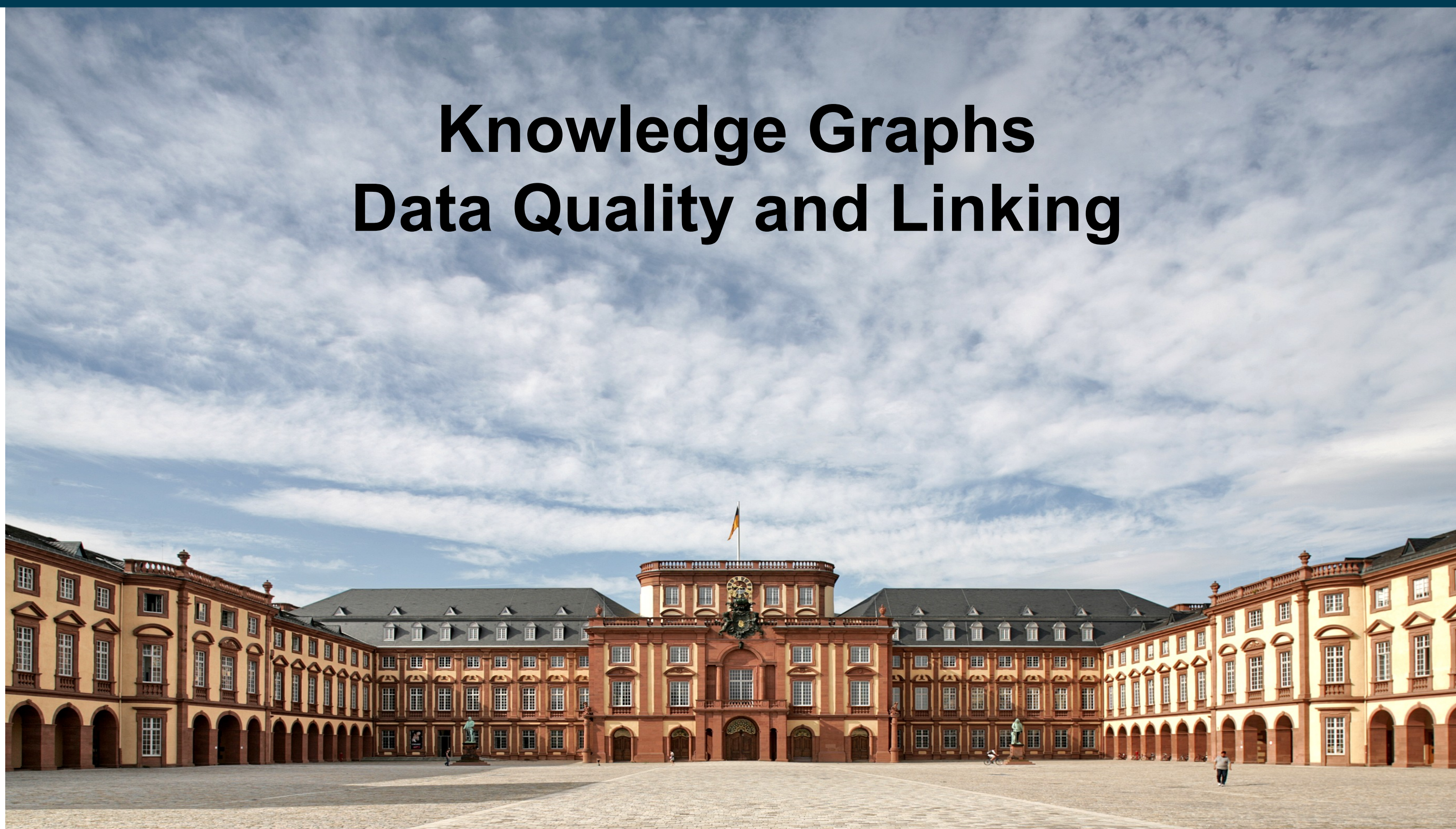
UNIVERSITÄT MANNHEIM

Knowledge Graphs
Data Quality and Linking



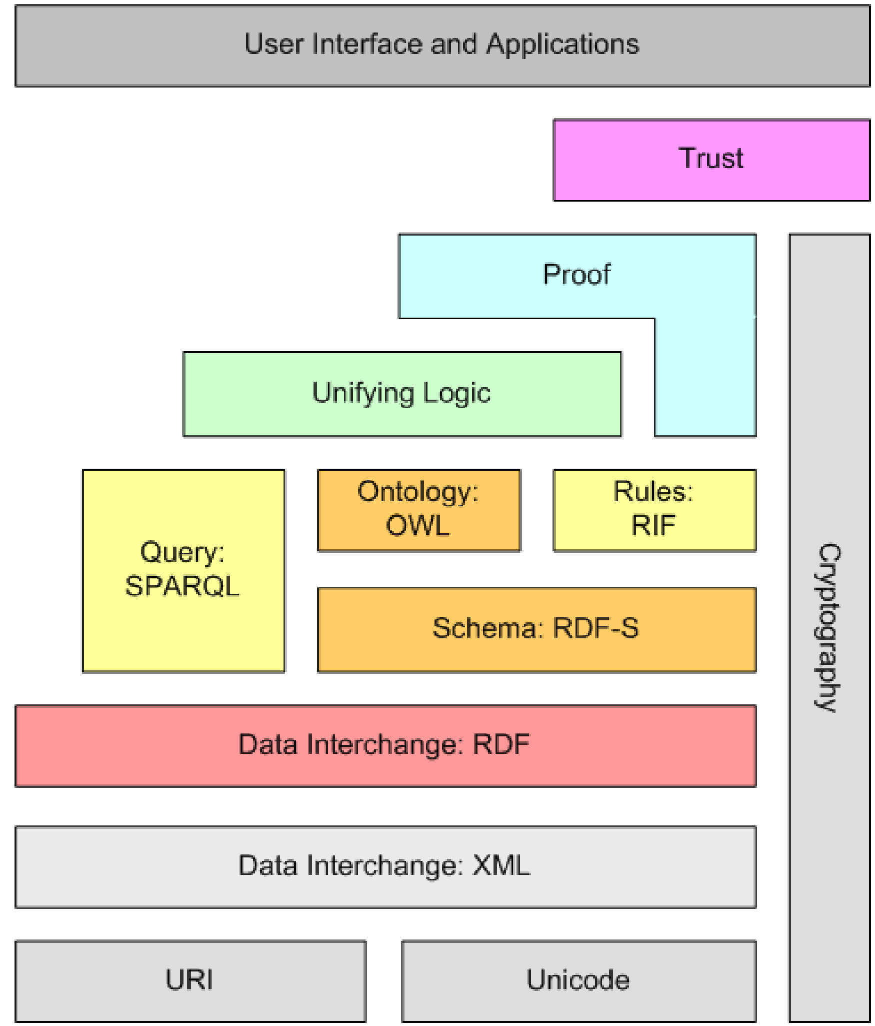
Heiko Paulheim

The Journey Ends Here

here be dragons...

Semantic Web
Technologies
(This lecture)

Technical
Foundations



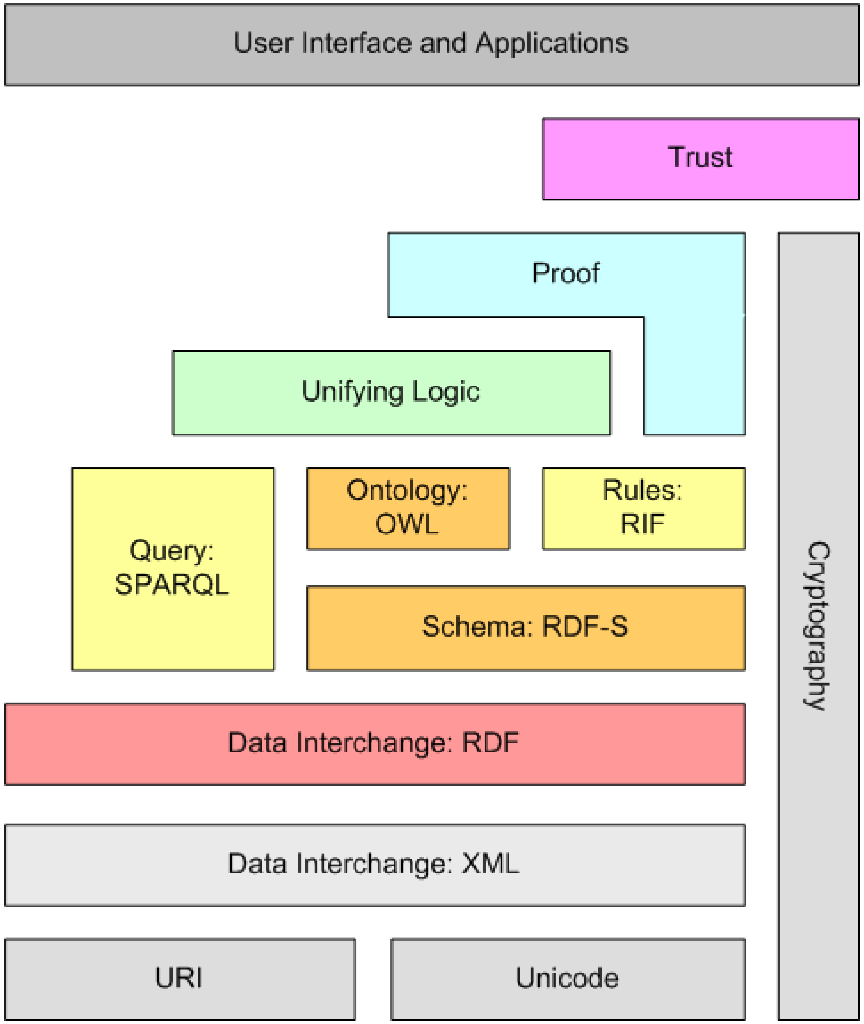
Berners-Lee (2009): *Semantic Web and Linked Data*
http://www.w3.org/2009/Talks/0120-campus-party-tbl/

Before You Go...

here be dragons...

Semantic Web
Technologies
(This lecture)

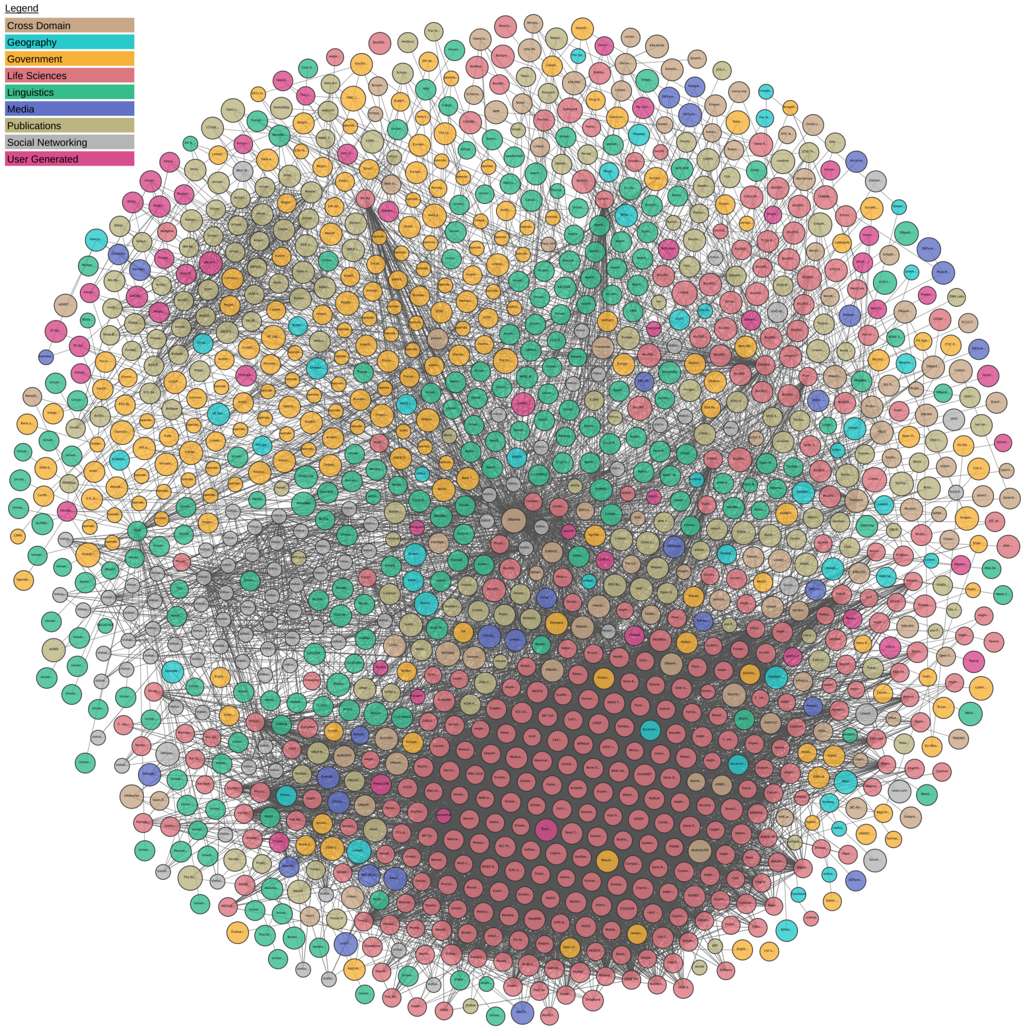
Technical
Foundations



Berners-Lee (2009): *Semantic Web and Linked Data*
http://www.w3.org/2009/Talks/0120-campus-party-tbl/

Before You Go...

- We've learned about
  - standards
  - methods
  - datasets
- You've played with
  - datasets
  - tools
- Now, let's be serious…
  - how good is that data, actually?



The Linked Open Data Cloud from lod-cloud.net

Previously on Knowledge Graphs

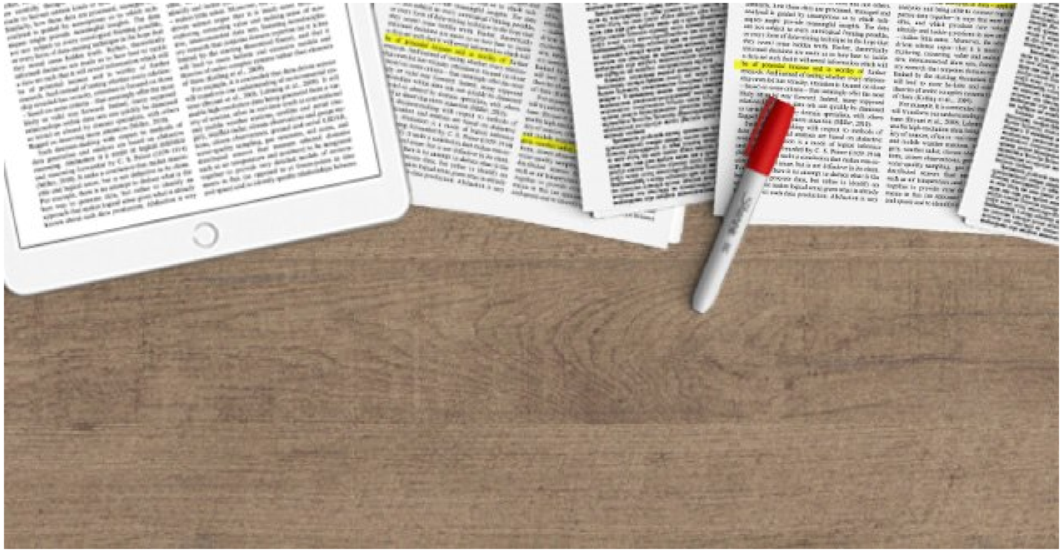
- Linked Open Data Best Practices (as defined by Heath and Bizer, 2011)

how well are they followed in practice?

1) Provide dereferencable URIs
2) Set RDF links pointing at other data sources
3) Use terms from widely deployed vocabularies
4) Make proprietary vocabulary terms dereferencable
5) Map proprietary vocabulary terms to other vocabularies
6) Provide provenance metadata
7) Provide licensing metadata
8) Provide data-set-level metadata
9) Refer to additional access methods

Studies of Best Practice Conformance

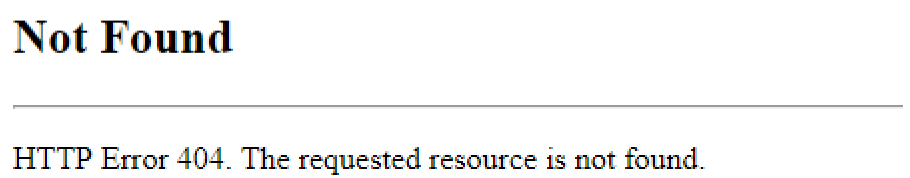
- Hogan et al.: *An empirical survey of Linked Data conformance,* 2012
  - top-level view
- Schmachtenberg et al.: *Adoption of the Linked Data Best Practicesin Different Topical Domains*, 2014
  - domain-specific view



1) Provide Derefencable URIs

- Metric: how many URIs used are actually derefencable?
  - i.e., do not link to HTTP 404 (possible bias: study time)
  - provide RDF

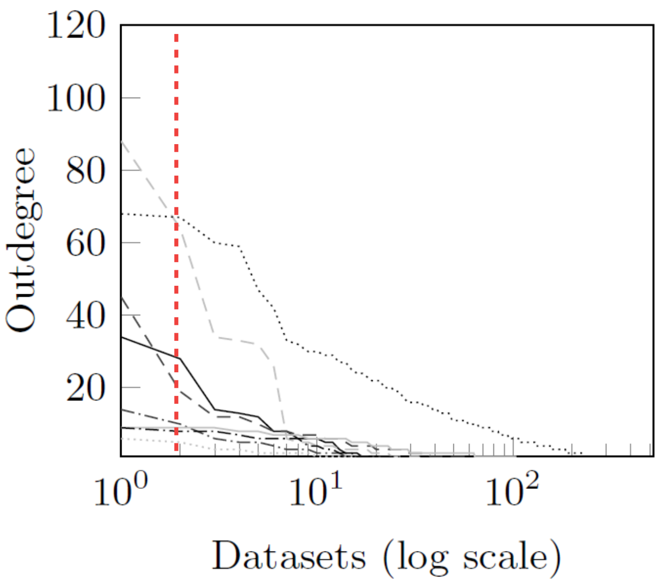
- Hogan et al.: ~70% of URIs are derefencable in above sense


Not Found

HTTP Error 404. The requested resource is not found.

2) Set RDF links pointing at other data sources

- Schmachtenberg et al.:
  - ~55% of all datasets link to at least one other dataset
  - There are some hubs as link targets
    - DBpedia (~200 datasets)
    - geonames.org (~140 datasets)
- Hogan et al.:
  - on average, a dataset links to 20.4 (±38.2) other datasets



2) Set RDF links pointing at other data sources

- Are all links owl:sameAs?
  - Schmachtenberg et al.: domain-specific differences

Table 3: Top three linking predicates per category. The percentage is relative to number of datasets within the category which set outgoing links

category	predicate	usage	category	predicate	usage
social networking	foaf:knows	59.87%	life sciences	owl:sameAs	57.69%
social networking	foaf:based_near	35.79%	life sciences	rdfs:seeAlso	38.46%
social networking	sioc:follows	34.11%	life sciences	dct:creator	19.23%
publications	owl:sameAs	32.20%	government	dct:publisher	47.12%
publications	dct:language	25.42%	government	dct:spatial	29.81%
publications	rdfs:seeAlso	23.73%	government	owl:sameAs	25.00%
user-generated content	owl:sameAs	52.94%	geographic	owl:sameAs	59.09%
user-generated content	rdfs:seeAlso	23.53%	geographic	skos:exactMatch	36.36%
user-generated content	dct:source	17.65%	geographic	skos:closeMatch	22.73%
media	owl:sameAs	76.47%	crossdomain	owl:sameAs	76.92%
media	rdfs:seeAlso	23.53%	crossdomain	rdfs:seeAlso	53.85%
media	foaf:based_near	17.65%	crossdomain	dct:creator	23.08%

3) Use terms from widely deployed vocabularies

- Schmachtenberg et al.: most used vocabularies

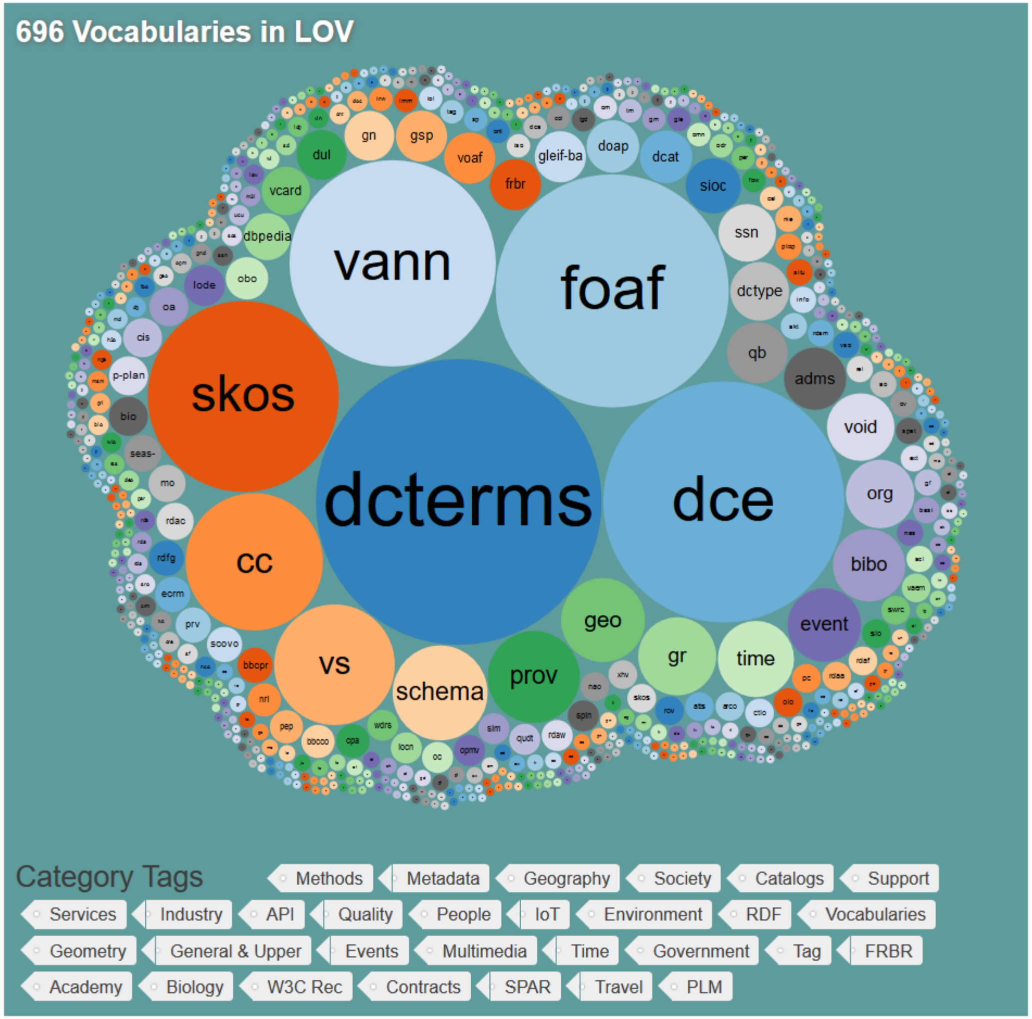
Table 5: Vocabularies used by more than 5% of all datasets.

prefix	occurrence	quota	prefix	occurrence	quota
rdf	1015	98.16%	void	137	13.25%
rdfs	740	71.57%	bio	125	12.09%
foaf	710	68.67%	cube	114	11.03%
dcterm	575	55.61%	rss	99	9.57%
owl	377	36.46%	odc	86	8.32%
wgs84	254	24.56%	w3con	77	7.45%
sioc	179	17.31%	doap	65	6.29%
admin	157	15.18%	bibo	64	6.19%
skos	145	14.02%	dcat	59	5.71%

- Hogan et al.: on average, 6.6k classes and properties are shared between at least two datasets

3) Use terms from widely deployed vocabularies

- Linked Open Vocabularies (LOV) project
  - analyze usage of vocabularies



4) Make proprietary vocabulary terms dereferencable

- Schmachtenberg et al.:
  - ~23% of all datasets use proprietary vocabularies
  - ~58% of all vocabularies are proprietary

Table 6: Proprietary vocabularies with dereferencability per category and quota of vocabularies linking to others

category	different prop. vocabs. used (% of all prop. vocab.)	# of datasets using prop. vocab. (% of all datasets)
social networking	128 (33.86%)	83 (15.99%)
publications	58 (15.34%)	35 (33.65%)
government	48 (12.70%)	35 (18.82%)
cross-domain	55 (14.55%)	16 (36.36%)
geographic	24 (6.34%)	16 (39.02%)
life sciences	35 (9.25%)	26 (29.21%)
media	22 (5.82%)	21 (56.76%)
user-gen. cnt.	30 (7.93%)	26 (47.27%)
Total	378 (58.24%)	241 (23.17%)

4) Make proprietary vocabulary terms dereferencable

- Schmachtenberg et al.:
  - less than 20% of all vocabularies are fully derefencable
- Common reasons:
  - use of deprecated terms
  - namespace hijacking

Table 6: Proprietary vocabularies with dereferencability per category and quota of vocabularies linking to others

category	different prop. vocabs. used (% of all prop. vocab.)	# of datasets using prop. vocab. (% of all datasets)	Dereferencability			#of vocabs linking(quota)
			full	partial	none	
social networking	128 (33.86%)	83 (15.99%)	16.41%	6.25%	77.78%	21 (16.41%)
publications	58 (15.34%)	35 (33.65%)	20.69%	6.90%	72.41%	14 (24.14%)
government	48 (12.70%)	35 (18.82%)	20.83%	12.50%	66.67%	16 (33.33%)
cross-domain	55 (14.55%)	16 (36.36%)	27.27%	10.91%	61.82%	14 (25.45%)
geographic	24 (6.34%)	16 (39.02%)	20.83%	4.17%	75.00%	5 (20.83%)
life sciences	35 (9.25%)	26 (29.21%)	28.57%	5.71%	65.71%	4 (11.43%)
media	22 (5.82%)	21 (56.76%)	0.00%	9.09%	90.91%	2 (9.09%)
user-gen. cnt.	30 (7.93%)	26 (47.27%)	13.33%	10.00%	76.67%	6 (20.00%)
Total	378 (58.24%)	241 (23.17%)	19.25%	8.00%	72.75%	78 (5.29%)

5) Map proprietary vocabulary terms to other vocabularies

- Schmachtenberg et al.:
  - only a small fraction of proprietary vocabularies are linked :-(

Table 7: Predicates used to link terms between different vocabularies.

term	% of vocabularies	term	% of vocabularies
rdfs:range	9.52%	rdfs:seeAlso	1.59%
rdfs:subClassOf	8.47%	owl:inverseOf	1.32%
rdfs:subPropertyOf	6.88%	owl:equivalentClass	1.32%
rdfs:domain	5.29%	swivt:type	1.06%
rdfs:isDefinedBy	3.70%	owl:equivalentProperty	0.79%

6) Provide provenance metadata

- Hogan et al.:
  - ~41% of all datasets provide (provenance) metadata
- Schmachtenberg et al.:
  - ~35% of all datasets provide provenance metadata
  - most used vocabulary is Dublin Core

Table 8: Provenance vocabulary usage and license vocabulary usage by category

Category	Any prov-vocab		Dublin Core	Admin	prv/prov
social networking	169	(32.56%)	56.21%	58.58%	1.18%
publications	39	(37.50%)	94.87%	5.13%	2.56%
government	77	(41.40%)	100.00%	0.00%	1.30%
life sciences	21	(23.60%)	100.00%	0.00%	2.56%
cross-domain	8	(18.18%)	100.00%	12.50%	0.00%
geographic	4	(9.76%)	100.00%	0.00%	25.00%
user-gen. content	11	(20.00%)	90.91%	54.55%	0.00%
media	5	(13.51%)	100%	0.00%	0.00%
Total	372	(35.77%)	28.37%	10.77%	0.77%

7) Provide licensing metadata

- Hogan et al.:
  - ~14% of all datasets provide licensing metadata
- Schmachtenberg et al.:
  - ~8% of all datasets provide licensing metadata

№	property	quads
1	xhtml:license	179,375
2	dc:licence	176,029
3	cc:license	59,790
4	dc:rights	7,007
5	sz:license_text	2,035
6	dbo:license	1,653
7	dct:licence	1,591
8	dbp:licence	383
9	wrcc:license	151
10	doap:license	92
–	dct:rights	23

Table 19
Top ten licencing properties according to use in our corpus

Table 8: Provenance vocabulary usage and license vocabulary usage by category

Category	Any prov-vocab		Dublin Core	Admin	prv/prov	Any license-vocab
social networking	169	(32.56%)	56.21%	58.58%	1.18%	5.20%
publications	39	(37.50%)	94.87%	5.13%	2.56%	3.85%
government	77	(41.40%)	100.00%	0.00%	1.30%	29.57%
life sciences	21	(23.60%)	100.00%	0.00%	2.56%	3.37%
cross-domain	8	(18.18%)	100.00%	12.50%	0.00%	11.36%
geographic	4	(9.76%)	100.00%	0.00%	25.00%	0.00%
user-gen. content	11	(20.00%)	90.91%	54.55%	0.00%	10.91%
media	5	(13.51%)	100%	0.00%	0.00%	5.41%
Total	372	(35.77%)	28.37%	10.77%	0.77%	7.85%

8) Provide data-set-level metadata

- Schmachtenberg et al.:
  - Issue: referral and discovery
  - methods: inline, link, /.well-known/void
  - in total, ~14% provide data-set-level metadata

Table 9: Percentage of datasets using the VoID vocabulary and percentage of datasets offering alternative access methods

Category	VOID	link	well-known	inline
social networking	6 (1.16%)	0.58%	0.19%	0.58%
publications	14 (13.46%)	6.73%	2.88%	5.77%
life sciences	29 (32.58%)	19.10%	4.49%	12.36%
government	75 (40.32%)	6.99%	3.23%	31.18%
user-gen. content	6 (10.91%)	5.45%	0.00%	5.45%
geographic	15 (36.59%)	14.63%	12.20%	12.20%
cross-domain	5 (11.36%)	9.09%	2.27%	2.27%
media	2 (5.41%)	2.70%	0.00%	2.70%
Total	140 (13.46%)	4.62%	1.44%	8.27%

9) Refer to additional access methods

- Schmachtenberg et al.:
  - SPARQL and dump download are rarely *referred to*
  - This does not mean that they don't exist...

Table 9: Percentage of datasets using the VoID vocabulary and percentage of datasets offering alternative access methods

Category	VOID	link	well-known	inline	alt. access	SPARQL	Dump
social networking	6 (1.16%)	0.58%	0.19%	0.58%	6 (1.16%)	1.16%	0.39%
publications	14 (13.46%)	6.73%	2.88%	5.77%	10 (10.58%)	9.62%	3.85%
life sciences	29 (32.58%)	19.10%	4.49%	12.36%	19 (21.35%)	20.22%	16.85%
government	75 (40.32%)	6.99%	3.23%	31.18%	61 (32.80%)	30.11%	30.65%
user-gen. content	6 (10.91%)	5.45%	0.00%	5.45%	3 (5.45%)	5.45%	1.82%
geographic	15 (36.59%)	14.63%	12.20%	12.20%	8 (19.51%)	12.20%	12.20%
cross-domain	5 (11.36%)	9.09%	2.27%	2.27%	4 (9.09%)	4.55%	6.82%
media	2 (5.41%)	2.70%	0.00%	2.70%	1 (2.70%)	0.00%	2.70%
Total	140 (13.46%)	4.62%	1.44%	8.27%	48 (5.89%)	4.54%	3.80%

9) Refer to additional access methods

- Study by Hertling & Paulheim (2013)
  - sample random URIs from large Linked Data corpus
  - try to discover a SPARQL endpoint, e.g., by
    - using /.well-known/void
    - using inline links
    - using external catalogs (!)

Table 1. Results on different strategies for finding SPARQL endpoints on 10,000 random URIs, reporting both the number of URIs for which *any* SPARQL endpoint was found, as well as the number of URIs for which a *valid* SPARQL endpoint was found. The numbers in parantheses denote the total number of endpoints found.

Strategy	Datahub Catalog	`/.well-known/void` (all)	`/.well-known/void` (standard)	Link to VoID
# found	7,389 (26,124)	110 (392)	94 (288)	9 (9)
# valid	1,375 (2,978)	53 (106)	53 (72)	0 (0)

More Indicators

- Zaveri et al.: *Quality Assessment for Linked Open Data: A Survey.* SWJ 7(1), 2016
  - also includes performance
  - latency, throughput, ...

Table 2

Data quality metrics related to accessibility dimensions (type QN refers to a quantitative metric, QL to a qualitative one).

Dimension	Abr	Metric	Description	Type
Availability	A1	accessibility of the SPARQL end-point and the server	checking whether the server responds to a SPARQL query [18]	QN
	A2	accessibility of the RDF dumps	checking whether an RDF dump is provided and can be downloaded [18]	QN
	A3	dereferenceability of the URI	checking (i) for dead or broken links i.e. when an HTTP-GET request is sent, the status code `404 Not Found` is not be returned (ii) that useful data (particularly RDF) is returned upon lookup of a URI, (iii) for changes in the URI i.e the compliance with the recommended way of implementing redirections using the status code `303 See Other` [18,30]	QN
	A4	no misreported content types	detect whether the HTTP response contains the header field stating the appropriate content type of the returned file e.g. `application/rdf+xml` [30]	QN
	A5	dereferenced forward-links	dereferenceability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [31]	QN
Licensing	L1	machine-readable indication of a license	detection of the indication of a license in the VoID description or in the dataset itself [18,31]	QN
	L2	human-readable indication of a license	detection of a license in the documentation of the dataset [18, 31]	QN
	L3	specifying the correct license	detection of whether the dataset is attributed under the same license as the original [18]	QN
Interlinking	I1	detection of good quality inter-links	(i) detection of (a) interlinking degree, (b) clustering coefficient, (c) centrality, (d) open sameAs chains and (e) description richness through sameAs by using network measures [25], (ii) via crowdsourcing [1,65]	QN
	I2	existence of links to external data providers	detection of the existence and usage of external URIs (e.g. using `owl:sameAs` links) [31]	QN
	I3	dereferenced back-links	detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object [31]	QN
Security	S1	usage of digital signatures	by signing a document containing an RDF serialization, a SPARQL result set or signing an RDF graph [13,18]	QN
	S2	authenticity of the dataset	verifying authenticity of the dataset based on a provenance vocabulary such as author and his contributors, the publisher of the data and its sources (if present in the dataset) [18]	QL
Performance	P1	usage of slash-URIs	checking for usage of slash-URIs where large amounts of data is provided [18]	QN
	P2	low latency	(minimum) delay between submission of a request by the user and reception of the response from the system [18]	QN
	P3	high throughput	(maximum) no. of answered HTTP-requests per second [18]	QN
	P4	scalability of a data source	detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [18]	QN

Linked Data Conformance vs. Quality

- So far, we've looked at conformance
  - i.e., following standards and best practices
  - technical dimension
  - can be evaluated automatically
- Quality
  - i.e., how complete/correct/… is the data
  - content dimension
  - hard to evaluate automatically

Quality of Knowledge Graphs

- Färber et al.: *Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO.* SWJ 9(1), 2018
  - internal validation
    - e.g., schema violations
  - proxy metrics
    - e.g., timeliness measured by frequency of dataset updates

      → does not necessarily imply more recent data
  - manual evaluation
    - e.g., semantic validity

Table 14
Framework with an example weighting which would be reasonable for a user setting as given in [30].

Dimension	Metric	DBpedia	Freebase	OpenCyc	Wikidata	YAGO	Example of User Weighting w_i
Accuracy	m_{synRDF}	1	1	1	1	1	1
	m_{synLit}	0.994	1	1	1	0.624	1
	$m_{semTriple}$	1	1	1	1	1	1
Trustworthiness	m_{graph}	0.5	0.5	1	0.75	0.25	1
	m_{fact}	0.5	1	0	1	1	2
	m_{NoVal}	0	1	0	1	0	1
Consistency	$m_{checkRestr}$	0	1	0	1	0	1
	$m_{conClass}$	0.875	1	0.999	1	0.333	1
	$m_{conRelat}$	0.991	0.45	1	0	0.992	1
Relevancy	$m_{Ranking}$	0	0	0	1	0	1
Completeness	$m_{cSchema}$	0.905	0.762	0.921	1	0.952	1
	m_{cCol}	0.402	0.425	0	0.285	0.332	1
	m_{cPop}	0.93	0.94	0.48	0.99	0.89	3
Timeliness	m_{Freq}	0.5	0	0.25	1	0.25	3
	$m_{Validity}$	0	1	0	1	1	1
	m_{Change}	0	1	0	0	0	1
Ease of understanding	m_{Descr}	0.704	0.972	1	0.9999	1	3
	m_{Lang}	1	1	0	1	1	2
	m_{uSer}	1	1	0	1	1	1
	m_{uURI}	1	0.5	1	0	1	2
Interoperability	m_{Reif}	1	0.5	0.5	0	0.5	1
	$m_{iSerial}$	1	0	0.5	1	1	2
	m_{extVoc}	0.61	0.108	0.415	0.682	0.134	2
	$m_{propVoc}$	0.15	0	0.513	0.001	0	1
Accessibility	m_{Deref}	1	0.437	1	0.414	1	2
	m_{Avai}	0.9961	0.9998	1	0.9999	0.7306	2
	m_{SPARQL}	1	0	0	1	1	1
	m_{Export}	1	1	1	1	1	0
	m_{Negot}	0.5	0	0	1	1	1
	m_{HTML_RDF}	1	1	0	1	1	0
	m_{Meta}	1	0	1	0	0	1
Licensing	$m_{macLicense}$	1	0	0	1	0	1
Interlinking	m_{Inst}	0.592	0.018	0.443	0	0.305	2
	m_{URIs}	0.929	0.954	0.894	0.957	0.956	1
Unweighted Average		0.708	0.605	0.498	0.738	0.625	
Weighted Average		0.718	0.575	0.516	0.742	0.646	

Issues with Automatic Evaluation

- Where to find a gold standard?
  - e.g., sample 1k population figures from DBpedia
  - check whether they are correct

- Open World Assumption
  - ~60% of all persons in DBpedia do not have a deathDate
  - so?

- …

Issues with Automatic Evaluation

- So, we need human experts!
  - however, human evaluation is often expensive
  - more complex problems
    are hard to specify as microtasks

amazon
mechanical turk

Regarding Amazon Mechanical Turk Project (HIT Type) 3SKITTYVZ4PF45V9L1Q331NN98X367

Von: [redacted]

Message from [redacted]

Worker ID: A2OEIM01BOUDEB
HIT Set ID: 3SKITTYVZ4PF45V9L1Q331NN98X367
HIT Title: Decide if two wiki pages describe the same thing
HIT Description: The wiki topics are Runescape(Gaming), Marvel (Comics) and Star Trek(TV)

hello, i believe you must be off a decimal point. you mean 1.50 not .15 right?

Greetings from Amazon Mechanical Turk,

The message above was sent by an Amazon Mechanical Turk user.
Please review the message and respond to it as you see fit.

Sincerely,
Amazon Mechanical Turk
https://requester.mturk.com

Example: Crowd Evaluation of DBpedia

- Acosta et al. *Detecting linked data quality issues via crowdsourcing: A DBpedia study.* Semantic web 9.3 (2018): 303-335.

About: **Lhoumois**
GO TO WIKIPEDIA ARTICLE: Lhoumois

1

Wikipedia The Free Encyclopedia	DBpedia	2 Type of Errors
elevation max m: 172 (3)	**elevation max m:** 172 Data type: Integer	☐Value ☐Data type ☐Link
Name: Lhoumois	**Name:** Lhoumois Data type: English	☐Value ☐Data type ☐Link
Type: *Not specified*	**Type:** populated place	☐Value ☐Data type ☐Link
arrondissement: Parthenay	**arrondissement:** Parthenay Data type: English	☐Value ☐Data type ☐Link
Label: *Not specified*	**Label:** Lhoumois Data type: French	☐Value ☐Data type ☐Link
Type: *Not specified*	**Type:** http://dbpedia.org/class/yago/Region108630985	☐Value ☐Data type ☐Link
Same As: *Not specified*	**Same As:** http://sws.geonames.org/6444136/	☐Value ☐Data type ☐Link

Example: Crowd Evaluation of DBpedia

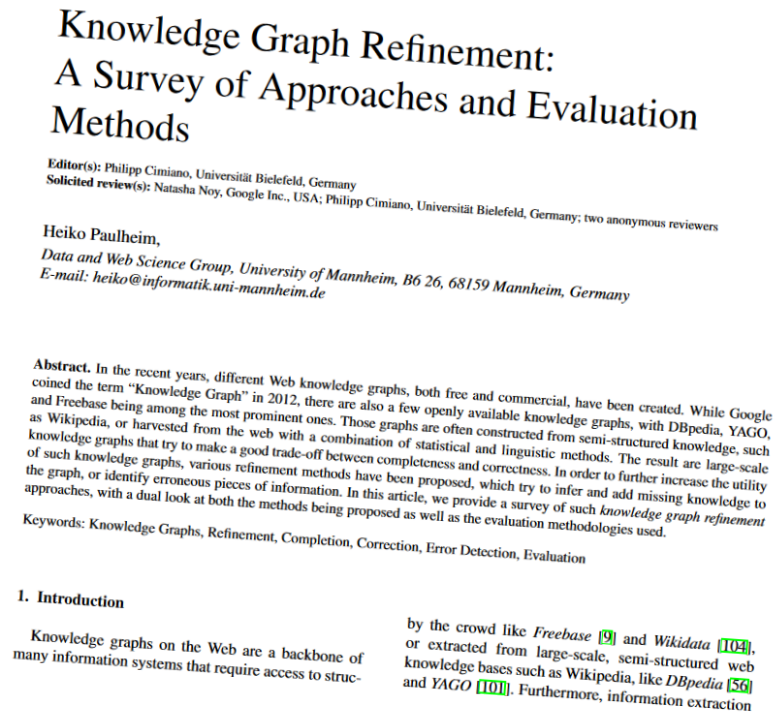
- Acosta et al. *Detecting linked data quality issues via crowdsourcing: A DBpedia study.* Semantic web 9.3 (2018): 303-335.

- From the paper: “Considering the HIT granularity, we paid 0.04 US dollar per 5 triples.”

- DBpedia (en): 176M statements
- Total cost of validation with this approach: 1.4M USD!

Intermediate Summary

- The Quality of Linked Open Data is far from perfect
  - conformance
  - content
- Improving the quality is an active field of research
  - Survey 2017: >40 approaches
  - since then: a lot of work
    in KG embeddings



Knowledge Graph Refinement:
A Survey of Approaches and Evaluation Methods

Editor(s): Philipp Cimiano, Universität Bielefeld, Germany
Solicited review(s): Natasha Noy, Google Inc., USA; Philipp Cimiano, Universität Bielefeld, Germany; two anonymous reviewers

Heiko Paulheim,
Data and Web Science Group, University of Mannheim, B6 26, 68159 Mannheim, Germany
E-mail: heiko@informatik.uni-mannheim.de

Abstract. In the recent years, different Web knowledge graphs, both free and commercial, have been created. While Google coined the term "Knowledge Graph" in 2012, there are also a few openly available knowledge graphs, with DBpedia, YAGO, and Freebase being among the most prominent ones. Those graphs are often constructed from semi-structured knowledge, such as Wikipedia, or harvested from the web with a combination of statistical and linguistic methods. The result are large-scale knowledge graphs that try to make a good trade-off between completeness and correctness. In order to further increase the utility of such knowledge graphs, various refinement methods have been proposed, which try to infer and add missing knowledge to the graph, or identify erroneous pieces of information. In this article, we provide a survey of such *knowledge graph refinement* approaches, with a dual look at both the methods being proposed as well as the evaluation methodologies used.

Keywords: Knowledge Graphs, Refinement, Completion, Correction, Error Detection, Evaluation

1. Introduction

Knowledge graphs on the Web are a backbone of many information systems that require access to struc- by the crowd like *Freebase* [9] and *Wikidata* [104], or extracted from large-scale, semi-structured web knowledge bases such as Wikipedia, like *DBpedia* [56] and *YAGO* [101]. Furthermore, information extraction

And now for something completely different

- Let’s jump back to the best practices one last time

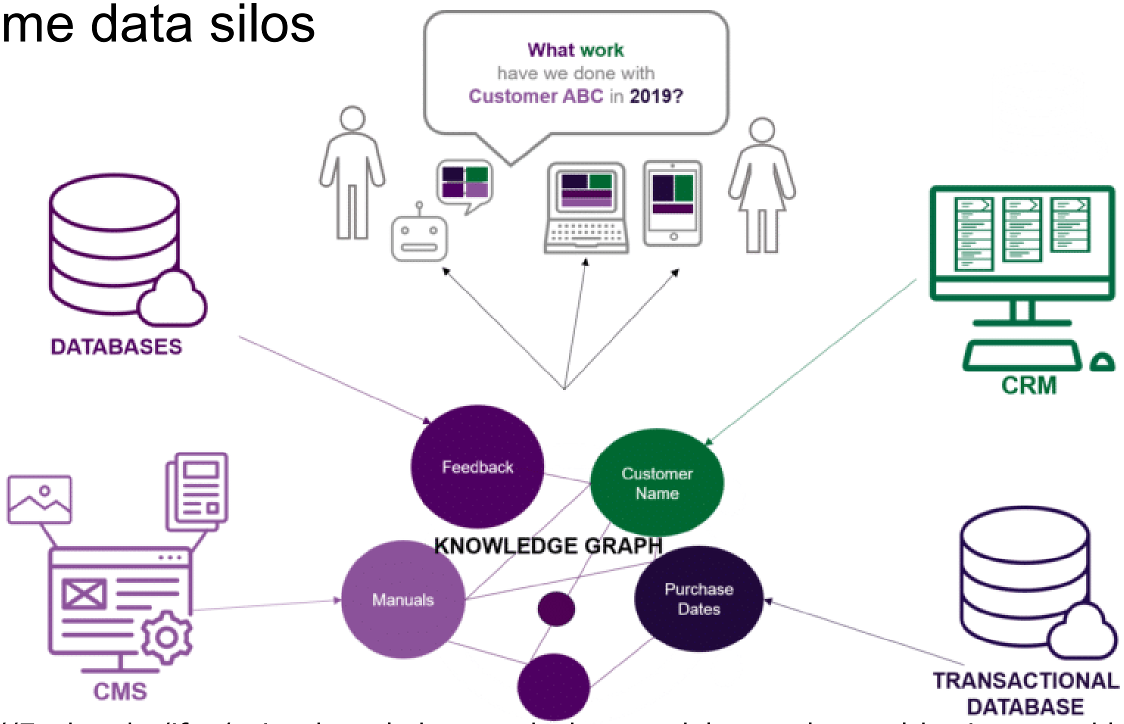
Previously on Knowledge Graphs

- Linked Open Data Best Practices
  (as defined by Heath and Bizer, 2011)

  1) Provide dereferencable URIs
  2) Set RDF links pointing at other data sources
  3) Use terms from widely deployed vocabularies
  4) Make proprietary vocabulary terms dereferencable
  5) Map proprietary vocabulary terms to other vocabularies
  6) Provide provenance metadata
  7) Provide licensing metadata
  8) Provide data-set-level metadata
  9) Refer to additional access methods

Previously on Knowledge Graphs

- Integrate data from different sources
- Make connections between entities in those sources
- Facilitate cross data source queries
- Overcome data silos



https://7wdata.be/ifttt/using-knowledge-graph-data-models-to-solve-real-business-problems/

Why do we need Links?

- Task:
  - Find contact data for Dr. Mark Smith
  - Input: various datasets

- Problems:
  - Every dataset uses its own identifiers (by design)
  - Every dataset may use its own vocabulary
  - Some reuse vocabularies, some don't

```
:p a foaf:Person .
:p foaf:name "Mark Smith" .
:p bar:profession bar:Physician .
...
```

Linked Data Set 1

```
:q a foo:Human .
:q foo:called "Mark Smith" .
:q foo:worksAs foo:MedDoctor .
...
```

Linked Data Set 2

How do we Create the Links?

- Technically, links can be added with OWL statements
- We know:
  – owl:sameAs, owl:equivalentClass, owl:equivalentProperty

```
:p a foaf:Person .
:p owl:sameAs foo:q .
:p foaf:name "Mark Smith" .
:p bar:profession bar:Physician .
bar:profession
  owl:equivalentProperty
  foo:worksAs .
bar:Physician owl:equiavlentClass
  foo:MedDoctor  .
...
```

Linked Data Set 1

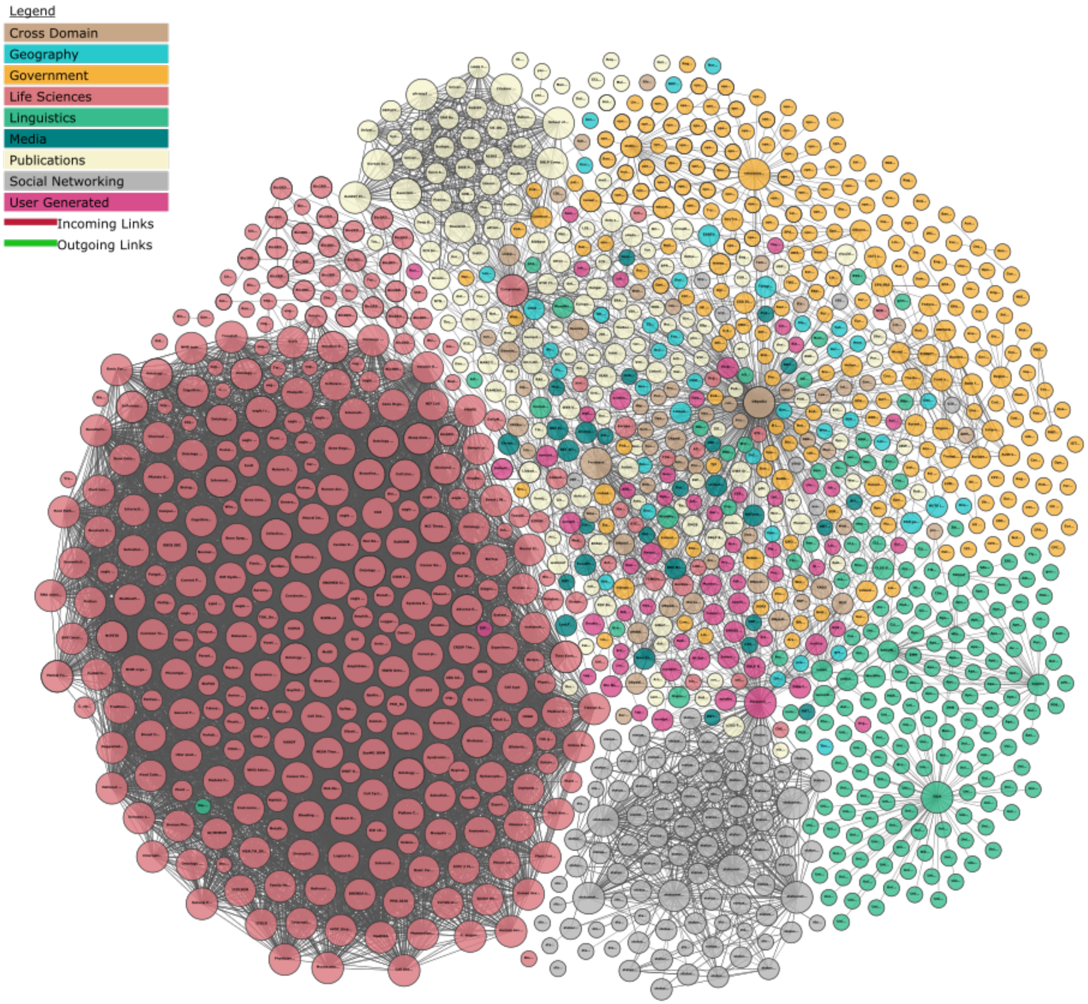
```
:q a foo:Human .
:q owl:sameAs bar:p .
:q foo:called "Mark Smith" .
foo:called
owl:EquivalentProperty
  foaf:name .
:q foo:worksAs foo:MedDoctor .
...
```

Linked Data Set 2

How do we Create the Links?

- Remember
  - The LOD cloud
  - >1,200 datasets

- Pairwise interlinking?



How do we Create the Links?

- Datasets with millions of entities...

Data Facts

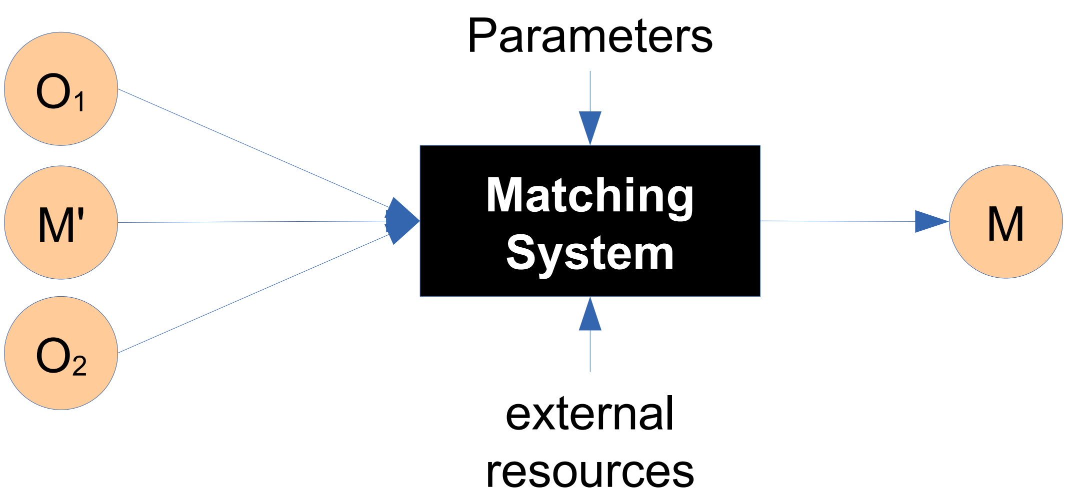
Total size	9,500,000,000 triples
Namespace	http://dbpedia.org/resource/
Links to 2000-us-census-rdf	12,529 triples
Links to dbtune-musicbrainz	22,981 triples
Links to education-data-gov-uk	1,697 triples
Links to eunis	3,600 triples
Links to flickr-wrappr	8,800,000 triples
Links to freebase	3,400,000 triples
Links to fu-berlin-dailymed	43 triples
Links to fu-berlin-dblp	196 triples
Links to fu-berlin-diseasome	1,943 triples
Links to fu-berlin-drugbank	729 triples
Links to fu-berlin-eurostat	137 triples
Links to fu-berlin-project-gutenberg	2,510 triples
Links to fu-berlin-sider	751 triples
Links to geonames-semantic-web	86,547 triples
Links to geospecies	15,972 triples
Links to italian-public-schools-linkedopendata-it	5,822 triples
Links to linkedgeodata	99,075 triples
Links to linkedmdb	13,800 triples
Links to nytimes-linked-open-data	10,359 triples
Links to opencyc	20,362 triples
Links to rdf-book-mashup	9,078 triples
Links to reference-data-gov-uk	22 triples
Links to revyu	6 triples

Tool Support

- A plethora of names
- Mostly used for schema level:
  - Ontology matching/alignment/mapping
  - Schema matching/mapping
- Mostly used for the instance level:
  - Instance matching/alignment
  - Interlinking
  - Link discovery

Automating Interlinking

- Given two input ontologies/datasets
  - And optional: a set of existing interlinks/mappings
- Provide a target set of interlinks/mappings



Automatic Interlinking

- Automatic interlinking is usually *heuristic*
  - i.e., not exact

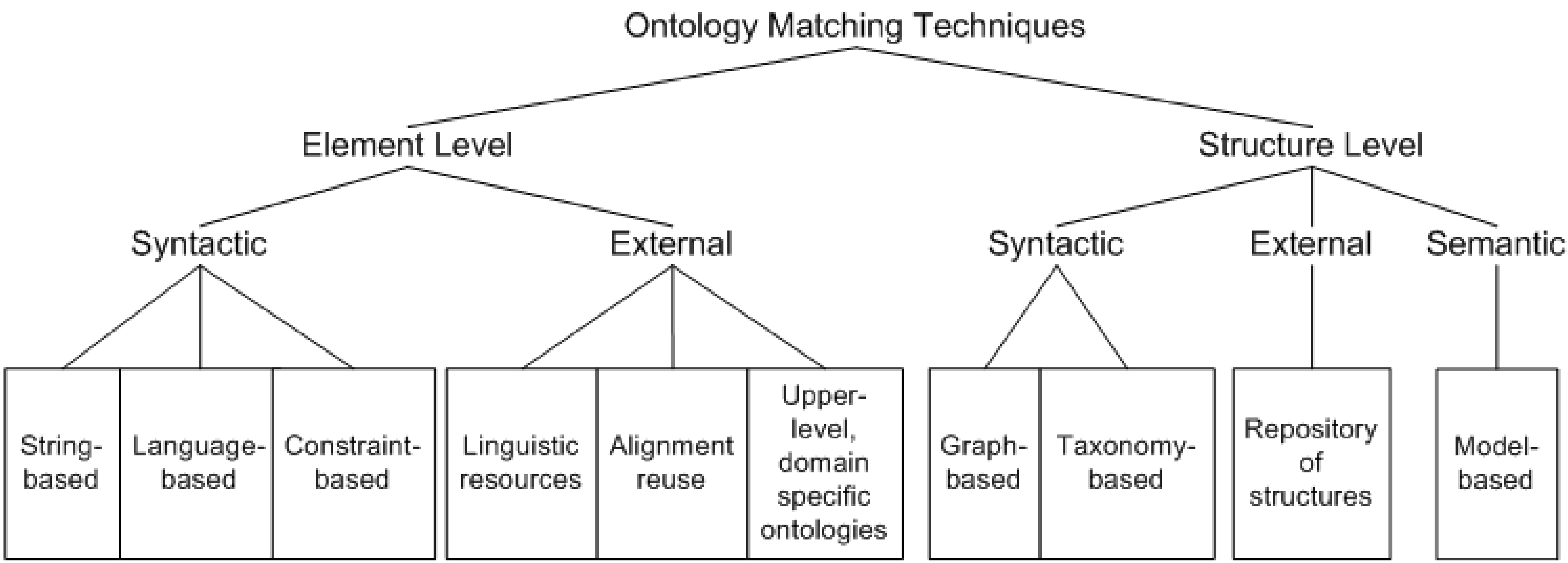
- Most approaches provide confidence scores
- General format: <e_1, e_2, relation, score>

  <dbpedia:University_of_Mannheim, wd:Q317070, owl:sameAs, 0.96>

- Relations may include
  - equality (owl:sameAs, owl:equivalentClass, owl:equivalentProperty)
  - specialization (rdfs:subClassOf, rdfs:subPropertyOf)
- Actively researched, but not yet finally solved
  - complex relations

Summary and Takeaways

- Over the years, a large variety of approaches has been developed



Euzenat & Shvaiko: Ontology Matching (2007)

Basic Interlinking Techniques

- Element vs. structural
  - Element level: only consider single elements in isolation
  - Structure based: exploit structure
    - e.g., class/property inheritance

- Syntactic vs. external vs. semantic
  - Syntactic: only use datasets themselves
  - External: use external sources of knowledge (e.g., dictionaries)
  - Semantic: exploit ontology semantics, e.g., by reasoning

Sources for Interlinking Signals

- Some datasets have “speaking” URIs, some don’t
  - http://dbpedia.org/resource/Germany, but
  - https://www.wikidata.org/wiki/Q183

- Most datasets have labels and textual descriptions
  - rdfs:label
  - skos:preferredLabel, skos:altLabel, …
  - rdfs:comment

- Proprietary string labels
  - dbo:abstract
  - https://www.wikidata.org/wiki/Property:P2561 (“name”)
  - ...

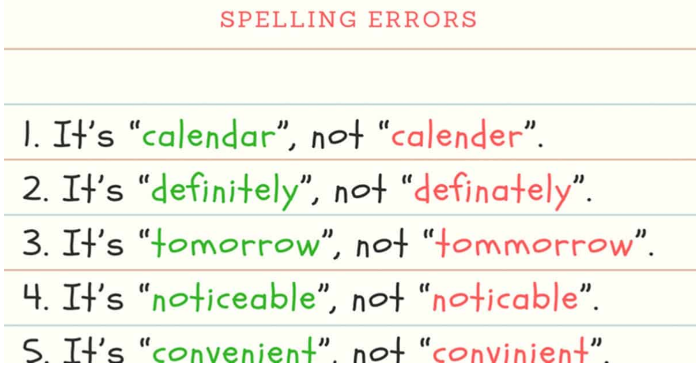
Simple String Based Metrics

- String equality
  - e.g. foo:University_of_Mannheim, bar:University_of_Mannheim
- Common prefixes
  - e.g. foo:United_States, bar:United_States_of_America
- Common postfixes
  - e.g. foo:Barack_Obama, bar:Obama

- Typical usage of prefixes/postfixes: |common|/max(length)
  - foo:United_States, bar:United_States_of_America → 12/22
  - foo:Barack_Obama, bar:Obama → 5/12

Edit Distance

- Notion: minimal number of basic edit operations needed to get from one string to the other
  - insert character
  - delete character
  - change character



- Can handle:
  - alternate spellings, small typos and variations
  - matches in different, but similar languages

- Example:
  - Universität Mannheim, University of Mannheim
  - Universitäy of Mannheim

    → edit distance 5/20 → similarity score = 3/4

N-gram based Similarity

- Problem: word order
  - e.g., University_of_Mannheim vs. Mannheim_University
  - prefix/postfix similarity: 0, edit distance similarity 5/11
- n-gram similarity
  - how many substrings of length n are common?
  - divided by no. of n-grams in longer string
- Example above with n=3
  - common: Uni, niv, ive, ver, ers, rsi, sit, ity, Man, ann, nnh, nhe, hei, eim
  - not common: ty_, y_o, _of, of_, f_M, _Ma, im_, m_U, _Un
- Similarity: 14/(14+9) = 14/25

Typical Preprocessing Techniques

- Unifying whitespace
  - University_of_Mannheim → University of Mannheim
  - UniversityOfMannheim → University Of Mannheim
- Unifying capitalization
  - University of Mannheim → university of mannheim
- Tokenization
  - university of mannheim → {university, of, mannheim}
  - similarity then becomes (average, maximum, …) similarity among token sets
  - also allows for other metrics, such as Jaccard overlap

Language-specific Preprocessing

- Stopword Removal
  - University of Mannheim → University Mannheim
- Stemming
  - German Universities → German Universit
  - Universities in Germany → Universit in German

- Usually, whole preprocessing pipelines are applied
  - e.g., stemming, stopword removal, tokenization, averaging the maximum edit distance similarity
- As above:
  - avg (max(similarity))({German, Universit}, {Universit, German}) = 1.0

Using External Knowledge

- e.g., linguistic resources (Wiktionary, BabelNet, ...)

Proper noun [edit]

New York

1. The largest city in New York State, a metr
   New York is a former capital of the US
2. A state of the United States of America ar
   The capital of New York is Albany, not
3. A county of New York State, coterminous

Synonyms [edit]

- (*state*): the Empire State, New York State, NY,
- (*city*): Big Apple (*informal*), New Amsterdam (

English · Arabic · Chinese · French · German · Greek · Hebrew · all preferred languages

• bn:00041611n • NOUN • Named Entity • Categories: 1624 establishments in North America, 1624 establishments in the Dutch Empire, 1898 establishments in New York, Cities in New York...

EN **New York City** · **New York** · **Greater New York** · **Big Apple** · **the five boroughs**

The largest city in New York State and in the United States; located in southeastern New York at the mouth of the Hudson river; a major financial and cultural center

WordNet *More definitions*

Lagos and New York City are both the largest cities in their respective countries. *Wiktionary*

More examples

IS A	metropolis • City • World city
PART OF	New York • New York metropolitan area
HAS PART	Bronx • Bronx-Whitestone Bridge • Brooklyn
CAPITAL OF	United States
CATALOG	vital articles level 3
CONTAINS ADMINISTRA...	Bronx • Brooklyn • Manhattan
COUNTRY	United States
DESCRIBED BY SOURCE	Brockhaus and Efron Encyclopedic Dictionary • Otto's encyclopedia • 1911 Encyclopedia Britannica

More relations

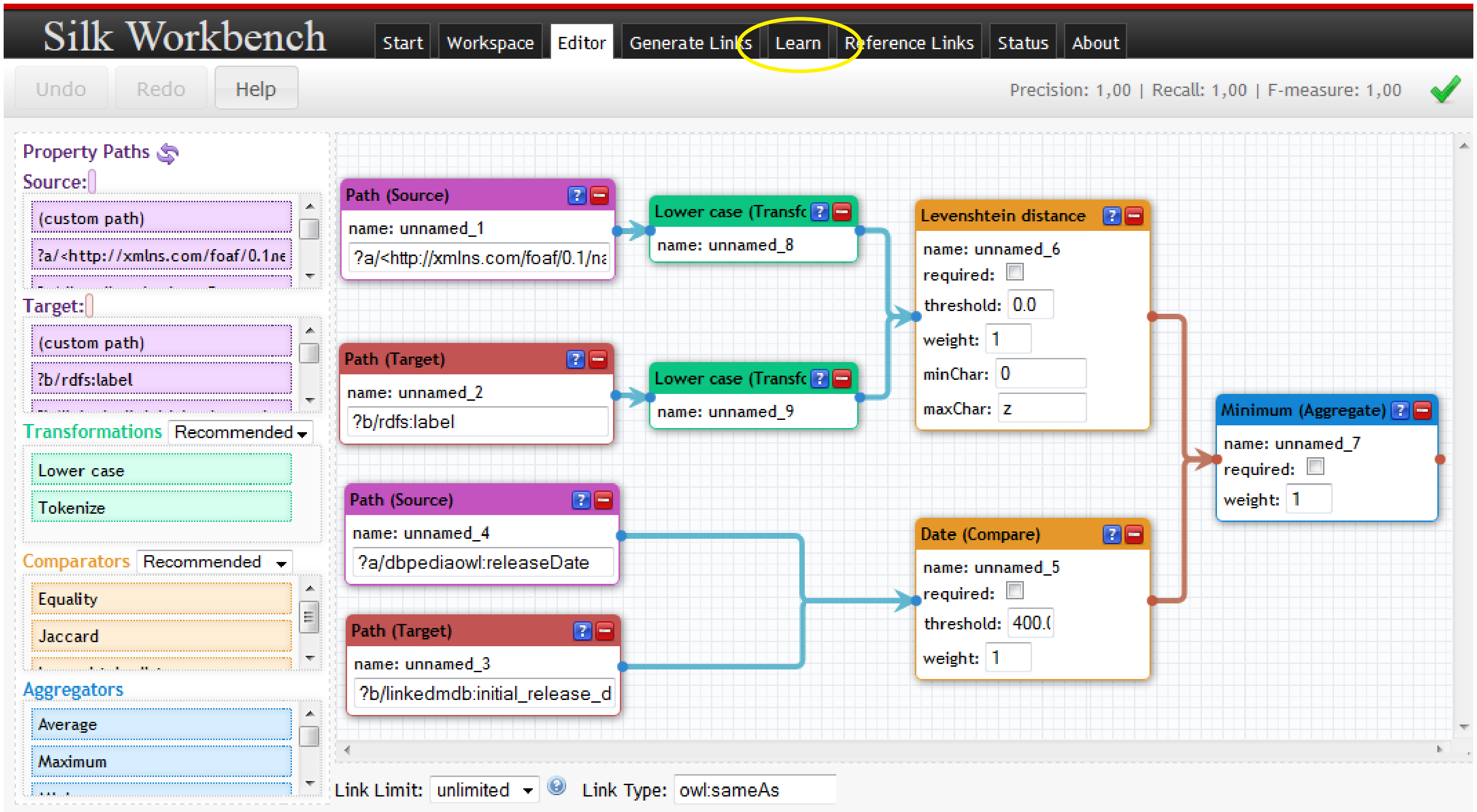
From Matching Literals to Matching Entities

- Exploiting properties
  - e.g., person: birth date
  - e.g., place: coordinates
  - e.g., movie: director
  - …

- Usually, a mix of measures
  - e.g., person: name similarity + equal birthdate
  - e.g., place: name similarity + coordinates w/in range
  - e.g., movie: name similarity + director name similarity
  - ...

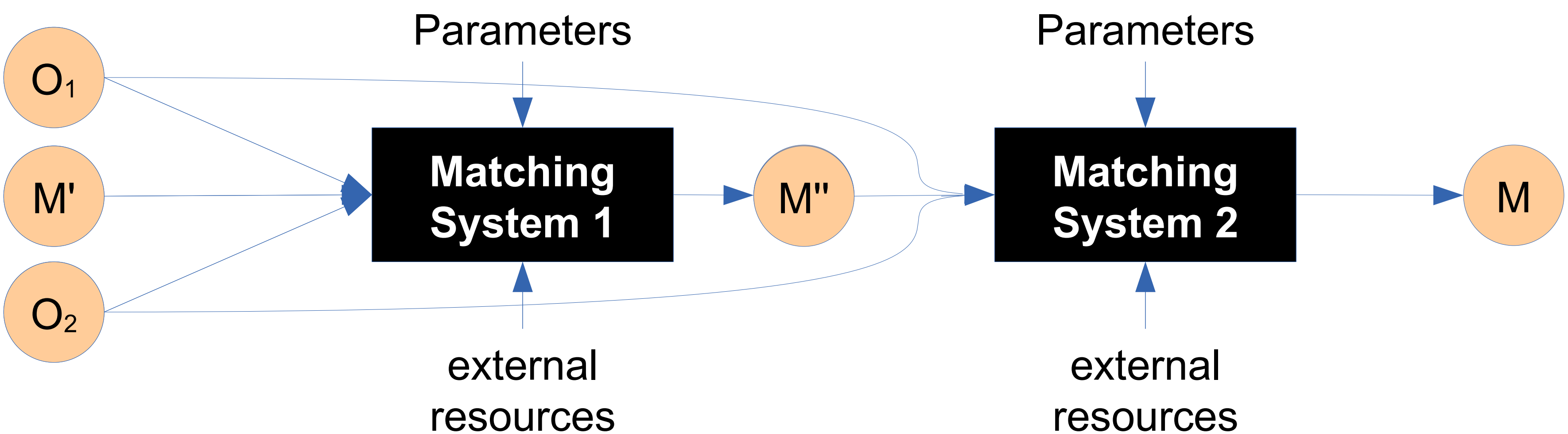
Preprocessing and Matching Pipelines

- Example tool: Silk Workbench



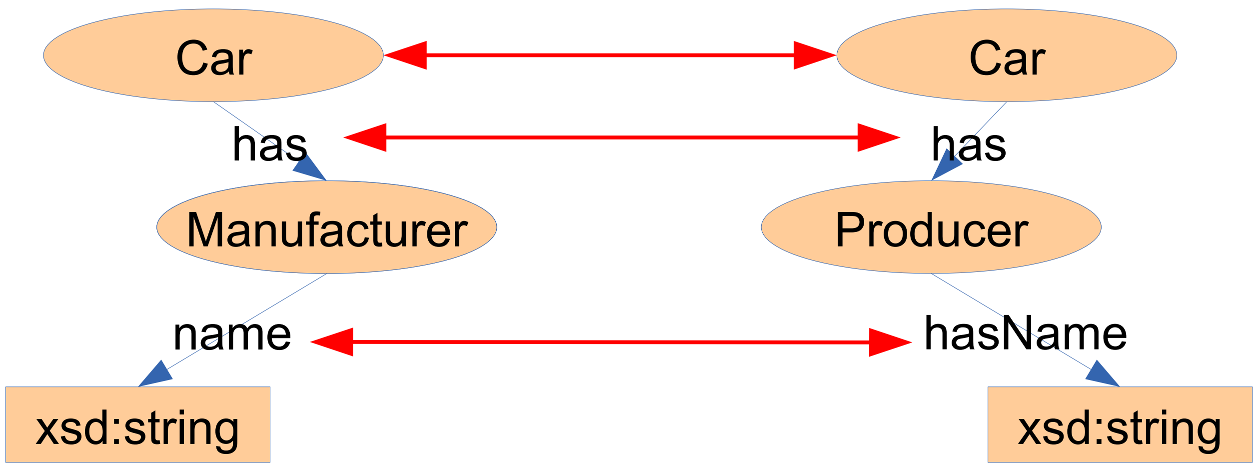
Schema Matching

- similar to interlinking
- typical approach: start with anchors based on string matching
- other signals
  - e.g., exploiting class/subclass similarity
  - e.g., exploiting property domain/range
  - using reasoning to determine validity



Schema Matching

- similar to interlinking
- typical heuristics include
  - classes appearing in the domain/range of matched properties are similar

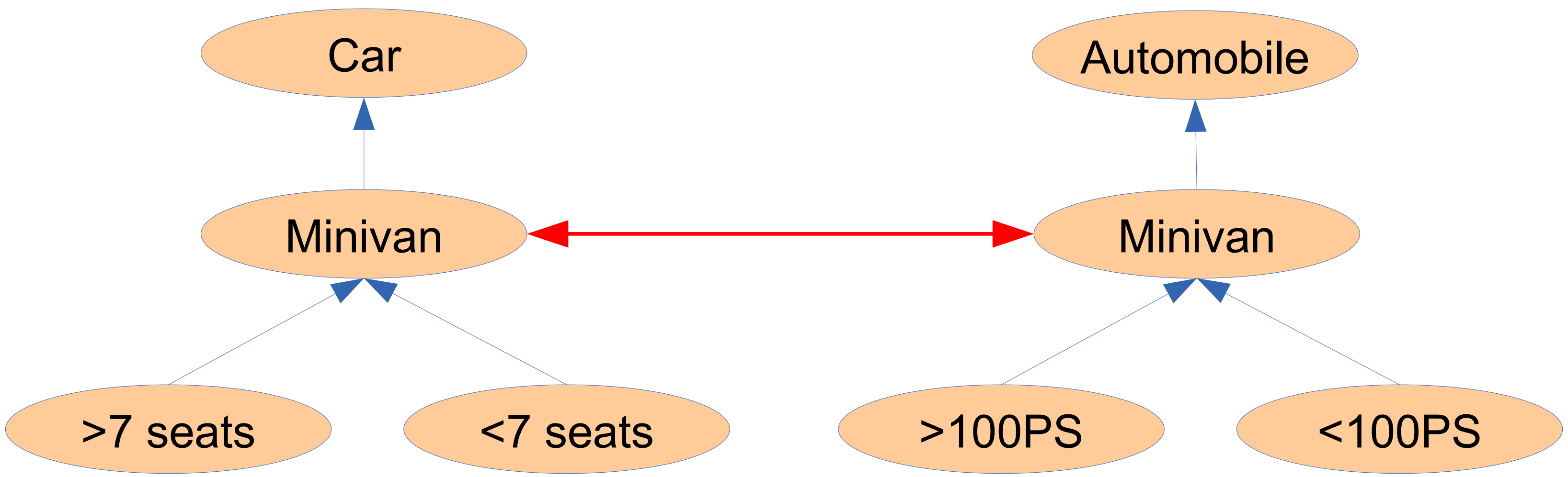


Schema Matching

- similar to interlinking
- typical heuristics include
  - properties having matched domains/ranges are similar

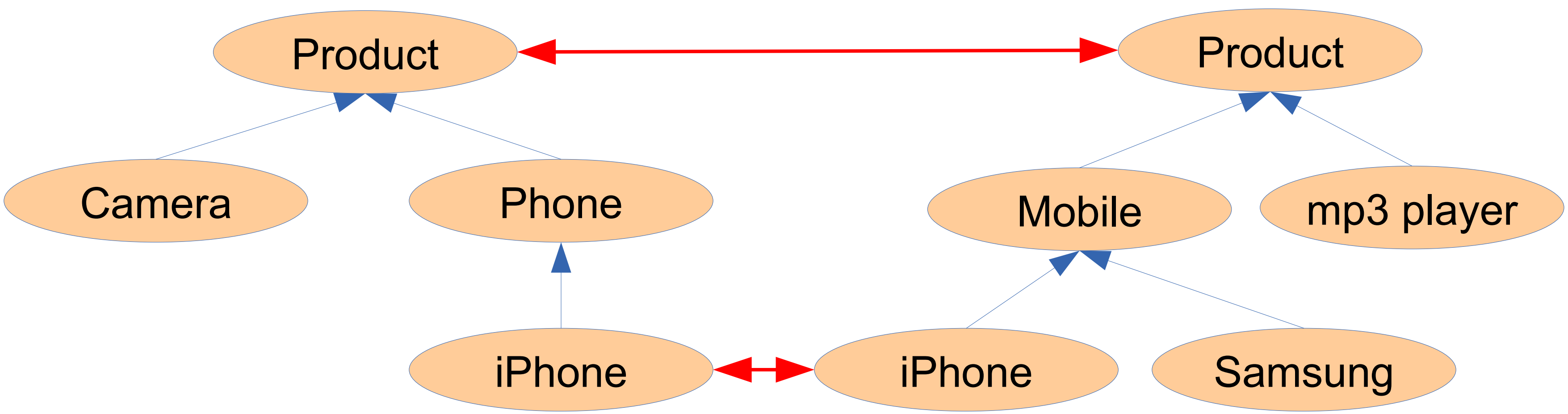
Schema Matching

- similar to interlinking
- typical heuristics include
  - superclasses of mapped classes are similar



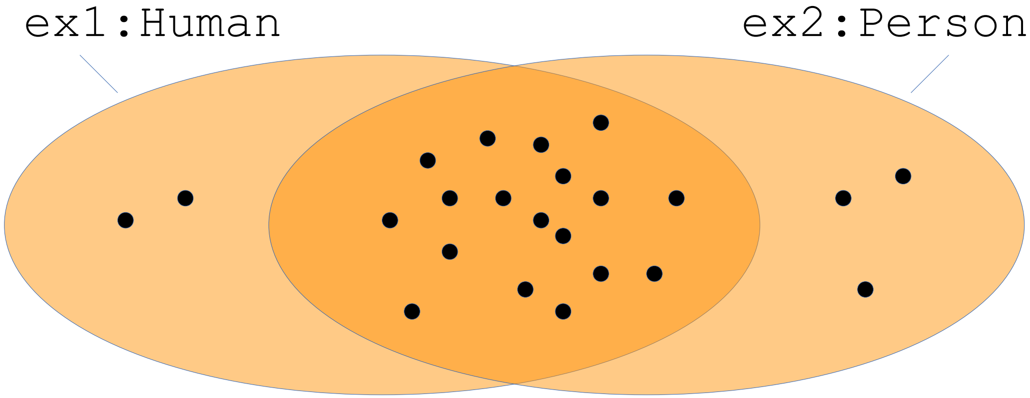
Schema Matching

- similar to interlinking
- typical heuristics include
  - pairs of classes along paths are similar (bounded path matching)



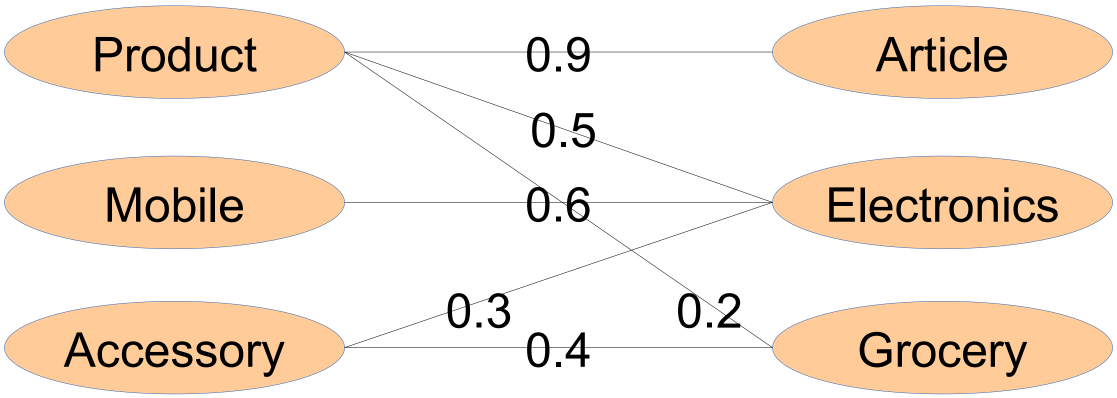
Instance based Matching

- Assumption: instances are already matched
  - either explicitly or heuristically
- Using, e.g., Jaccard
  - |ex1:Human ⊓ ex2:Person| / |x1:Human ⊔ ex2:Person|
  - example below: 18/23 → confidence ~0.78
- Finds non-trivial matches
  - e.g., dbpedia:Park ↔ yago:ProtectedArea



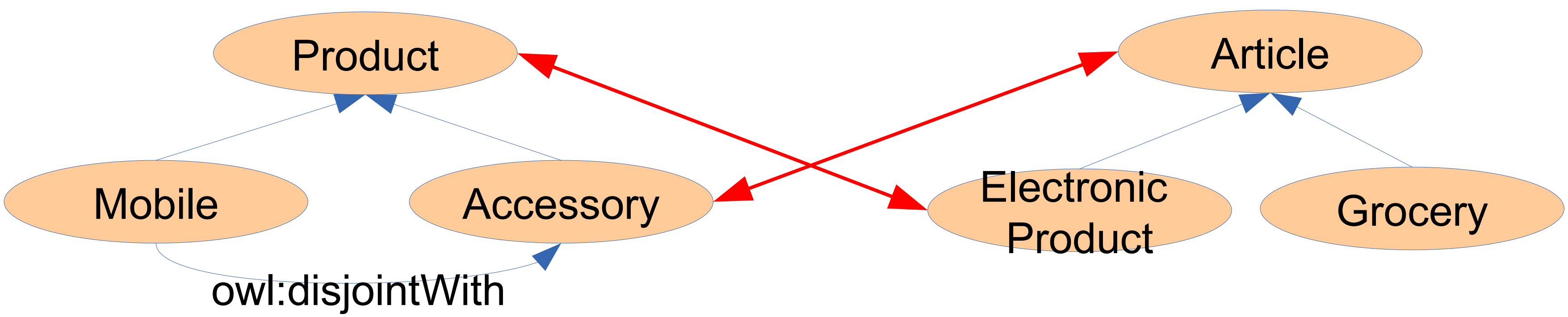
Enforcing 1:1 Mappings

- Assumption
  - each element can only be mapped to one other element
  - very often used in matching and linking
- Example:
  - stable marriage problem
  - try to find best matching partner for each element



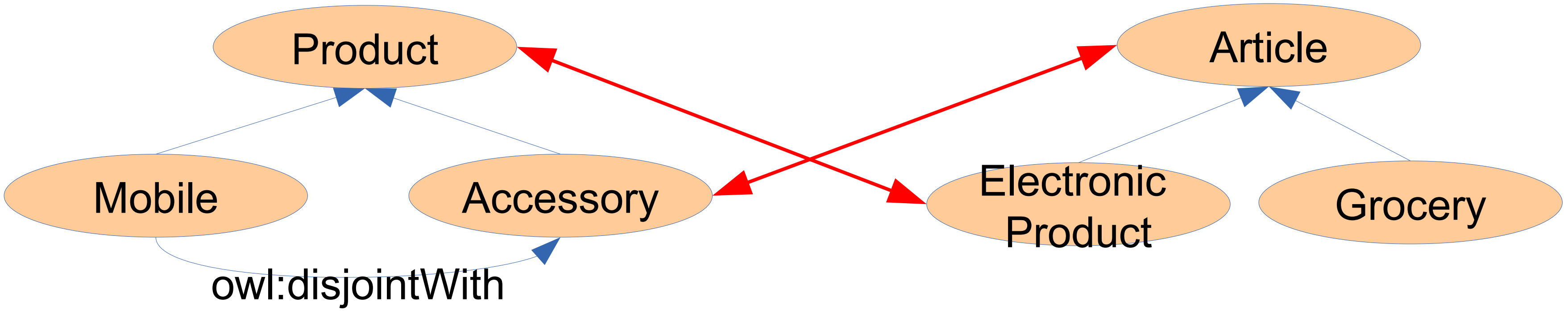
Schema Matching

- Refining a matching with reasoning
  - i.e., is the matching consistent with the ontology



Schema Matching

- Refining a matching with reasoning
  - i.e., is the matching consistent with the ontology




```
:Mobile
  rdfs:subClassOf :Product .
:Accessory
  rdfs:subClassOf :Product .
:Mobile owl:disjointWith
  :Accessory .
```

```
:ElectronicProduct
  rdfs:subClassOf :Article .
:Grocery
  rdfs:subClassOf :Article .
```

```
ex1:Product owl:equivalentClass
  ex2:ElectronicProduct.
ex1:Accessory
  owl:equivalentClass ex2:Article .
```

Reasoning on Mappings

- Reasoning:

```
ex1:Mobile rdfs:subClassOf ex1:Product .
+ ex1:Product owl:equivalentClass ex2:ElectronicProduct
.
→ ex1:Mobile rdfs:subClassOf ex2:ElectronicProduct .

+ ex2:ElektronicProduct rdfs:subClassOf ex2:Article .
→ ex1:Mobile rdfs:subClassOf ex2:Article .

+ ex2:Article owl:equivalentClass ex1:Accessory .
→ ex1:Mobile rdfs:subClassOf ex1:Accessory .
```

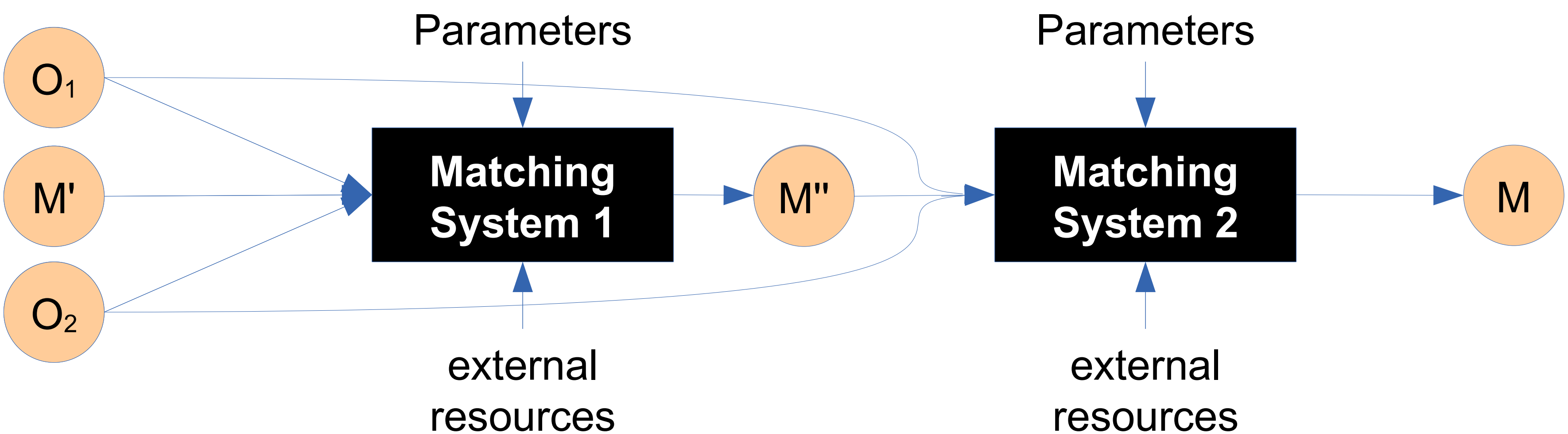
- And

```
ex1:Mobile owl:disjointWith ex1:Accessory .
```

- The mapping is contradictory!
  - Solution: remove a mapping element
  - e.g. by lowest confidence

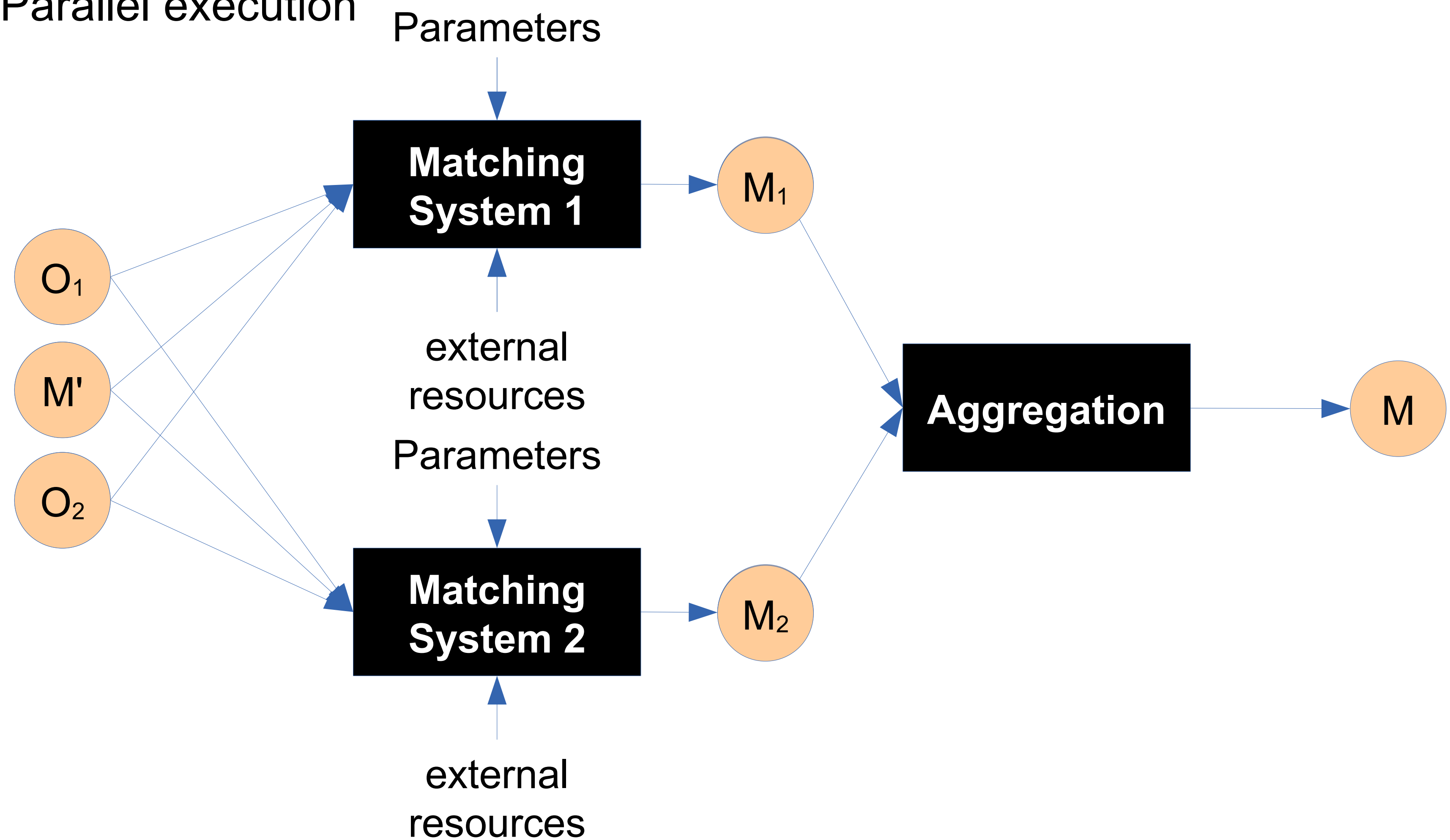
Matcher Combination

- Chaining



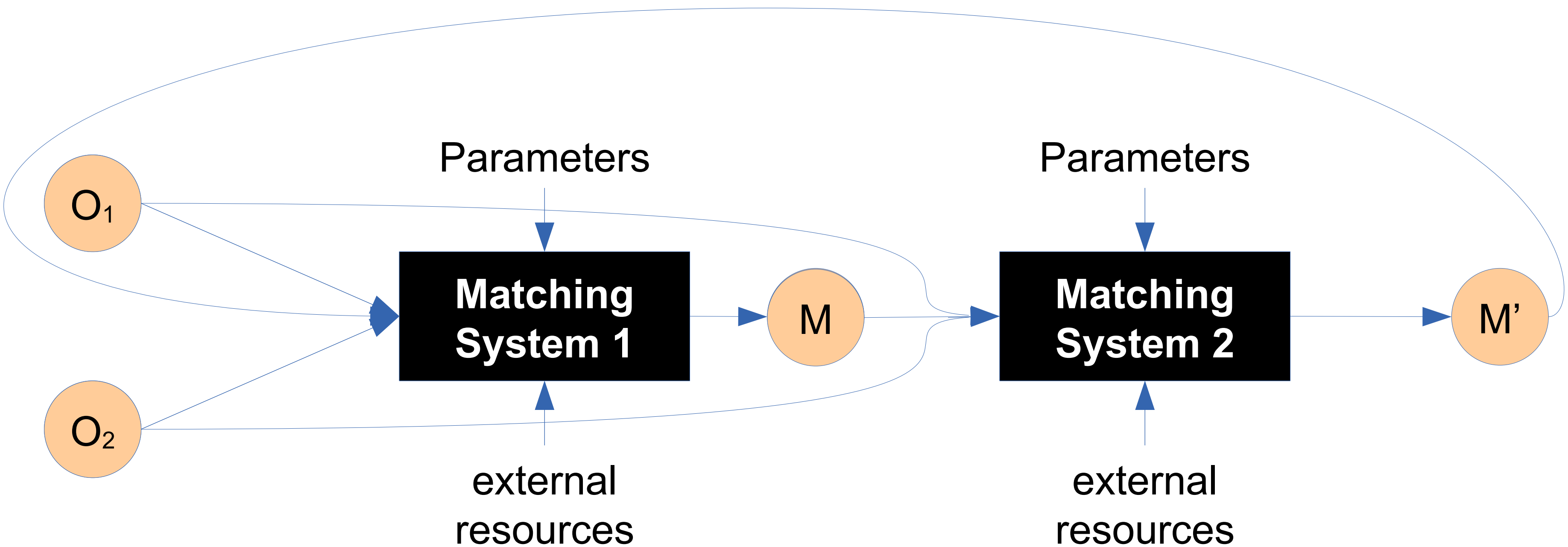
Matcher Combination

- Parallel execution



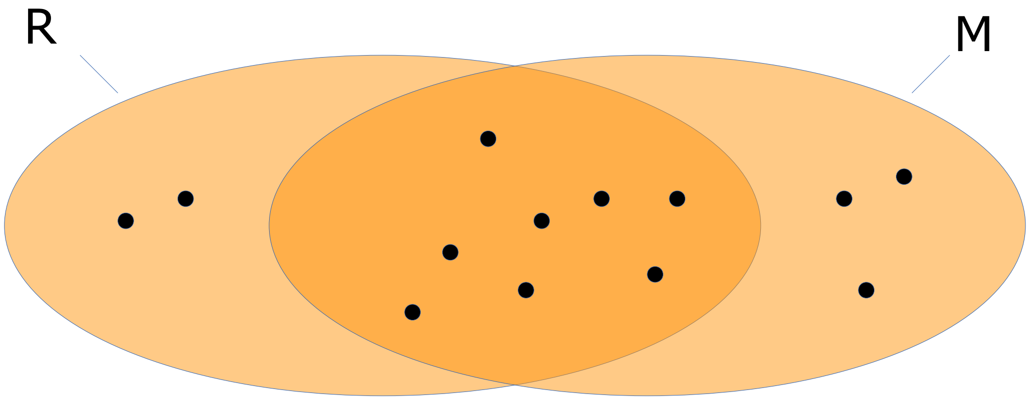
Matcher Combination

- Iterative execution



Evaluating Matchers

- Typical measures: recall, precision, F1
  - Scenario: reference alignment (gold standard) R, matcher found M
- Recall $r = |R \cap M| / |R|$
- Precision $p = |R \cap M| / |R|$
- F1 = harmonic mean of r and p
  - i.e., $2*r*p / (r+p)$



OAEI: an Annual Competition for Matching

- Different Tracks
  - started 2014
  - tracks usually repeated over the years
    - track progress in the field
- Different focus
  - domains
  - scalability
  - schema/instance
  - interactive

Track Example: Knowledge Graphs

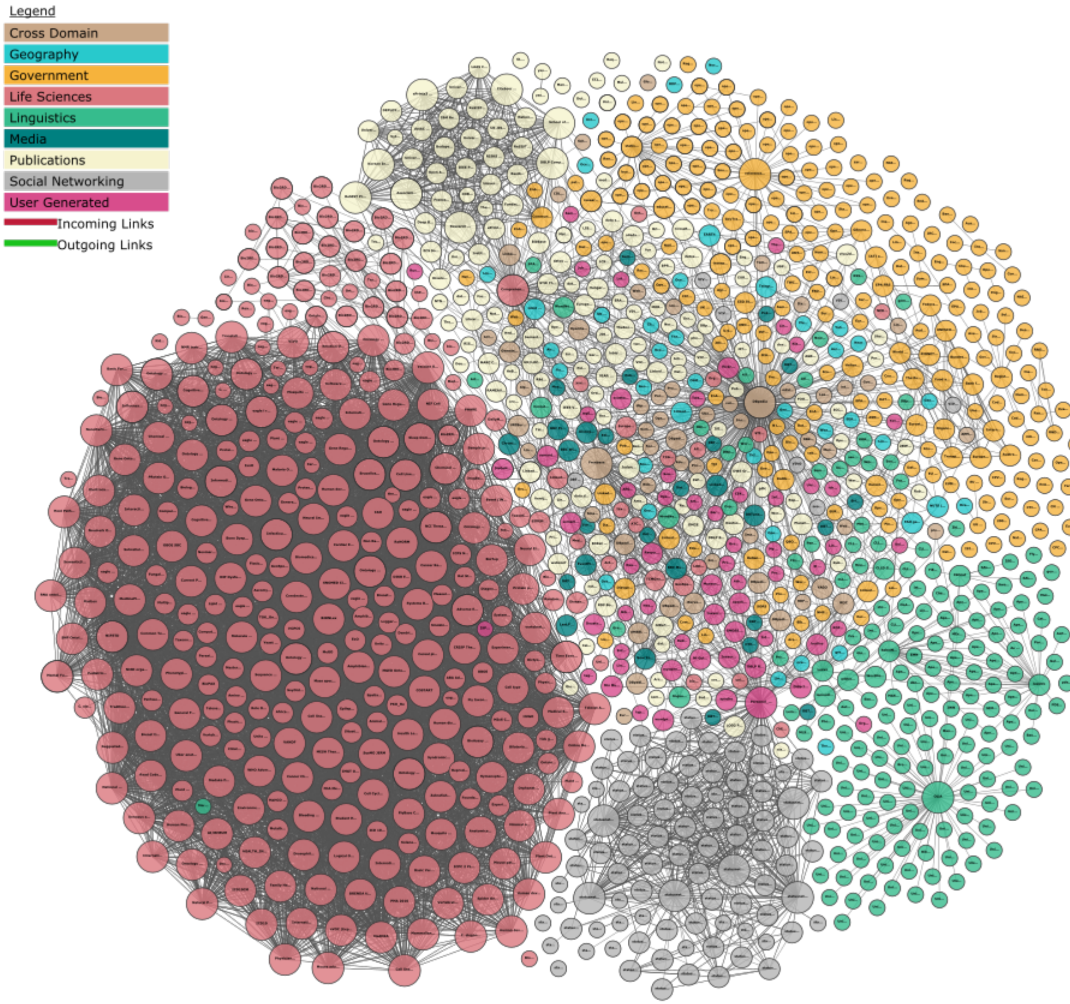
- Uses data from DBkWik
  - different graphs extracted from Wikis
  - (partial) gold standard: explicit links

			class				property				instance				overall			
System	**Time**	**#testcases**	**Size**	**Prec.**	**F-m.**	**Rec.**	**Size**	**Prec.**	**F-m.**	**Rec.**	**Size**	**Prec.**	**F-m.**	**Rec.**	**Size**	**Prec.**	**F-m.**	**Rec.**
ALOD2Vec	0:13:24	5	20.0	1.00 (1.00)	0.80 (0.80)	0.67 (0.67)	76.8	0.94 (0.94)	0.95 (0.95)	0.97 (0.97)	4893.8	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)	4990.6	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)
AML	0:50:55	5	23.6	0.98 (0.98)	0.89 (0.89)	0.81 (0.81)	48.4	0.92 (0.92)	0.70 (0.70)	0.57 (0.57)	6802.8	0.90 (0.90)	0.85 (0.85)	0.80 (0.80)	6874.8	0.90 (0.90)	0.85 (0.85)	0.80 (0.80)
ATBox	0:16:22	5	25.6	0.97 (0.97)	0.87 (0.87)	0.79 (0.79)	78.8	0.97 (0.97)	0.96 (0.96)	0.95 (0.95)	4858.8	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)	4963.2	0.89 (0.89)	0.85 (0.85)	0.81 (0.81)
baselineAltLabel	0:10:57	5	16.4	1.00 (1.00)	0.74 (0.74)	0.59 (0.59)	47.8	0.99 (0.99)	0.79 (0.79)	0.66 (0.66)	4674.8	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)	4739.0	0.89 (0.89)	0.84 (0.84)	0.80 (0.80)
baselineLabel	0:10:44	5	16.4	1.00 (1.00)	0.74 (0.74)	0.59 (0.59)	47.8	0.99 (0.99)	0.79 (0.79)	0.66 (0.66)	3641.8	0.95 (0.95)	0.81 (0.81)	0.71 (0.71)	3706.0	0.95 (0.95)	0.81 (0.81)	0.71 (0.71)
DESKMatcher	0:13:54	5	91.4	0.76 (0.76)	0.71 (0.71)	0.66 (0.66)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	3820.6	0.94 (0.94)	0.82 (0.82)	0.74 (0.74)	3912.0	0.93 (0.93)	0.81 (0.81)	0.72 (0.72)
LogMap	2:55:14	5	24.0	0.95 (0.95)	0.84 (0.84)	0.76 (0.76)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	29190.4	0.40 (0.40)	0.54 (0.54)	0.86 (0.86)	29214.4	0.40 (0.40)	0.54 (0.54)	0.84 (0.84)
LogMapBio	4:35:29	5	24.0	0.95 (0.95)	0.84 (0.84)	0.76 (0.76)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	24.0	0.95 (0.95)	0.01 (0.01)	0.00 (0.00)
LogMapIM	2:49:34	5	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	29190.4	0.40 (0.40)	0.54 (0.54)	0.86 (0.86)	29190.4	0.40 (0.40)	0.54 (0.54)	0.84 (0.84)
LogMapKG	2:47:51	5	24.0	0.95 (0.95)	0.84 (0.84)	0.76 (0.76)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	29190.4	0.40 (0.40)	0.54 (0.54)	0.86 (0.86)	29214.4	0.40 (0.40)	0.54 (0.54)	0.84 (0.84)
LogMapLt	0:07:19	4	23.0	0.80 (1.00)	0.56 (0.70)	0.43 (0.54)	0.0	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	6653.8	0.73 (0.91)	0.67 (0.84)	0.62 (0.78)	6676.8	0.73 (0.92)	0.66 (0.83)	0.61 (0.76)
Wiktionary	0:30:12	5	22.4	1.00 (1.00)	0.80 (0.80)	0.67 (0.67)	80.0	0.94 (0.94)	0.95 (0.95)	0.97 (0.97)	4893.8	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)	4996.2	0.91 (0.91)	0.87 (0.87)	0.83 (0.83)

Aggregated results per matcher, divided into class, property, instance, and overall alignments. Time is displayed as HH:MM:SS. Column *#testcases* indicates the number of testcases where the tool is able to generate (non empty) alignments. Column *size* indicates the averaged number of system correspondences. Two kinds of results are reported: (1) those not distinguishing empty and erroneous (or not generated) alignments, and (2) those considering only non empty alignments (value between parenthesis).

Track Example: Knowledge Graphs

- Generally, performance is high (F1>0.9) on many OAEI tracks
- So, what keeps us from interlinking the entire LOD cloud?
  - Performance is one issue, but...



The Golden Hammer Bias

- Challenge:
  - OAEI setup expect two **related** KGs
  - in the general case,
    this cannot be taken for granted
    - manual pre-inspection for
      every pair is infeasible
  - Experiments with unrelated KGs:

	mcu lyrics		memoryalpha lyrics		starwars lyrics	
Matcher	matches	precision	matches	precision	matches	precision
AML	2,642	0.12	7,691	0.00	3,417	0.00
baselineAltLabel	588	0.44	1,332	0.02	1,582	0.04
baselineLabel	513	0.54	1,006	0.06	1,141	0.06
FCAMap-KG	755	0.40	2,039	0.14	2,520	0.02
LogMapKG	29,238	0.02	-	-	-	-
LogMapLt	2,407	0.08	7,199	0.00	2,728	0.04
Wiktionary	971	0.12	3,457	0.02	4,026	0.00

See: ESWC 2020 Paper on OAEI Knowledge Graph Track

Challenges in Matching

- Usage of external resources
  - which are useful for which task? automatic selection?
  - embeddings?
- Automatic matcher combination & parameterization
  - analogy: AutoML
- Scalability
  - more or less solved for large **pairs**
  - open for large number of datasets
- Robustness
  - almost all of the OAEI tasks have a positive outcome bias (aka as “Golden Hammer Bias”)

Summary and Takeaways

- Data Quality on the Semantic Web
  - Conformance and Content
  - Both have weak spots
  - An active research area

- Matching
  - Schema and instance matching
  - Typical measures, heuristics, preprocessing
  - Still: no one size fits all matcher
    - we are far from full automation
    - deep learning and embeddings
      have also not brought the ultimate weapon

Recommendations for Upcoming Semesters

- Information Retrieval and Web Search (FSS), Prof. Ponzetto
- Web Mining (FSS), Prof. Bizer

- Web Data Integration (HWS), Prof. Bizer
- Relational Learning (HWS), Prof. Stuckenschmidt
- Text Analytics (HWS), Prof. Strohmaier

Coming up Next

- ...your presentations!
- Mind the submission of your reports (9.12.)
- Prepare for (max) 10 minutes presentation + 2 minutes questions
  - present a consistent story
  - focus on key issues and lessons learned
  - demonstrations are appreciated
    - but make sure you stick to the time limit!

Questions?