO/+1sqW+/rCZe3eeqVz6+4XQbI8Uil0bTk5/3rShFQgjwmQo/aD9d/ry5C4xpyWWJHMb5qeSkwmyh6eeMc8WLoJSGqYVacik1wiXcUF9Pr77O
cMYOoQHWMdwtIWoX0Gm07B+ItwosKQOCNpRuRK0+2iffN8Ashaz0BmynLgInilSQymZktf4oxDzLtyLeExCqPYsdjUbAuI4JebmrtGF6ciDOBuWSjH1OLkAOfe/I2sG73yYCg6XgqhpBkmBJUcF95sxpIbDcjOblA==" ;
        dataid:sorted                  false ;
        dataid:uncompressedByteSize    "16243387"^^xsd:decimal ;
        dataid-cv:lang                 "el" ;
        dct:conformsTo                 "http://dataid.dbpedia.org/ns/core#" ;
        dct:hasVersion                 "2019.10.01" ;
        dct:issued                     "2019-10-01T00:00:00Z"^^xsd:dateTime ;
        dct:license                    <http://purl.oclc.org/NET/rdflicense/cc-by3.0> ;
        dct:modified                   "2019-10-10T11:42:41Z"^^xsd:dateTime ;
```

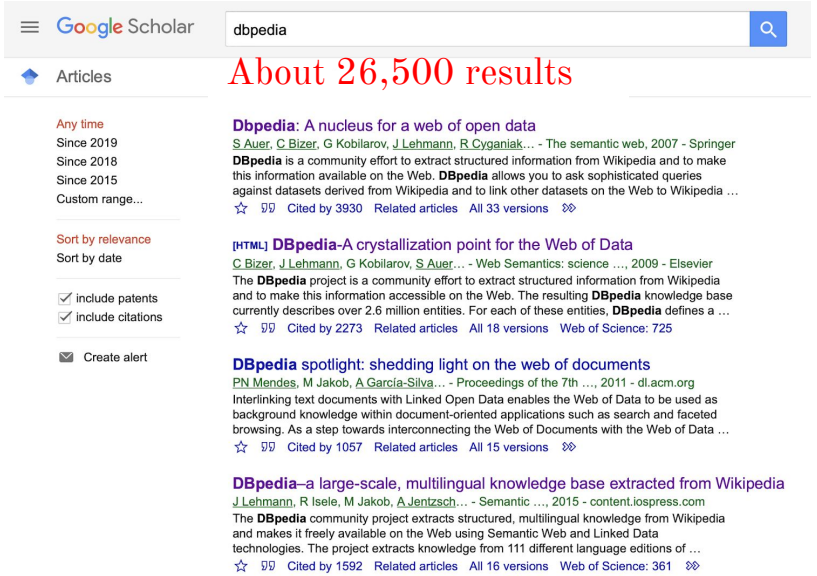
The Community

- Regular meetings
  - 14 community meetings since 2014 + several meetups
  - Usually co-located with SEMANTiCS or LDK conferences
  - https://wiki.dbpedia.org/join/community-meetings
- Language chapters
  - German, Czech, Polish, Dutch, Italian, Catalan, Spanish, ...

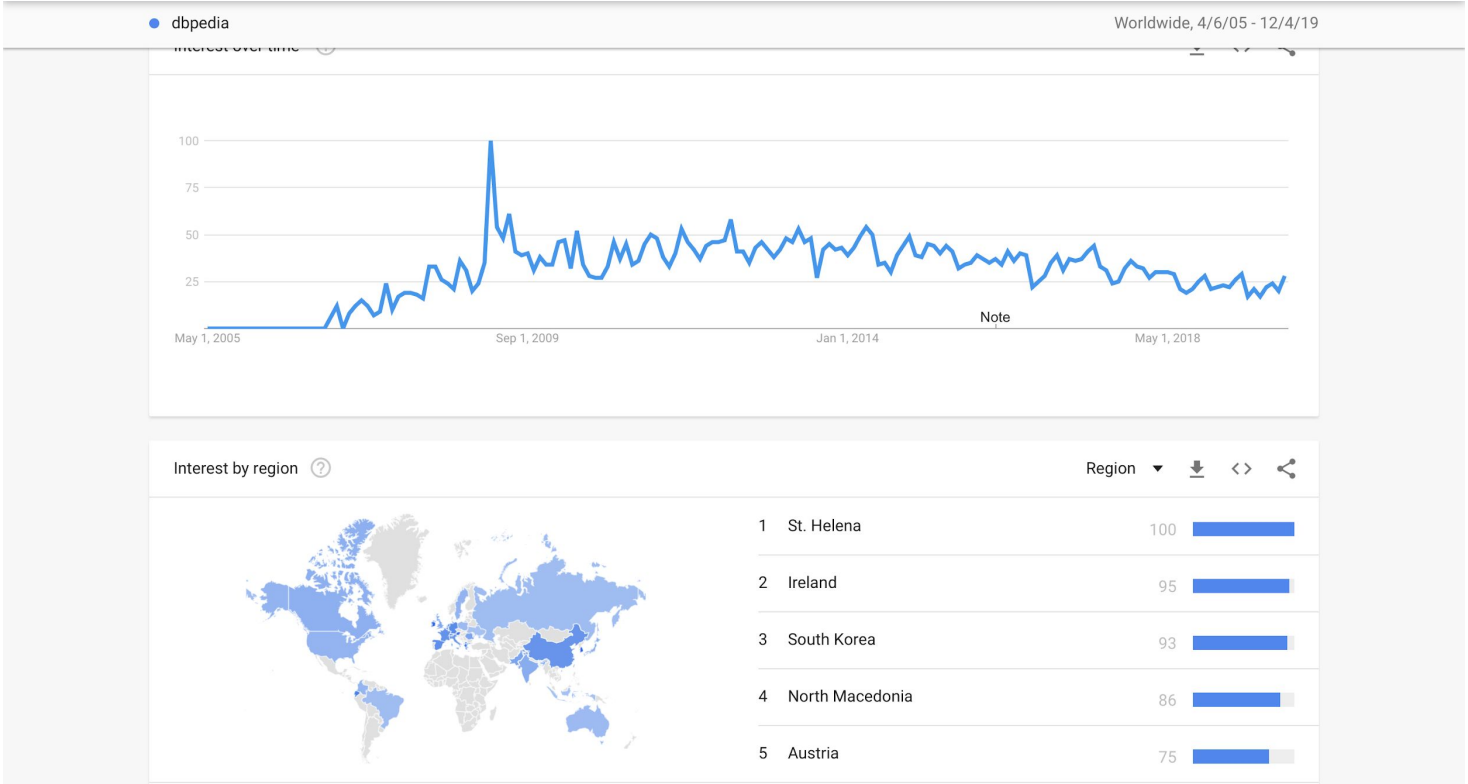
Impact

- Over 26,000 DBpedia related papers
- 400 developers/researchers
- 600K file downloads per year
- 20 million hits daily (all APIs)
- 400 developers/researchers
- Inlinks by all major datasets
- 4000 website visitors weekly

+ and growing



Google Trends: DBpedia



DBpedia Association members

Key Links

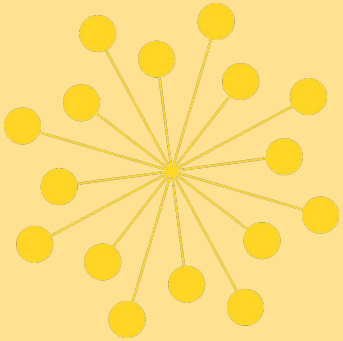
https://wiki.dbpedia.org - main website, general info

https://databus.dbpedia.org -> Databus interface, find data

https://forum.dbpedia.org -> DBpedia forum, Q&A

https://downloads.dbpedia.org/repo/lts/ -> data location, but better use the databus interface

http://dbpedia.slack.com -> slack channel



DBpedia

Part 1: Latest developments in DBpedia

Part 2: Entityclassfier.eu NER

Entityclassifier.eu NER

Recognition, Classification, Linking of Salient Named Entities

Background

- Project initiated in May 2012
- Developed during the LinkedTV EU FP7 project
- Some parts developed as part of the LOD2 project

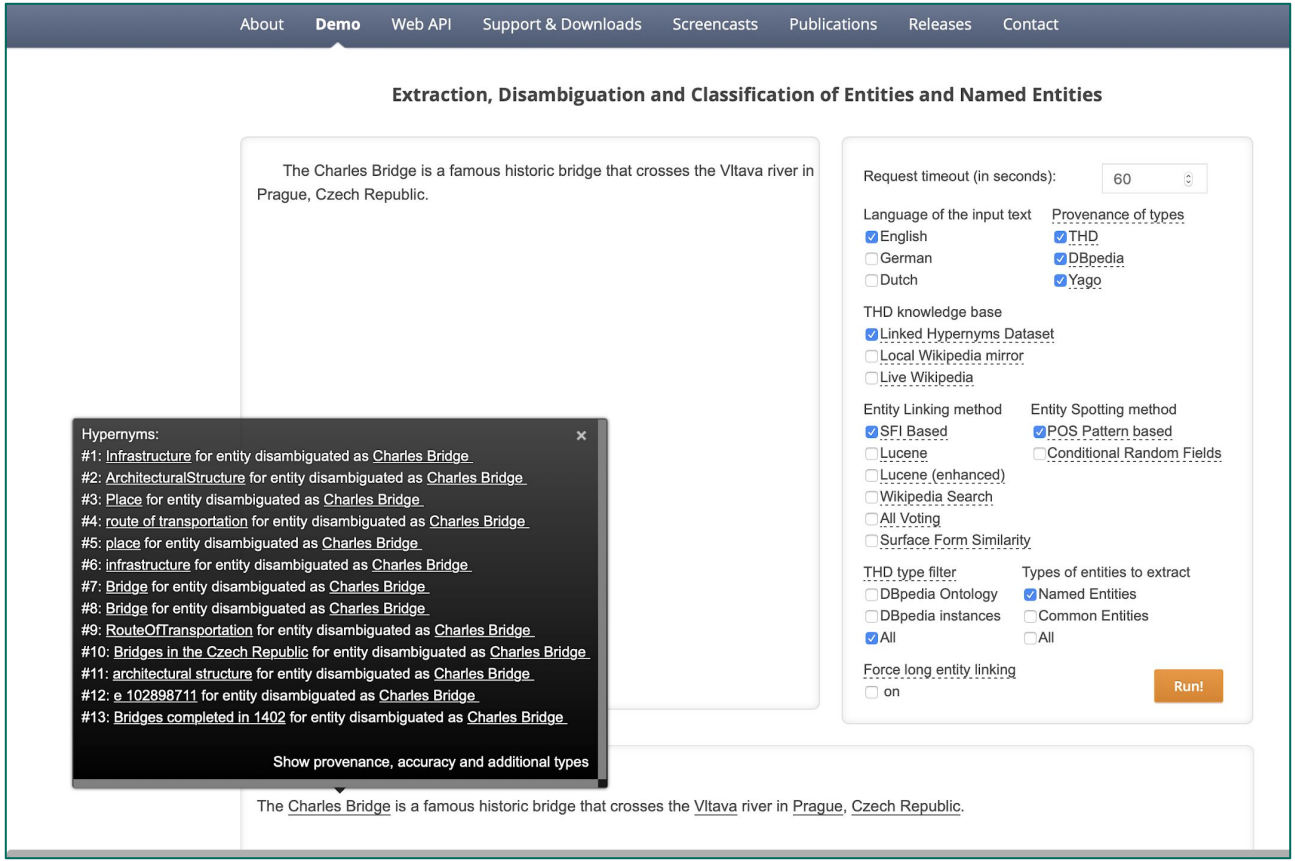
A System for Recognition, Classification and Linking

of Salient Named Entities

http://Entityclassifier.eu

Entityclassifier.eu NER

Recognition, Classification, Linking of Salient Named Entities



Core Features

- Implemented in Java and GATE: https://gate.ac.uk
- Entity spotting -> JAPE grammars -> POS tags based
- Entity classification/typing -> JAPE grammars
  - Hypernym extraction of from the respective Wikipedia article
  - “Prague is the capital and largest city in the Czech Republic…”
  - Hypernum linked to DBpedia entity,http://dbpedia.org/resource/Capital
- Entity linking
  - Most Frequent Sense based -> Wikipedia Lucene index
  - Surface Form Similarity
  - Context based (DBpedia abstract as entity description)
  - … several additional modifications/combinations

Core Features (cont.)

- Entity Salience
  - Extraction of the most important entities
  - Entities that play an important role in the document
- Machine learning
  - Training data via crowdsourcing (Figure Eight, ex Crowdflower)
    - most salient, less salient, not salient annotations
  - Experimented with several ML algorithms
  - Random tree forest best performing algorithm

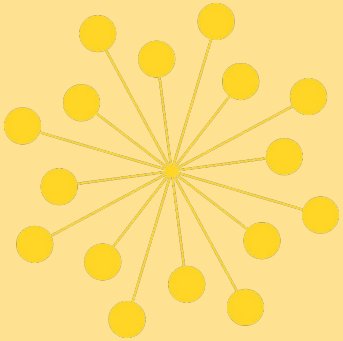
Unique Features

- Realtime entity classification
  - Executed against live Wikipedia API
- Temporal type validity
  - Diego maradona: now “football manager”, past “footballer”
  - Content is Wikipedia is regularly updated (for popular entities)
- Extraction of salient entities
  - Documents contain important and unimportant entities
  - Filter only important entities

Lessons learned

- Simple approaches perform relatively well
  - POS based recognition, most-frequent-sense based linking
- Provide an API
  - People would rather use the API then install the tool on local host
- Measure the impact
  - Collect data via API key request form
  - Jan 2015 to Aug 2017 = 59 API request, 42 unique users, 18 countries, 22 institutions, 59 English, 11 Dutch, 9 German
- Publish a paper with title the name of the tool (select keywords)

"Entityclassifier.eu: Real-Time Classification of Entities in Text with Wikipedia"



Thank you

Milan Dojčinovski

KEG Seminar
5 December, 2019