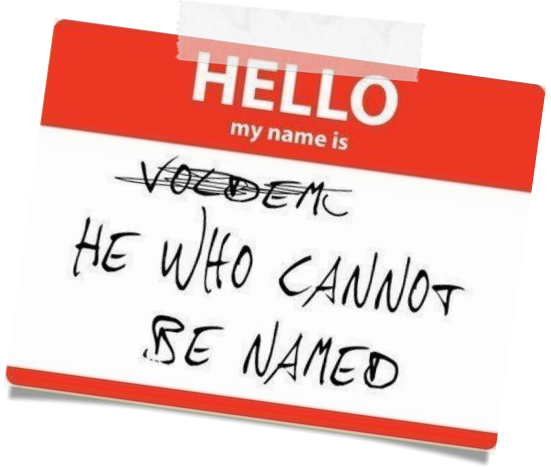
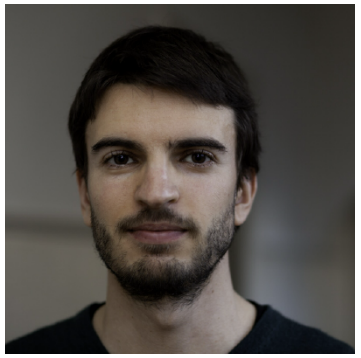


VoldemortKG*: Mapping `schema.org` and Web Entities to Linked Open Data

**The knowledge graph that everybody knows exists, but no one talks about*

Alberto Tonon, Victor Felder, Djellel E. Difallah, Philippe Cudré-Mauroux

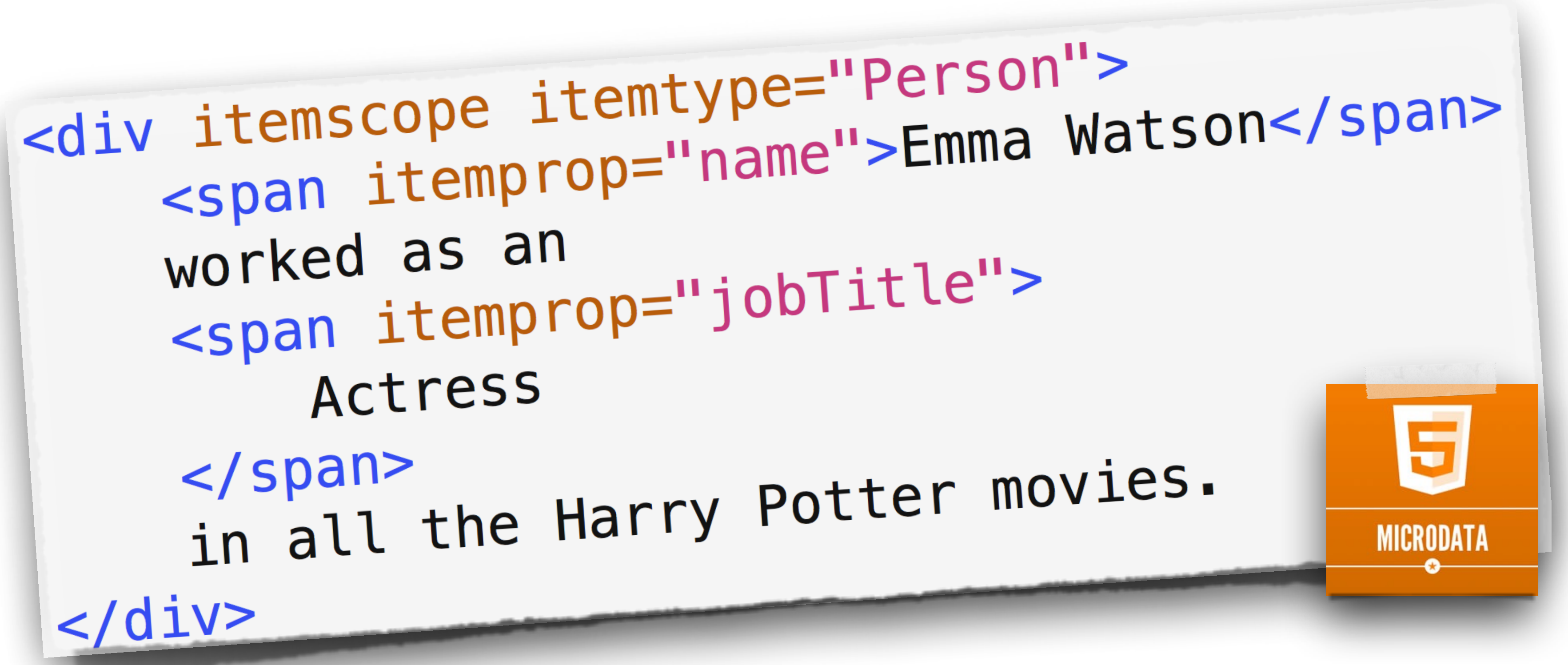


eXascale Infolab

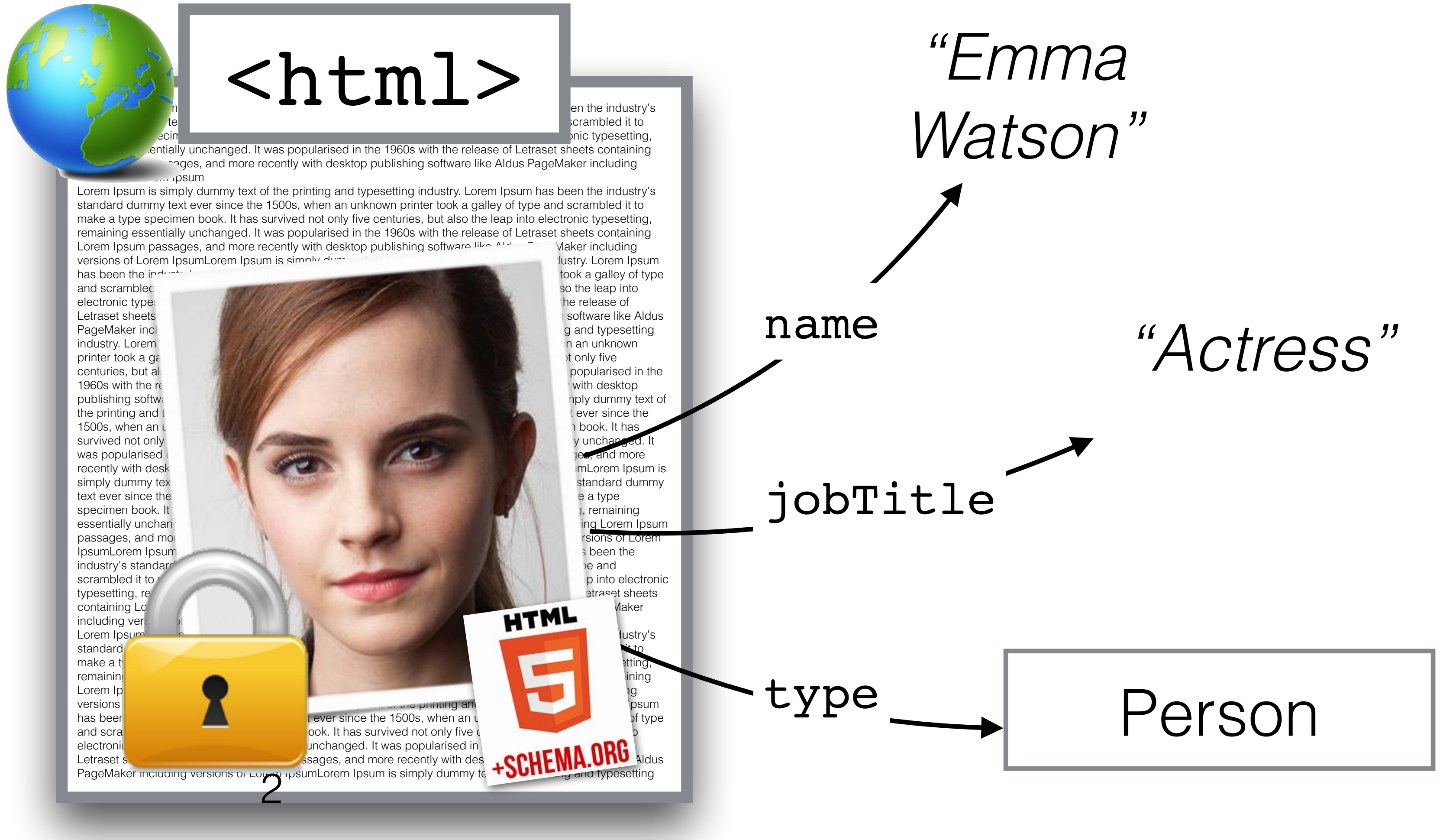
DAPLAB
Data Analysis and Processing Lab

UNI
FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Web Entities



```
<div itemscope itemtype="Person">
    <span itemprop="name">Emma Watson</span>
    worked as an
    <span itemprop="jobTitle">
        Actress
    </span>
    in all the Harry Potter movies.
</div>
```



“Emma Watson”

name

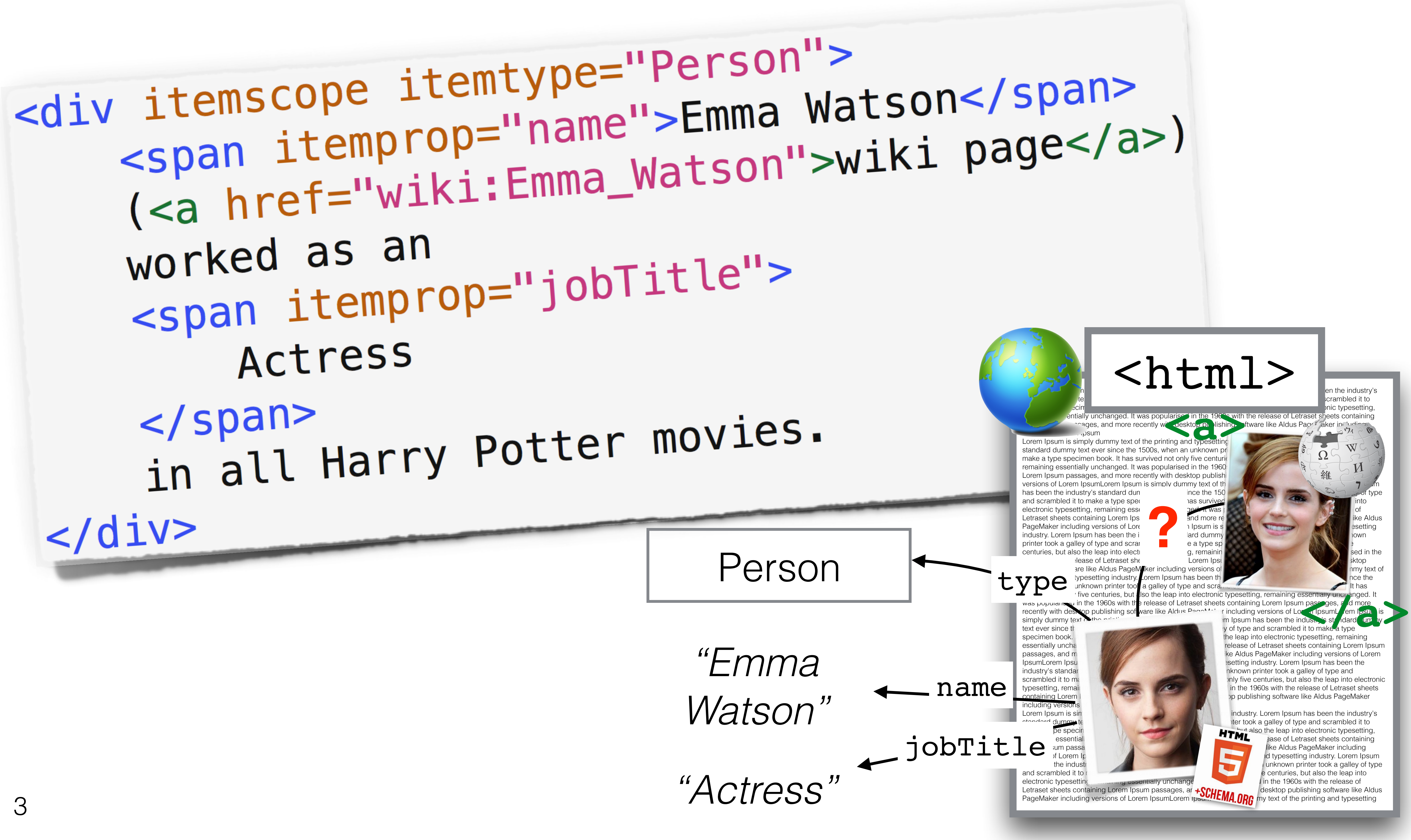
“Actress”

jobTitle

type

Person

Web Entity + Wikipedia Entities



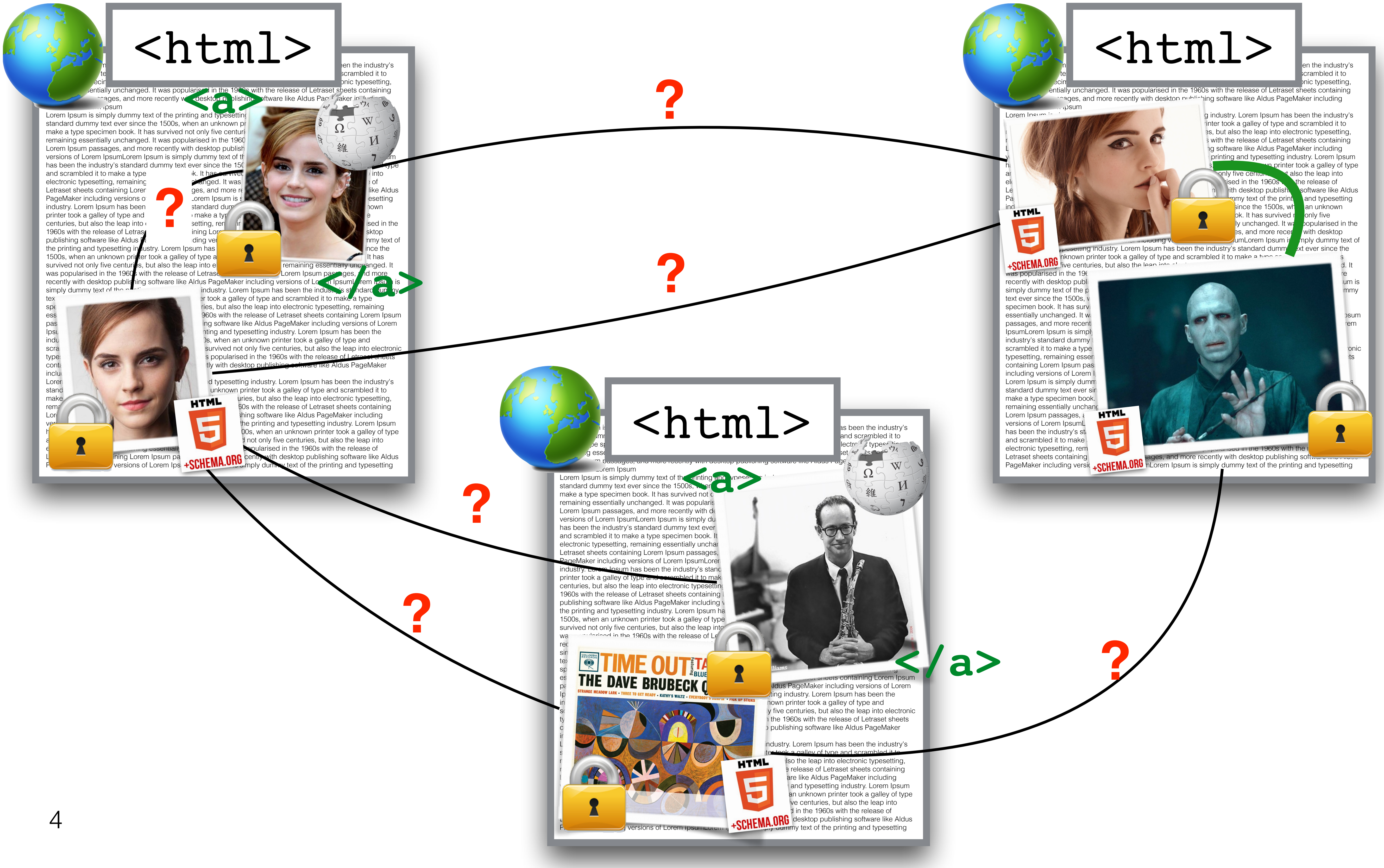
```
<div itemscope itemtype="Person">
    <span itemprop="name">Emma Watson</span>
    (<a href="wiki:Emma_Watson">wiki page</a>)
    worked as an
    <span itemprop="jobTitle">
         Actress
    </span>
    in all Harry Potter movies.
</div>
```

Person

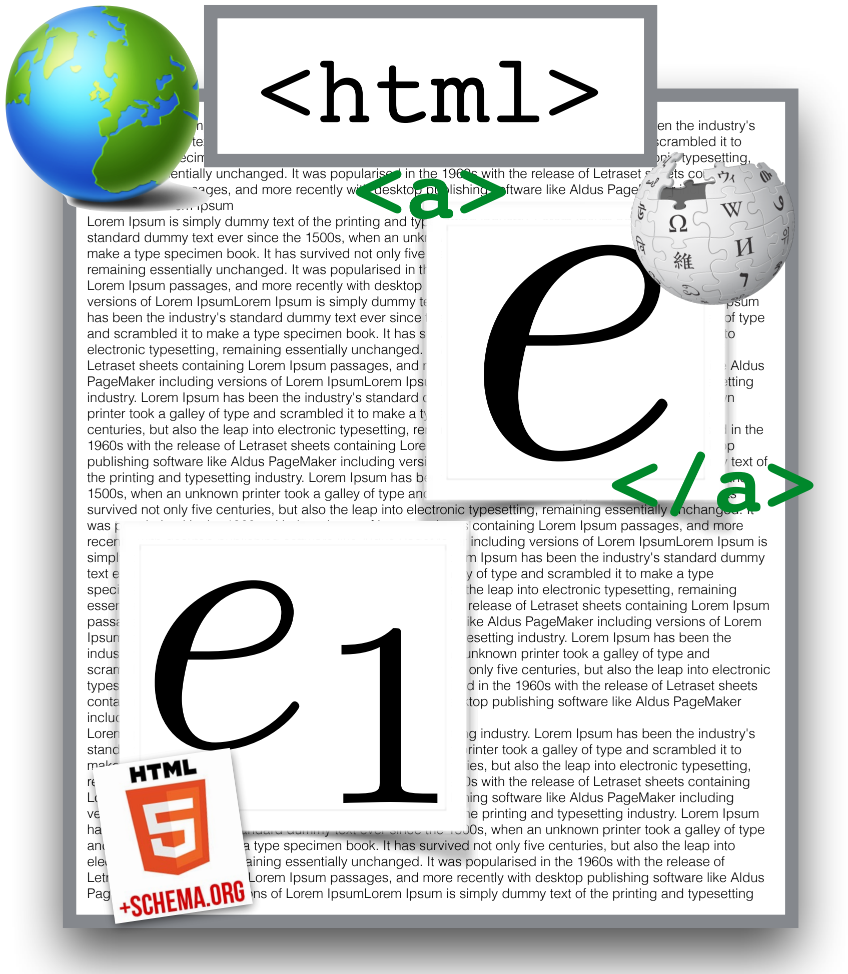
“Emma Watson”

“Actress”

VoldemortKG



Our Resource



x 7.8M

Our Resource

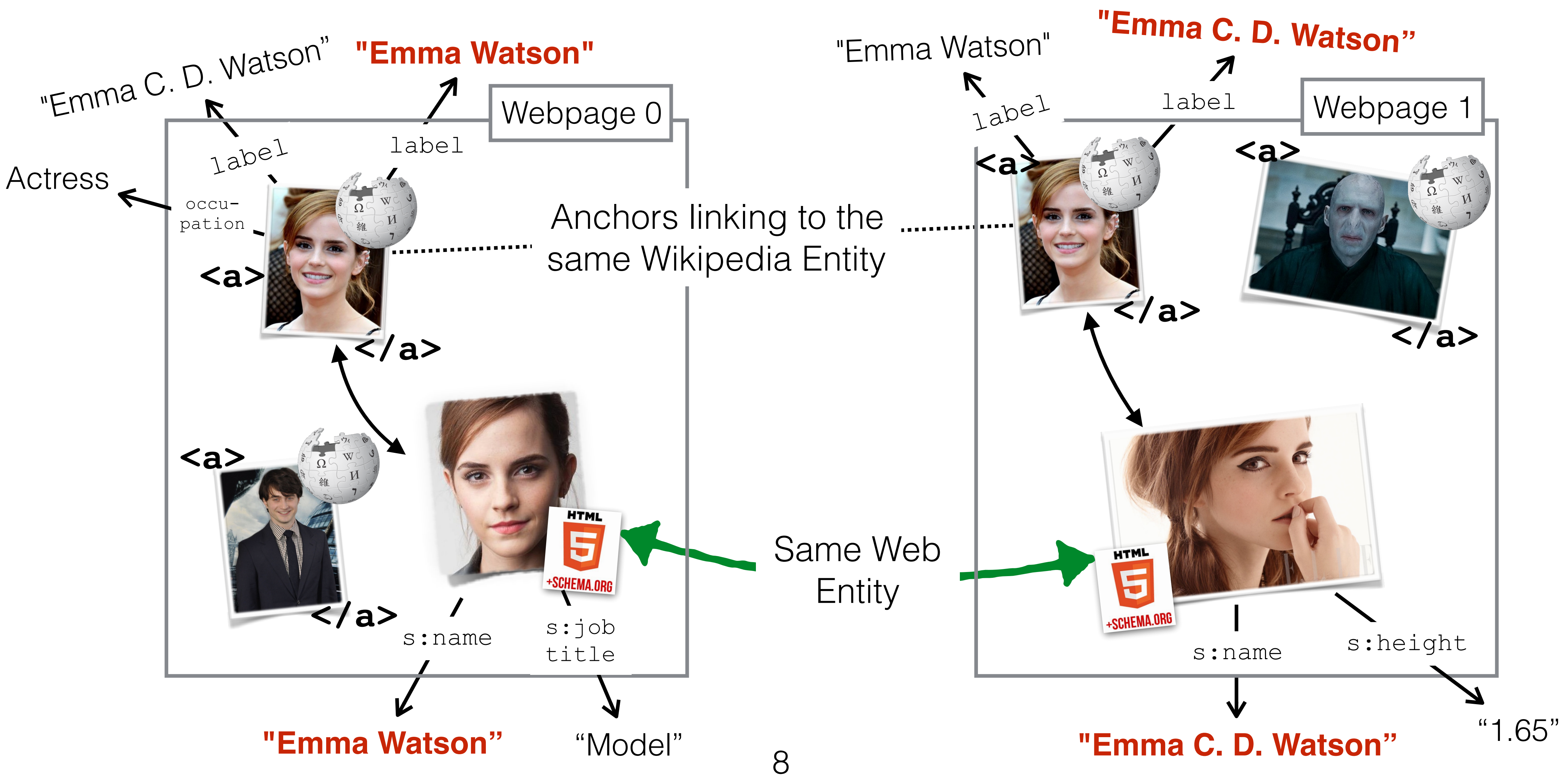
- permalink: **http://w3id.org/voldemortkg**
- 7,818,314 webpages containing both **Web Entities** and **links to Wikipedia entities**
- Structured data contained in the pages
- Wikipedia anchor's text
- Crawled from the Common Crawl (Nov. 2015) by a variant of the WDC Framework

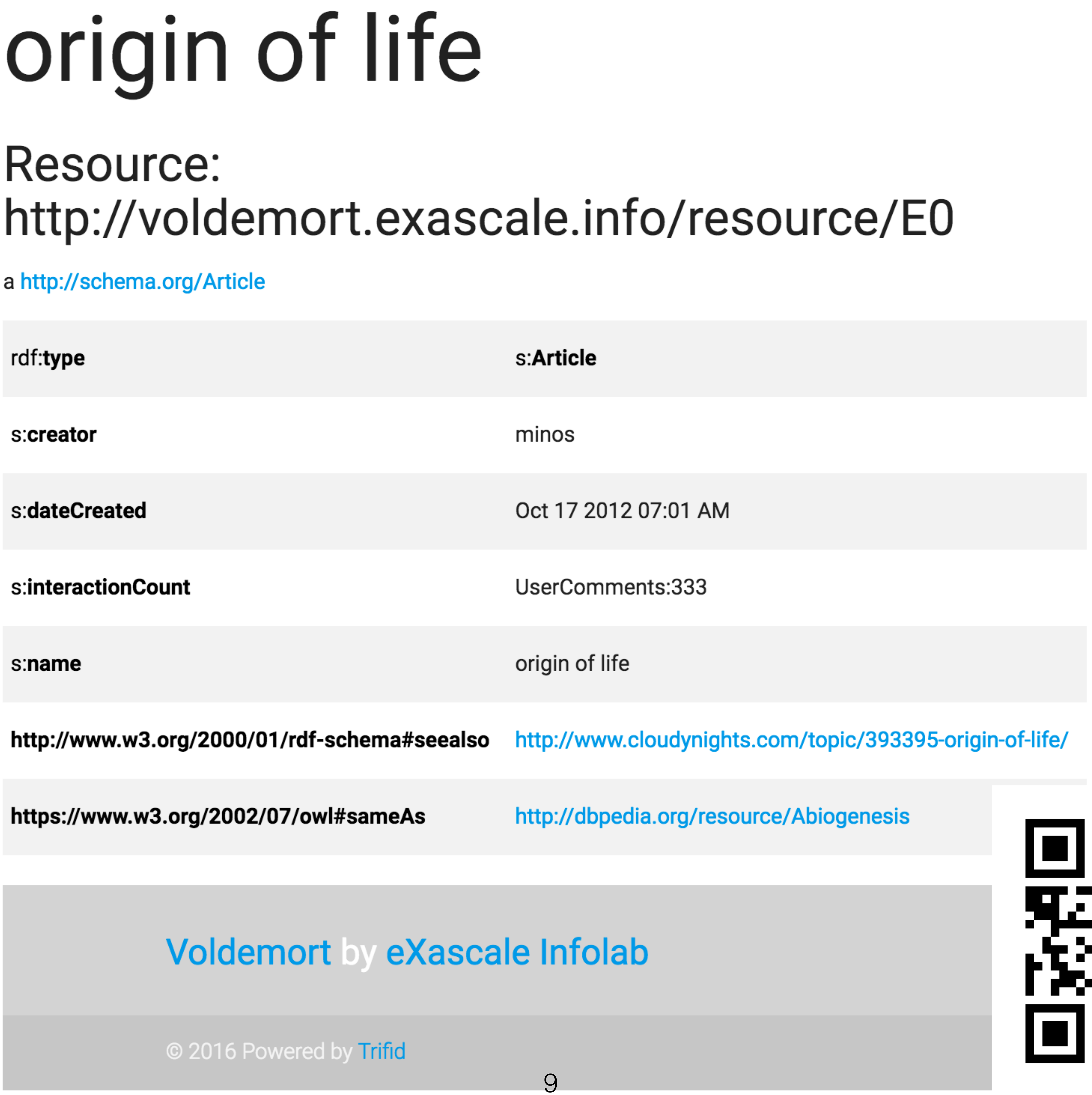
Dataset Stats

- 54% of the pages contains μF, 28% microdata, 18% RDFa (JSON-LD, future work)
- annotations use > 10 ontologies
  - in particular, 2.5M pages with annotations using > 2 ontologies
- Good playground for instance/ontology matchers

A First Prototype of VoldemortKG

- Preliminary version of VoldemortKG based on simple label matching.





origin of life

Resource: http://voldemort.exascale.info/resource/E0

a http://schema.org/Article

rdf:**type**	s:**Article**
s:**creator**	minos
s:**dateCreated**	Oct 17 2012 07:01 AM
s:**interactionCount**	UserComments:333
s:**name**	origin of life
http://www.w3.org/2000/01/rdf-schema#seealso	http://www.cloudynights.com/topic/393395-origin-of-life/
https://www.w3.org/2002/07/owl#sameAs	http://dbpedia.org/resource/Abiogenesis

Voldemort by eXascale Infolab

© 2016 Powered by Trifid

VoldemortKG: Few Stats

- 2.8M triples extracted from 202K webpages
- 134 different types
- Information on the same entity scattered across several pages
    - s:alternateName and owl:sameAs appear on avg. 367 and 11 pages per entity

Still Much Work to Do

- The instance matching strategy can be improved
  - e.g. E13140 in Voldemort is a Person b
    DBpedia is an Organisation
- https://en.wikipedia.org/wiki/Drowned_in
  scattered across all pages of the websit
- There is still much potential in the dataset:
  you’re welcome to explore and use it!

Research Challenges

- Match DBpedia entries to Web Entities (we provide a simple but reasonable baseline)
- Match Web Entities from different pages
- Ontology matching
- Data Fusion: how to merge different values of the same property?
- Knowledge Graph augmentation: verify if additional properties can be added to DBpedia entities (à la Knowledge Vault)

Future Work

- Data for evaluating DBpedia <—> Web Entity mappings

- Extraction of JSON-LD structured data

- Any idea?

Conclusions

x 7.8M

- Challenging for matching DBpedia entities to Web Entities (baseline provided)

- Schema matching across all schemata used to describe Web Entities and DBpedia

- Data fusion: different values for the same property?

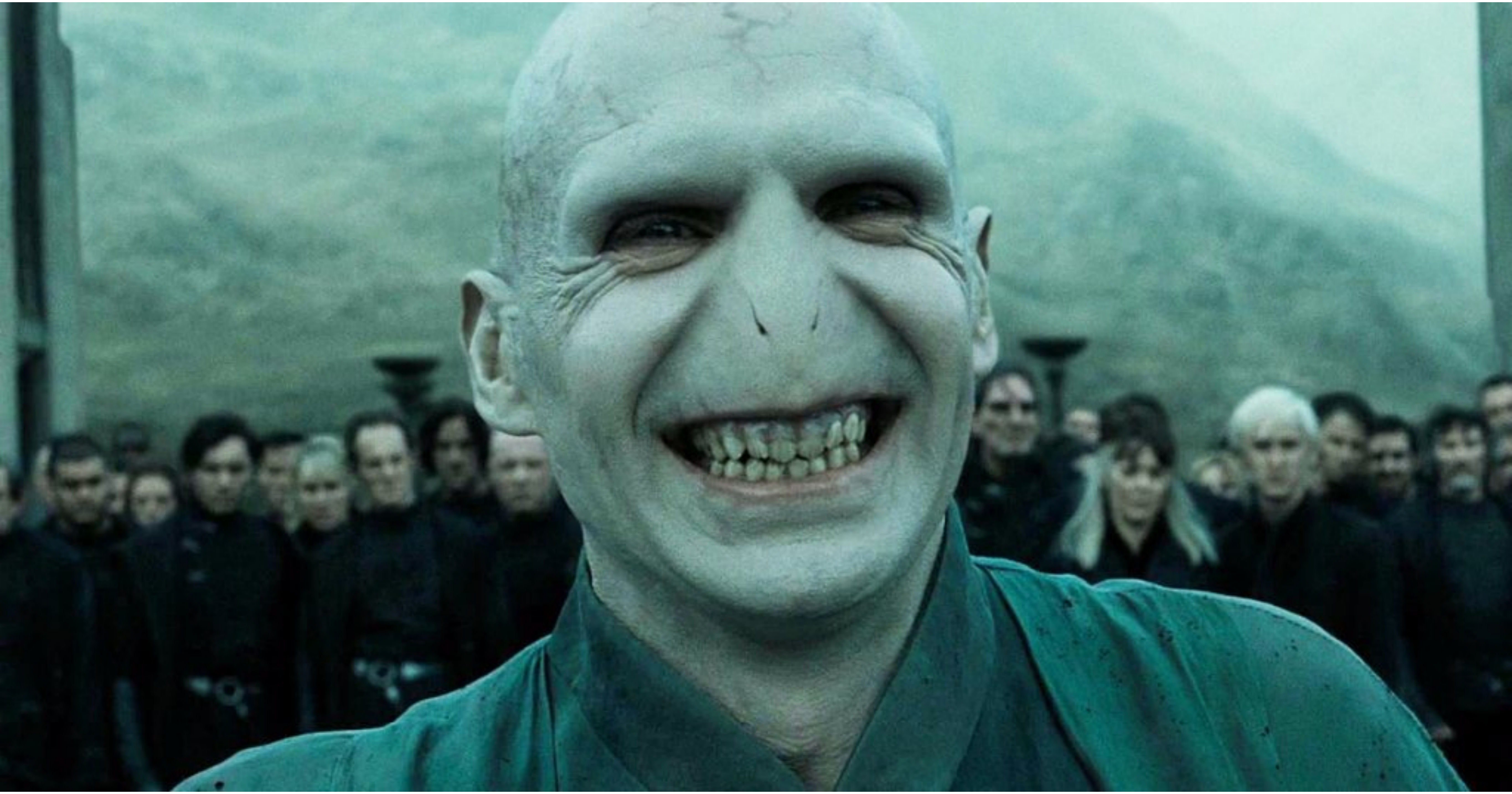
Acknowledgements

- Thanks, **Semantic Web Science Association** and **U.S. National Science Foundation**, for my Travel Grant!

- Dataset created with **DapLab**'s cluster (http://daplab.ch)

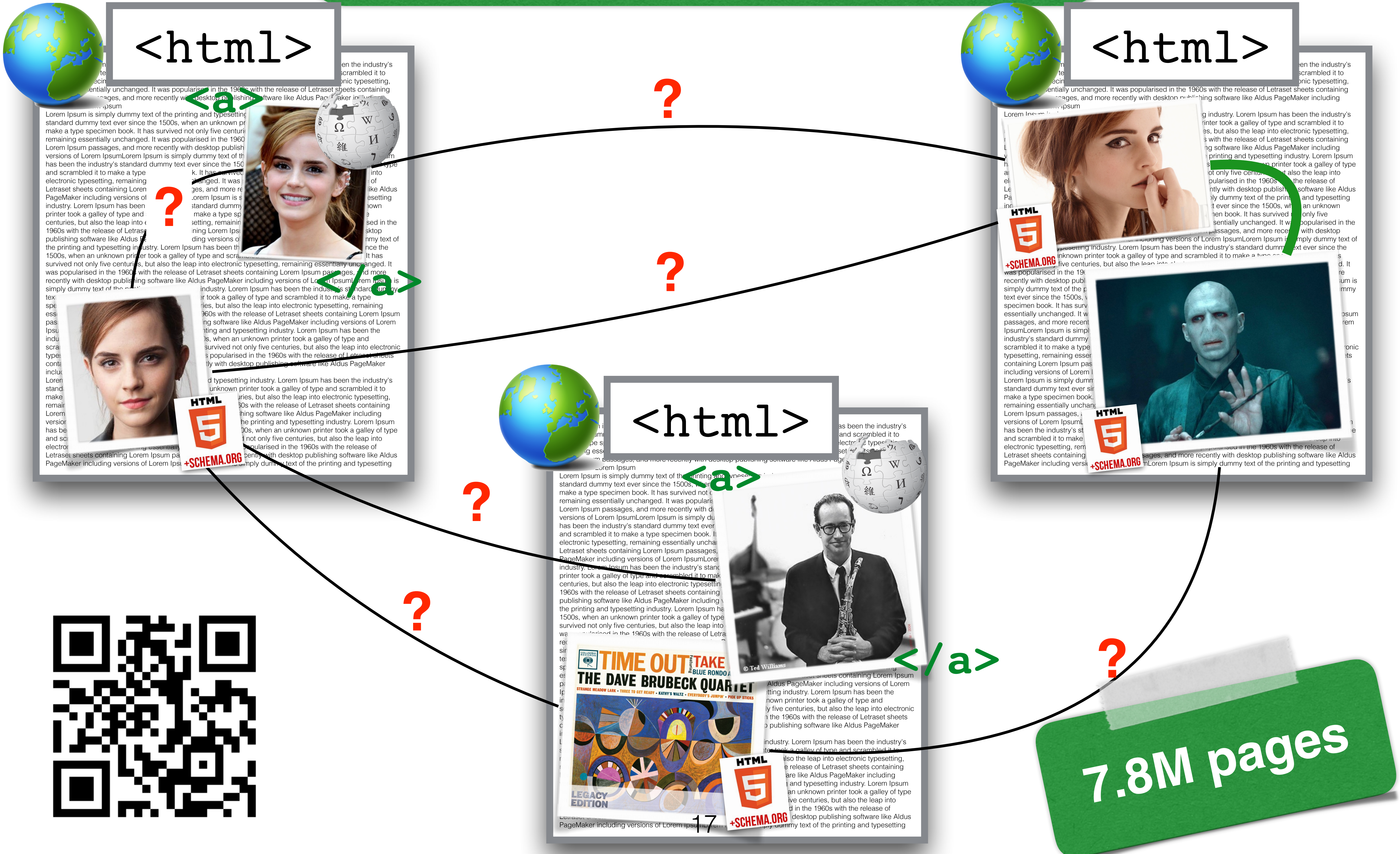
- VoldemortKG's triples are proudly displayed by **Trifid-LD** (https://github.com/zazukoians/trifid-ld)

Thanks for your Attention



VoldemortKG

http://w3id.org/voldemortkg



7.8M pages