universität innsbruck

Department of Computer Science

STI INNSBRUCK

How To Build A Knowledge Graph

Semantics Conference 2019, Tutorial

Elias Kärle & Umutcan Simsek

STI2, University of Innsbruck, September 9th, 2019

About Us

Elias Kärle
PhD Student
elias.kaerle@sti2.at
Twitter: @eliaska

Umutcan Simsek
PhD Student
umutcan.simsek@sti2.at
Twitter: @umutsims

Acknowledgement

This tutorial is based on the work being done in the MindLab, an industrial research project for building knowledge graphs to be consumed by conversational agents in domains like tourism.

An extensive version of the content of this tutorial can be found in our upcoming book “Knowledge Graphs in Use” (working title)

mindLab https://mindlab.ai

feratel window to the world

ONLIM

universität innsbruck

About the Tutorial

The tutorial aims to introduce our take on the knowledge graph lifecycle

Tutorial website: https://stiinnsbruck.github.io/kgt/

For industry practitioners:

An entry point to knowledge graphs. Several pointers for tackling different tasks on knowledge graph lifecycle

For academics:

A brief overview of the literature, introduction of some tools, especially in knowledge curation.

Relevant Literature:

https://mindlab.ai/en/publications/ - An extensive list of the literature on knowledge graphs and their applications with conversational agents

Outline and Agenda

13:30 – 15:00 Part 1

1) Introduction
2) Knowledge Creation
3) Knowledge Hosting

15:00 – 15:30 Coffee Break

15:30 – 17:30 Part 2

4) Knowledge Curation
5) Knowledge Deployment & Discussion

1. What is a Knowledge Graph?

TL;DR:

very large semantic nets that integrate various and heterogeneous information sources to represent knowledge about certain domains of discourse.

Term coined by Google in 2012.

1. What is a Knowledge Graph?

- A graph is a mathematical structure in which some pairs in a set of objects are somehow related. See https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)

- Knowledge: knowledge level vs symbol level

  We ascribe knowledge to the actions of an agent.

  At the symbol level resides implementations like graph-databases.

- An agent would interpret a knowledge graph to make rational decisions to take actions to reach its goals

1. What is a Knowledge Graph?

But wait, aren't knowledge bases already doing this?

There are certain characteristic differences between KBs and KGs:

- KBs have a strict separation of TBox and Abox
- KGs do not have a big TBox, but have a very large ABox. There is not much to reason.
- No strict schema: Good for integrating heterogeneous sources, not so much in terms of data quality.

1. Knowledge Graphs in the Wild

Name	Instances	Facts	Types	Relations
DBpedia (English)	4,806,150	176,043,129	735	2,813
YAGO	4,595,906	25,946,870	488,469	77
Freebase	49,947,845	3,041,722,635	26,507	37,781
Wikidata	15,602,060	65,993,797	23,157	1,673
NELL	2,006,896	432,845	285	425
OpenCyc	118,499	2,413,894	45,153	18,526
Google´s Knowledge Graph	570,000,000	18,000,000,000	1,500	35,000
Google´s Knowledge Vault	45,000,000	271,000,000	1,100	4,469
Yahoo! Knowledge Graph	3,443,743	1,391,054,990	250	800

Numerical Overview of some Knowledge Graphs, taken from [Paulheim, 2017]

1. What is a Knowledge Graph?

- Knowledge graphs are not the first attempt for making data useful for automated agents by integrating and enriching data from heterogeneous sources.
- Building knowledge graphs are expensive. Scaling them is challenging.
- A knowledge graph may cost 0,1 - 6 USD per fact [Paulheim, 2018]

1. What is a Knowledge Graph?

Two main entry points for improving the quality of knowledge graphs:

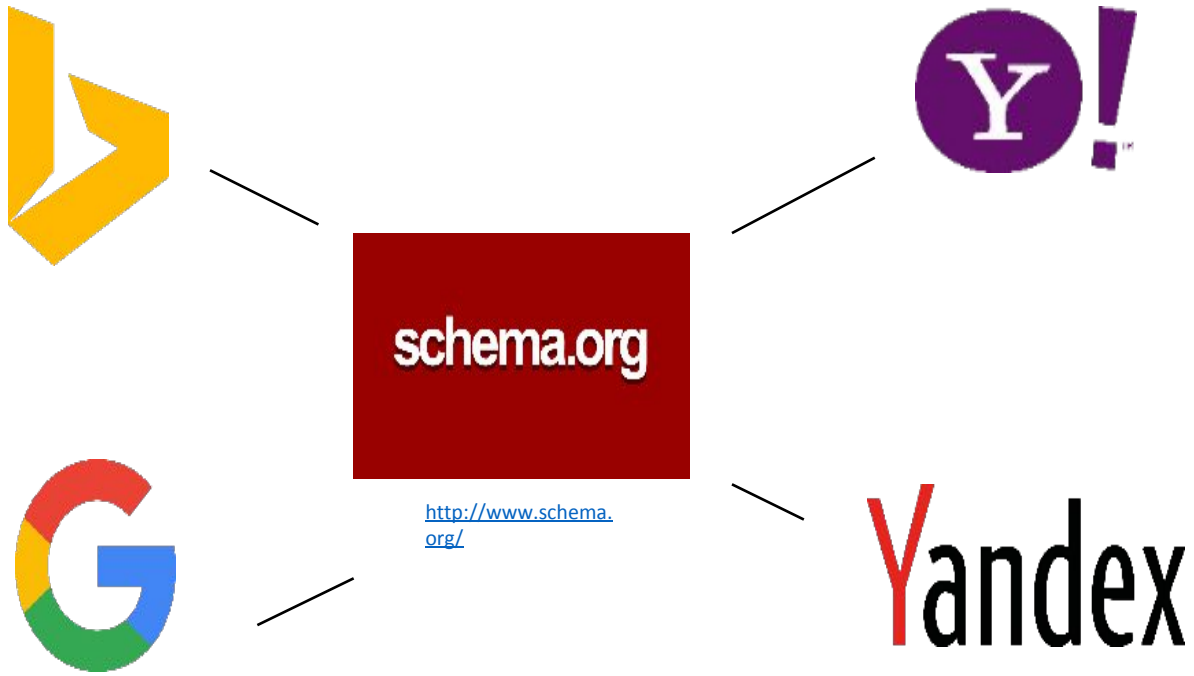
Fixing TBox

- We accept schema.org (and its extensions) as golden standard. No problem here.

Fixing ABox

- This is where knowledge curation comes in.

1. Schema.org



Created, recommended and maintained by four major search engines providers

1. Schema.org

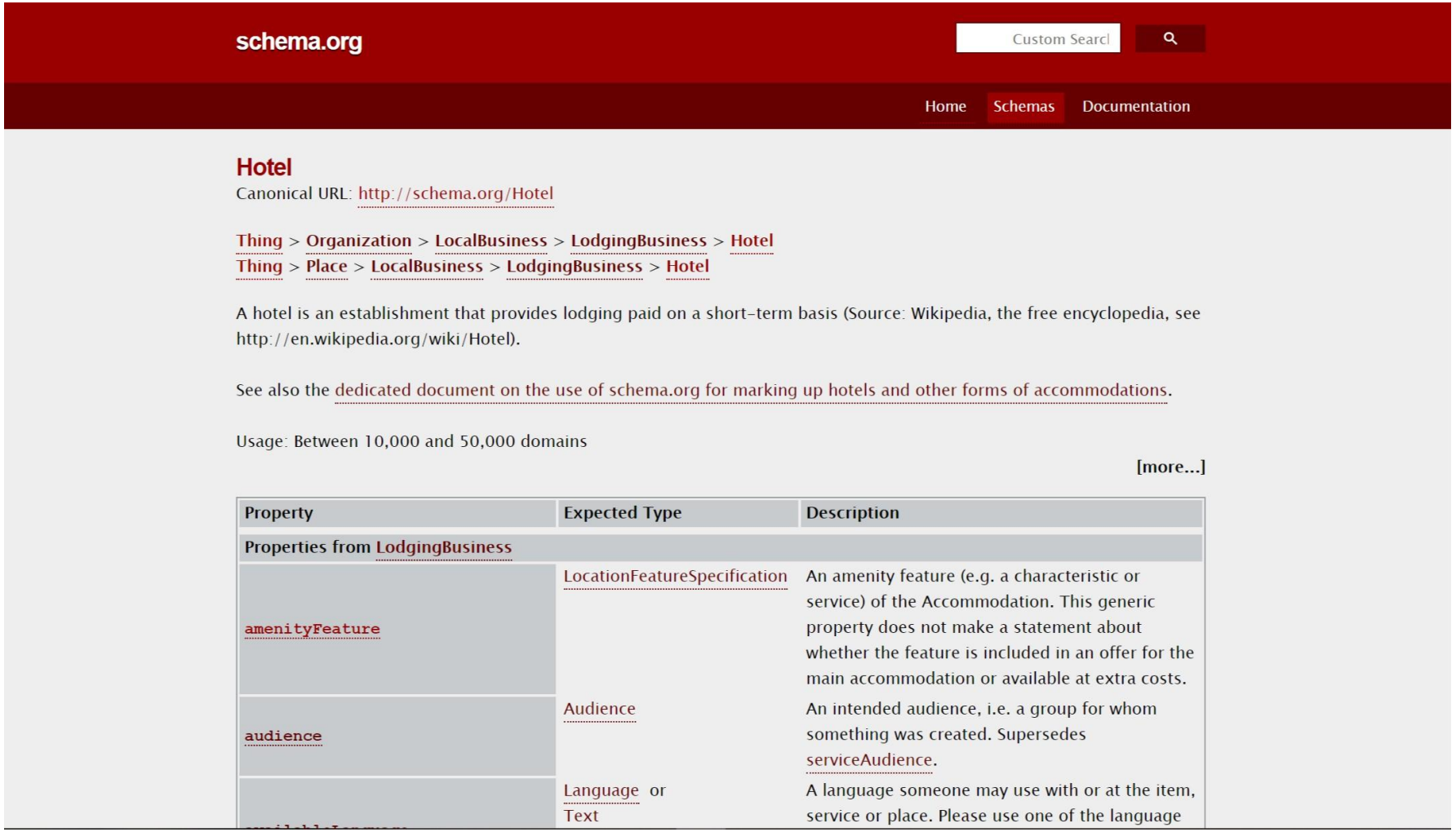
```
<div vocab="http://schema.org/" typeof="Movie">
  <h1 property="name">Avatar</h1>
  <div property="director" typeof="Person">
  Director: <span property="name">James Cameron</span>
(born <time property="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span property="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" property="trailer">Trailer</a>
</div>
```

```
<div itemscope itemtype="http://schema.org/Movie">
  <h1 itemprop="name">Avatar</h1>
  <div itemprop="director" itemscope itemtype="http://schema.org/Person">
  Director: <span itemprop="name">James Cameron</span>
(born <time itemprop="birthDate" datetime="1954-08-16">August 16, 1954</time>)
  </div>
  <span itemprop="genre">Science fiction</span>
  <a href="../movies/avatar-theatrical-trailer.html" itemprop="trailer">Trailer</a>
</div>
```

```
<script type="application/ld+json">
{
  "@context": "http://schema.org/",
  "@type": "Movie",
  "name": "Avatar",
  "director":
    {
       "@type": "Person",
       "name": "James Cameron",
       "birthDate": "1954-08-16"
    },
  "genre": "Science fiction",
  "trailer": "../movies/avatar-theatrical-trailer.html"
}
</script>
```

- Embedded in HTML source
  - Microdata
  - RDFa
  - JSON-LD

1. Schema.org



schema.org

Custom Searcl

Home Schemas Documentation

Hotel

Canonical URL: http://schema.org/Hotel

Thing > Organization > LocalBusiness > LodgingBusiness > Hotel
Thing > Place > LocalBusiness > LodgingBusiness > Hotel

A hotel is an establishment that provides lodging paid on a short-term basis (Source: Wikipedia, the free encyclopedia, see http://en.wikipedia.org/wiki/Hotel).

See also the dedicated document on the use of schema.org for marking up hotels and other forms of accommodations.

Usage: Between 10,000 and 50,000 domains

[more...]

Property	Expected Type	Description
Properties from LodgingBusiness		
amenityFeature	LocationFeatureSpecification	An amenity feature (e.g. a characteristic or service) of the Accommodation. This generic property does not make a statement about whether the feature is included in an offer for the main accommodation or available at extra costs.
audience	Audience	An intended audience, i.e. a group for whom something was created. Supersedes serviceAudience.
	Language or Text	A language someone may use with or at the item, service or place. Please use one of the language

1. Schema.org

```
{
  "@context": "http://schema.org",
  "@type": "LocalBusiness",
  "name": "Imbiss-Stand \"Wurscht & Durscht\"",
  "geo": {
    "@type": "GeoCoordinates",
    "latitude": "47.3006092921797",
    "longitude": "10.9136698539673"
  },
  "address": {
    "@type": "PostalAddress",
    "streetAddress": "Unterer Mooswaldweg 2",
    "addressLocality": "Obsteig",
    "postalCode": "6416",
    "addressCountry": "AT",
    "telephone": "+43 664 / 26 32 319",
    "faxNumber": "",
    "email": "info@wudu-imbiss.at",
    "url": "www.wudu-imbiss.at"
  },
  "description": "Der Imbisstand direkt an der Bundesstraße B 189 in Obsteig verwöhnt die Gäste mit qualitativ hochwertigen \"Würschtln\" (Wurst) aller Art.",

}
```

1. Schema.org

Event

Thing > Event

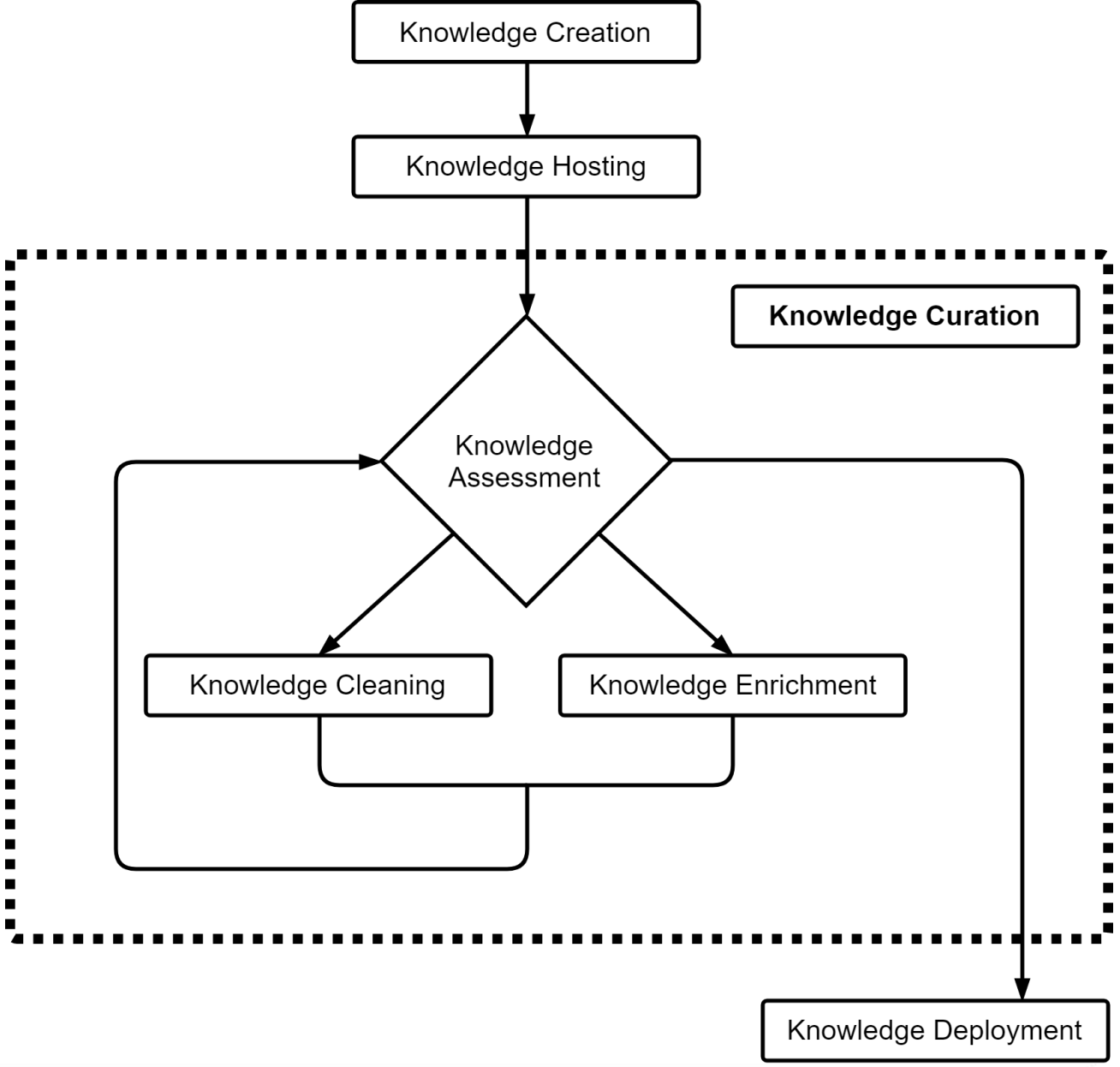
An event happening at a certain time and location, such as a concert, lecture, or festival. Ticketing information may be added via the offers property. Repeated events may be structured as separate Event objects.

[more...]

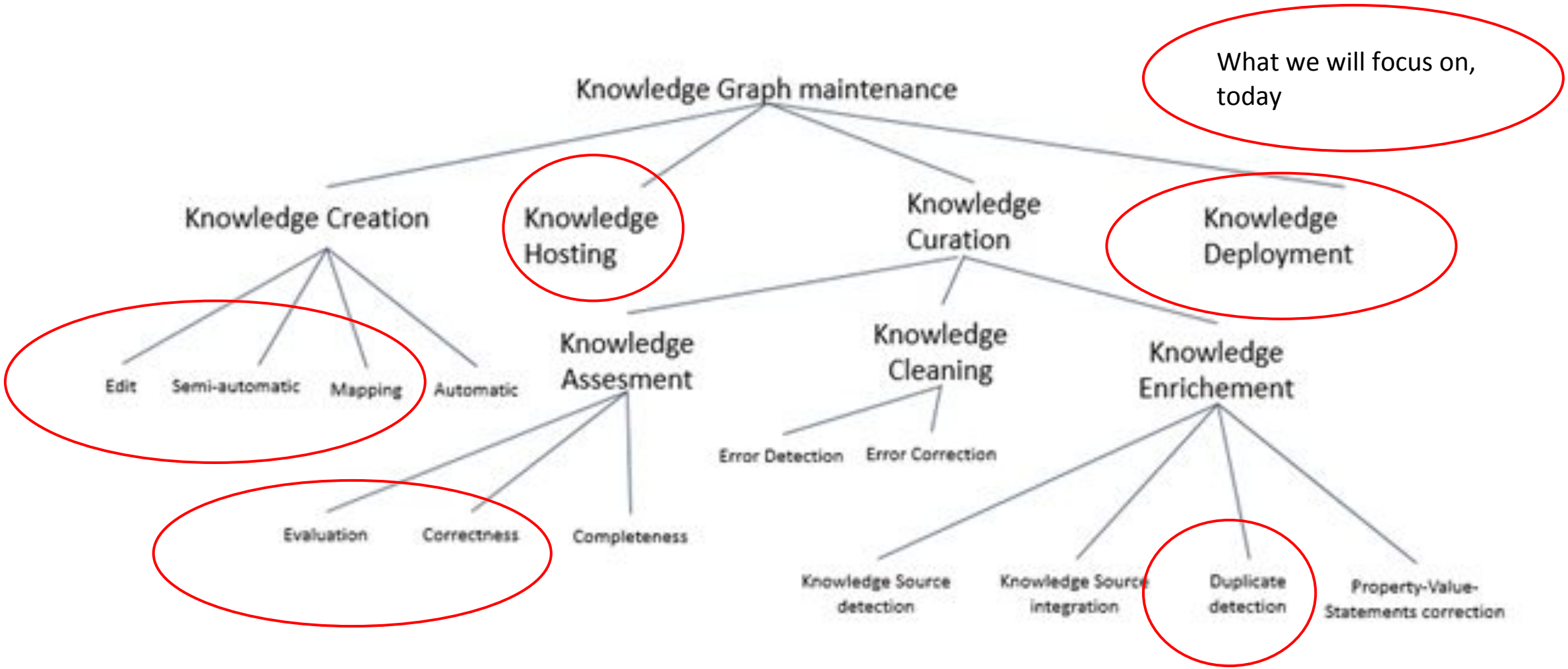
Property	Expected Type	Description
Properties from Event		
about	Thing	The subject matter of the content. Inverse property: subjectOf.
actor	Person	An actor, e.g. in tv, radio, movie, video games etc., or in an event. Actors can be associated with individual items or with a series, episode, clip. Supersedes actors.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item. The overall rating, based on a collection of reviews or ratings, of the item.
attendee	Organization or Person	A person or organization attending the event. Supersedes attendees.

- schema.org is organized as a hierarchy of types and properties
- the data model is derived from RDFS
- domainIncludes, rangeIncludes instead of rdfs:domain, rdfs:range
- The ranges are disjunctive
- Types are arranged in multiple inheritance hierarchy

Knowledge Graph Building Process Model

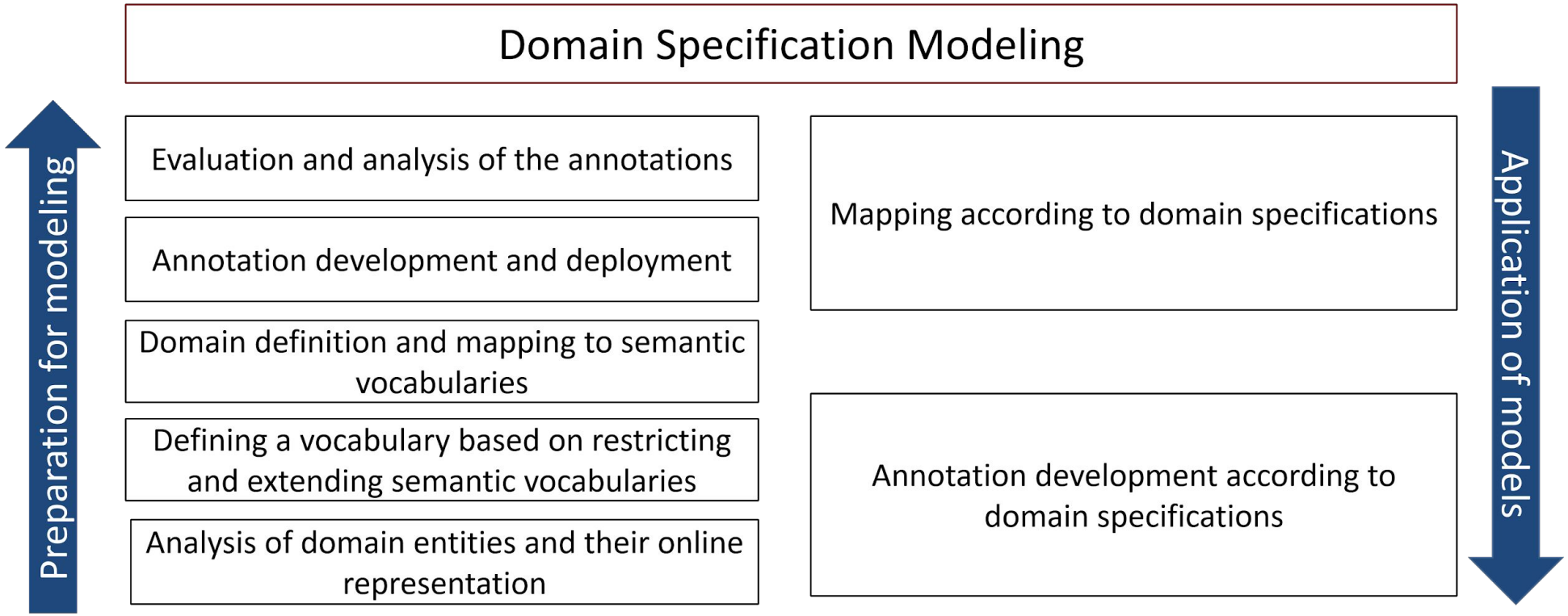


1. Knowledge Graph Building: Task Model

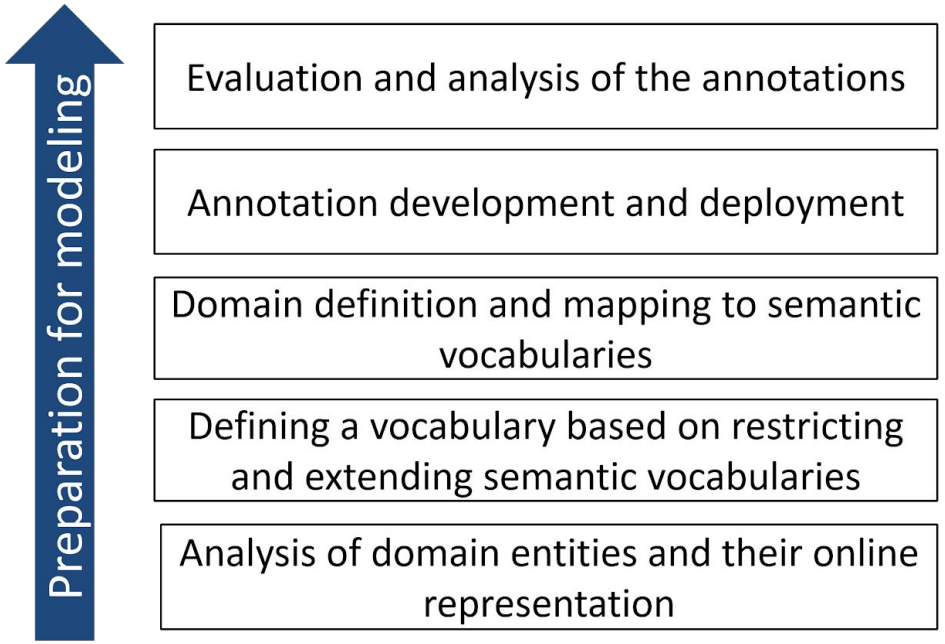


2. Knowledge Creation - Methodology

a.k.a Knowledge Acquisition: “...describes the process of extracting information from different sources, structuring it, and managing established knowledge” - Schreiber et al.



2. Knowledge Creation - Methodology



1) **bottom-up**: describes a first annotation process
   a) analysis of a domain's entities and their (online) representation
   b) defining a vocabulary (potentially by restricting and/or extending an already existing voc.)
   c) "domain definition", mapping to semantic vocabularies
   d) annotation
   e) evaluation and analysis of annotations

2. Knowledge Creation - Methodology

Domain Specification Modeling

2) **domain specification modeling:** reflects the results of step 1)

formalize the findings of step 1) in a

- unified
- exchangable
- machine-read and understandable way

⇒ **Domain Specifications**

2. Knowledge Creation - Domain Specifications

"A **domain specification** is a document, defining **syntactic** and **semantic** constraints for schema.org* annotations regarding a **specific domain** or **application**" [Holzknecht, 2019]

"[A] **domain specification** [is] a(n) (extended) **subset of properties** and **restrict[s]** the **range** of those properties to a **subset of subclasses** of the range defined by schema.org*" [Simsek et al., 2017]

*or any other ontology

(extended: because we not only use schema.org, but also extensions of it if necessary)

Domain Specification are:

- annotation patterns
- a best practice for annotation users
- a "crutch" for annotation laymen
- a means of sharing a common understanding about a domain's annotation application

2. Knowledge Creation - Domain Specifications

- DSs are serialized in SHACL

Hotel

A hotel is an establishment that provides lodging paid on a short-term basis (Source: Wikipedia, the free encyclopedia, see http://en.wikipedia.org/wiki/Hotel).

See also the dedicated document on the use of schema.org for marking up hotels and other forms of accommodations.

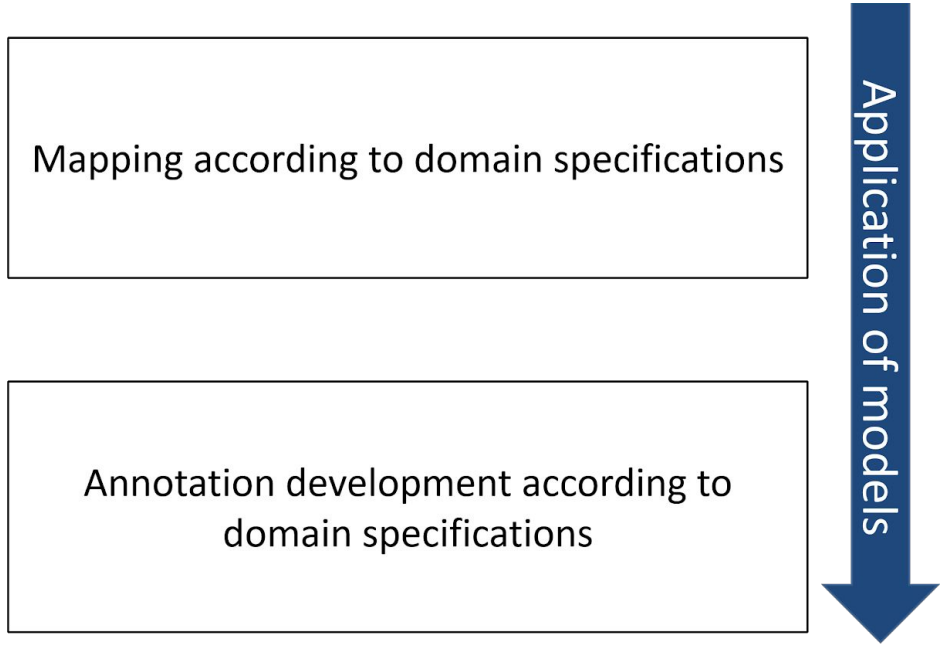
External link External link to schema.org

Property	Expected Type	Description	Cardinality
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.	1
availableLanguage	Text	A language someone may use with the item. Please use one of the language codes from the IETF BCP 47 standard. See also inLanguage	0..N
checkinTime	DateTime	The earliest someone may check into a lodging establishment.	1
checkoutTime	DateTime	The latest someone may check out of a lodging establishment.	1
contactPoint	ContactPoint	A contact point for a person or organization.	0..1
containsPlace	Place Accommodation	The basic containment relation between a place and another that it contains.	0..N
currenciesAccepted	Text	The currency accepted (in ISO 4217 currency format).	0..N
department	Organization	A relationship between an organization and a department of that organization, also described as an organization (allowing different urls, logos, opening hours). For example: a store with a pharmacy, or a bakery with a cafe.	0..N
description	Text	A description of the item.	1

```
{
  "@context": {
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "sh": "http://www.w3.org/ns/shacl#",
    "xsd": "http://www.w3.org/2001/XMLSchema#",
    "schema": "http://schema.org/",
    "sh:targetClass": {
      "@id": "sh:targetClass",
      "@type": "@id"
    },
    "sh:property": {
      "@id": "sh:property",
      "@type": "@id"
    },
    "sh:path": {
      "@id": "sh:path",
      "@type": "@id"
    },
    "sh:nodeKind": {
      "@id": "sh:nodeKind",
      "@type": "@id"
    },
    "sh:datatype": {
      "@id": "sh:datatype",
      "@type": "@id"
    },
    "sh:node": {
      "@id": "sh:node",
      "@type": "@id"
    },
```

2. Knowledge Creation - Methodology

3) **top-down:** applies models for further knowledge acquisition
   a) mapping according to domain specifications
   b) annotation development according to domain specifications



2. Knowledge Creation - tools - semantify.it

In the “early days” of our KG building efforts: three core questions (by our show-case users*) arised

* our efforts were always driven by educating people (real users, outside of academia, mostly from the industry/tourism) to create their own semantically rich content

1) which vocabulary to use
2) how to create JSON-LD files
3) how to publish those annotations (schema.org in JSON-LD files)

Tool, developed as a research project, grown to a full-stack annotation creation, validation and publication framework!

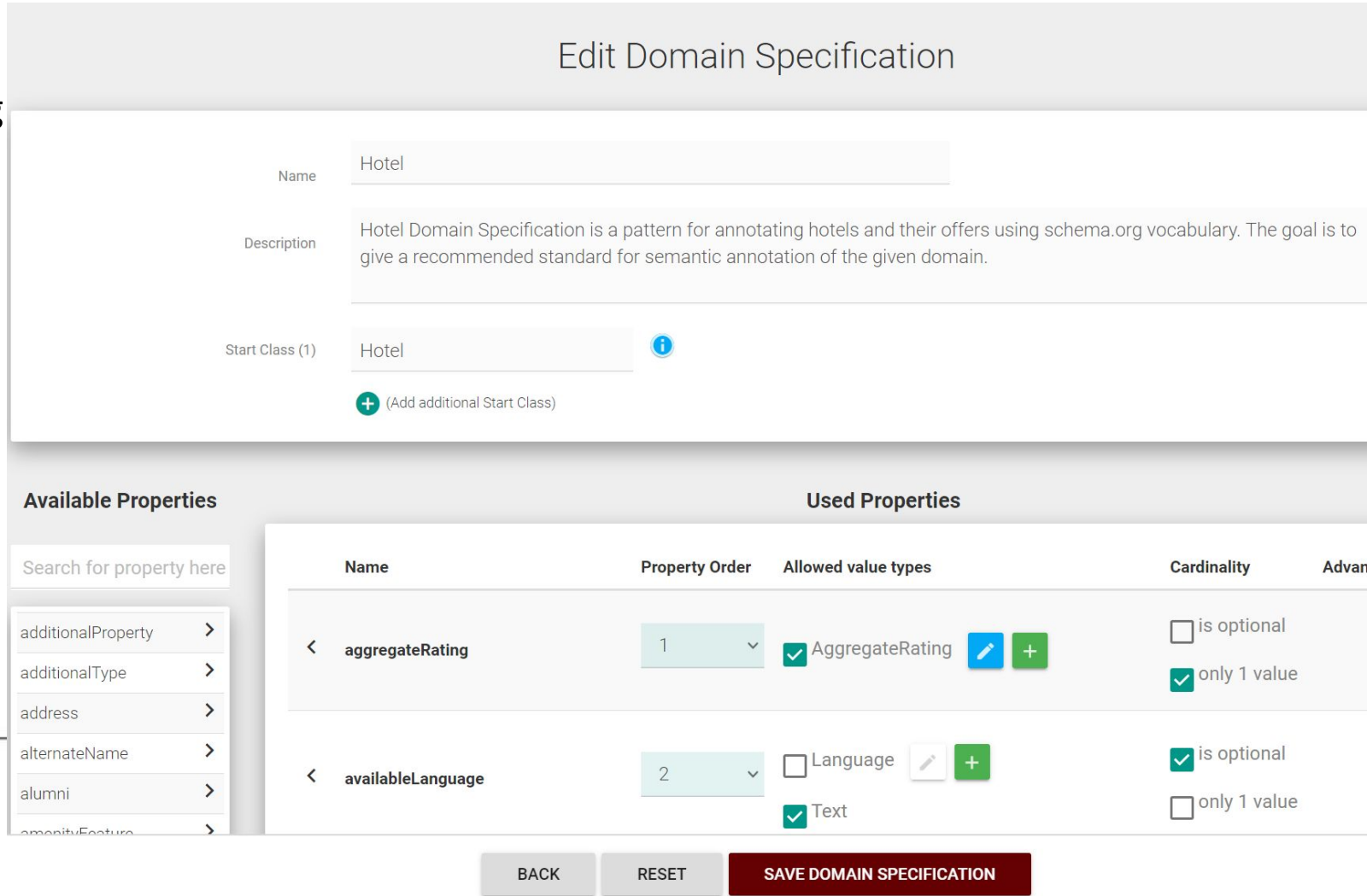
2. Knowledge Creation - tools - semantify.it

1) Which vocabulary to choose? ⇒ schema.org

Still hundreds of classes and properties in schema.org?

Domain Specifications

- (Extended) subset of schema.org
- Domain expert builds
  DS files as templates for editor
- Easy to use DS editor



universität innsbruck

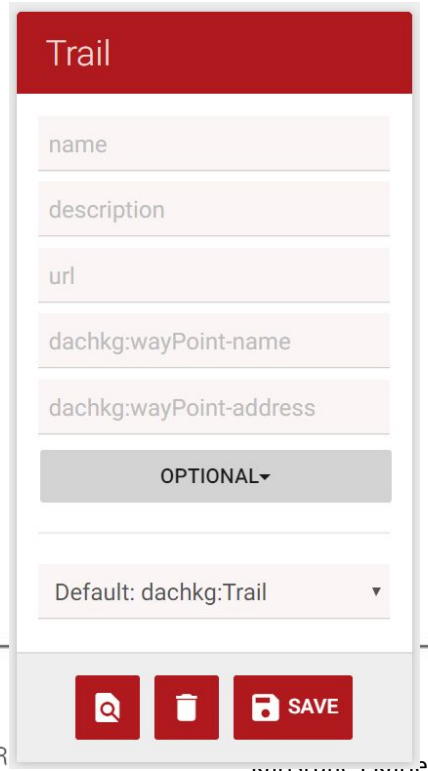
STI · INNSBRUCK

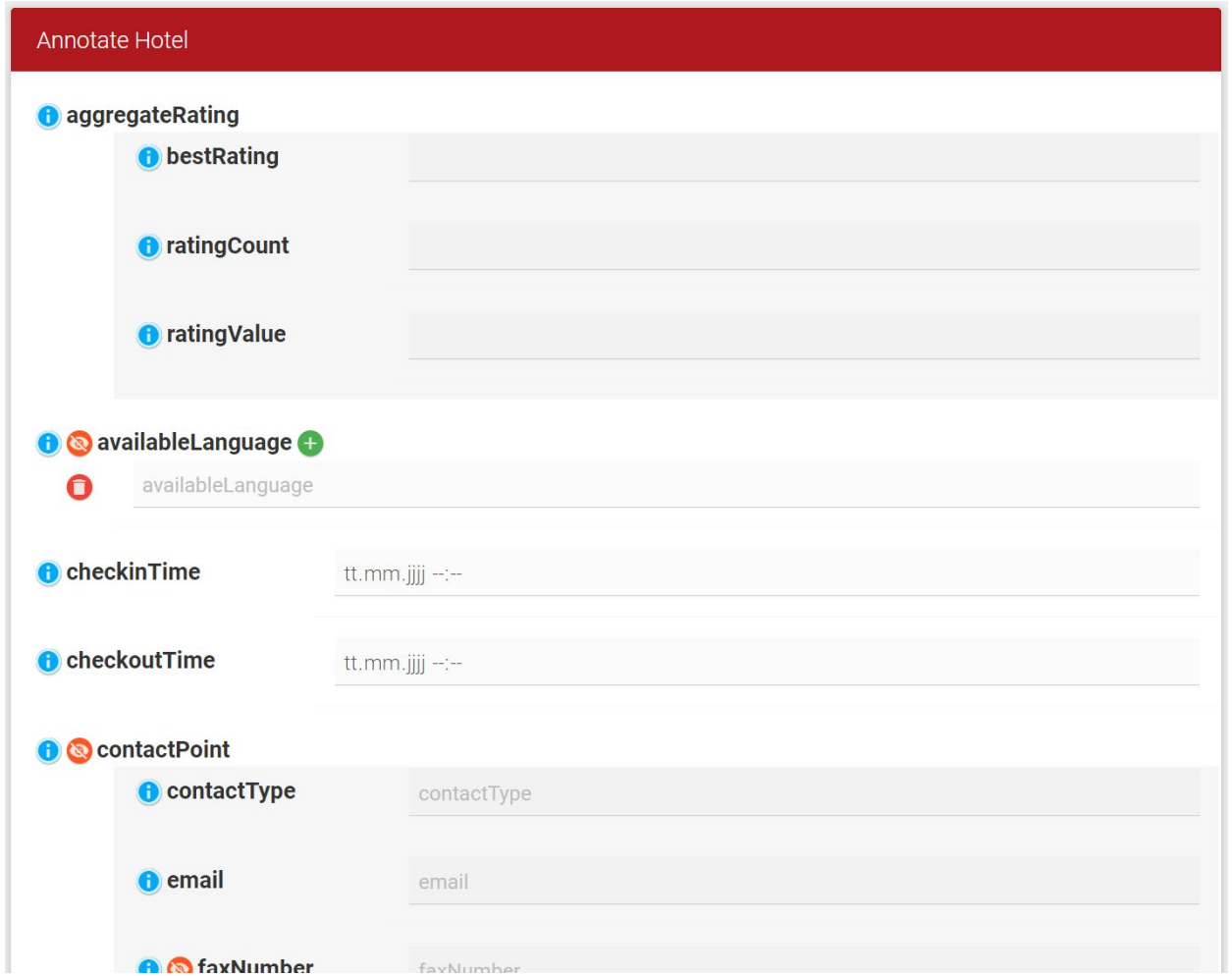
2. Knowledge Creation - tools - semantify.it

2) How to create those JSON-LD files?

- Semantify.it editor & instant annotations
- based on DS
- Inside platform (big DS files)
- or Instant Annotations (IA)
  portable to every website (based on JS)
- mappers (RocketRML)
- wrapper framework
- semi-automatic

RocketRML ⇒

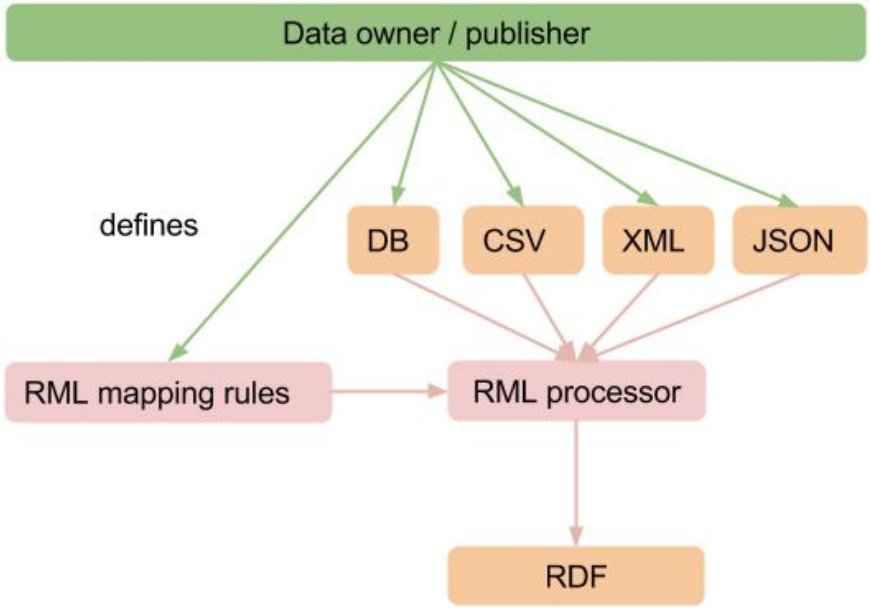




2. RocketRML - A Quite Scalable RML Mapper [Simsek et al., 2019]

Based on RML [Dimou et al., 2014]:

- Easier to learn RML than a programming language
- Easy sharing
- Mapping can be visualized
- Mapfiles can be faster to write than code
- Easily change mappings



UNIVERSITEIT GENT

RML

RML **YARRRML** **Matey**

2. RocketRML - A Quite Scalable RML Mapper

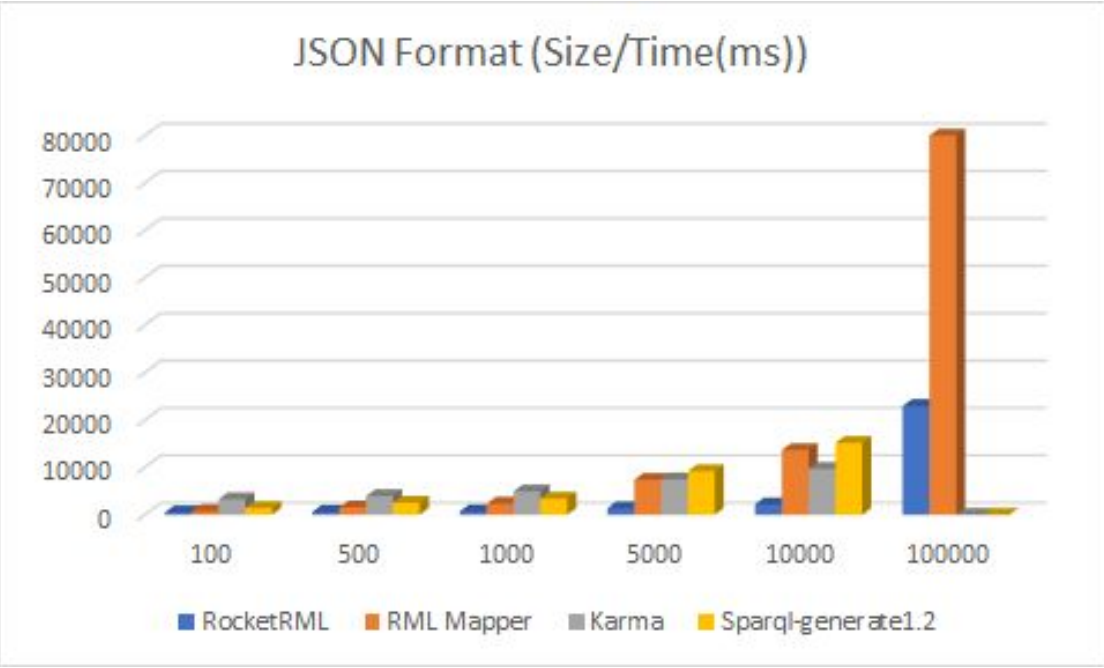
- Resolving JOINs is the main bottleneck when it comes to mapping large input files.
- Each TriplesMap is iterated once
- Before starting the mapping process for a TriplesMap, we check whether the TriplesMap is in the join condition of another TriplesMap. If it is, then we get the parent path of the join condition and evaluate it. The value then is cached as path - value pair

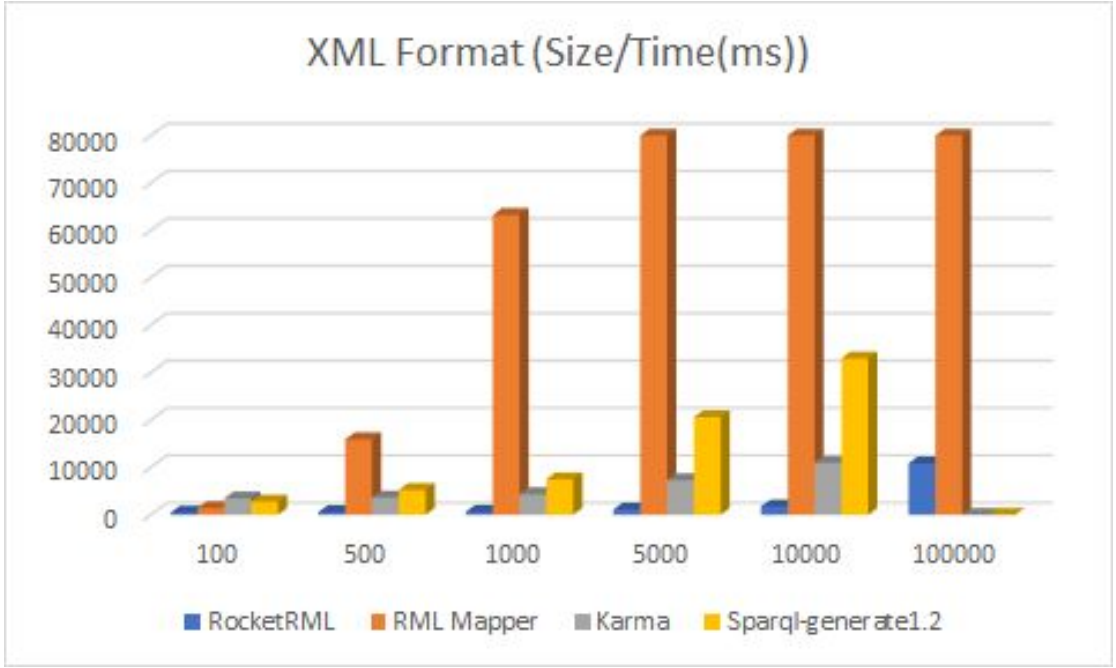
2. RocketRML - A Quite Scalable RML Mapper

- Then we map the data based on the TriplesMap as usual. If there is a join condition encountered during the mapping, then value of the child and path to the parent is cached in the child

- After everything is mapped, we go through the two caches and join the objects with matching child and parent values.

2. RocketRML - Performance





2. RocketRML - Source Code

RocketRML - An RML Mapper

View the Project on GitHub
semantifyit/RocketRML

RocketRML

For the legacy version with the different behavior of the iterator please see this version.

This is a javascript RML-mapper implementation for the RDF mapping language (RML).

Install

```
npm install rocketrml
```

Quick-start

After installation you can to copy index.js into your current working directory. Also the mapfile.ttl and the input is needed.

```
node index.js
```

Starts the execution and the output is then written to ./out.n3.

Also an example Dockerfile can be seen here.

This project is maintained by semantifyit

Hosted on GitHub Pages — Theme by orderedlist

https://semantifyit.github.io/RocketRML/

Node.js implementation

Also available as Docker container

2. RocketRML - A Quite Scalable RML Mapper

- Quick demo (https://semantifyit.github.io/rml):

Raw data set (JSON):

```json
{
  "persons": [
      {
          "firstname": "Elias",
          "lastname": "Kärle",
          "speaks": [
              "de",
              "en",
              "it",
              "fr",
              "Tyrolean"
          ]
      },
      {
          "firstname": "Umutcan",
          "lastname": "Simsek",
          "speaks": [
              "tr",
              "en",
              "de",
              "Hessisch"
          ]
      }
  ]
}
```

Mapping file (YARRRML*):

```yaml
prefixes:
  schema: "http://schema.org/"
  myfunc: "http://myfunc.com/"
mappings:
  person:
    sources:
      - ['input~jsonpath', '$.persons[*]']
    s: http://example.com/$(firstname)
    po:
      - [a, schema:Person]
      - [schema:name, $(firstname)]
      - [schema:language, $(speaks.*)]
```

* YARRRML is the yaml-based, human readable, translation of the actual turtle-based RML syntax. (http://rml.io/yarrrml/matey/)

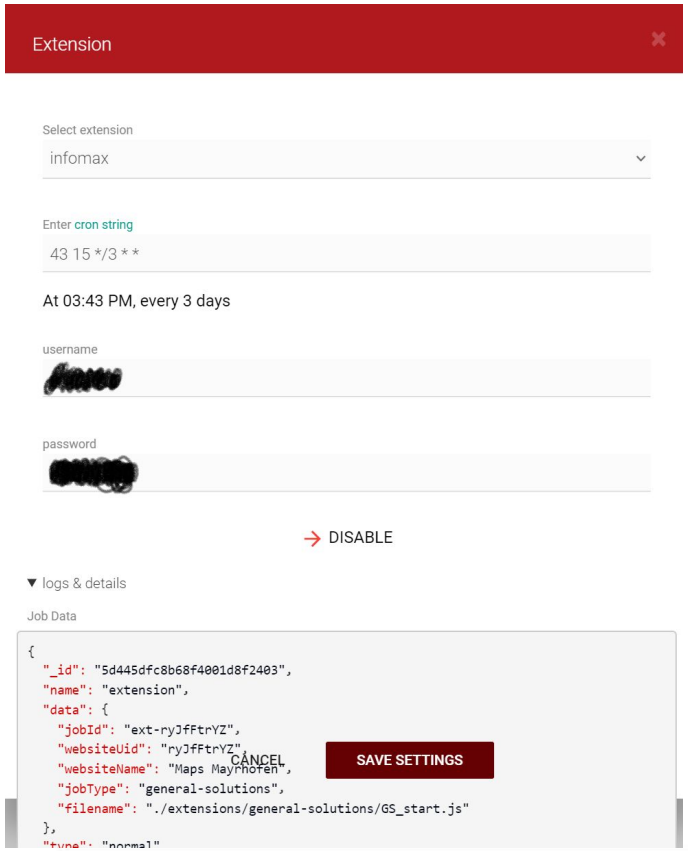
Mapping result:

```json
[
  {
    "@id": "http://example.com/Elias",
    "@type": "Person",
    "language": [
      "de",
      "en",
      "it",
      "fr",
      "Tyrolean"
    ],
    "name": "Elias",
    "@context": {
      "@vocab": "http://schema.org/"
    }
  },
  {
    "@id": "http://example.com/Umutcan",
    "@type": "Person",
    "language": [
      "tr",
      "en",
      "de",
      "Hessisch"
    ],
    "name": "Umutcan",
    "@context": {
      "@vocab": "http://schema.org/"
    }
  }
]
```

2. Knowledge Creation - tools - semantify.it

2) How to create those JSON-LD files?

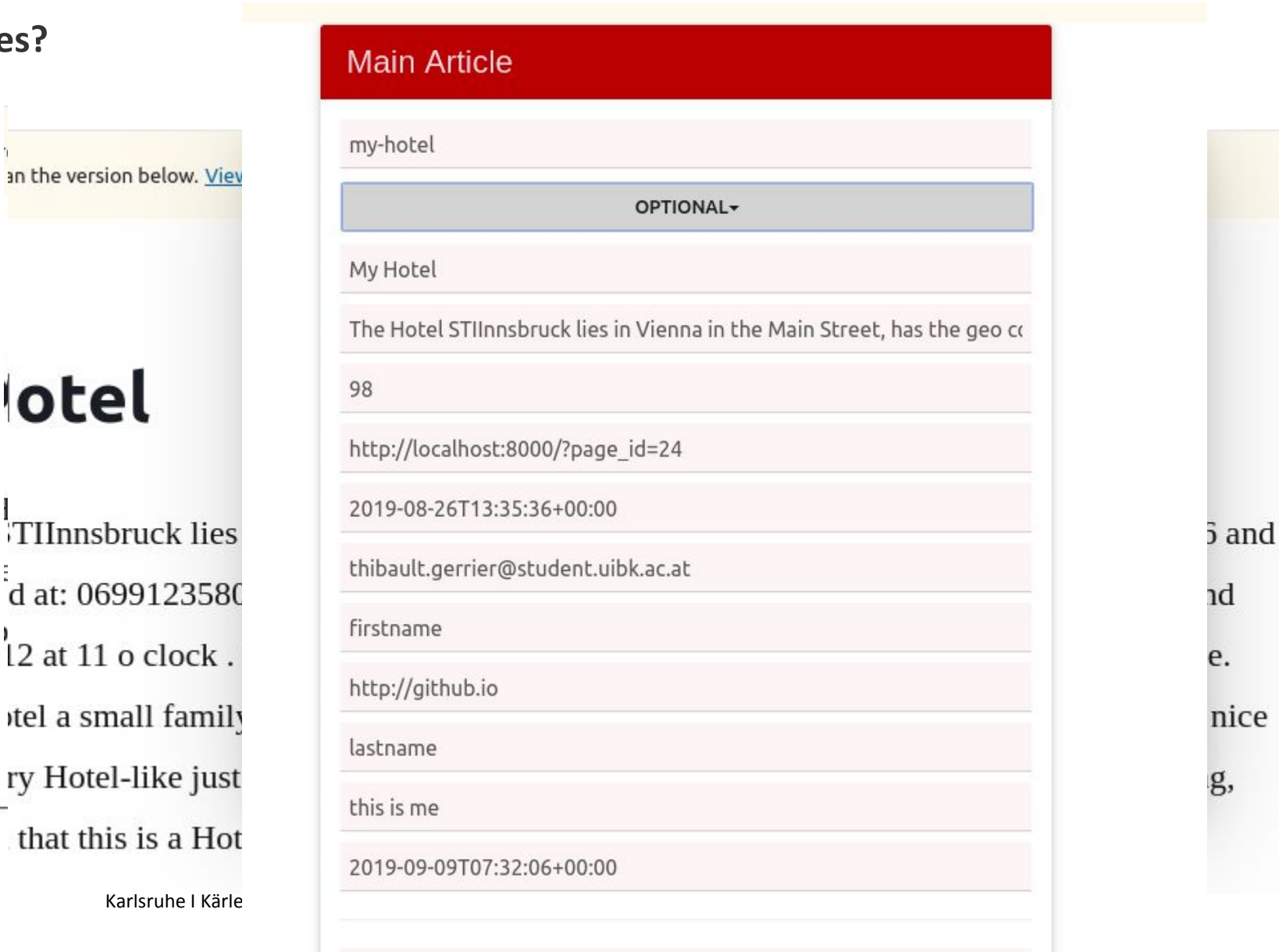
- wrapper framework



2. Knowledge Creation - tools - semantify.it

2) How to create those JSON-LD files?

- semi automatic generation
  - WordPress plugin
  - “guess” the entities of the web page through machine learning
  - model trained on entities in our knowledge graph



2. Knowledge Creation - tools - semantify.it

3) How to publish annotations (schema.org in JSON-LD files)?

- copy&paste?
  → pasting content to website is no option for inexperienced users and does not scale

- semantify.it **stores** all created annotations and **provides** them over an **API**

(**http://smtfy.it/sj7Fie2** OR **http://smtfy.it/url/http//**... OR **http://smtfy.it/cid/374fm38dkgi**...)

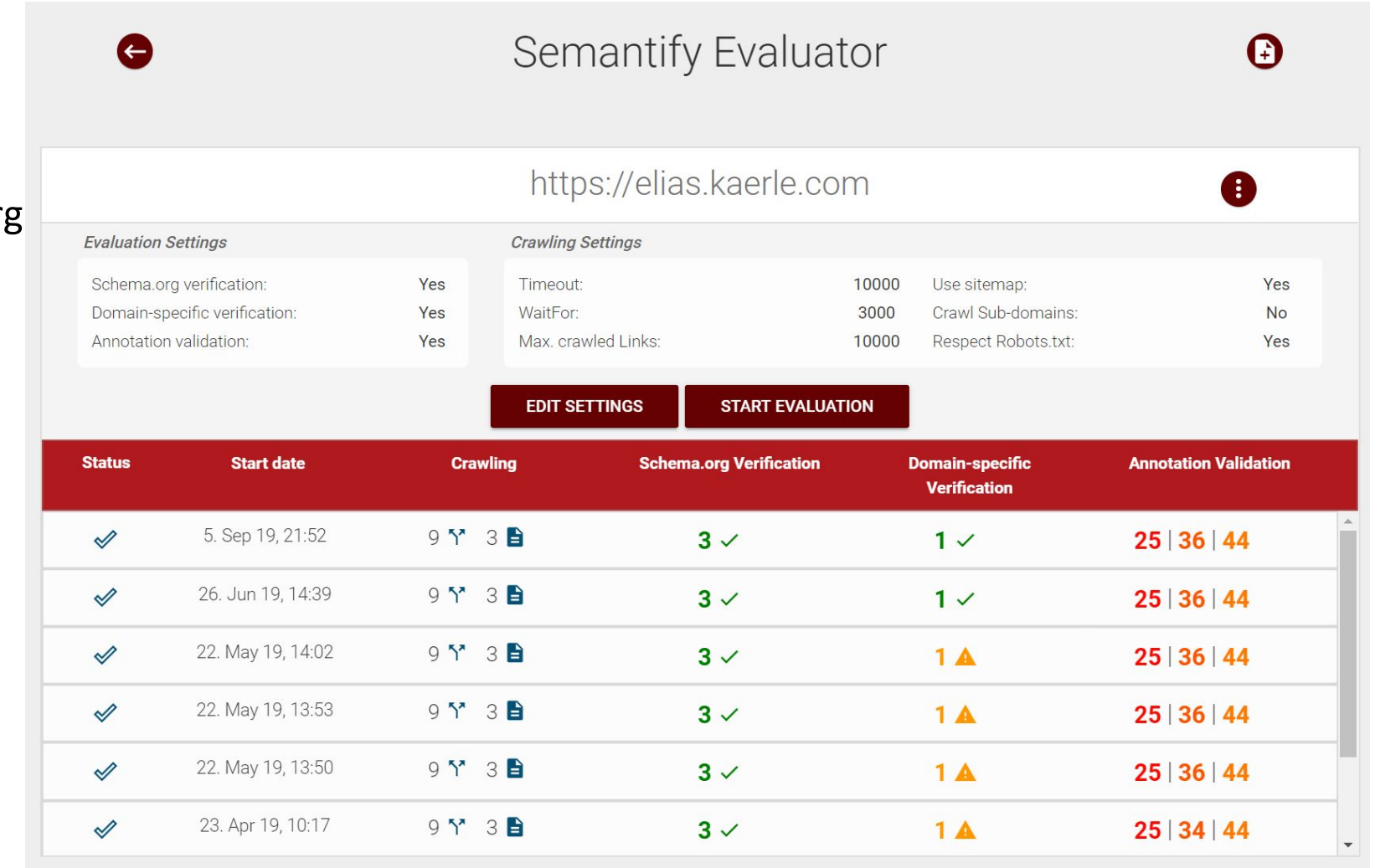
- publication of annotations over JS or into popular CMSs trough plugins (Wordpress, TYPO3 etc.)

2. Knowledge Creation - tools - semantify.it

Evaluator:

validation & verification

- **verification** against schema.org
- **verification** against DS
- **validation** against website →



2. Knowledge Creation - tools - semantify.it

Evaluator:

- validation against content of website

Property	Value	Score
givenName	Elias Kärle	100
email	elias.kaerle@sti2.at	0
telephone	+4351250753738	0
image	https://elias.kaerle.com/elias.jpg	0
jobTitle	Scientific Assistant	85
worksFor.department.name	Semantic Technology Institute (STI)	68
worksFor.name	University of Innsbruck	95
worksFor.url	https://www.sti-innsbruck.at/	0
faxNumber	+4351250753738	0

3. Knowledge Hosting

In our context:

"Knowledge is represented in the form of semantically enriched data"

→ **metadata** is added to **describe** the data by using a (de-facto) **standard vocabulary**, according to the principles of **RDF**

1) identify resource with URI: e.g. http://fritz.phantom.com

2) describe s, p, o

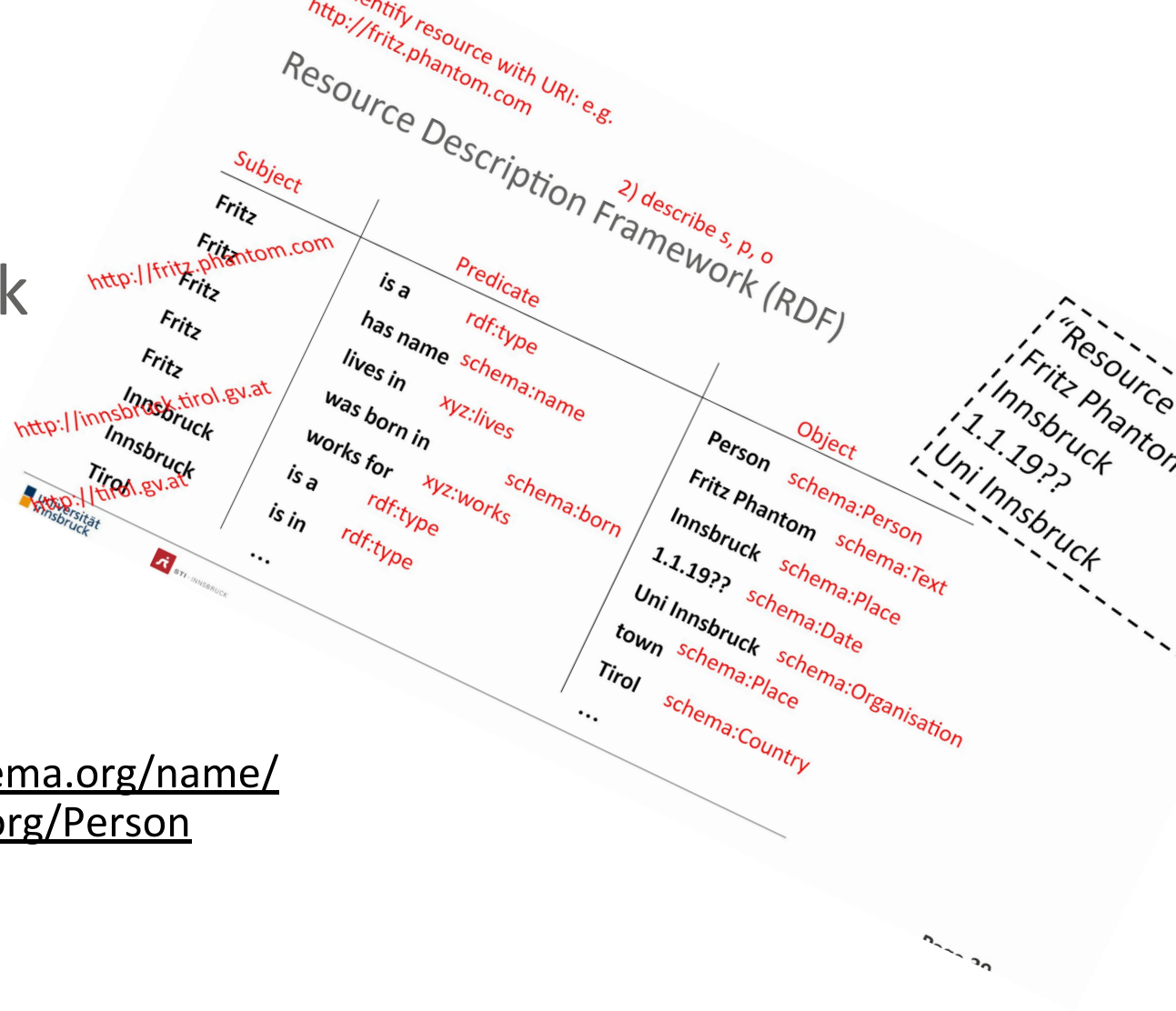
Resource Description Framework (RDF)

"Resource":
Fritz Phantom
Innsbruck
1.1.19??
Uni Innsbruck

Subject	Predicate	Object
Fritz	is a rdf:type	Person schema:Person
Fritz	has name schema:name	Fritz Phantom schema:Text
Fritz	lives in xyz:lives	Innsbruck schema:Place
Fritz	was born in schema:born	1.1.19?? schema:Date
Fritz	works for xyz:works	Uni Innsbruck schema:Organisation
Innsbruck	is a rdf:type	town schema:Place
Innsbruck	is in rdf:type	Tirol schema:Country
Tirol	…	…

http://fritz.phantom.com
http://innsbruck.tirol.gv.at
http://tirol.gv.at

Resource Description Framework



what actually are the s, p, o?

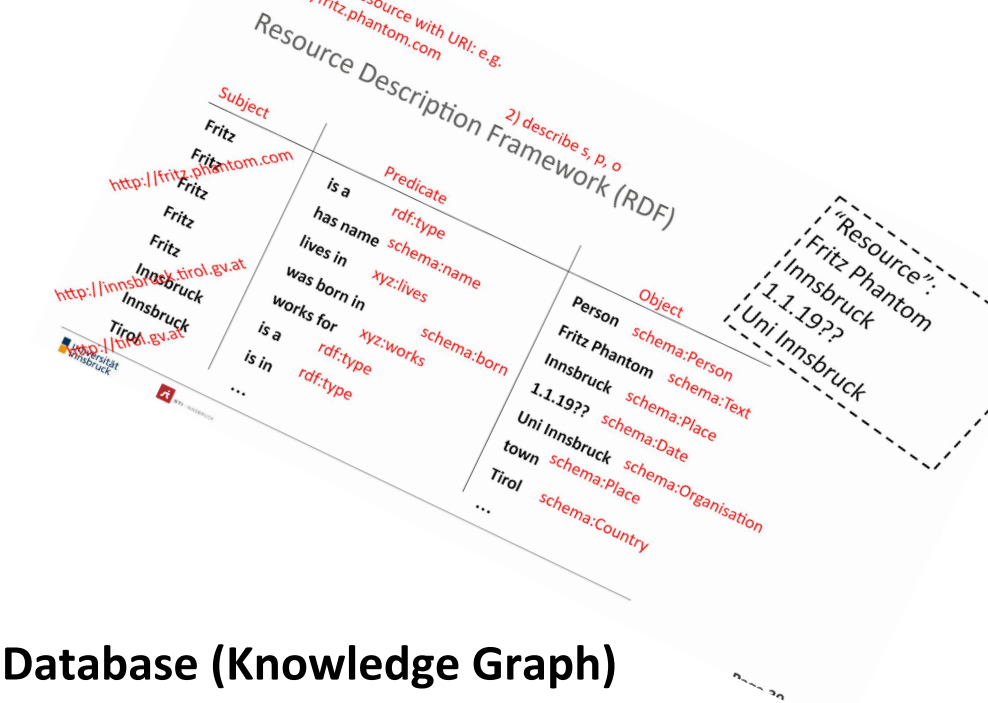
Either a URL:

- to identify resources http://fritz.phantom.com
- to refer to properties of an ontology http://schema.org/name/
- to refer to types of an ontology http://schema.org/Person

or a literal

- String: “Fritz Phantom”
- Date: “1.1.19??”
- Number: 42

Resource Description Framework



» 2 ways of representation (at least):

1. JSON-LD (for websites)

{“@context”:”http://schema.org”

“@type”:”Person”

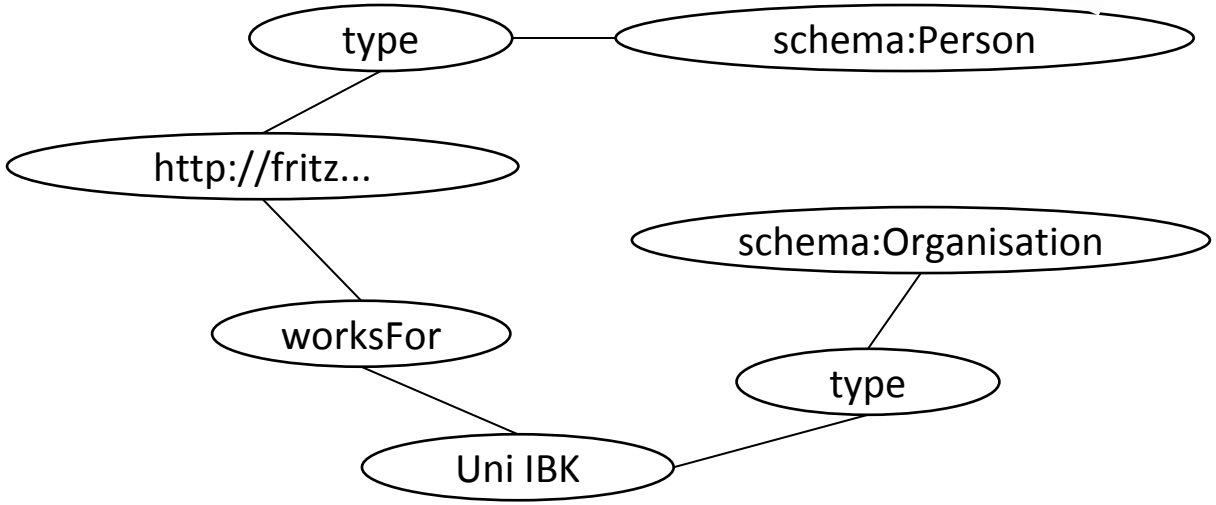
“@id”: “https://fritz.phantom.com”,

“livesIn”:”Innsbruck”

“born”:”19??-01-01”

“worksFor”:{“@type”:”Organisation”,
“name”:”Uni Innsbruck”}}

2. Graph Database (Knowledge Graph)



3. Knowledge Hosting

Two different approaches for storing semantically annotated data, depending on the use case:

Either as

1) **JSON-LD**

or as

2) **Knowledge Graph**

3. Knowledge Hosting

1) Storing as **JSON-LD**:

Use-case: storing semantically annotated data for usage on websites

→ the classical semantify.it use-case

→ many people use semantic annotations exclusively for website for SEO

Collection/creation: manual or semi-automatic editing, mapping, wrapper framework (was covered in previous section) or even crawling of annotated web-sites

Storage: JSON-based document database, e.g. MongoDB

(JSON-LD is in fact JSON)

3. Knowledge Hosting

1) Storing as **JSON-LD**:

Pros:

- seamless and lightning-fast storage and retrieval (through advanced JSON indexing)
- lightweight (little processing power overhead)
- cost effective (starts with powerful free versions)
- good framework integration for web-development
- well documented
- huge community

Cons:

- no native RDF reasoning
- reasonig requires extensive programming and processing power overhead

3. Knowledge Hosting

1) Storing as **JSON-LD**:

Query:

- over an API, through GET request

GET https://smtfy.it/BJgn06IHNb	{ "@context": "http://schema.org", "@type": "LodgingBusiness", "name": "Haus Olmarausch", "disambiguatingDescription": "Unser Haus liegt in schöner, sonniger Lage inmitten von Leutasch. Wir bieten ein gut ausgestattetes heimeliges Haus und herzliche Gastfreundschaft. Wir wollen vor allem eines: Dass Sie sich von Anfang an wie zu Hause fühlen. \nDer Loipeneinstieg und befestigte Winterwanderwage sind direckt vis a vis vom Haus. \nIm Sommer Ausgangspunkt für herrliche Wanderungen und Radtouren auf schönen und sicheren Wander - und Radwegen in den Bergen von Leutasch. Das Ortszentrum, Gasthöfe und Bäckerei sind in kurzer Zeit erreichbar.", "@description": " Unser Haus liegt in schöner, sonniger Lage inmitten von Leutasch. Wir bieten ein gut ausgestattetes heimeliges Haus und herzliche Gastfreundschaft. Wir wollen vor allem eines: Dass Sie sich von Anfang an wie zu Hause fühlen. Der Loipeneinstieg und befestigte Winterwanderwage sind direckt vis a vis vom Haus. Im Sommer Ausgangspunkt für herrliche Wanderungen und Radtouren auf schönen und sicheren Wander - und Radwegen in den Bergen von Leutasch. Das Ortszentrum, Gasthöfe und Bäckerei sind in kurzer Zeit erreichbar. "

Summary:

- works very well with tens of millions of JSON-LD files
- we replicate this data periodically into a graph database for “real” Knowledge Graph usage

3. Knowledge Hosting

2) Storing as **Knowledge Graph:**

Use-case: storing semantically annotated data as a full-fledged Knowledge Graph

⟶ Open Data repositories in tourism

⟶ enterprise Knowledge Graphs

⟶ advanced reasoning needs

⟶ AI, intelligent assistants

Collection/creation: due to potentially millions of annotation files: mapping, wrapper framework or also crawling of annotated web-sites ⟶ semantify.it-broker

3. Knowledge Hosting

semantify.it-broker:

- crawling platform to collect annotated data in JSON-LD, Microdata, RDFa
- storage in graph database
- provision of SPARQL UI

SPARQL EDITOR

(Graph: https://broker.semantify.it/graph/OryoBrBiiM/WHJLA8uQh8/latest)

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 SELECT * WHERE {
4   ?sub ?pred ?obj .
5 }
6 LIMIT 10
```

FILTERS

Blacklist sdoType	BREADCRUMBLIST
Whitelist markup	JSONLD

CRAWLING STATISTICS

CRAWLING TIME

Crawling took	10 minutes
Crawling started	Friday, April 27th 2018, 21:40:16
Crawling ended	Friday, April 27th 2018, 21:50:46
Crawled pages	3480

CRAWLING FILTERS

Blacklist sdoType	BREADCRUMBLIST
Blacklist markup	MICRODATA, RDFA
Whitelist markup	JSONLD

FOUND ANNOTATIONS

sdo Types	BREADCRUMBLIST - 2209, PLACE - 23, ARTICLE - 26234, FOODEVENT - 6, MUSICEVENT - 18, BUSINESSEVENT - 8, EVENT - 4, DANCEEVENT - 6, POSTALADDRESS - 153, SPORTSEVENT - 2, LOCALBUSINESS - 44, LODGINGBUSINESS - 2, NEWSARTICLE - 367, PERSON - 10, TOURISTATTRACTION - 77, GEOCOORDINATES - 77, LISTITEM - 77
Markup	MICRODATA - 28873, JSONLD - 444
Total	29317

SAVED ANNOTATIONS

sdo Types	PLACE - 8, FOODEVENT - 6, MUSICEVENT - 18, BUSINESSEVENT - 8, EVENT - 4, DANCEEVENT - 6, SPORTSEVENT - 2, LOCALBUSINESS - 23, LODGINGBUSINESS - 2, NEWSARTICLE - 367
Markup	JSONLD - 444
Total	444

3. Knowledge Hosting

2) Storing as **Knowledge Graph:**

Storage: due to RDF-nature, storage in graph database

with respect to:

- provenance
- historical data
- data duplication

In our current setting:

- historical data is kept im named graphs
- ~13 Billion statements

3. Knowledge Hosting

2) Storing as **Knowledge Graph:**

Storage: popular triple stores (https://www.w3.org/wiki/LargeTripleStores)

#	Name	# triples tested with
1	Oracle Spatial and Graph with Oracle Database 12c	1.08 T
2	AnzoGraph DB by Cambridge Semantics	1.065 T
3	AllegroGraph	1+ T
4	Stardog	50 B
5	OpenLink Virtuoso v7+	39.8 B
6	GraphDB™ by Ontotext	17 B

3. Knowledge Hosting

2) Storing as **Knowledge Graph:**

Pros:

- querying through native SPARQL endpoint

Cons:

- resource intensive
- expensive

3. Knowledge Hosting

2) Storing as **Knowledge Graph:**

Query:

- SPARQL

http://graphdb.sti2.at:8080/sparql

Summary:

- overhead aside: great for big knowledge graphs

```
PREFIX schema: <http://schema.org/>
SELECT DISTINCT ?name ?street ?location ?zip WHERE {
    ?s a schema:LodgingBusiness;
    schema:name ?name;
    schema:address ?address.
    ?address schema:addressLocality ?location;
        schema:streetAddress ?street;
        schema:postalCode ?zip.
    FILTER (regex(str(?location), "Mayrhofen") || regex(str(?location),
"Ginzling") || regex(str(?location), "Ramsau") || regex(str(?location),
"Schwendau") || regex(str(?location), "Hippach") ||
regex(str(?location), "Brandberg"))
}
```

4. Knowledge Curation

- Knowledge Assessment
- Knowledge Cleaning
- Knowledge Enrichment

4. Knowledge Curation - A Simple KR Formalism - TBox

1. Two disjoint and finite sets of type and property names T and P.

2. A finite number of type definitions isA(t1,t2) with t1 and t2 are elements of T. isA is reflexive and transitive.

3. A finite number of property definitions:
   - hasDomain(p,t) with p is an element of P and t an element of T.

   - Range definition for a property p with p is an element of P, t1 and t2 are Elements of T. Simple definition: Global property definition: hasRange(p,t2)
     - Refined definition: Local property definition: hasRange(p,t2) for domain t1, short: hasLocalRange(p,t1,t2)

4. Knowledge Curation - A Simple KR Formalism - ABox

1. A countable set of instance identifiers I. i, i1, and i2 are elements of I.

2. Instance assertions: isElementOf(i,t).isElementOf is a special property with build-in semantics. If isA(t1,t2) AND isElementOf(i,t1) THEN isElementOf(i,t2).

3. Property value assertions: p(i1,i2).

4. Equality assertions: isSameAs(i1,i2). We allow another build-in property to express identity of instances. It is symmetric, reflexive, and transitive.

4. Knowledge Curation - Knowledge Assessment

- First step to improve the quality of a KG: Assess the situation
- Closely related to data quality literature
- Various dimensions for data quality assessment introduced [Batini & Scannapieco, 2006], [Färber et al., 2018], [Pipino et al., 2002], [Wang, 1998], [Wang & Strong, 1996], [Wang et al., 2001], [Zaveri et al., 2016])

4. Knowledge Curation - Knowledge Assessment

1. accessibility
2. accuracy (veracity)
3. appropriate amount
4. believability
5. completeness
6. concise representation
7. consistent representation
8. cost-effectiveness
9. easy of manipulating
10. easy of operation
11. easy of understanding
12. flexibility
13. free-of-error
14. interoperability
15. objectivity
16. relevancy
17. reputation,
18. security,
19. timeliness (velocity),
20. traceability,
21. understandability,
22. value-added, and
23. variety

****fitness for use****

4. Knowledge Curation - Knowledge Assessment Tasks

- Two core assessment dimensions for Knowledge Graphs
  - Correctness
  - Completeness

- Three quality issue sources:
  - Instance assertions
  - Property value assertions
  - Equality assertions

4. Knowledge Curation - Knowledge Assessment Tools

- WIQA (Web Information Quality Assessment Framework) http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/ [Bizer and Cyganiak, 2009]:

  Allows defining policies to filter triples in a graph

- SWIQA (Semantic Web Information Quality Assessment Framework) [Fürber & Hepp, 2011]:

  A set of SPARQL-based rules to assess data quality

4. Knowledge Curation - Knowledge Assessment Tools

- LINK-QA [Guéret et al., 2012]

  Benefits from network features to assess data quality (e.g. counting open chains to find wrongly asserted isSameAs relationships)

- Sieve [Mendes et al., 2012] https://github.com/wbsg/ldif/

  Uses data quality indicators, scoring functions and assessment metrics

4. Knowledge Curation - Knowledge Assessment Tools

- Validata [Hansen et al., 2015] https://github.com/HW-SWeL/Validata

  An online tool check the conformance of RDF graphs against ShEx (Shape Expressions)

- Luzzu (A Quality Assessment Framework for Linked Open Datasets) [Debattista et al., 2016] https://eis-bonn.github.io/Luzzu/downloads.html

  Allows declarative definitions of quality metrics and produces machine-readable assessment reports based on Dataset Quality Vocabulary

4. Knowledge Curation - Knowledge Assessment Tools

- RDFUnit [Kontokostas et al., 2014] https://github.com/AKSW/RDFUnit/ :

  A framework that assesses linked data quality based on test cases defined in various ways (e.g. RDFS/OWL axioms can be converted into constraints)

- SDType [Paulheim & Bizer, 2013] https://github.com/HeikoPaulheim/sd-type-validate

  Uses statistical distributions to predict the types of instances. Incoming and outgoing properties are used as indicators for the types of resources.

4. Knowledge Curation - Knowledge Assessment Tools - Example

- Sieve for Data Quality Assessment
  - **Data Quality Indicators**: Various type of (meta)data that can be used to assess data quality e.g. data about the dataset provider, user ratings
  - **Scoring Functions:** A set of functions that help the calculation of assessment metrics based on the indicators
  - **Assessment Metrics:** Metrics like relevancy, timeliness that help users to assess the quality for an intended use
  - **Aggregate Metrics:** Allow users to aggregate new metrics based on simple assessment metrics.

4. Knowledge Curation - Knowledge Assessment Tools - Example

SCORING FUNCTION	EXAMPLE
TimeCloseness	measures the distance from the input date to the current (system) date. Dates outside the range receive value 0, and dates that are more recent receive values closer to 1.
Preference	assigns decreasing, uniformly distributed, real values to each graph URI provided as a space-separated list.
SetMembership	assigns 1 if the value of the indicator provided as input belongs to the set informed as parameter, 0 otherwise.
Threshold	assigns 1 if the value of the indicator provided as input is higher than a threshold informed as parameter, 0 otherwise.
IntervalMembership	Assigns 1 if the value of the indicator provided as input is within the interval informed as parameter, 0 otherwise.

Assessment Metrics in Sieve

4. Knowledge Curation - Knowledge Cleaning

- The actions taken to improve the correctness of a knowledge graph.
- Two major steps:
  - Error detection
  - Error correction

4. Knowledge Curation - Knowledge Cleaning Tasks

Detection and correction of wrong instance assertions: isElementOf(i.t)

Error	Correction
i is not a proper instance identifier	Delete assertion or correct i
i1 is not a valid instance identifier	Delete assertion or correct t.
Instance assertion is semantically incorrect	Delete assertion or find proper t.

4. Knowledge Curation - Knowledge Cleaning Tasks

Detection and correction of wrong property value assertions p(i1,i2)

Error	Correction
p is not a valid property	Delete assertion or correct p
i1 is not a valid instance identifier	Delete assertion or correct i1
i1 is not in any domain of p	Delete assertion or add assertion isElementOf(i1,t) where t is in a domain of p

4. Knowledge Curation - Knowledge Cleaning Tasks

Detection and correction of wrong property value assertions p(i1,i2)

Error	Correction
i2 is not a valid instance identifier	delete assertion or correct i2
i2 is not in any range of p where i1 is an element of a domain of p.	Delete assertion or Add assertion isElementOf(i1,t1) given that hasLocalRange(t1,p,t2) and isElementOf(i2,t2) or Add assertion isElementOf(i2,t2) given that hasLocalRange(t1,p,t2) and isElementOf(i1,t1)
Property assertion is semantically incorrect.	Delete assertion or define a proper i2 or find a better p or better i1

4. Knowledge Curation - Knowledge Cleaning Tasks

Detection and correction of wrong equality assertions isSameAs(i1,i2)

Error	Correction
i1 is not a valid instance identifier	Delete assertion or correct i1
i2 is not a valid instance identifier	Delete assertion or correct i2
Equality assertion is semantically wrong	Delete assertion or loosen the semantics (e.g. replace by a skos operator)

4. Knowledge Curation - Knowledge Cleaning Tools

- HoloClean [Rekatsinas et al., 2017] https://hazyresearch.github.io/snorkel/blog/holoclean.html

An error detection and correction tool based on integrity constraints to identify conflicting and invalid values, external information to support the constraints, and quantitative statistics to detect outliers.

- KATARA [Chu et al., 2015]

Learns the relationships between data columns and validate the learn patterns with the help of existing Knowledge Bases and crowd, in order to detect errors in the data. Afterwards it also suggests possible repairs.

4. Knowledge Curation - Knowledge Cleaning Tools

- SDValidate [Paulheim & Bizer, 2014] https://github.com/HeikoPaulheim/sd-type-validate

  Uses statistical distribution to detect erroneous statements that connect two resources. The statements with less frequent predicate-object pairs are selected as candidates for being wrong.

- SHACL https://www.w3.org/TR/shacl/ and ShEx https://shex.io/shex-semantics/index.html

  Two approaches that aim to verify RDF graphs against a specification (so called shapes).
  For a comparison of two approaches, see Chapter 7 in [Gayo et al., 2017]

4. Knowledge Curation - Knowledge Cleaning Tools

- LOD Laundromat [Beek et al., 2014] http://lodlaundromat.org/

  Detects and corrects syntactic errors (e.g. bad encoding, broken IRIs), replaces blank nodes with IRIs, removes duplicates in dirty linked open data and re-publishes it in a canonical format.

- TISCO [Rula et al., 2019]

  A framework that tries to identify the time interval where a statement was correct. It uses external knowledge bases and the web content to extract evidence to assess the validity of a statement for a time interval.

4. Knowledge Curation - Knowledge Enrichment

Improve the completeness of a knowledge graph by adding new statements

- Consists of following steps
  - Identifying new knowledge sources
  - Integration of TBox
  - Integration of Abox

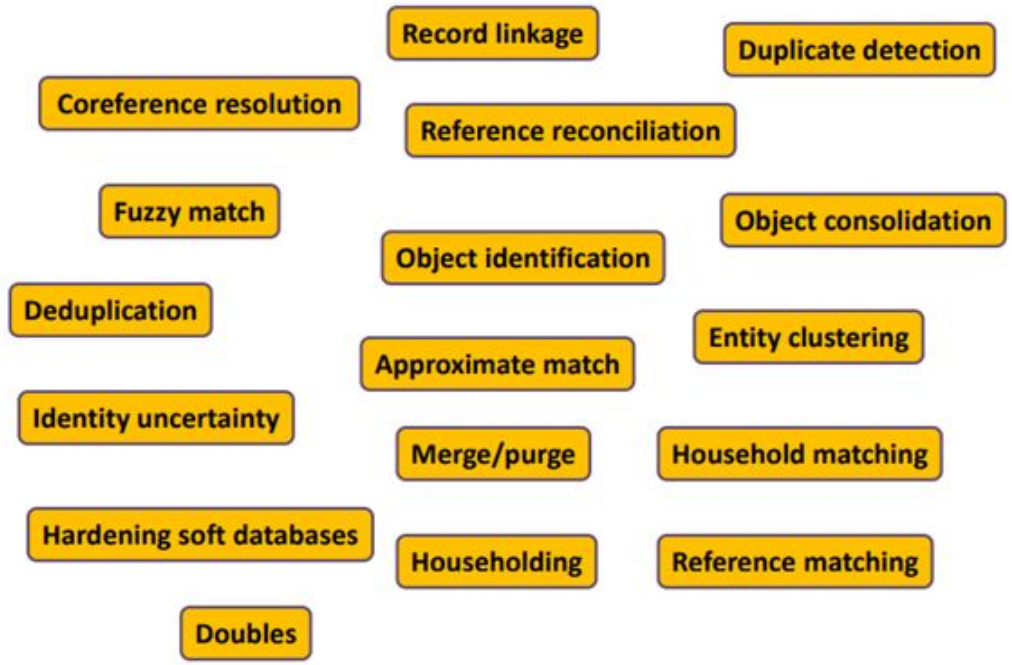
4. Knowledge Curation - Knowledge Enrichment

- Identifying knowledge sources
  - Open sources (e.g. LOD) - may be automated to some extent
  - Proprietary sources - usually very hard automate
- Integration of TBox
  - We assume that all data sources are mapped to schema.org
  - Non-RDF sources can be also mapped with the techniques described in Knowledge Creation

4. Knowledge Curation - Knowledge Enrichment

- Integration of ABox
  - Issue-1: Identifying and resolving duplicates
  - Issue-2: Invalid property assertions (e.g. multiple disjoint values for unique properties, domain and range violations)

4. Knowledge Curation - Knowledge Enrichment



Different names for the same problem! [Getoor et al., 2012]

Tackling issues:

- Entity resolution: Derive new isSameAs(i1,i2) assertions and aligning their property assertions
- Conflict resolution: Resolve conflicting property assertions
- Enrichment also has implications towards cleaning!

4. Knowledge Curation - Knowledge Enrichment Tasks

- Identifying and resolving duplicates
- Resolving conflicting property assertions

can be realized by

- addition of missing instance assertions: isElementOf(i,t)
- addition or deletion of property value assertions: p(i1,i2)
- addition of missing equality assertions: isSameAs(i1,i2)

4. Knowledge Curation - Knowledge Enrichment Tools

Duplication detection and resolution tools

- Dedupe: https://github.com/dedupeio/dedupe

  A python library that uses machine learning to find duplicates in a dataset and to link two datasets.

- Duke [Garshol & Borge, 2013]: https://github.com/larsga/Duke

  Uses various similarity metrics to detect duplicates in a dataset or link records between two datasets based on a given configuration. The configuration parameters can be

4. Knowledge Curation - Knowledge Enrichment Tools

Duplication detection and resolution tools

- Legato [Achichi et al., 2017] https://github.com/DOREMUS-ANR/legato

  A recording linkage tool that utilizes *Concise Bounded Description** of resources for comparison.
  *https://www.w3.org/Submission/2004/SUBM-CBD-20040930/#r6

- LIMES [Ngomo & Auer, 2011] https://github.com/dice-group/LIMES

A link discovery approach that benefits from the metric spaces (in particular triangle inequality) to reduce the amount of comparisons between source and target dataset.

4. Knowledge Curation - Knowledge Enrichment Tools

Duplication detection and resolution tools

- SERIMI [Araújo et al., 2011] https://github.com/samuraraujo/SERIMI-RDF-Interlinking

  A link discovery tool that utilizes string similarity functions on “label properties” without a prior knowledge of data or schema

- SILK [Volz et al., 2009] http://silkframework.org/

  A link discovery tool with declerative linkage rules applying different similarity metrics (e.g. string, taxonomic, set) that also supports policies for the notification of datasets when one of them publishes new links to others.

4. Knowledge Curation - Knowledge Enrichment Tools

Conflict resolution tools

- FAGI [Giannopoulos et al., 2014] https://github.com/GeoKnow/FAGI-gis

  A framework for fusing geospatial data. It suggests fusion strategies based on two datasets with geospatial data and a set of linked entities.

- KnoFuss [Nikolov et al., 2008] http://technologies.kmi.open.ac.uk/knofuss/

  A framework that allows the application of different methods on different attributes in the same dataset for identification of duplicates and resolves inconsistencies caused by the fusion of linked instances.

4. Knowledge Curation - Knowledge Enrichment Tools

Conflict resolution tools

- ODCleanStore [Knap et al., 2012]

  A framework that contains a fusion module that allows users to configure conflict resolution policies based on different functions (e.g. AVG, MAX, CONCAT) that can be applied on conflicting property values.

- Sieve [Mendes et al., 2012]

  Sieve has a data fusion module that supports different fusion functions on selected property values. It also utilizes the assessment values from the assessment module in the fusion process.

4. Knowledge Curation - Knowledge Enrichment Tools - Demo

Duplication detection and resolution with Duke

5. Knowledge Deployment

- training of ML models based on KGs
  - due to the RDF nature data in KGs is semantically described
  - good training data for ML models
- conversational agents
  - chatbots
  - intelligent personal assistants
  - **question answering over LinkedData**
- OpenData sharing platforms
  - currently Open(Government)Data often makes little sense (scanned pdfs, weird spreadsheets,csv, …)
  - LinkedData is self explaining (see lod-cloud https://lod-cloud.net)

5. Knowledge Deployment - discussion

- are you using KGs in your enterprise / research already?
- are you planning to?
- where do you see the potential
- where do you see challenges / risks?

References

[Achichi et al., 2017] Achichi, M., Bellahsene, Z., Todorov, K.: Legato results for OAEI 2017. In: Proceedings of the 12th International Workshop on Ontology Matching (OM2017) co-located with the 16th International Semantic Web Conference (ISWC2017), Vienna, Austria, October 21, 2017. CEUR Workshop Proceedings, vol. 2032, pp. 146–152. CEUR-WS.org (2017)

[Araújo et al., 2011] Araújo, S., Hidders, J., Schwabe, D., de Vries, A.P.: SERIMI - resource description similarity, RDF instance matching and interlinking. In: Proceedings of the 6th International Workshop on Ontology Matching (OM2011), Bonn, Germany, October 24, 2011. CEUR Workshop Proceedings, vol. 814. CEUR-WS.org (2011), http://ceur-ws.org/Vol-814/om2011_poster6.pdf

[Batini & Scannapieco, 2006] Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Data-Centric Systems and Applications, Springer (2006). https://doi.org/10.1007/3-540-33173-5

[Beek et al., 2014] Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD laundromat: A uniform way of publishing other people's dirty data. In: Proceedings of the 13th International Semantic Web Conference (ISWC2014), Riva del Garda, Italy, October 19-23, 2014. Lecture Notes in Computer Science, vol. 8796, pp. 213–228. Springer (2014). https://doi.org/10.1007/978-3-319-11964-9_14

[Bizer and Cyganiak, 2009] Bizer, C., Cyganiak, R.: Quality-driven information filtering using the WIQA policy framework. Journal of Web Semantics 7(1), 1–10 (2009). https://doi.org/10.1016/j.websem.2008.02.005

[Chu et al., 2015] Chu, X., Ouzzani, M., Morcos, J., Ilyas, I.F., Papotti, P., Tang, N., Ye, Y.: KATARA: reliable data cleaning with knowledge bases and crowdsourcing. Proceedings of the 41st International Conference on Very Large Data Bases (PVLDB2015), VLDB Endowment, Hawaii, August 31- September 4, 2015 8(12), 1952–1955 (2015). https://doi.org/10.14778/2824032.2824109, http://www.vldb.org/pvldb/vol8/p1952-chu.pdf

[Debattista et al., 2016] Debattista, J., Auer, S., Lange, C.: Luzzu - A methodology and framework for linked data quality assessment. Journal of Data and Information Quality (JDIQ) 8(1), 4:1–4:32 (2016). https://doi.org/10.1145/2992786

References

[Dimou et al.,, 2014] Dimou, A., Sande, M.V., Colpaert, P., Verborgh, R., Mannens, E., de Walle, R.V.: RML: A generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the Workshop on Linked Data on the Web (LDOW2014) colocated with the 23rd International World Wide Web Conference (WWW2014), Seoul, Korea, April 8, 2014. CEUR Workshop Proceedings, vol. 1184. CEUR- WS.org (2014), http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf

[Färber et al., 2018] Farber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of dbpedia, freebase, opencyc, wikidata, and YAGO. Semantic Web Journal 9(1), 77–129 (2018). https://doi.org/10.3233/SW-170275

[Fürber & Hepp, 2011] Fürber, C., Hepp, M.: Swiqa - a semantic web information quality assessment framework. In: Proceedings of the 19th European Conference on Information Systems (ECIS2011), Helsinki, Finland, June 9-11, 2011. p. 76. Association for Information Systems (AIS e Library) (2011), http://aisel.aisnet.org/ecis2011/76

[Garshol & Borge, 2013] Garshol, L.M., Borge, A.: Hafslund sesam - an archive on semantics. In: Proceedings of the 10th Extending Semantic Web Conference (ESWC2013): Semantics and Big Data, Montpellier, France, May 26-30, 2013. Lecture Notes in Computer Science, vol. 7882, pp. 578–592. Springer (2013). https://doi.org/10.1007/978-3-642-38288-8_39

[Gayo et al., 2017] Gayo, J. E. L., Prud'hommeaux, E., Boneva, I.,, Kontokostas, D. Validating RDF Data. Morgan & Claypool Publishers, (2017).

[Getoor et al., 2012] Getoor, L., Machanavajjhala, A.: Entity resolution: Theory, practice & open challenges. Proceedings of the 38th International Conference on Very Large Data Bases 5(12), 2018–2019 (2012). https://doi.org/10.14778/2367502.2367564

[Giannopoulos et al., 2014] Giannopoulos, G., Skoutas, D., Maroulis, T., Karagiannakis, N., Athanasiou, S.: FAGI: A framework for fusing geospatial RDF data. In: Proceedings of the Confederated International Conferences ”On the Move to Meaningful Internet Systems” (OTM2014), Amantea, Italy, October 27-31, 2014. Lec- ture Notes in Computer Science, vol. 8841, pp. 553–561. Springer (2014). https://doi.org/10.1007/978-3-662-45563-0_33

References

[Guéret et al., 2012] Guéret, C., Groth, P.T., Stadler, C., Lehmann, J.: Assessing linked data mappings using network measures. In: Proceedings of the 9th Extended Semantic Web Conference (ESWC2012), Heraklion, Greece, May 27-31, 2012. Lecture Notes in Computer Science, vol. 7295, pp. 87–102. Springer (2012). https://doi.org/10.1007/978-3-642-30284-8_13

[Hansen et al., 2015] Hansen, J.B., Beveridge, A., Farmer, R., Gehrmann, L., Gray, A.J.G., Khutan, S., Robertson, T., Val, J.: Validata: An online tool for testing RDF data conformance. In: Proceedings of the 8th International Conference on Semantic Web Applications and Tools for Life Sciences (SWAT4LS2015), Cambridge, UK, December 7-10, 2015. CEUR Workshop Proceedings, vol. 1546, pp. 157–166. CEUR-WS.org (2015), http://ceur-ws.org/Vol-1546/paper_3.pdf

[Knap et al., 2012] Knap, T., Michelfeit, J., Necaský, M.: Linked open data aggregation: Conflict resolution and aggregate quality. In: Proceedings of the 36th Annual IEEE Computer Software and Applications Conference Workshops (COMPSAC2012), Izmir, Turkey, July 16-20, 2012. pp. 106–111. IEEE Computer Society (2012). https://doi.org/10.1109/COMPSACW.2012.29

[Kontokostas et al., 2014] Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd International Conference on World Wide Web (WWW2014), Seoul, Korea, April 07 - 11, 2014. pp. 747–758. ACM (2014). https://doi.org/10.1145/2566486.2568002

[Mendes et al., 2012] Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of 2nd International Workshop on Linked Web Data Management (LWDM 2012), in conjunction with the 15th International Conference on Extending Database Technology (EDBT2012): Workshops, Berlin, Germany, March 30, 2012. pp. 116–123. ACM (2012). https://doi.org/10.1145/2320765.2320803

[Ngomo & Auer, 2011] Ngomo, A.N., Auer, S.: LIMES - A time-efficient approach for large-scale link discovery on the web of data. In: Walsh, T. (ed.) Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI2011), Barcelona, Spain, July 1622, 2011. pp. 2312–2317. AAAI Press (2011). https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-385

References

[Nikolov et al., 2008] Nikolov, A., Uren, V.S., Motta, E., Roeck, A.N.D.: Integration of semantically annotated data by the knofuss architecture. In: Gangemi, A., Euzenat, J. (eds.) Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008): Practice and Patterns, Acitrezza, Italy, September 29 - October 2, 2008. Lecture Notes in Computer Science, vol. 5268, pp. 265–274. Springer (2008). https://doi.org/10.1007/978-3-540-87696-0_24

[Paulheim & Bizer, 2013] Paulheim, H., Bizer, C.: Type inference on noisy RDF data. In: Proceedings of the 12th International Semantic Web Conference (ISWC2013), Sydney, Australia, October 21-25, 2013. Lecture Notes in Computer Science, vol. 8218, pp. 510–525. Springer (2013). https://doi.org/10.1007/978-3-642-41335-3_32

[Paulheim & Bizer, 2014] Paulheim, H., Bizer, C.: Improving the quality of linked data using statistical distributions. International Journal on Semantic Web and Information Systems (IJSWIS) 10(2), 63–86 (2014). https://doi.org/10.4018/ijswis.2014040104

[Paulheim, 2017] Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web Journal 8(3), 489–508 (2017). https://doi.org/10.3233/SW-160218

[Paulheim, 2018] Paulheim, H.: How much is a triple? estimating the cost of knowledge graph creation. In: Proceedings of the 17th International Semantic Web Conference (ISWC2018): Posters & Demonstrations, Industry and Blue Sky Ideas Tracks, Monterey, USA, October 8-12, 2018. CEUR Workshop Proceedings, vol. 2180. CEUR-WS.org (2018), http://ceur-ws.org/Vol-2180/ISWC_2018_Outrageous_Ideas_paper_10.pdf

[Pipino et al., 2002] Pipino, L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Communications of the ACM 45(4), 211–218 (2002). https://doi.org/10.1145/505248.5060010

[Rekatsinas et al., 2017] Rekatsinas, T., Chu, X., Ilyas, I.F., R´e, C.: Holoclean: Holistic data repairs with probabilistic inference. Proceedings of the Very Large Data Bases Endowment 10(11), 1190–1201 (2017). https://doi.org/10.14778/3137628.3137631, http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf

References

[Rula et al., 2019] Rula, A., Palmonari, M., Rubinacci, S., Ngomo, A.N., Lehmann, J., Maurino, A., Esteves, D.: TISCO: temporal scoping of facts. Journal of Web Semantics 54, 72–86 (2019). https://doi.org/10.1016/j.websem.2018.09.002

[Simsek et al., 2019] Simsek, U., Kärle, E., Fensel, D.: Rocketrml - A nodejs implementation of a use-case specific RML mapper. In: Proceedings of 1st Knowledge Graph Building Workshop co-located with 16th Extended Semantic Web Conference (ESWC), Portoroz, Slovenia, June 3, 2019. vol. abs/1903.04969 (2019), http://arxiv.org/abs/1903.04969

[Volz et al., 2009] Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. In: Proceedings of the 8th International Semantic Web Conference (ISWC 2009), Chantilly, USA, October 25-29, 2009. Lecture Notes in Computer Science, vol. 5823, pp. 650–665. Springer (2009). https://doi.org/10.1007/978-3-642-04930-9_41

[Wang, 1998] Wang, R.Y.: A product perspective on total data quality management. Communication of the ACM 41(2), 58–65 (1998). https://doi.org/10.1145/269012.269022

[Wang & Strong, 1996] Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems 12(4), 5–33 (1996), http://www.jmis-web.org/articles/1002

[Wang et al., 2001] Wang, R.Y., Ziad, M., Lee, Y.W.: Data Quality, Advances in Database Systems, vol. 23. Kluwer Academic Publisher (2001). https://doi.org/10.1007/b116303

[Zaveri et al., 2016] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web Journal 7(1), 63–93 (2016)

Think **GREEN**
Only print if it's essential

www.uibk.ac.at