Sustainable Linked Data Generation

The case of DBpedia

Wouter Maroy, Anastasia Dimou, Dimitris Kontokostas,
Ben De Meester, Ruben Verborgh, Jens Lehmann,
Erik Mannens, Sebastian Hellman

GHENT UNIVERSITY

IDLab
INTERNET & DATA LAB

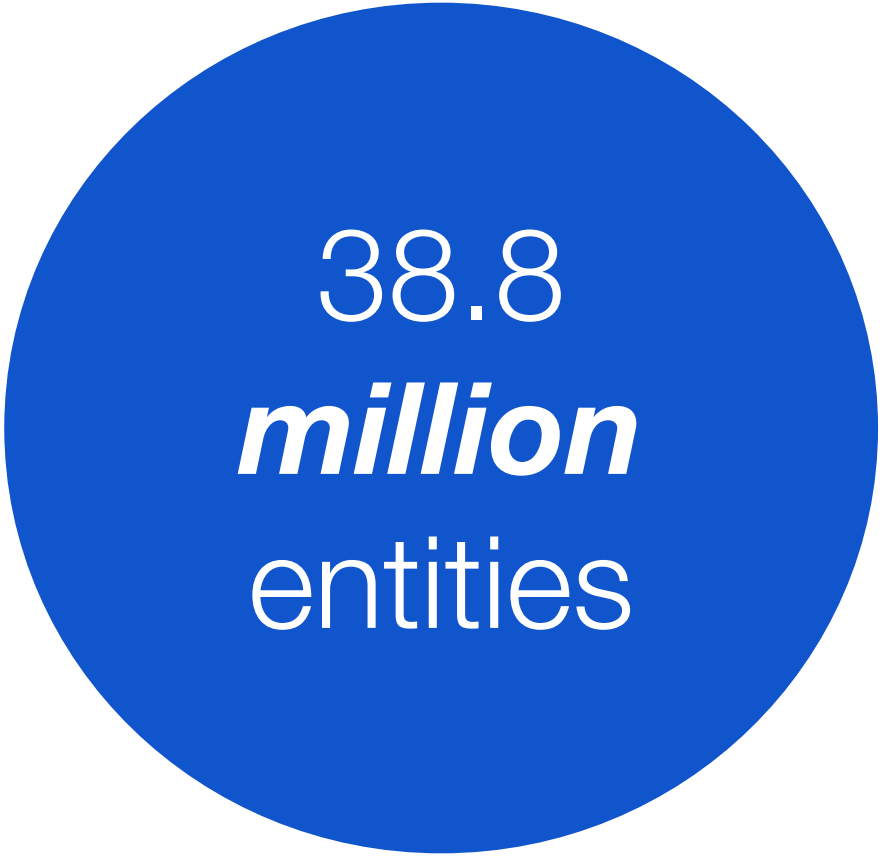
imec

wouter.maroy@ugent.be

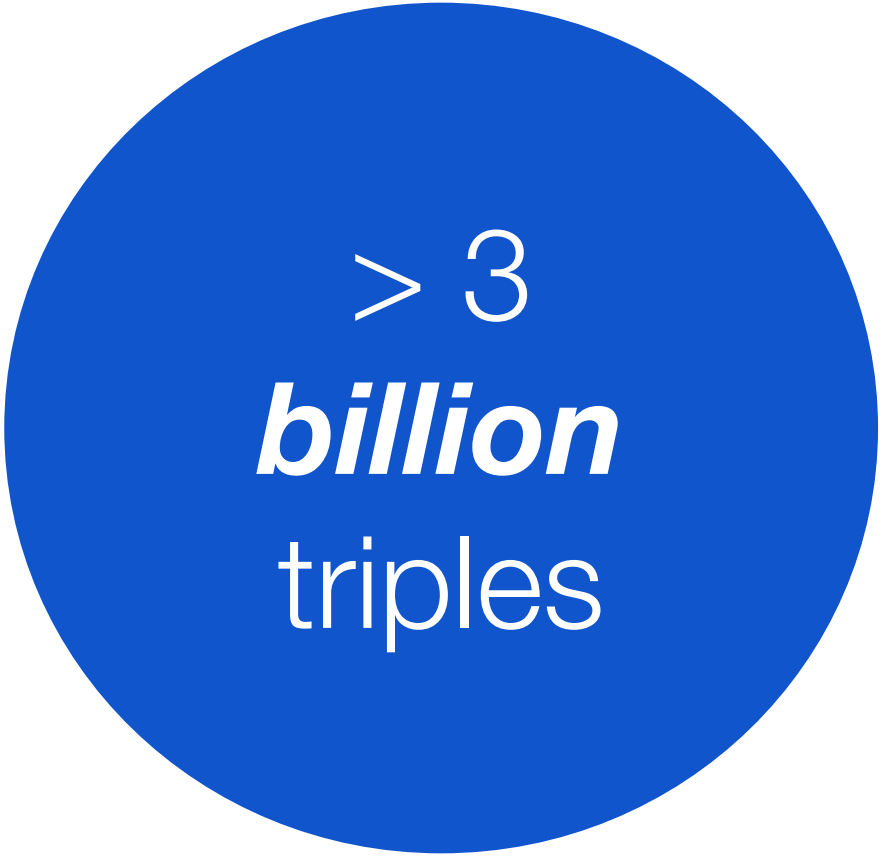
@wmaroy


DBpedia

DBpedia describes



DBpedia contains

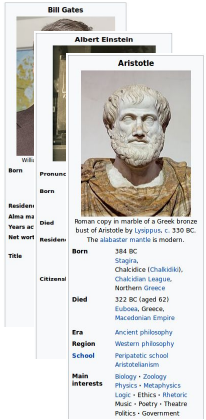


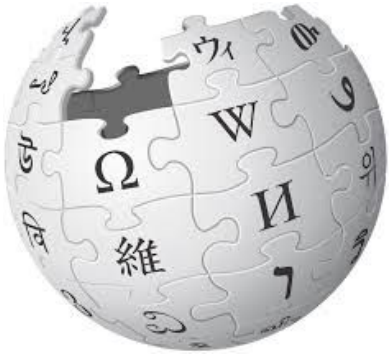
DBpedia keeps growing,
but there are quality issues

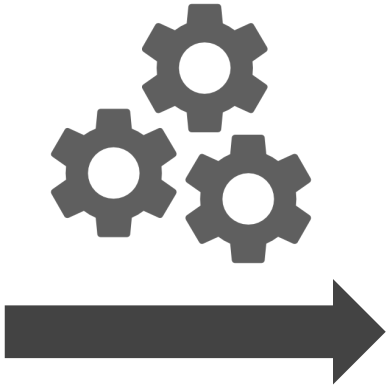
There are 2 types of quality issues

- Schema-level
- Data-level

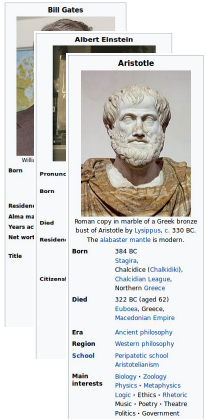
What are the causes?

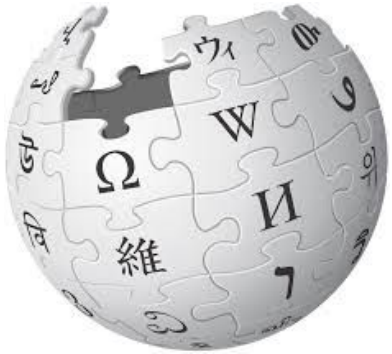


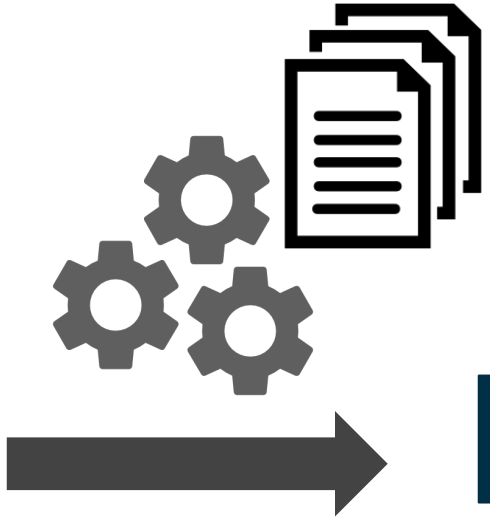




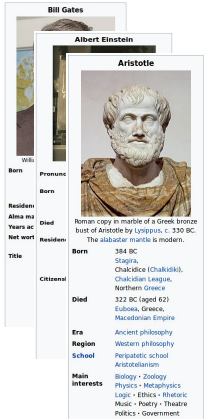
DBpedia's **Extraction Framework**
extracts Wikipedia's infoboxes

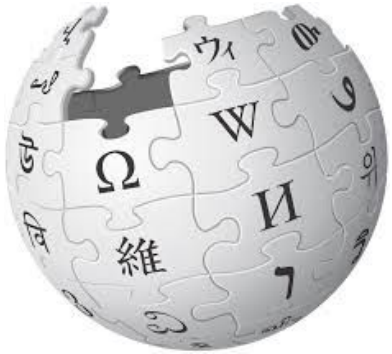


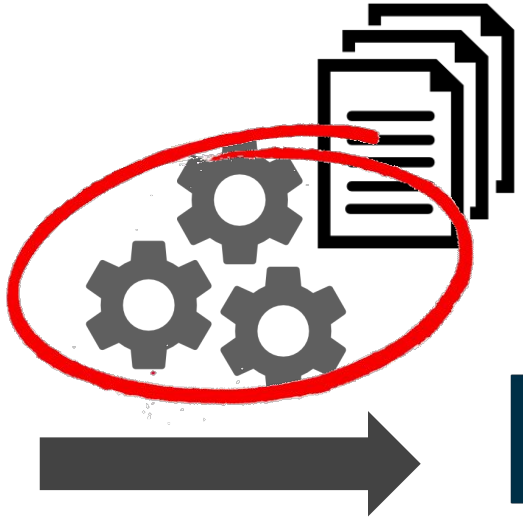




The community created **mapping rules** from infobox properties to a schema



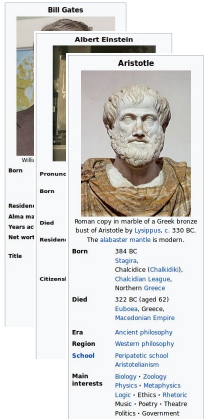


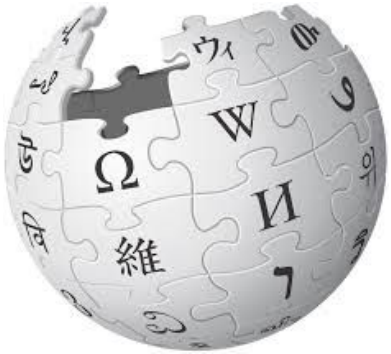


DBpedia

Causes for the issues can be found in the

Extraction Framework (EF)

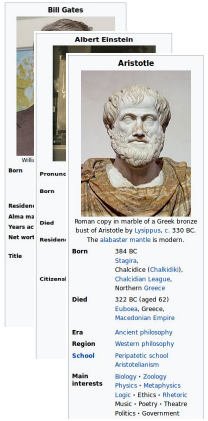


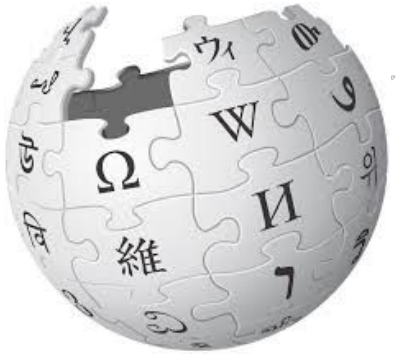


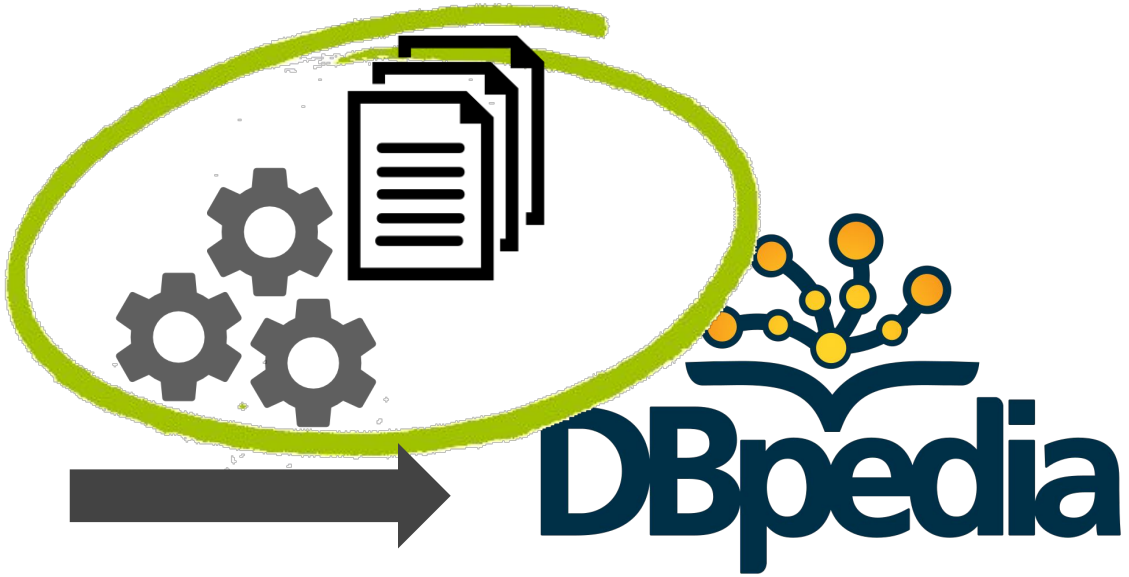
DBpedia

Causes for the issues can be found in the
mapping rules (MR)

Causes for the issues can be found in
Wikipedia itself





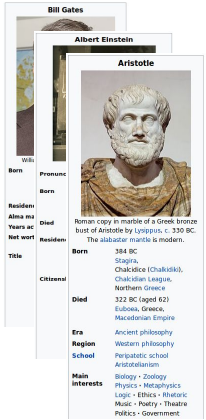


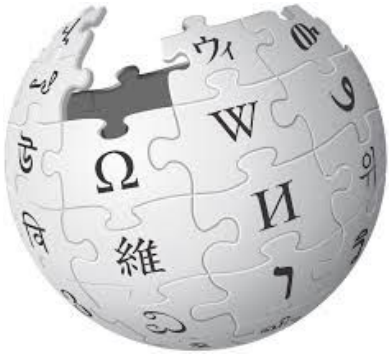
Our goal is to adjust the **EF & the MR**
to provide a more sustainable framework

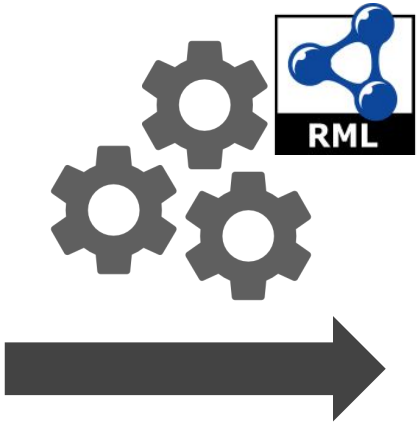
We integrated a generic, modular, and sustainable mapping language.

RML.io

RDF mapping language







The result is a framework that enables

sustainable Linked Data generation

DBpedia is making the switch!

Sustainable Linked Data Generation

The case of DBpedia

Before

After

Progress

Limitations of the EF

- **Hard-coded mapping rules**
- No machine-interpretable mapping rules
- No other ontology
- No schema-validation on mapping rules

Hard-coded mapping rules

Subject **Predicate** **Object**

Mapping rules define triple generation

Hard-coded mapping rules

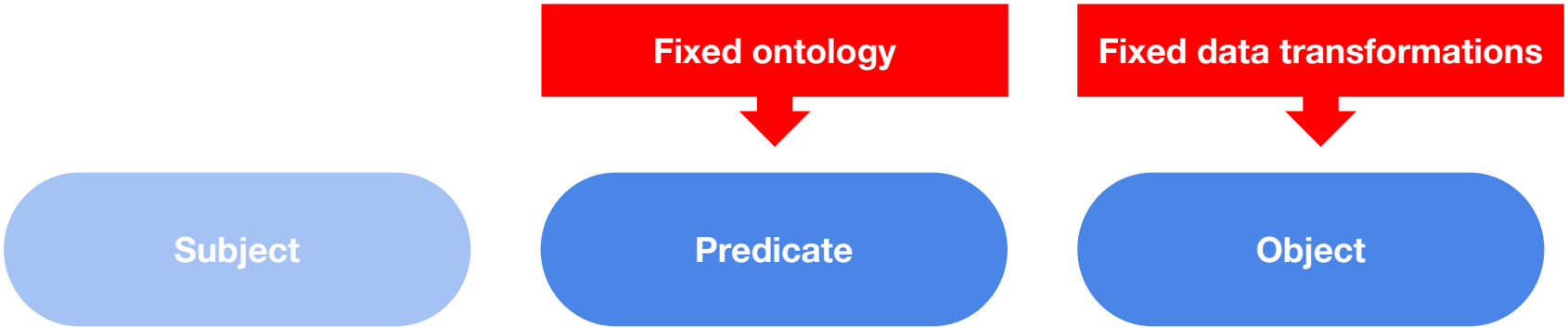
Subject

Predicate

Object

Mapping rules only influence the predicate and object due to ***implementation coupling***

Hard-coded mapping rules



This also limits defining the predicate and object

Limitations of the EF

- Hard-coded mapping rules
- **No machine-interpretable mapping rules**
- No other ontology
- No schema-validation on mapping rules

No machine-interpretable MR

Mapping rules are in **Wikitext format**

The same format that is used for defining Wikipedia articles

No machine-interpretable MR

Mapping rules are in **Wikitext format**

This format cannot be interpreted automatically

No machine-interpretable MR

Mapping rules are in **Wikitext format**

No querying

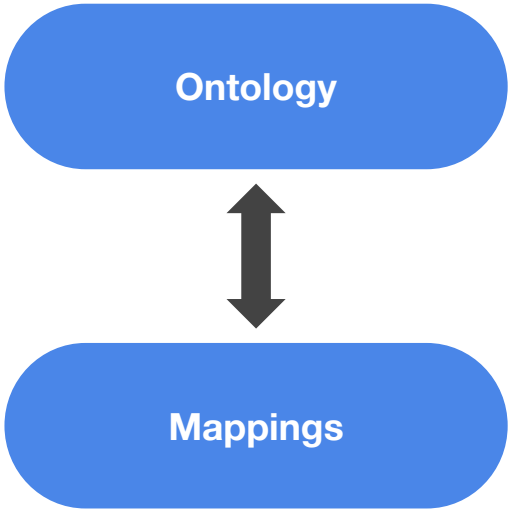
No schema-validation

No generation

Limitations of the EF

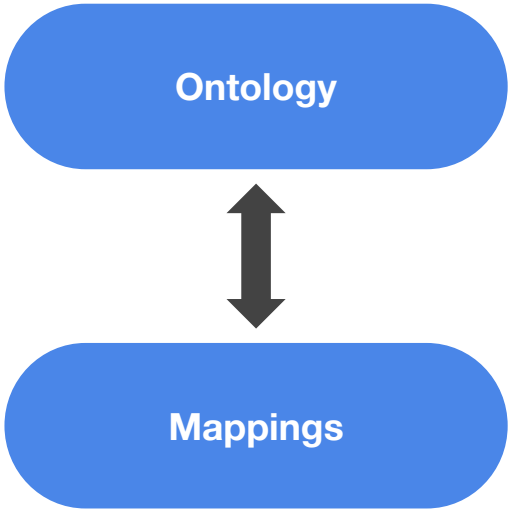
- Hard-coded mapping rules
- No machine-interpretable mapping rules
- **Restricted to the DBpedia ontology**
- No schema-validation on mapping rules

Restricted to the DBpedia ontology



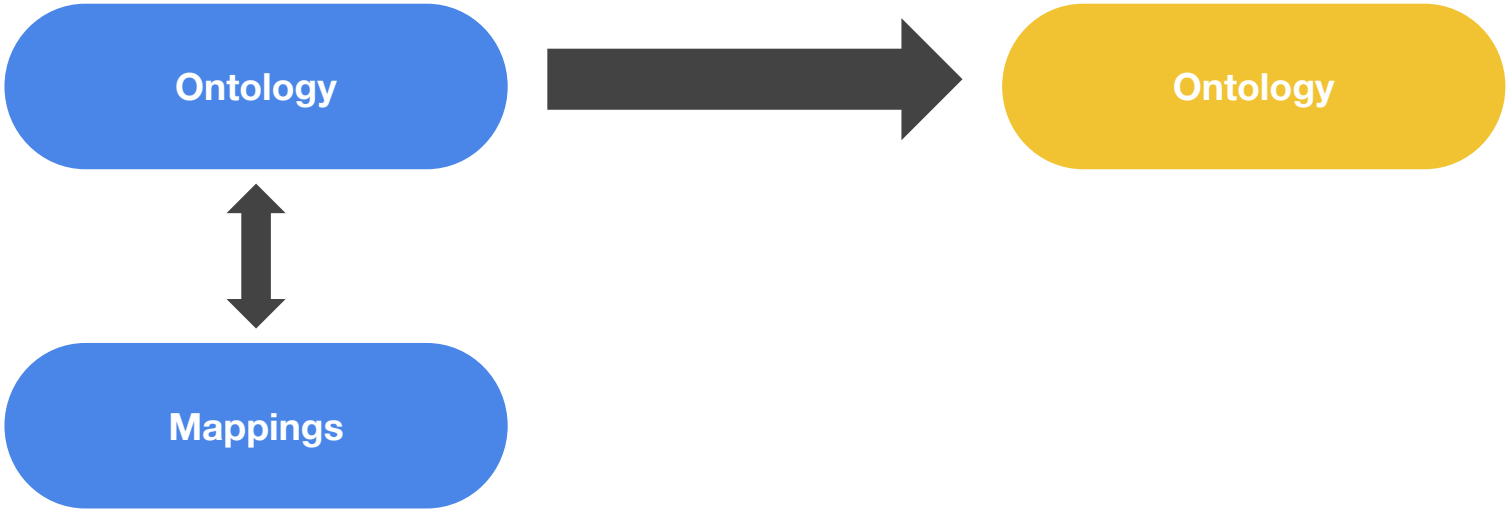
One ontology is used

Restricted to the DBpedia ontology



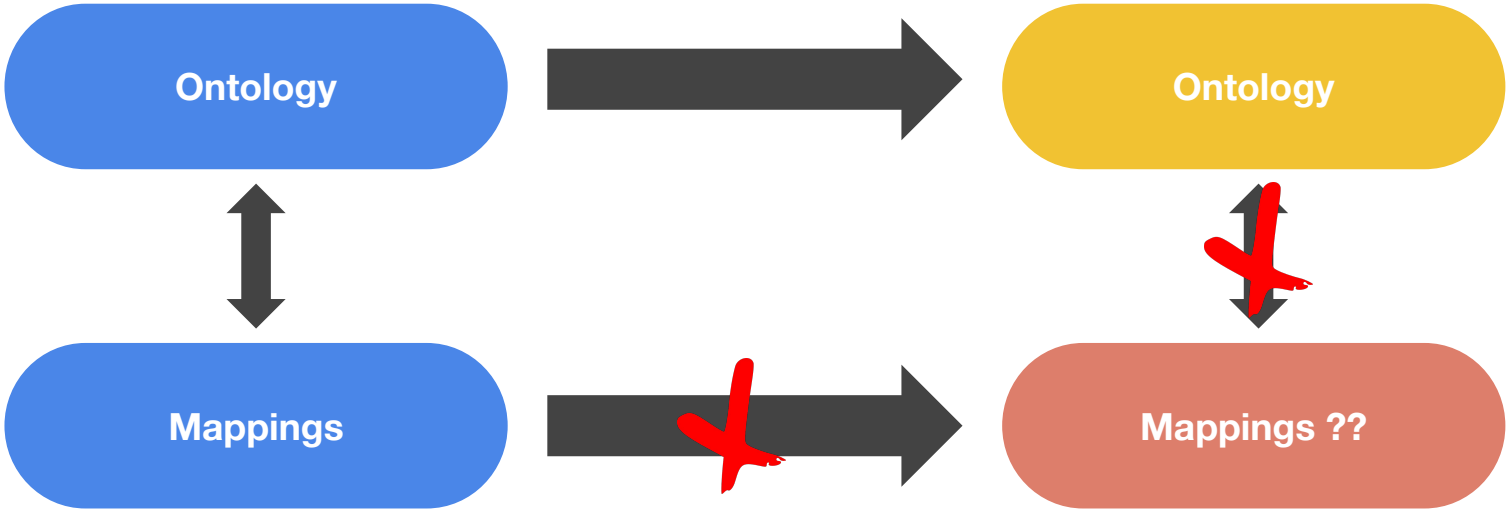
That is also coupled with its implementation

Restricted to the DBpedia ontology



What if we want to change the ontology?

Restricted to the DBpedia ontology



All mappings need to be changed manually

Limitations of the EF

- ✗ Hard-coded mapping rules
- ✗ No machine-interpretable mapping rules
- ✗ Restricted to the DBpedia ontology
- ✗ **No schema validation on mapping rules**

No schema validation

Validating the dataset
requires **many resources**

There are other options...

> 3
billion
triples

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date   =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

An infobox template for defining persons on Wikipedia

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

Bill Gates

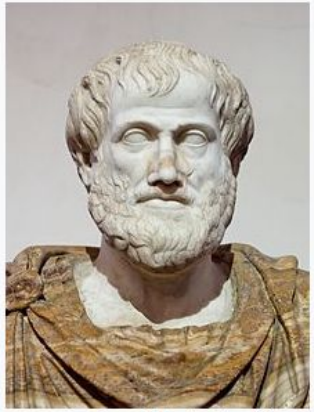
William H. Gates III in June 2015

Born	William Henry Gates III October 28, 1955 (age 61) Seattle, Washington, US
Residence	Medina, Washington, US
Alma mater	Harvard University
Years active	1975–present
Net worth	US$85.1 billion (September 2017)[1]
Title	Technology Advisor of Microsoft Co-Chairman of the Bill & Melinda Gates Foundation CEO of Cascade Investment Chairman of Branded Entertainment Network Chairman of TerraPower

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

Bill Gates

Aristotle



Roman copy in marble of a Greek bronze bust of Aristotle by Lysippus, c. 330 BC. The alabaster mantle is modern.

Born	384 BC Stagira, Chalcidice (Chalkidiki), Chalcidian League, Northern Greece
Died	322 BC (aged 62) Euboea, Greece, Macedonian Empire
Era	Ancient philosophy
Region	Western philosophy
School	Peripatetic school Aristotelianism
Main interests	Biology · Zoology Physics · Metaphysics Logic · Ethics · Rhetoric Music · Poetry · Theatre Politics · Government

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

> 253 000 pages
use the “person” infobox template

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```

Only **one** mapping
is responsible for extraction

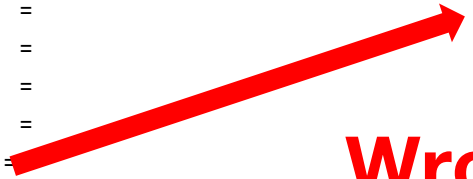
```
{{Infobox person
| name
| image          =
| alt            =
| caption        =
| birth_name     =
| birth_date     =
| death_date     =
| death_place    =
| nationality    =
| other_names    =
| occupation     =
| years_active   =
| known_for      =
| notable_works  =
...
}}
```

dbo:**name**

Changes at least **250 000** times

dbo:**givenName**

```
{{Infobox person
| name          =
| image         =
| alt           =
| caption       =
| birth_name    =
| birth_date    =
| death_date    =
| death_place   =
| nationality   =
| other_names   =
| occupation    =
| years_active  =
| known_for     =
| notable_works =
...
}}
```



dbo:**productionStartYear**

Wrong at least
250 000 times

No schema validation

Validating mappings is more **feasible**, more **efficient** & **sustainable**

DBpedia mappings quality assessment
Anastasia Dimou, Dimitris Kontokostas, Markus Freudenberg, Ruben Verborgh, Jens Lehmann, Erik Mannens, Sebastian Hellman and Rik Van de Walle

No schema validation

But there is no validation for the current mapping rules!

Sustainable Linked Data Generation

The case of DBpedia

Before

After

Progress

A sustainable framework is needed that provides

- ✓ Declarative mapping rules
- ✓ Machine-interpretable format
- ✓ Schema validation
- ✓ Usage of other ontologies

Our solution

The **RDF Mapping language (RML)**

A generic scalable mapping language
defined to express rules that map data
in heterogeneous structures and serializations
to the **RDF** data model

Our solution

The **RDF Mapping language (RML)**

Mapping rules in **RML** are **RDF**

A sustainable framework is needed that provides

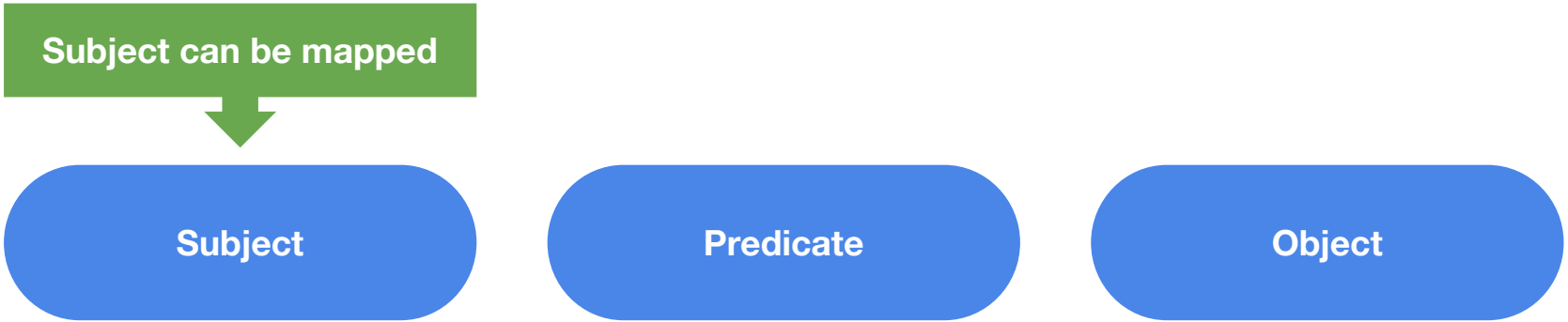
- ✓ **Declarative mapping rules**
- ✓ Machine-interpretable format
- ✓ Schema validation
- ✓ Allows alternative ontology

✓ Declarative mapping rules

Subject **Predicate** **Object**

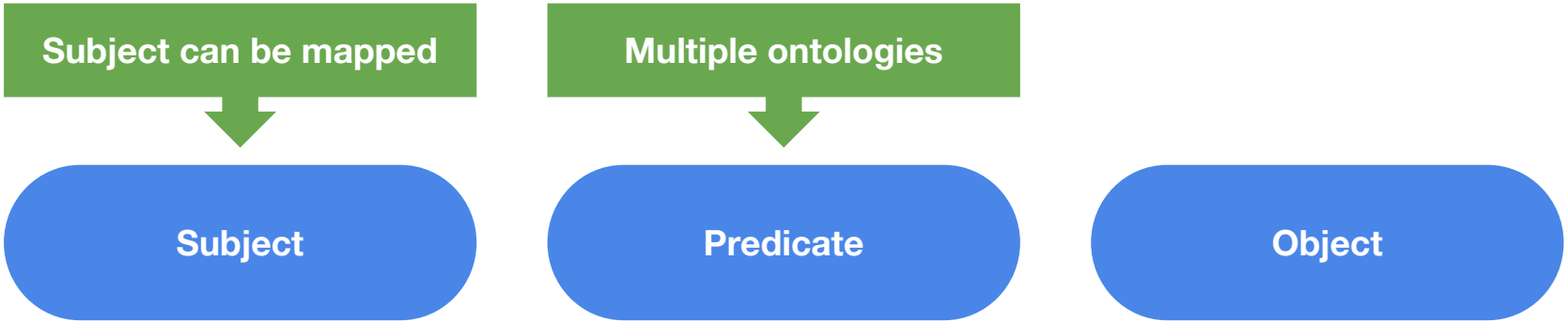
Mapping rules are decoupled from their implementation

✓ Declarative mapping rules



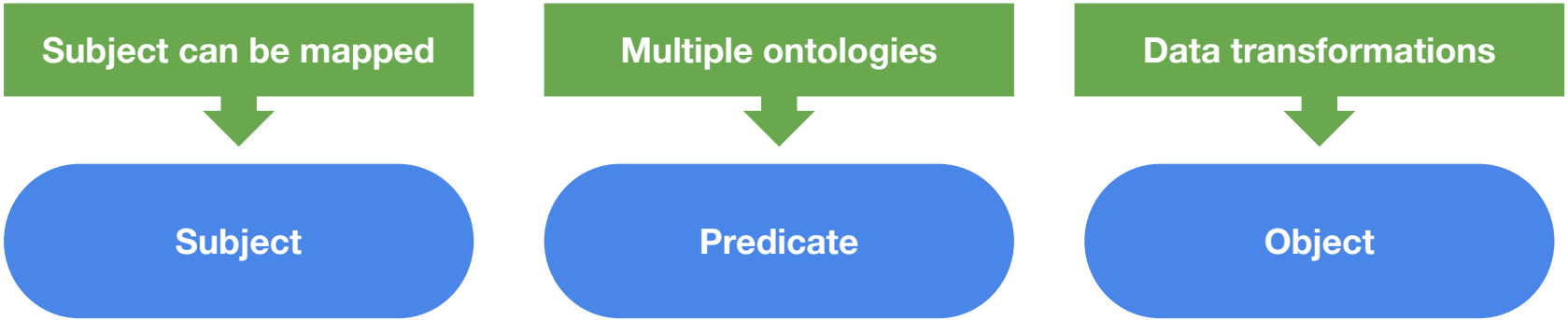
Subject can be defined

✓ Declarative mapping rules



Other ontologies can be defined

✓ Declarative mapping rules



Data transformations can be defined

Declarative data transformations for linked data generation : the case of DBpedia
Ben De Meester, Wouter Maroy, Anastasia Dimou, Ruben Verborgh and Erik Mannens

A sustainable framework is needed that provides

- ✓ Declarative mapping rules
- ✓ **Machine-interpretable format**
- ✓ Schema validation
- ✓ Usage of other ontologies

✓ Machine-interpretable format

✓ Machine-interpretable format

Querying

Schema validation

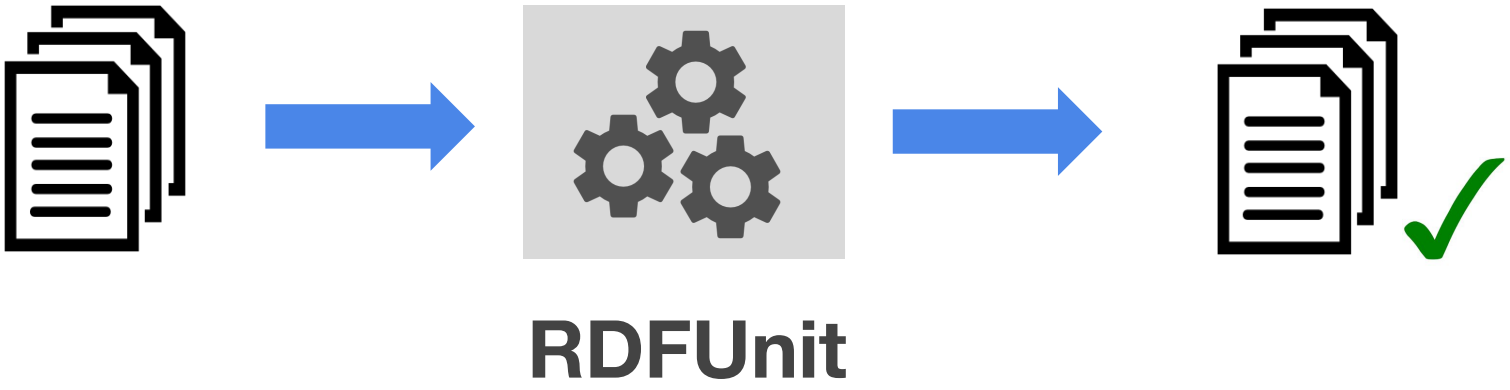
Generating

Automated processing with RDF tools

A sustainable framework is needed that provides

- ✓ Declarative mapping rules
- ✓ Machine-interpretable format
- ✓ **Schema validation**
- ✓ Usage of other ontologies

✓ Schema validation



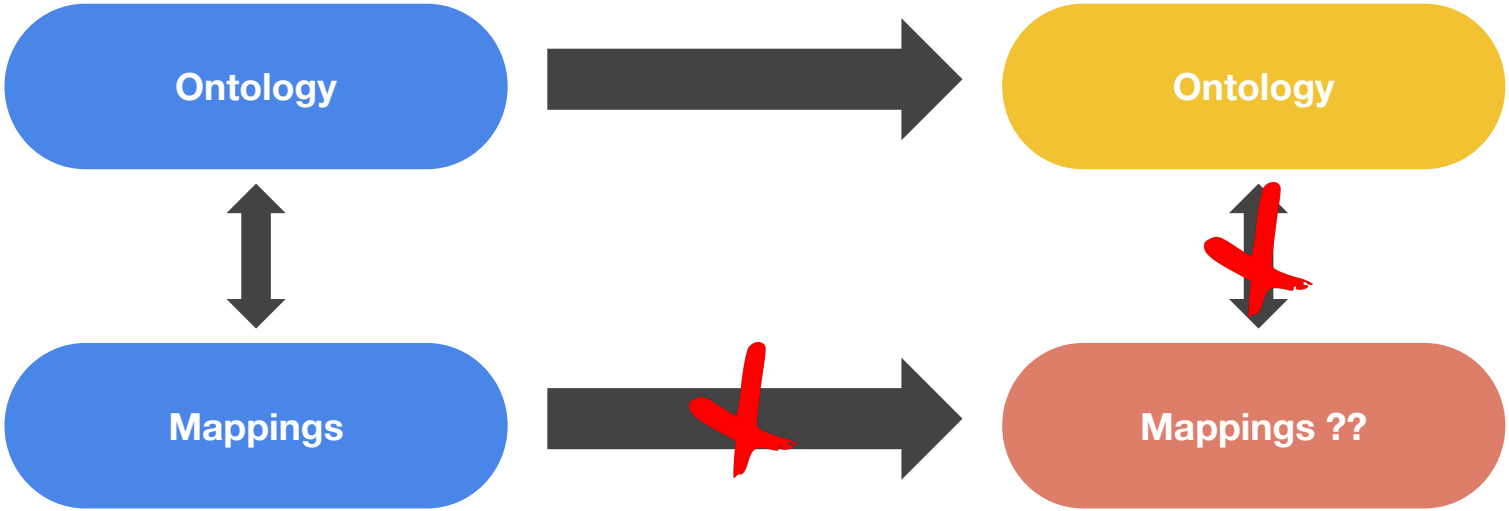
Test-driven Evaluation of Linked Data Quality
Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen

http://rdfunit.aksw.org/

A sustainable framework is needed that provides

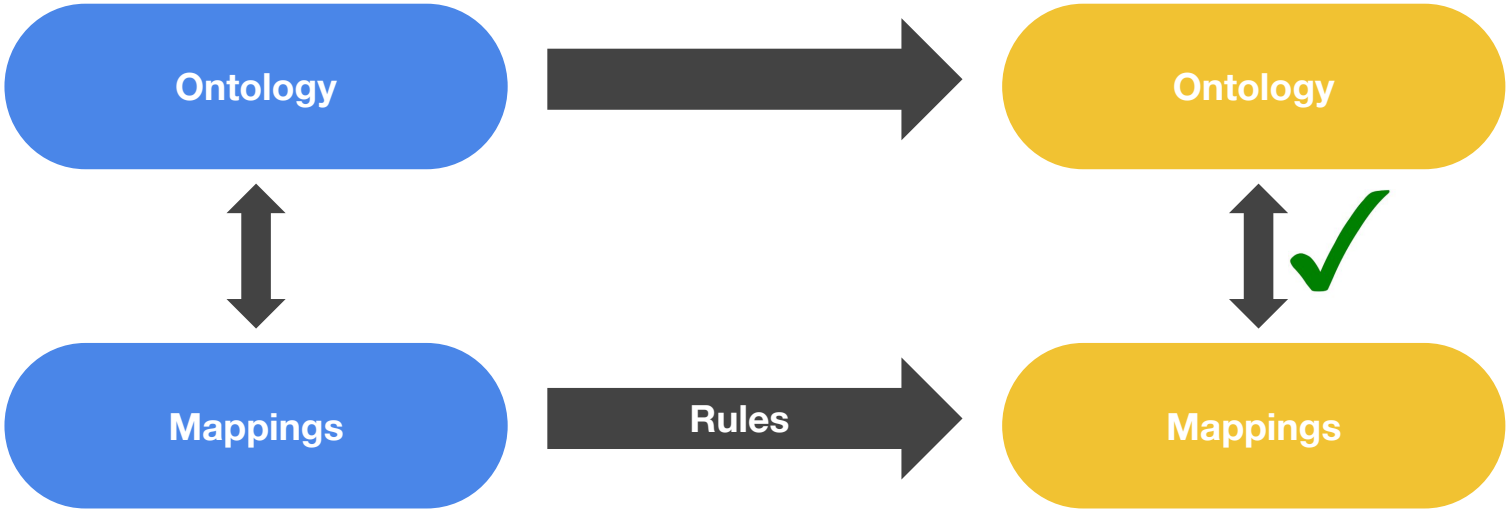
- ✓ Declarative mapping rules
- ✓ Machine-interpretable format
- ✓ Schema validation
- ✓ **Usage of other ontologies**

✓ Usage of other ontologies



Mappings don’t work for other ontologies

✓ Usage of other ontologies



Mappings can be changed automatically!

Sustainable Linked Data Generation

The case of DBpedia

Before

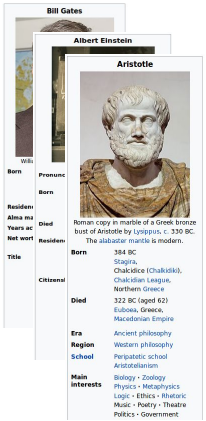
After

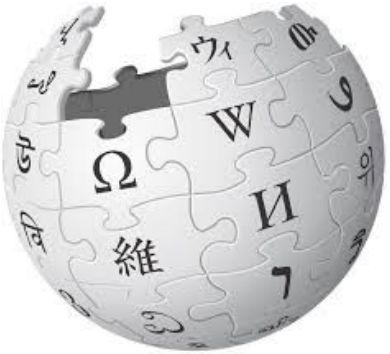
Progress

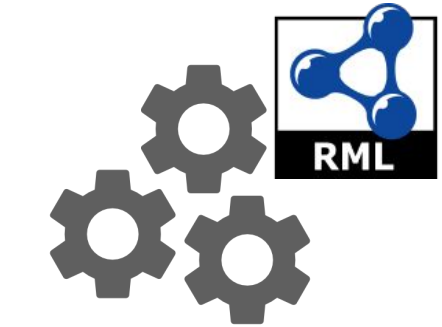
Switching to RML

Translating all DBpedia mappings to RML documents

Switching to RML







The Extraction Framework needed to process RML documents

Evaluation: coverage

Complete extraction was done on the English Wikipedia with ***98% coverage*** in comparison with the original dataset

Evaluation: performance

The framework offers ***more sustainable mapping rules*** at a performance cost of 35%

Next step is optimizing!

Evaluation: flexibility

Mapping rules in RDF can be (automatically) updated

Other datasets can be generated from Wikipedia because of the ontology independency

Evaluation: flexibility

We extracted a ***dataset of all persons*** on Wikipedia with the ***schema.org*** vocabulary

by only changing mapping rules

A sustainable framework that has

- ✓ Declarative mapping rules
- ✓ Machine-interpretable format
- ✓ Schema validation
- ✓ Usage of other ontologies

The future is bright!

Sustainable Linked Data Generation

The case of DBpedia

Wouter Maroy, Anastasia Dimou, Dimitris Kontokostas,
Ben De Meester, Ruben Verborgh, Jens Lehmann,
Erik Mannens, Sebastian Hellman

GHENT UNIVERSITY

IDLab
INTERNET & DATA LAB

imec

wouter.maroy@ugent.be

@wmaroy

Made possible by