

VoldemortKG*:

Mapping schema.org and Web Entities to Linked Open Data

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

<u>Alberto Tonon</u>, Victor Felder, Djellel E. Difallah, Philippe Cudré-Mauroux





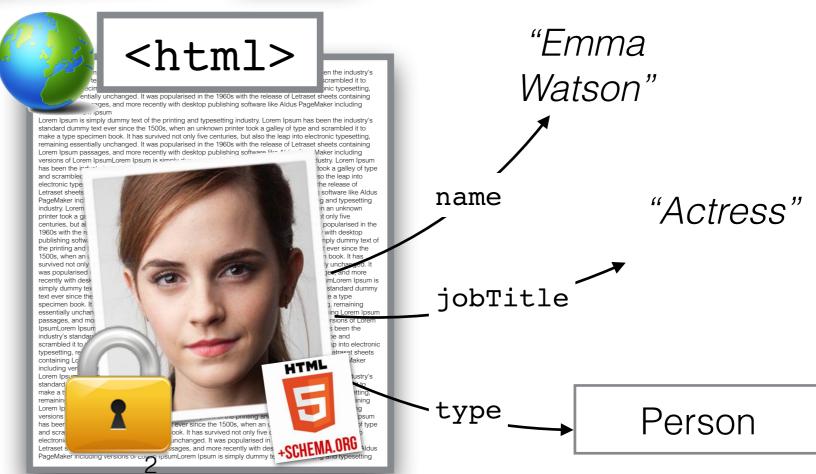








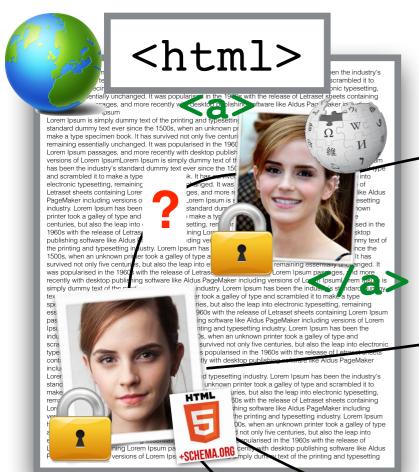
Web Entities

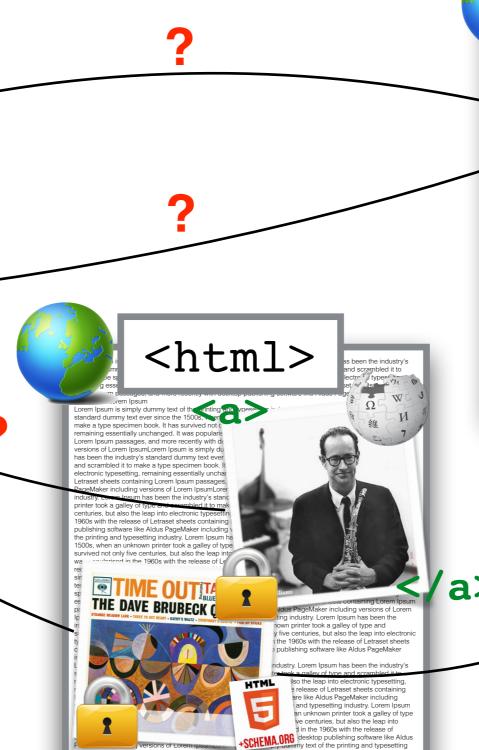


Web Entity + Wikipedia Entities

```
<div itemscope itemtype="Person">
    <span itemprop="name">Emma Watson
    (<a href="wiki:Emma_Watson">wiki page</a>)
    worked as an
    <span itemprop="jobTitle">
         Actress
                                              <html>
     </span>
     in all Harry Potter movies.
 </div>
                             Person
                             "Emma
                                       name
                            Watson"
                                     jobTitle sesser
                            "Actress"
3
```

VoldemortKG





<html> nic typesetting, a industry. Lorem Ipsum has been the industry's inter took a galley of type and scrambled it to es, but also the leap into electronic typesetting, with the release of Letraset sheets containing ng software like Aldus PageMaker including printing and typesetting industry. Lorem Ipsum oftware like Aldus and typesetting since the 1500s, w an unknown ok. It has survived ly unchanged. It w es, and more rece with desktop umLorem Ipsum i +SCHEMA.OR6 inknown printer took a galley of type and scrambled it to make a type was popularised in the 195. recently with desktop publ simply dummy text of the patent text ever since the 1500s, specimen book. It has surv ssentially unchanged. It v passages, and more recen Insuml orem Insum is simple scrambled it to make a type typesetting, remaining esser containing Lorem Ipsum pas including versions of Lorem Lorem Ipsum is simply dumn standard dummy text ever si

lummy text of the printing and typesetting

make a type specimen book.

remaining essentially unchan Lorem Ipsum passages, a

versions of Lorem IpsumL

has been the industry's stand scrambled it to make

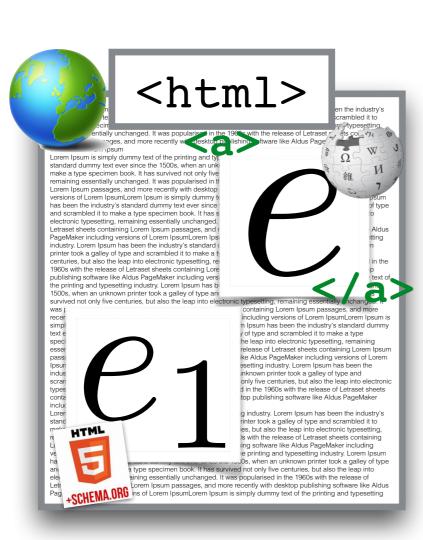
electronic typesetting, rem

HTML

PageMaker including versit +SCHEMA.ORG nLorem lpsum is simply

HTML

Our Resource





Our Resource



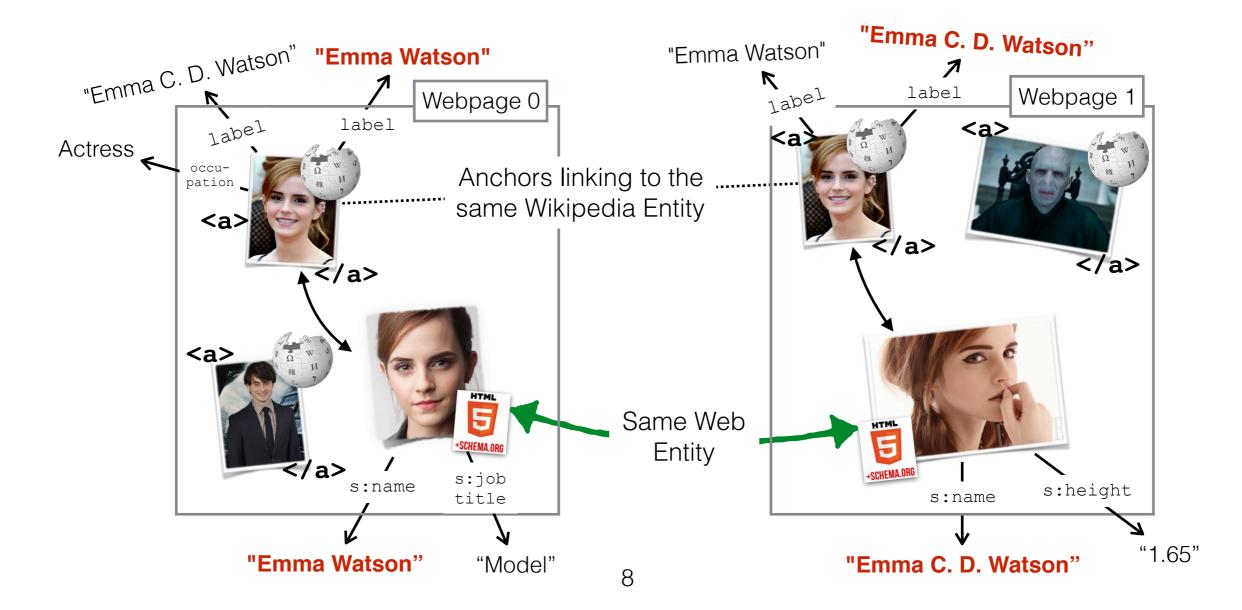
- permalink: http://w3id.org/voldemortkg
- 7,818,314 webpages containing both Web Entities and links to Wikipedia entities
- Structured data contained in the pages
- Wikipedia anchor's text
- Crawled from the Common Crawl (Nov. 2015) by a variant of the WDC Framework

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



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 but in DBpedia is an Organisation

 https://en.wikipedia.org/wiki/Drowned_ir scattered across all pages of the websit

 There is still much potential in the dataset: you're welcome to explore and use it!

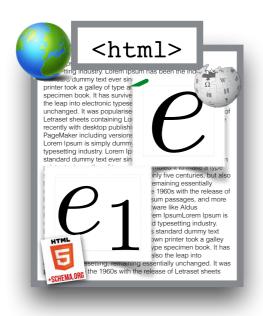
Research Challenges

- 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!

15

- Dataset created with **DapLab**'s cluster (http:// daplab.ch)
- VoldemortKG's triples are proudly displayed by Trifid-LD (https://github.com/zazukoians/trifid-ld)

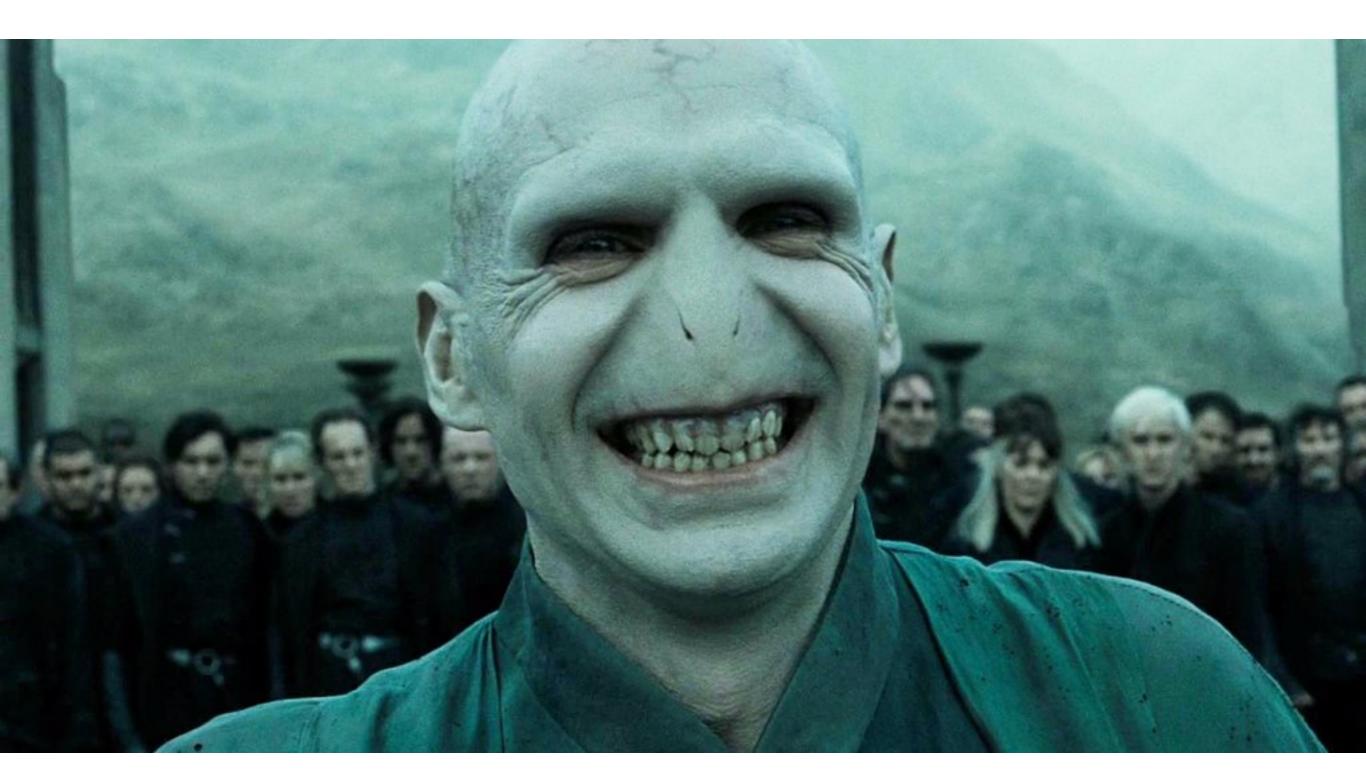








Thanks for your Attention



VoldemortKG

http://w3id.org/voldemortkg

<html>

em Ipsum is simply dummy text of the printing an standard dummy text ever since the 1500s, when an unknown p make a type specimen book. It has survived not only five centuri remaining essentially unchanged. It was popularised in the 1960 Lorem Ipsum passages, and more recently with desktop publist versions of Lorem IpsumLorem Ipsum is simply dummy text of the has been the industry's standard dummy text ever since the

and scrambled it to make a type electronic typesetting, remaining Letraset sheets containing Loren PageMaker including versions of industry. Lorem Ipsum has been printer took a galley of type and centuries, but also the leap into 6 1960s with the release of Letrase

Lorem versior has be

publishing software like Aldus the printing and typesetting 1500s, when an unknown p r took a galley of type and scr

survived not only five cent. recently with desktop publis ng software like Aldus PageMaker including versions of Lo

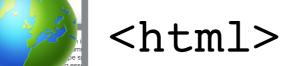
industry. Lorem Ipsum has been the in spe ess pas lpsu indu scral types conta include Loren stand make remain ries, but also the leap into electronic typesetting, remaining 960s with the release of Letraset sheets containing Lorem Ipsun ng software like Aldus PageMaker including versions of Loren nting and typesetting industry. Lorem Ipsum has been the

when an unknown printer took a galley of type and survived not only five centuries, but also the leap into spopularised in the 1960s with the release of Letrage

rambled it to

typesetting industry. Lorem Ipsum has been the industry's known printer took a galley of type and scrambled it to uries, but also the leap into electronic typesetting, HTML 30s with the release of Letraset sheets containing ning software like Aldus PageMaker including e printing and typesetting industry. Lorem Ipsum 00s, when an unknown printer took a galley of type

desktop publishing software like Aldus



Lorem I psum is simply dummy text of the centing and vicesses standard dummy text ever since the 1500s, representations of the centing and vicesses make a type specimen book. It has survived not a remaining essentially unchanned. It was possible. Lorem Ipsum passages, and more recently with diversions of Lorem IpsumLorem Ipsum is simply du has been the industry's standard dummy text eve and scrambled it to make a type specimen book. I electronic typesetting, remaining essentially uncha etraset sheets containing Lorem Ipsum passages

industry. Loren Insum has been printer took a galley of type and centuries, but also the leap into electronic typesetting 1960s with the release of Letraset sheets containing publishing software like Aldus PageMaker including the printing and typesetting industry. Lorem Ipsum ha 1500s, when an unknown printer took a galley of type survived not only five centuries, but also the leap into



ting industry. Lorem Ipsum has been the wn printer took a galley of type and the 1960s with the release of Letraset sheets publishing software like Aldus PageMake

nd scrambled it to

release of Letraset sheets containing are like Aldus PageMaker including and typesetting industry. Lorem Ipsum an unknown printer took a galley of type centuries, but also the leap into +SCHEMA.ORG desktop publishing software like Aldus <html>

ambled it to ing software like Aldus PageMaker including

> g industry. Lorem Ipsum has been the industry's rinter took a galley of type and scrambled it to es, but also the leap into electronic typesetting, with the release of Letraset sheets containing ng software like Aldus PageMaker including printing and typesetting industry. Lorem Ipsum nter took a galley of type ntly with desktop publis oftware like Aldus bly dummy text of the print t ever since the 1500s, v nen book. It has survived only five assages, and more red th desktop

ing versions of Lorem IpsumLorem Ipsum nly dummy text o nknown printer took a galley of type and scrambled it to make a five centuries, but also the lean in

recently with desktop publ simply dummy text of the r text ever since the 1500s, v specimen book. It has surv essentially unchanged. It w passages, and more recen IpsumLorem Ipsum is simp ndustry's standard dumn typesetting, remaining esse containing Lorem Ipsum pa ncluding versions of Lorer Lorem Ipsum is simply dumi standard dummy text ever : nake a type specimen book remaining essentially uncha HTML Lorem Ipsum passages, a versions of Lorem IpsumL has been the industry's st and scrambled it to ma electronic typesetting, rem Letraset sheets containing

HTML



es, and more recently with desktop publishing software like Aldus PageMaker including versit +SCHEMA.ORG nLorem Ipsum is simply

7.8M pages