

Instance Matching Benchmarks for Linked Data

<http://www.ics.forth.gr/isl/BenchmarksTutorial/>

Evangelia Daskalaki,
Institute of Computer Science – FORTH , Greece

Tzanina Saveta,
Institute of Computer Science – FORTH , Greece

Irini Fundulaki,
Institute of Computer Science – FORTH , Greece

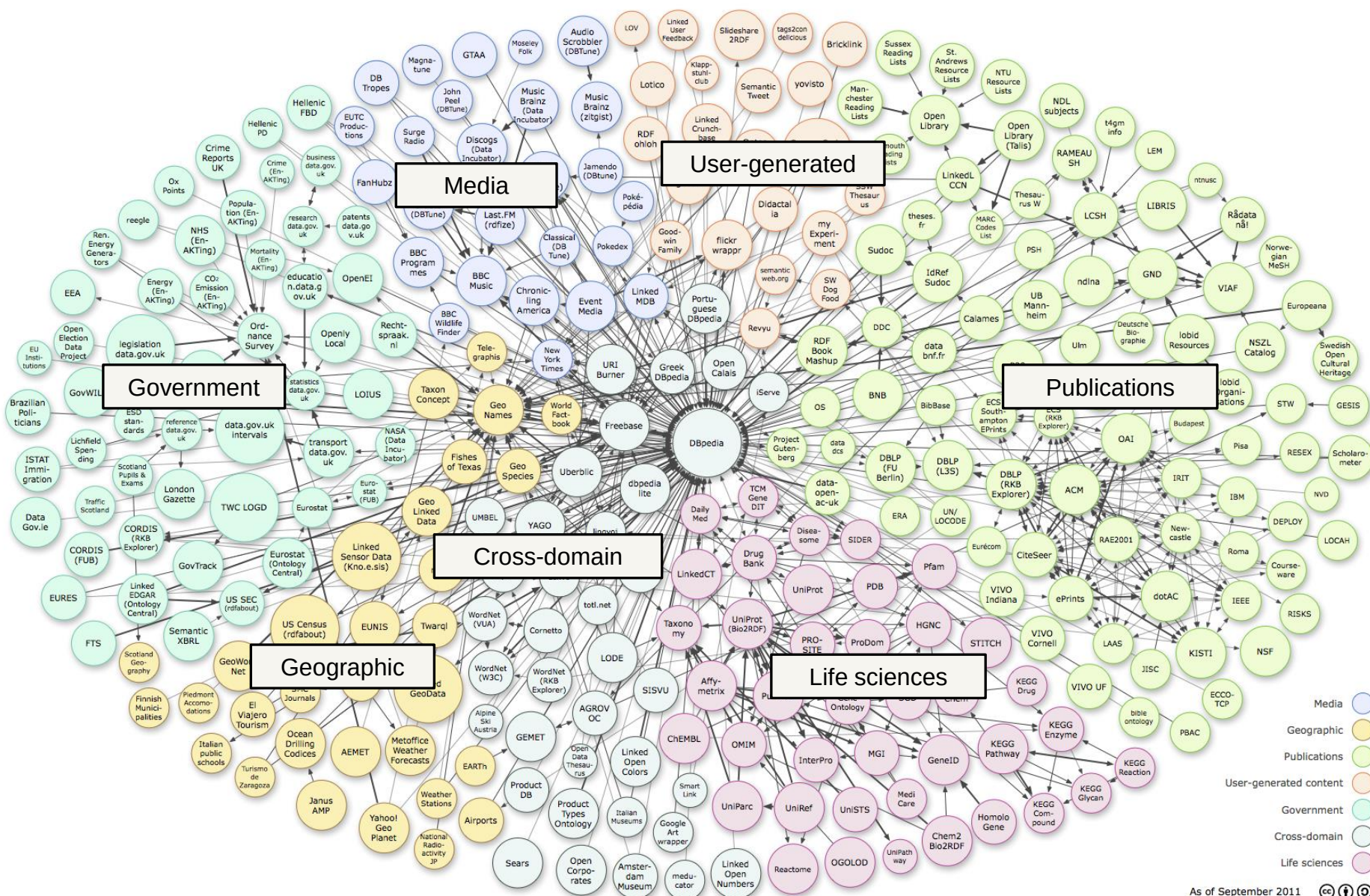
Melanie Herschel,
Inria

Teaser Slide

- We will talk about Benchmarks
- Benchmarks are generally a set of tests to assess computer systems' performances
- Specifically we will talk about: **Instance Matching (IM) Benchmark for Linked Data.**

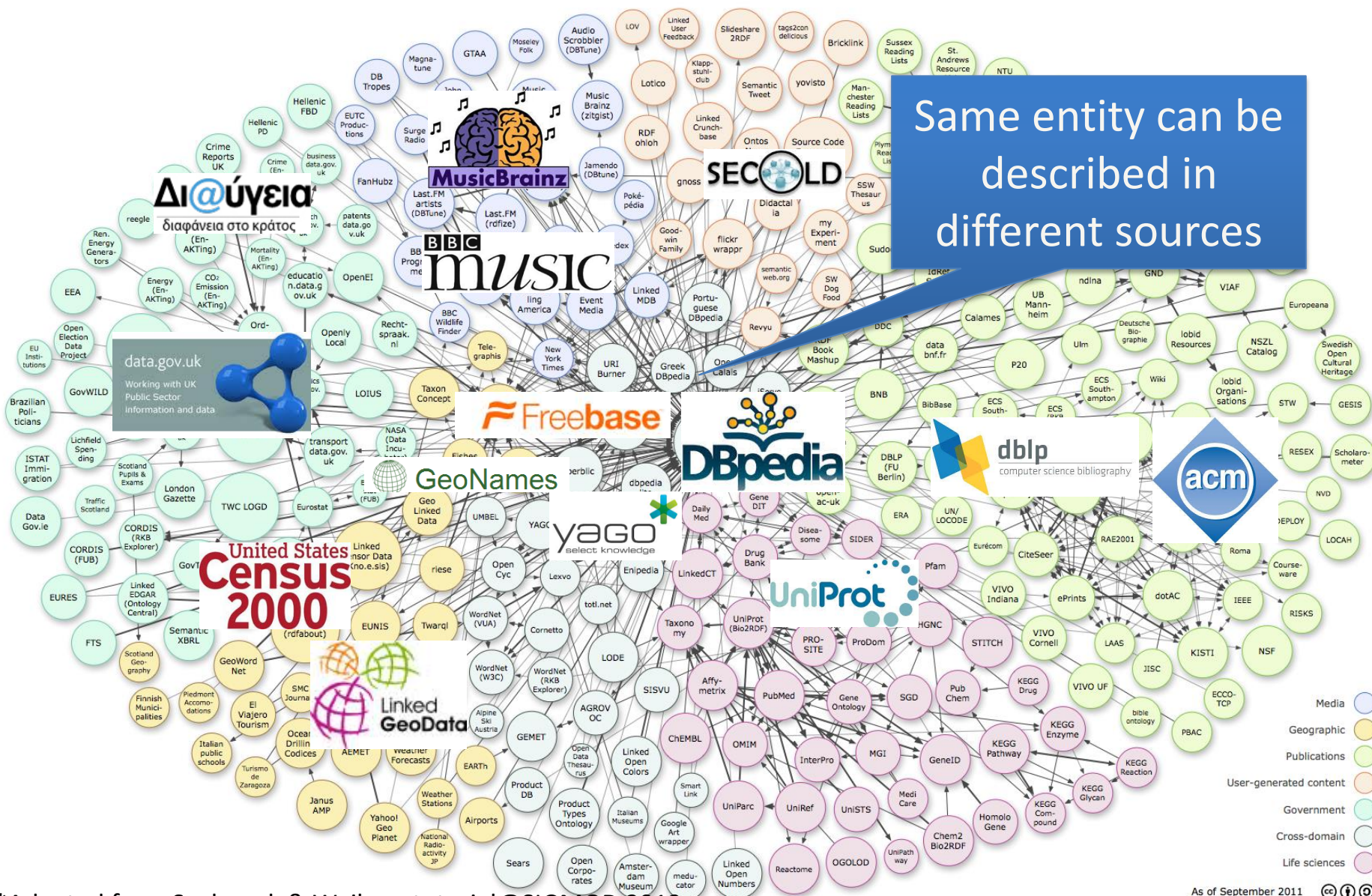
- **Introduction into Linked Data**
- Instance Matching
- Benchmarks for Linked Data
 - Why Benchmarks?
 - Benchmarks Characteristics
 - Benchmarks Dimensions
- Benchmarks in the literature
 - Synthetic Benchmarks
 - Real Benchmarks
 - Isolated Benchmarks
- Outcomes & Conclusions

Linked Data - The LOD Cloud



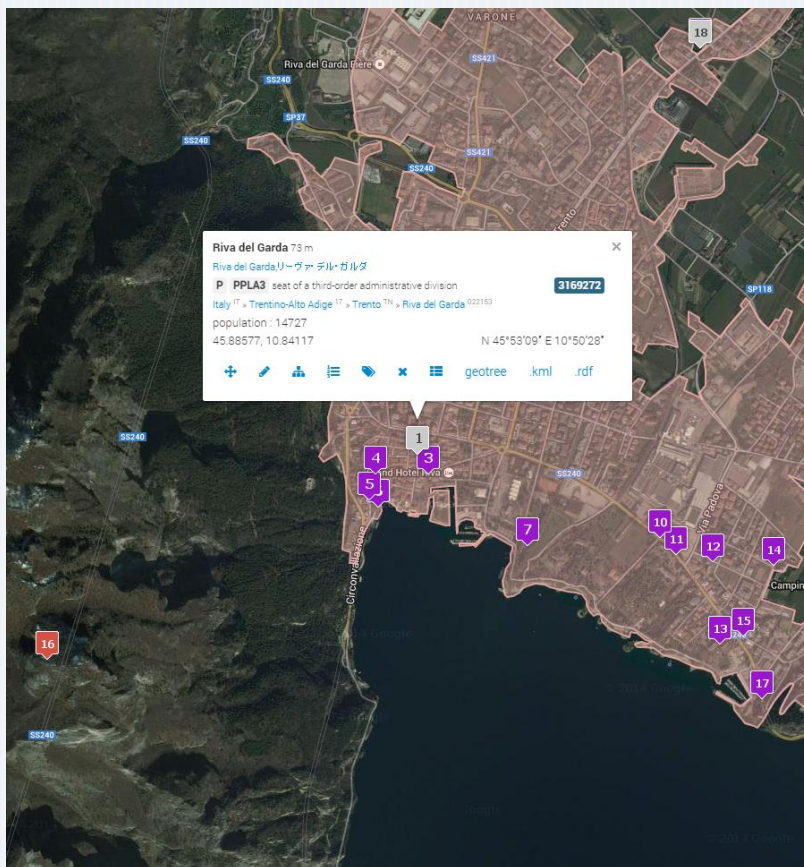
Linked Data – The LOD Cloud

Same entity can be described in different sources



Different Descriptions of Same Entity in Different Sources

"Riva del Garda description in GeoNames"



"Riva del Garda description in DBPedia"

About: Riva del Garda

An Entity of Type : [municipality](#), from Named Graph : <http://dbpedia.org>, within Data Space : [dbpedia.org](#)

Riva del Garda is a town and comune in the northern Italian province of Trento of the Trentino Alto Adige region. It is also known simply as Riva. The estimated pop

Property	Value
dbpedia-owl:PopulatedPlace/areaTotal	42.0
dbpedia-owl:abstract	Riva del Garda is a town and comune in the northern Italian province of Trento of the Trentino Alto
dbpedia-owl:areaCode	0464
dbpedia-owl:areaTotal	42000000.000000 (xsd:double)
dbpedia-owl:country	dbpedia:Italy
dbpedia-owl:elevation	70.000000 (xsd:double)
dbpedia-owl:populationTotal	15151 (xsd:integer)
dbpedia-owl:postalCode	38066
dbpedia-owl:province	dbpedia:Trento
dbpedia-owl:region	dbpedia:Trentino-Alto_Adige/Südtirol
dbpedia-owl:thumbnail	http://commons.wikimedia.org/wiki/Special:FilePath/Gardesana01.jpg?width=300
dbpedia-owl:twinCity	dbpedia:Bensheim
dbpedia-owl:twinCountry	dbpedia:Germany
dbpedia-owl:wikiPageExternalLink	http://www.discoverlakegarda.net/riva-lake-garda.asp http://www.rivadelgardaweb.com/en http://www.gardatrentino.it/en/riva-del-garda http://www.altogardameteo.com http://www.comune.rivadelgarda.tn.it
dbpedia-owl:wikiPageID	2724897 (xsd:integer)
dbpedia-owl:wikiPageRevisionID	596924125 (xsd:integer)
dbpprop:areaCode	464 (xsd:integer)
dbpprop:areaTotalKm	42 (xsd:integer)
dbpprop:coordinatesDisplay	title
dbpprop:coordinatesType	region:IT_type:city
dbpprop:day	0001-08-15 (xsd:date)
dbpprop:elevationM	70 (xsd:integer)
dbpprop:frazioni	Campi, Pregasina, Rifugio Capanna Grassi, Rifugio Nino Pernici, San Nazzaro, Sant'Alessandro, V
dbpprop:hasPhotoCollection	http://wifo5-03.informatik.uni-mannheim.de/flickrwrappi/photos/Riva_del_Garda
dbpprop:imageCaption	Riva del Garda view with Lake Garda on the background
dbpprop:imageSkyline	Gardesana01.jpg
dbpprop:latd	45 (xsd:integer)
dbpprop:latm	53 (xsd:integer)
dbpprop:latns	N
dbpprop:longd	10 (xsd:integer)
dbpprop:longew	E
dbpprop:longm	51 (xsd:integer)
dbpprop:mayor	Adalberto Mosaner
dbpprop:name	Riva del Garda
dbpprop:officialName	Comune di Riva del Garda
dbpprop:populationAsOf	2004 (xsd:integer)

- Introduction into Linked Data
- **Instance Matching**
- Benchmarks for linked Data
 - Why Benchmarks?
 - Benchmarks Characteristics
 - Benchmarks Dimensions
- Benchmarks in the literature
 - Benchmarks with synthetic dataset
 - Benchmarks with real dataset
 - Individually created Benchmarks
- Outcomes & Conclusions

Instance Matching: the cornerstone for Linked Data



open/social data



data integration

How can we automatically recognize
multiple mentions of the same entity
across or within sources?

=

Instance Matching



data acquisition



data
evolution

Instance Matching

- Problem has been considered for more than half a decade in Computer Science [EIV07]
- Traditional instance matching over relational data (known as record linkage)

Title	Genre	Year	Director
Troy	Action	2004	Petersen
contradiction missing value			
Troj	History		Petersen

Nicely and homogeneously structured data.
→ Value variations

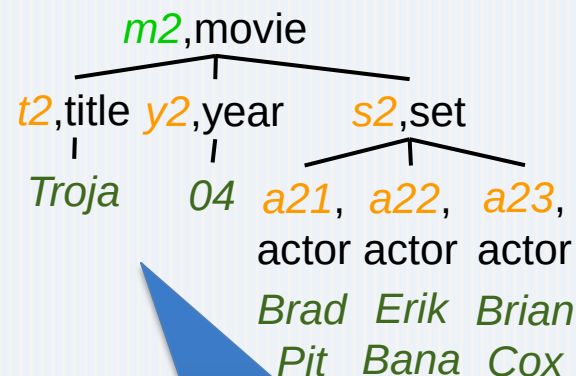
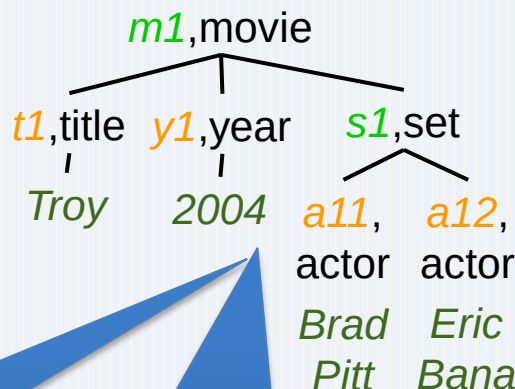
Dense data.

Typically few sources compared

Web Data Instance Matching

« The Early Days »

- IM algorithms for semi-structured XML model used to represent and exchange data.

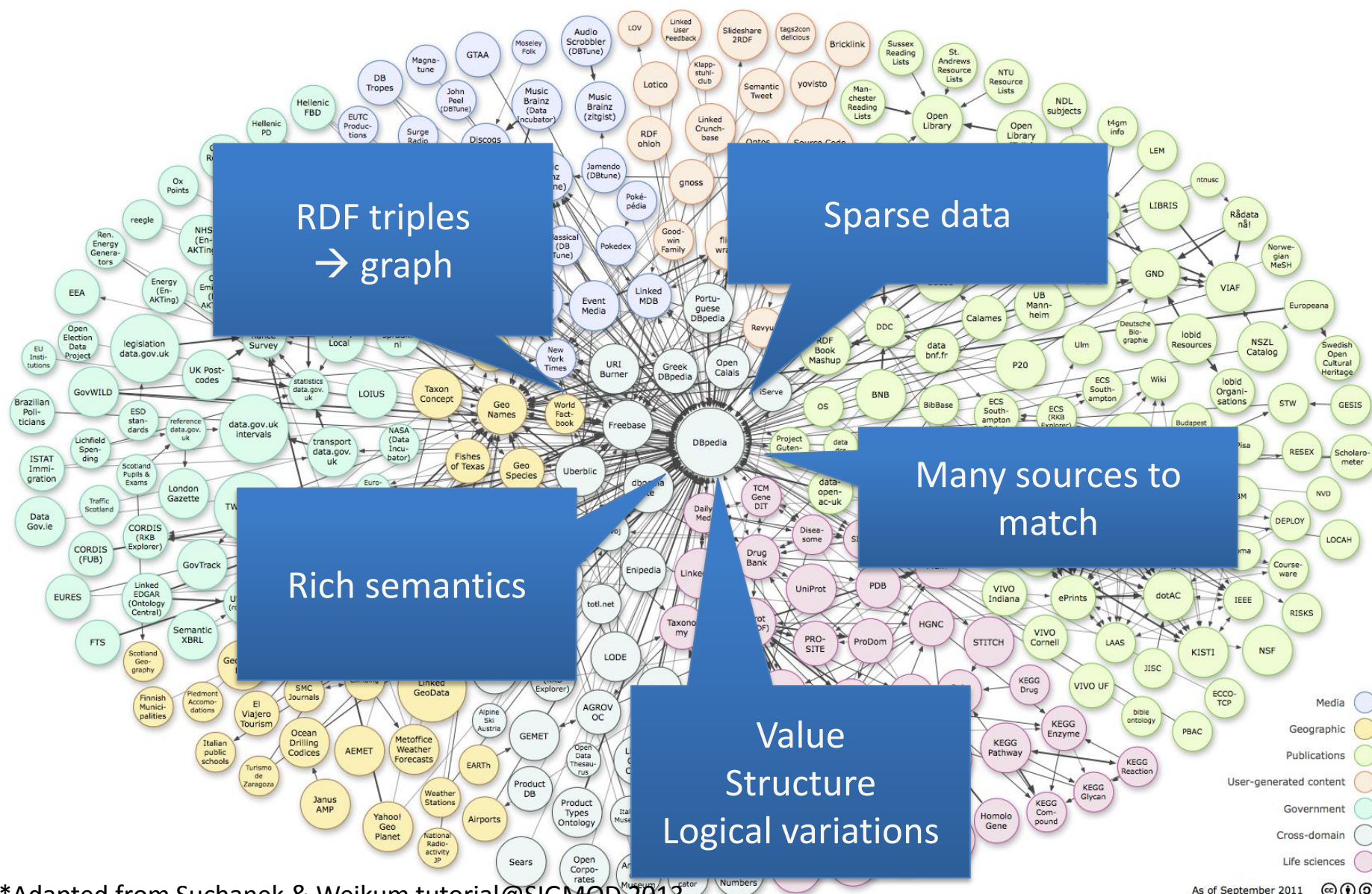


Solutions assume one common schema

Structural variation

Dense data

Instance Matching Today



Need for IM techniques

- Continuously **increasing number of datasets** published in the LOD Cloud
- People **interconnect** their dataset with existing ones.
 - These links are often **manually curated** (or semi-automatically generated).
- Size and number of data sets is **huge**, so it is vital to automatically detect additional links : **making the graph more dense**.

Benchmarking

Instance matching research has led to the development of various systems.

- How to compare these?
 - How can we assess their performance?
 - How can we push the systems to get better?
- These systems need to be benchmarked!

- Introduction into Linked Data
- Instance Matching
- **Benchmarks for linked Data**
 - **Why Benchmarks?**
 - **Benchmarks Characteristics**
 - **Benchmarks Dimensions**
- Benchmarks in the literature
 - Benchmarks with synthetic dataset
 - Benchmarks with real dataset
 - Individually created Benchmarks
- Outcomes & Conclusions

Benchmarking

- Benchmarking from a philosophical point of view is:
“the practice of being humble enough to admit that someone else is better at something, and wise enough to try to learn how to match and even surpass them at it.” [American Productivity & Quality Centre, 1993]
- A domain specific Benchmark is:
“A Benchmark specifies a workload characterizing typical applications in the specific domain. The performance of this workload of various computer systems gives a rough estimate of their relative performance on that problem domain”[G92]

Instance Matching Benchmark

Ingredients [FLM08]

- **Datasets**

The raw material of the benchmarks. These are the source and the target dataset that will be matched together to find the links

- **Ground Truth / Gold Standard / Reference Alignment**

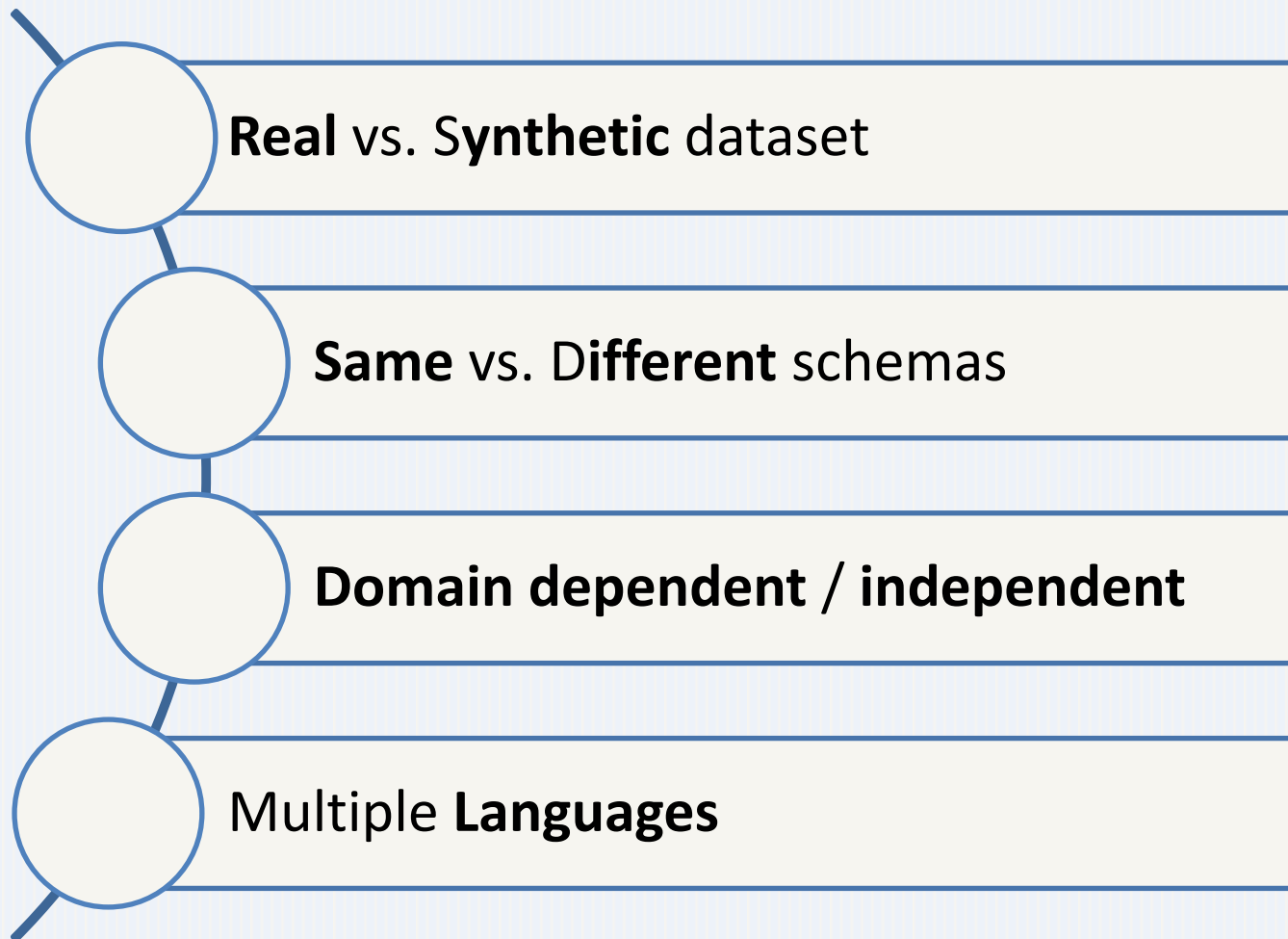
The “correct answer sheet” used to judge the completeness and soundness of the instance matching algorithms.

- **Metrics**

The performance metric(s) that determine the systems behavior and performance

- **Organized into test cases** each addressing different kind of requirements:
 - Source dataset
 - Target dataset
 - Ground Truth

Datasets



Real vs. Synthetic Benchmarks

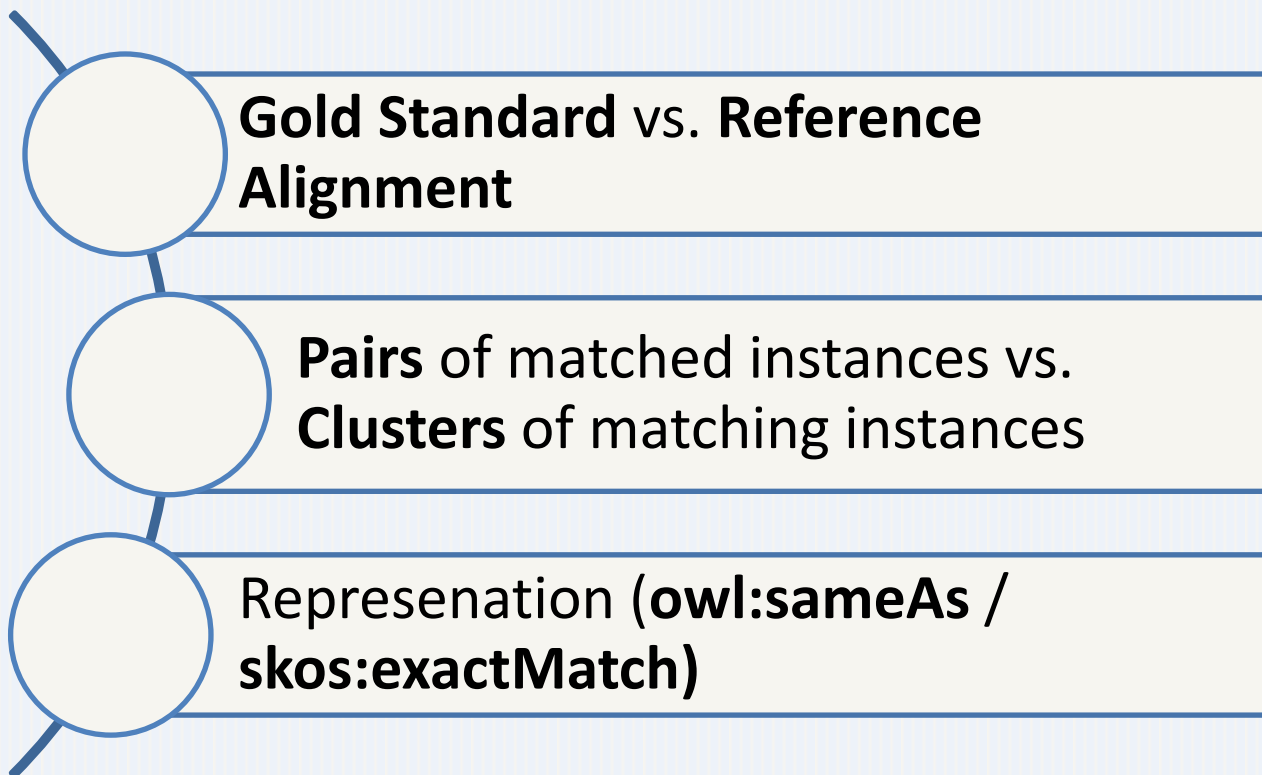
Real datasets (in whole or part of it):

- Real Realistic conditions for heterogeneity problems
- Realistic distributions
- Error prone Ground Truth

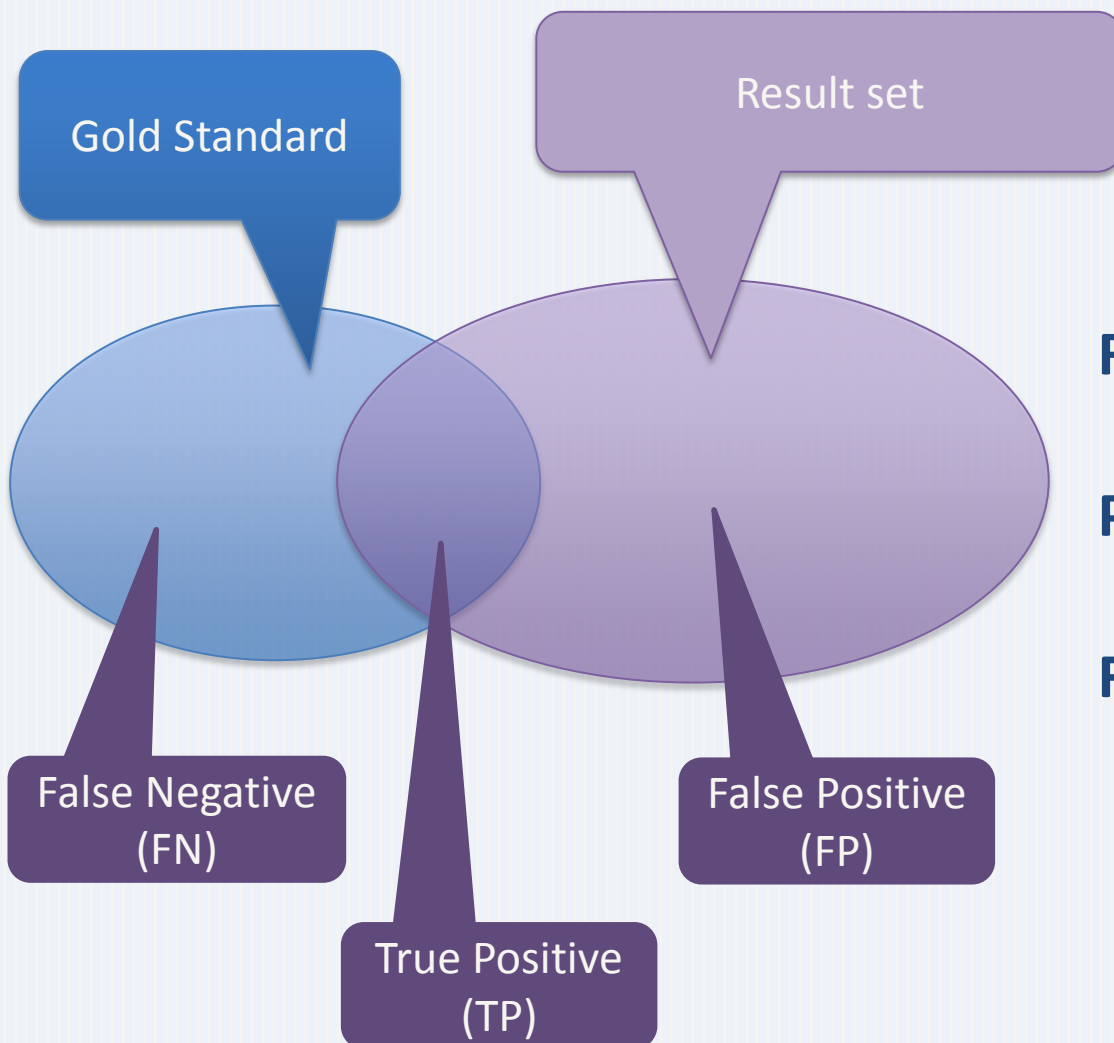
Synthetic (variations added into the datasets):

- Fully controlled test conditions
- Accurate Gold Standards
- Unrealistic distributions
- Systematic heterogeneity problems

Ground Truth



Recall / Precision / F-measure

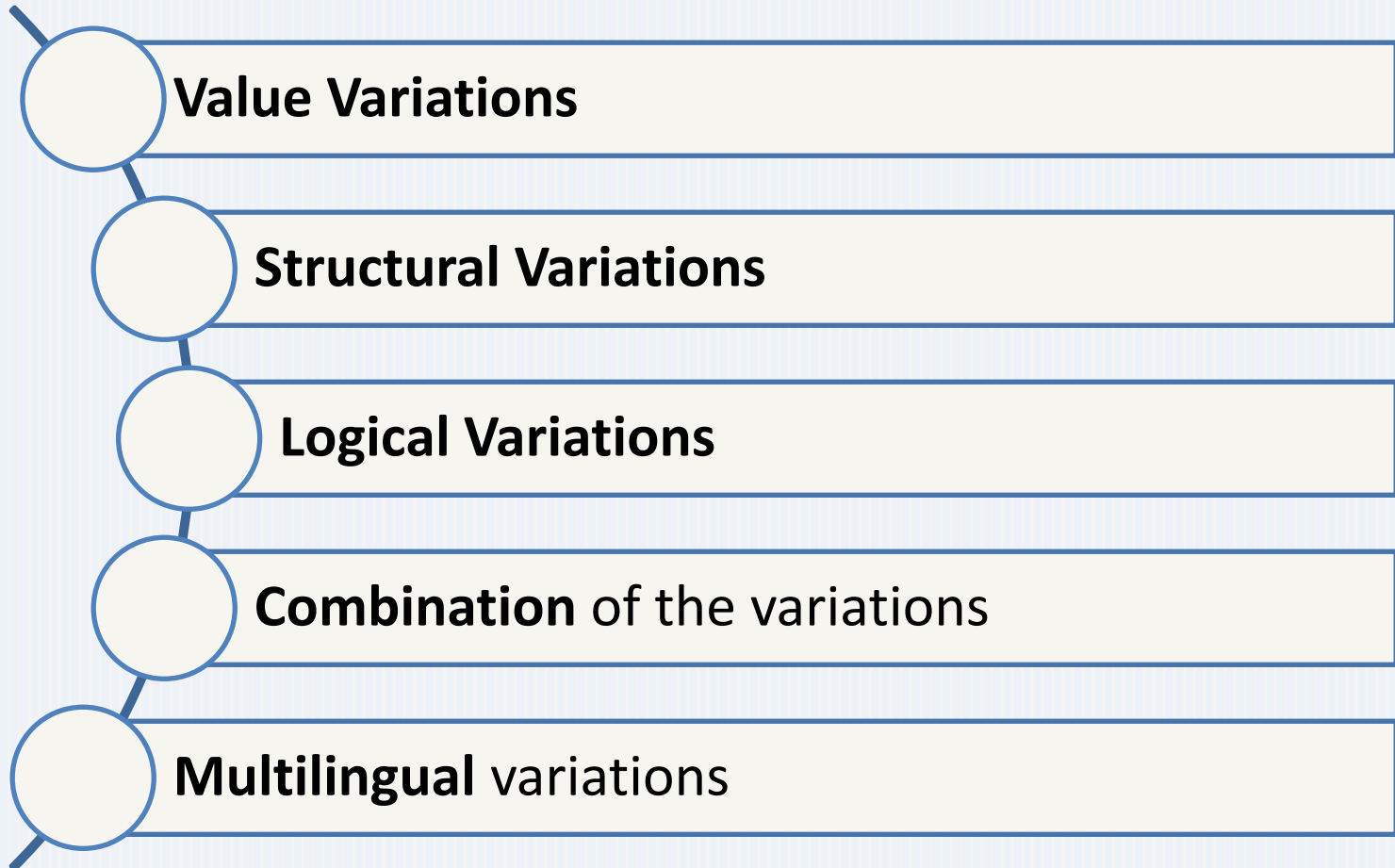


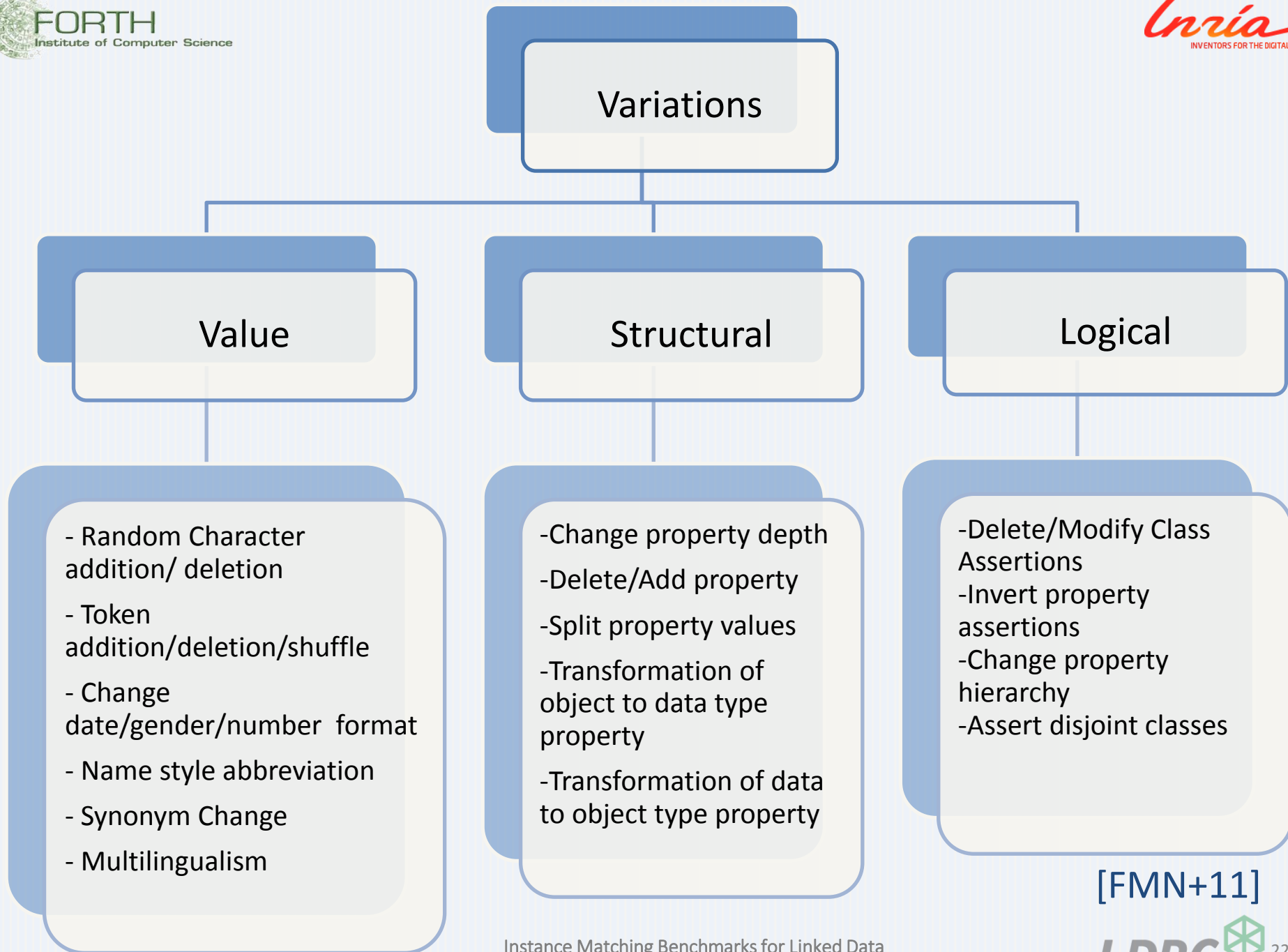
Recall $r = TP / (TP + FN)$

Precision $p = TP / (TP + FP)$

F-measure $f = 2 * p * r / (p + r)$

Data Variations





Benchmark Characteristics

Systematic Procedure	matching tasks are reproducible and the execution has to be comparable
Availability	related to the availability of the benchmark in time.
Quality	Precise evaluation rules and high quality ontologies
Equity	no system privileged during the evaluation process
Dissemination	How many systems have used this benchmark to be evaluated with
Volume	How many instances did the datasets contain
Ground Truth	existence of ground truth (Gold Standard/Reference Alignment) and it's accuracy.

Benchmarks Systems

- Instance matching techniques have, until recently, been benchmarked in an ad-hoc way.
- There does not exist a **standard** way of benchmarking the performance of the systems, when it comes to Linked Data.
- On the other hand, IM benchmarks have been mainly driven forward by the **Ontology Alignment Evaluation Initiative (OAEI)**

Ontology Alignment Evaluation Initiative

- OAEI provides a family of data integration benchmarks
- Since 2005, OAEI organizes an annual campaign aiming at evaluating ontology matching solutions
- In 2009, OAEI introduced the *Instance Matching (IM) Track*
 - focuses on the evaluation of different instance matching techniques and tools for Linked Data

- Introduction into Linked Data
- Instance Matching
- Benchmarks for linked Data
 - Why Benchmarks?
 - Benchmarks Characteristics
 - Benchmarks Dimensions
- **Benchmarks in the literature**
 - Synthetic Benchmarks
 - Real Benchmarks
 - Isolated Benchmarks
- Outcomes & Conclusions

Synthetic Benchmarks

OAEI IIMB
2009

OAEI IIMB
2010

OAEI Persons-
Restaurants
2010

OAEI IIMB
2011

Sandbox
2012

OAEI IIMB
2012

OAEI RDFT
2013

SWING

OAEI IIMB (2009) [EFH+09]

First attempt to create IM benchmark a with synthetic dataset

- **Datasets**

- OKKAM project containing actors, sport persons, and business firms
- Domain independent
- Number of instances up to ~200
- Shallow ontology max depth=2
- Small RDF /OWL ontology comprised of 6 classes, 47 data type properties

- **TestCases** (Divided into 37 test cases)

- Test case 2-10 including *value variations* (Typographical errors, Use of different formats)
- Test case 11-19 including *structural variations* (Property deletion, Change property types)
- Test case 20-29 including *logical variations* (subClass of assertions, Modify class assertions)
- Test case 30-37 including *Combination of the above*

- **Ground Truth**

- Automatically created gold standard

Typographical Errors

Property	Original Instance	Transformed Instance
type	"Actor"	"Actor"
Wikipedia-name	"James Anthony Church"	"qJaes Anthnodziurcdh"
name	"Tony Church"	"Toty fCurch"
description	"James Anthony Church (Tony Church) (May 11, 1930 - March 25, 2008) was a British Shakespearean actor, who has appeared on stage and screen"	"Jpes Athwobyi tuscr(nTons Courh)pMa y1sl1,9 3i- mrc 25, 200hoa s Bahirtishwaksepearna ctdor, woh hmwse appezrem yo nytmlaenn dscerepnq"

Original Instance	Transformed Instance
type (uri1, "Actor")	type (uri2, "Actor")
cogito-Name (uri1, "Wheeler Dryden")	cogito-Name (uri2, "Wheeler Dryden")
cogito-first_sentence (uri1, "George Wheeler Dryden (August 31, 1892 in London - September 30, 1957 in Los Angeles) was an English actor and film director, the son of Hannah Chaplin and" ...)	cogito-first_sentence (uri2,uri3)
	hasDataValue (uri3, "George Wheeler Dryden (August 31, 1892 in London - September 30, 1957 in Los Angeles) was an English actor and film director, the son of Hannah Chaplin and" ...)
cogito-tag (uri1, "Actor")	cogito-tag (uri2,uri4)
	hasDataValue (uri4, "Actor")

**Triples in the form of property (subject ,object)*

Logical Variations IIMB 2009

Sportsperson subClassOf Thing

Property name	Original instance	Transformed instance
type	"Sportsperson"	owl:Thing
wikipedia-name	"Sammy Lee"	"Sammy Lee"
cogito-first_sentence	"Dr. Sammy Lee (born August 1, 1920 in Fresno, California) is the first Asian American to win an Olympic gold..."	"Dr. Sammy Lee (born August 1, 1920 in Fresno, California) is the first Asian American to win an Olympic gold ..."
cogito-tag	"Sportsperson"	"Sportsperson"
cogito-domain	"Sport"	"Sport "

**Triples in the form of property, object*

Gold Standard IIMB 2009

- RDF/XML file
- Pairs of mapped instances
- Contains mappings in the form of <Cell>

```
<Cell>  
<entity1 rdf:resource="http://www.okkam.org/ens/id1"/>  
<entity2 rdf:resource="http://islab.dico.unimi.it/iimb/abox.owl#ID3"/>  
<measure rdf:datatype="http://www.w3.org/2001/XMLSchema#float">1.0</measure>  
<relation>=</relation>  
</Cell>
```

Systems- Results IIMB 2009

System	AFlood			ASMOV			DSSim		
Test	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
002 - 010	1.00	0.99	0.99	1.00	1.00	1.00	1.00	0.37	0.54
011 - 019	0.90	0.72	0.80	0.99	0.92	0.96	0.99	0.28	0.43
020 - 029	0.85	1.00	0.92	1.00	1.00	1.00	0.85	0.99	0.91
030 - 037	0.94	0.75	0.83	1.00	0.98	0.99	1.00	0.30	0.46
H-means	0.92	0.87	0.89	1.00	0.98	0.99	0.92	0.48	0.63

System	HMatch			FBEM			RiMOM		
Test	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
002 - 010	0.97	0.98	0.97	0.95	0.93	0.94	1.00	1.00	1.00
011 - 019	0.88	0.83	0.85	0.78	0.52	0.62	1.00	0.93	0.97
020 - 029	0.78	1.00	0.88	0.08	1.00	0.15	0.85	1.00	0.92
030 - 037	0.94	0.89	0.92	0.10	0.53	0.16	1.00	0.99	0.99
H-means	0.89	0.93	0.91	0.16	0.75	0.27	0.96	0.98	0.97

Balanced benchmark - shows both good and bad results from systems.

*Source OAEI 2009 <http://oei.ontologymatching.org/2009/results/oei2009.pdf>

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~200
Dissemination	6
Availability	+
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	+ (limited)
Multilinguality	-

- **Datasets**

- Freebase Ontology- *Domain independent*.
- Implemented in small version with ~ 350 instances and large version with ~ 1400 instances
- OWL ontologies consisting of 29 classes (81 for large), 32 object prop, 13 data prop.
- Shallow ontology with max depth=3

- **Test cases** (divided into 80 test cases)

- *Test cases 1-20 containing Value variations* (all types of variations)
- *Test cases 21-40 containing Structural variations* (all types of variations)
- *Test cases 41-60 containing Logical variations* (all types of variations)
- *Test cases 61-80 Combination of the above*

- **Ground Truth**

- Automatically created Gold Standards (same format as IIMB 2009)
- Created using the SWING Tool [FMN+11]

Value Variations IIMB (2010)

Variation	Original Instance	Transformed instance
Typographical errors	"Luke Skywalker"	"L4kd Skiwaldek"
Date Format	1948-12-21	December 21, 1948
Name Format	"Samuel L. Jackson"	"Jackson, S.L."
Gender Format	"Male"	"M"
Synonyms	"Jackson has won multiple awards(...)."	"Jackson has gained several prizes (...)."
Integer	10	110
Float	1.3	1.30

Structural Variations

IIMB (2010)[FMN+11]

Original Instance	Transformed Instance
name (uri1, "Natalie Portman")	name (uri3, "Natalie")
	name (uri3, "Portman")
born_in (uri1, uri2)	born_in (uri3, uri4)
name (uri2, "Jerusalem")	name (uri4, "Jerusalem")
	name (uri4, "Aukland")
gender (uri1, "Female")	obj_gender(uri3 , uri5)
date_of_birth(uri1, "1981-06-09")	has_value(uri5, "Female")

****Triples in the form of property(subject, object)***

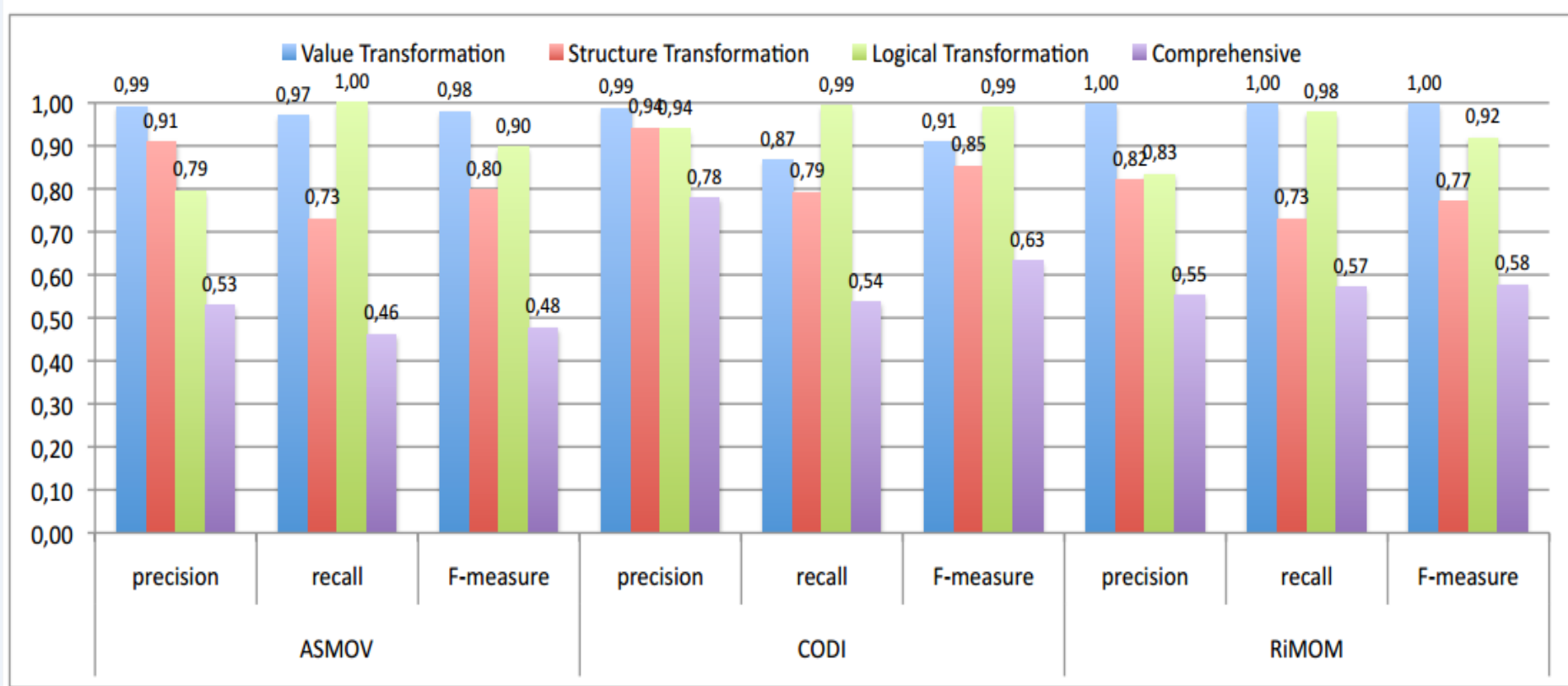
```

Character subClassOf Creature
created_by inverseOf creates
acted_by subPropertyOf featuring
Creature subClassOf Thing
    
```

Original Values	Transformed values
Character(uri1)	Creature(uri4)
Creature(uri2)	Creature(uri5)
Creature(uri3)	Thing(uri6)
created_by(uri1,uri2)	creates(uri5,uri4)
acted_by(uri1,uri3)	featuring(uri4,uri6)
name(uri1, "Luke Skywalker")	name(uri4, "Luke Skywalker")
name(uri1, "George Lucas")	name(uri4, "George Lucas")
name(uri1, "Mark Hamill")	name(uri4, "Mark Hamill")

****Triples in the form of property(subject, object)***

Systems Results OAEI 2010 (large version)



The closer to the reality it comes, the more challenging it gets.

*Source OAEI 2010 Results http://disi.unitn.it/~p2p/OM-2010/oei10_paper0.pdf

Characteristics

Variations

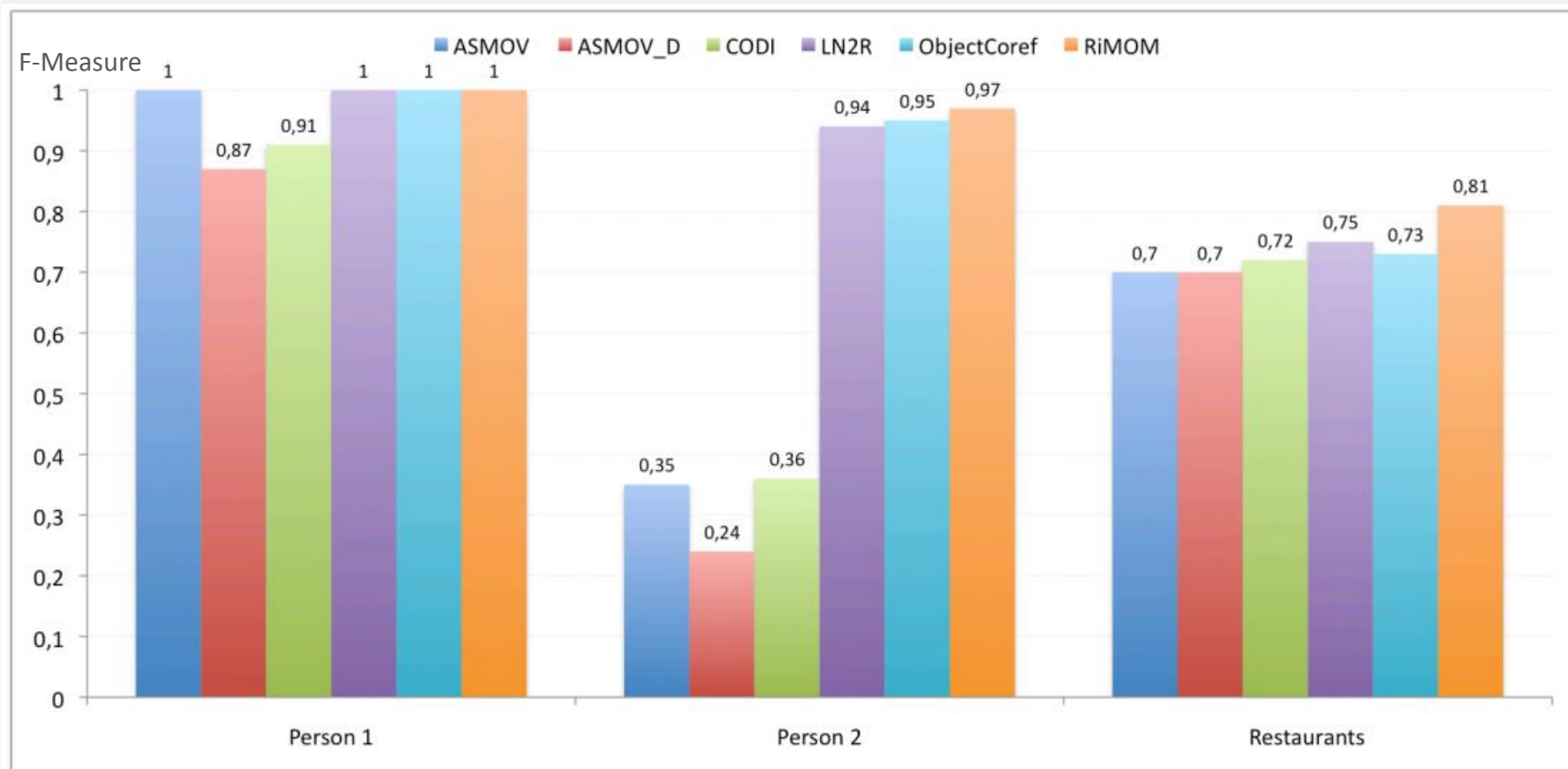
Systematic Procedure	+
Quality	+
Equity	+
Volume	~ 1400
Dissemination	3
Availability	+
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	+
Multilinguality	-

OAEI Persons & Restaurants Benchmark (2010) [EFM+10]

First Benchmark that includes the clustering matchings (1-n matchings)

- **Datasets**
 - Febrl project about Persons
 - Fodor's and Zagat's restaurant guides about Restaurants
 - Domain specific Datasets
 - Same Schemata
- **TestCases** (Small number of instances)
 - Person 1 ~500 instances (Max. 1 mod./property)
 - Person 2 ~600 instances (Max 3 mod./property and max 10 mod./instance)
 - Restaurant ~860 instances (no known number of modifications)
- **Variations**
 - *Combination of Value and Structural variations* (all types of variations)
- **Ground Truth**
 - Automatically created gold standard (same format as IIMB 2009)
 - 1-N matching in Person 2

Systems Results PR 2010



1. The more variations are added the worse the systems perform
2. Some systems could not cope with 1-n mappings requirement

*Source OAEI 2010 Results http://disi.unitn.it/~p2p/OM-2010/oeai10_paper0.pdf

Overview PR 2010

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~860
Dissemination	6
Availability	+
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	-

OAEI IIMB (2011) [EHH+11]

- **Datasets**

- Freebase Ontology- *Domain independent*.
- OWL ontologies consisting of 29 concepts, 20 object properties, 12 data properties
- ~4000 instances

- **Testcases** (Divided into 80 test cases)

- Divided into 80 test cases
- *Test cases 1-20 containing **Value variations** (all types of variations)*
- *Test cases 21-40 containing **Structural variations** (all types of variations)*
- *Test cases 41-60 containing **Logical variations** (all types of variations)*
- *Test cases 61-80 Combination of the above*

- **Ground Truth**

- Automatically created Gold Standard (same format as IIMB 2009)
- Created using the SWING Tool

System Results IIMB 2011

CODI system results

Test	Precision	F-measure	Recall
001–010	0.94	0.84	0.76
011–020	0.94	0.87	0.81
021–030	0.89	0.79	0.70
031–040	0.83	0.66	0.55
041–050	0.86	0.72	0.62
051–060	0.83	0.72	0.64
061–070	0.89	0.59	0.44
071–080	0.73	0.33	0.21

The closer to the reality it comes, the more challenging it gets.

Overview IIMB 2011

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~4000
Dissemination	1
Availability	+
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	+
Multilinguality	-

OAEI Sandbox (2012) [AEE+12]

- **Datasets**
 - Freebase Ontology- *Domain independent*
 - Collection of OWL files consisting of 31 concepts, 36 object properties, 13 data properties
 - ~375 instances
- **Test cases** (Divided into 10 test cases)
 - Divided into 10 test cases containing *Value Variations*
- **Ground Truth**
 - Automatically created Gold Standard (same format as IIMB 2009)

Attracted new systems to participate in instance matching task

Systems Results Sandbox 2012

Simple tests – Very good Results

Systems/Results	Precision	Recall	F- Measure
LogMap	0.94	0.94	0.94
LogMap Lite	0.95	0.89	0.92
SBUEI	0.95	0.98	0.96

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~375
Dissemination	3
Availability	+
Ground Truth	+
Value Variations	+
Structural Variations	-
Logical Variations	-
Multilinguality	-

OAEI IIMB (2012) [AEE+12]

Enhanced Sandbox Benchmarks

- **Datasets**
 - Freebase Ontology- *Domain independent*
 - No information about classes and instances
- **Test Cases**
 - Divided into 80 test cases
 - *Test cases 1-20 containing **Value variations***
 - *Test cases 21-40 containing **Structural variations***
 - *Test cases 41-60 containing **Logical variations***
 - *Test cases 61-80 Combination of the above*
- **Ground Truth**
 - Automatically created Gold Standard (same format as IIMB 2009)
 - Generated using the SWING Tool

IIMB 2012 Systems & Results

test	001–020			021–040			041–060			061–080		
	P	F	R	P	F	R	P	F	R	P	F	R
LogMap	.94	.90	.87	.93	.96	1.0	.93	.96	1.0	.95	.86	.79
LogMap_lite	.95	.78	.66	.93	.95	.97	.74	.84	.98	.77	.73	.69
SBUEI	.96	.97	.98	.97	.98	.99	.91	.90	.89	.58	.53	.48
semsim	.93	.93	.93	.91	.91	.91	.94	.94	.94	.66	.66	.66

Slight drop on F-measure when combination of variations occur

*Source OAEI 2012 Results <http://oei.ontologymatching.org/2012/results/oei2012.pdf>

Overview IIMB 2012

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	?
Dissemination	4
Availability	-
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	+
Multilinguality	-

OAEI RDFT (2013) [GDE+13]

First synthetic Benchmark with language variations

First synthetic Benchmark with Blind Evaluation

- **Datasets**

- RDF benchmark created by extracting data from **DBPedia** – **Domain independent**
- 430 instances, 11 RDF properties and 1744 triples
- Use of same schemata

- **Test Cases**

- Divided into 5 test cases
- Test case 1 contains *Value* variations
- Test case 2 contains *Structural* variations
- Test case 3 contains *Language* variations for comments and labels (English – French)
- Test case 4 contains *combinations* of the above variations
- Test case 5 contains *combinations* of the above variations

- **Ground Truth**

- Automatically created Gold Standard (same format as IIMB 2009)
- Cardinality 1-n matchings for test case 5

RDFT Systems - Results

system	tescase01			tescase02			tescase03			tescase04			tescase05		
	Prec.	F-m.	Rec.	Prec.	F-m.	Rec.	Prec.	F-m.	Rec.	Prec.	F-m.	Rec.	Prec.	F-m.	Rec.
LilyIOM	1.00	1.00	0.99	0.74	0.74	0.74	0.94	0.93	0.92	0.71	0.72	0.73	0.71	0.58	0.49
LogMap	0.97	0.80	0.69	0.79	0.88	0.99	0.98	0.84	0.73	0.95	0.80	0.70	0.92	0.74	0.62
RiMOM2013	1.00	1.00	1.00	0.95	0.97	0.99	0.96	0.98	0.99	0.94	0.96	0.98	0.93	0.96	0.99
SLINT+	0.98	0.98	0.98	1.00	1.00	1.00	0.94	0.92	0.91	0.91	0.91	0.91	0.87	0.88	0.88

1. Systems can cope with multilingualism
2. Slight drop of the F-measure for cluster mappings (apart from RiMOM)

*Source OAEI 2013 Results http://ceur-ws.org/Vol-1111/oaei13_paper0.pdf

Overview RDFT 2013

Characteristics

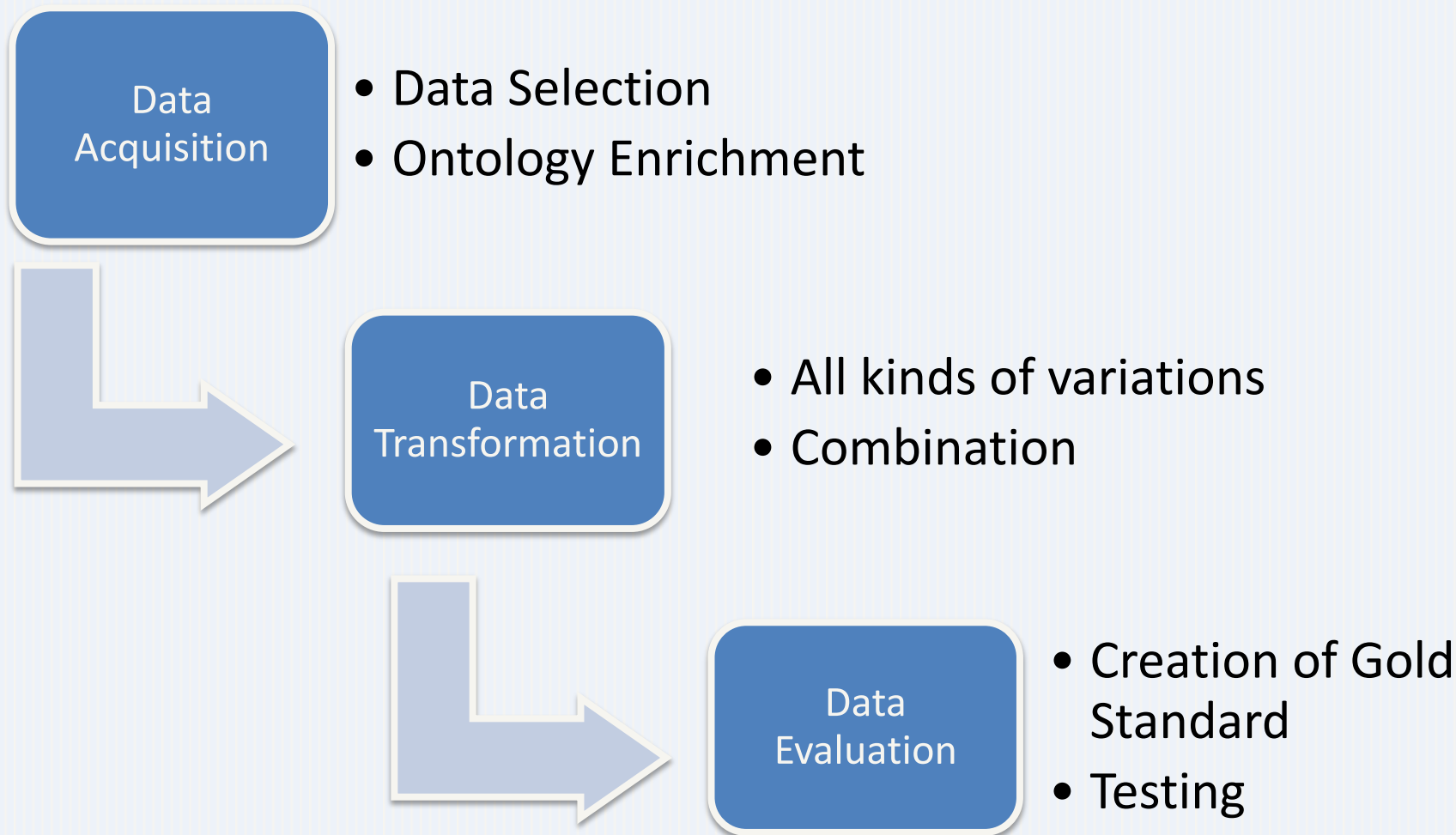
Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~430
Dissemination	4
Availability	-
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	+

Semi-automatic generator of IM Benchmarks

- Contributed in the generation of IIMB Benchmarks of OAEI in 2010, 2011 and 2012
- Freely available (<https://code.google.com/p/swing-generator/>)
- **Variations allowed**
 - *All kind of variations (apart from Multilingualism)*
- **Ground Truth**
 - Automatically created Gold Standard

SWING phases



Overview SWING

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	n/a
Dissemination	3
Availability	+
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	+
Multilinguality	-

Comparison of synthetic Benchmarks

	IIMB 2009	IIMB 2010	PR 2010	IIMB 2011	Sandbox 2012	IIMB 2012	RDFT 2013	SWING 2010
Systematic Procedure	+	+	+	+	+	+	+	+
Quality	+	+	+	+	+	+	+	+
Equity	+	+	+	+	+	+	+	+
Availability	+	+	+	+	+	—	—	+
Volume	~200	~ 1400	~860	~4000	~375	?	~430	n/a
Dissemination	6	3	6	1	3	4	4	3
Ground Truth	+	+	+	+	+	+	+	+
Value variations	+	+	+	+	+	+	+	+
Structural variations	+	+	+	+	—	+	+	+
Logical variations	+	+	—	+	—	+	—	+
Multilinguality	—	—	—	—	—	—	+	—
Blind Evaluations	—	—	—	—	—	—	+	n/a
1-n Mappings	—	—	+	—	—	—	+	n/a

- Introduction into Linked Data
- Instance Matching
- Benchmarks for linked Data
 - Why Benchmarks?
 - Benchmarks Characteristics
 - Benchmarks Dimensions
- **Benchmarks in the literature**
 - Synthetic Benchmarks
 - Real Benchmarks
 - Isolated Benchmarks
- Outcomes & Conclusions

Real Benchmarks

ARS
(OAEI 2009)

VLCR
(OAEI 2009)

DI
(OAEI 2010)

DI-NYT
(OAEI 2011)

(ARS - OAEI 2009) [EFH+09]

- **Datasets**

- **AKT-Eprints archive** - information about papers produced within the AKT project.
- **Rexa dataset**- computer science research literature, people, organizations, venues and research communities data
- **SWETO-DBLP dataset** - publicly available dataset listing publications from the computer science domain.
- All three datasets were structured using the same schema - **SWETO-DBLP ontology**
- Domain *dependent*

- **Test cases** (Value/Structural variations)

- AKT / Rexa
- AKT /DBLP
- Rexa / DBLP

- **Challenges**

- Many instances (almost 1M instances)
- Ambiguous labels (person names and paper titles) and
- Noisy data (some sources contained incorrect information)

ARS Data Statistics

- **Dataset Statistics**

- AKT-Eprints: **564**-foaf: Persons and **283**-sweto:Publications
- Rexa : **11.050**-foaf: Persons and **3.721**-sweto:Publications
- SWETO-DBLP : **307.774**-foaf: Persons and **983.337**-sweto:Publications

- **Ground Truth**

- Manually constructed - Error prone Reference Alignment
- AKT-REXA contains 777 overall mappings
- AKT-DBLP contains 544 overall mappings
- REXA-DBLP contains 1540 overall mappings

ARS Systems & Results

System	sweto:Publication			foaf:Person			Overall		
	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.	Prec.	Rec.	FMeas.
AKT/Rexa									
DSSim	0.15	0.16	0.16	0.81	0.30	0.43	0.60	0.28	0.38
RiMOM	1.00	0.72	0.84	0.92	0.70	0.79	0.93	0.70	0.80
FBEM	0.99	0.61	0.76	0.73	0.02	0.03	0.94	0.10	0.18
HMatch	0.97	0.89	0.93	0.94	0.39	0.56	0.95	0.46	0.62
ASMOV	0.32	0.79	0.46	0.76	0.24	0.37	0.52	0.32	0.39
AKT/DBLP									
DSSim	0	0	0	0.15	0.19	0.17	0.11	0.15	0.13
RiMOM	0.96	0.97	0.96	0.93	0.50	0.65	0.94	0.59	0.73
FBEM	0.98	0.80	0.88	0	0	0	0.98	0.16	0.28
HMatch	0.93	0.97	0.95	0.58	0.57	0.57	0.65	0.65	0.65
Rexa/DBLP									
DSSim	0	0	0	0	0	0	0	0	0
RiMOM	0.94	0.95	0.94	0.76	0.66	0.71	0.80	0.72	0.76
FBEM	0.98	0.15	0.26	1.00	0.11	0.20	0.99	0.12	0.21
HMatch	0.45	0.96	0.61	0.40	0.34	0.37	0.42	0.48	0.45

1. Scalability issues from some the systems
2. Structural variations in names of Persons lower the F-measure of systems

*Source OAEI results 2009 http://ceur-ws.org/Vol-551/oeai09_paper0.pdf

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~1M
Dissemination	5
Availability	+
Ground Truth	Reference Alignment
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	-

Very Large Crosslingual Resources (OAEI 2008-2009) [EFH+09]

First attempt to interlink sources with different languages

- **Datasets**

- Thesaurus of the Netherlands Institute for Sound and Vision (**GTAA**- National television thesaurus) in **SKOS representation**
- English **WordNet** from Princeton University (Lexical database of English. Nouns, verbs, adjectives and adverbs) in **RDF/OWL representation**
- **DBPedia** - Extracted structured information from Wikipedia - **RDF/OWL representation**

- **Dataset Statistics**

- GTAA : **27.000** Names, **14.000** Locations, **97.000** Persons, and **3.800** Subject keywords
- WordNet : **117.000** synsets
- DBPedia: **2.18 M** "things"

VLCR Test cases

- **Test Cases**

- GTAA Names
 - GTAA Locations
 - GTAA Persons
 - GTAA Subject keywords
 - GTAA Names
 - GTAA Locations
 - GTAA Persons
 - GTAA Subject keywords
- DBPedia Things
- Wordnet synsets

- **Ground Truth**

- Manually curated (links in the form of <skos:exactMatch>)
- Small and error prone Reference Alignment
- Precision: random sample of 71-97 mappings from each GTAA facet in each alignment manually assessed
- Recall: Reference Alignment of 100 mappings for Subject keywords per alignment

VCRL Results

	Precision		Recall	
Alignment	GG2WW	DSSim	GG2WW	DSSIM
name-dbp	0.63	0.64	0.62	0.30
location-dbp	0.94	0.80		
person-dbp	0.91	0.79		
subject-dbp	0.86	0.70		
name-wn	.	0.44		
location-wn	.	0.61	0.59	0.19
person-wn	.	0.07		
subject-wn	0.59	0.77		
total	0.78	0.62		

Difficult to judge whether the problem of the bad results is due to the systems or because of the small and error prone Reference Alignment.

*Source OAEI results 2009 http://ceur-ws.org/Vol-551/oaiei09_paper0.pdf

Overview VLCR 2009

Characteristics

Systematic Procedure	+
Quality	+
Equity	+
Volume	~2M
Dissemination	2
Availability	+
Ground Truth	Small Reference Alignment
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	+

Data Interlinking

(OAEI 2010) [EFM+10]

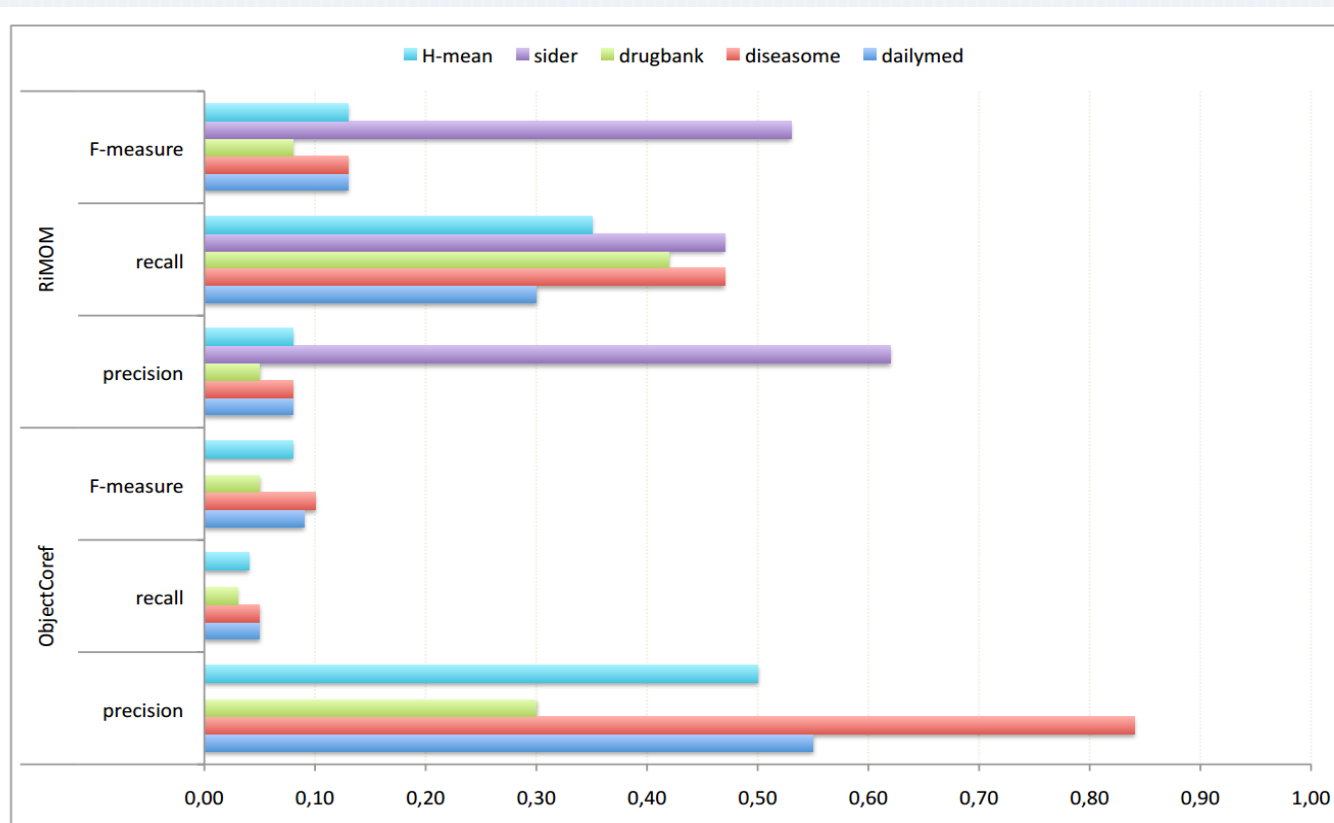
The first real Benchmark that contained semi-automatically created reference alignments

- **Datasets**

- **DailyMed** - Provides marketed drug labels containing 4308 drugs
- **Diseasome** - Contains information about 4212 disorders and genes
- **DrugBank** - Is a repository of more than 5900 drugs approved by the US Federal Drugs Agency
- **SIDER** - Contains information on marketed medicines (996 drugs) and their recorded adverse drug reaction (4192 side effects).

- **Reference Alignments**

- Semi-automatically created reference alignments
- Running the test with Silk and LinQuer systems
- In the form of pairs of matched instances (same as in IIMB 2009)



1. Providing a reliable mechanism for systems' evaluation
2. Improving the performances of matching systems

*Source OAEI 2010 Results http://disi.unitn.it/~p2p/OM-2010/oei10_paper0.pdf

Overview DI 2010

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	~6000
Dissemination	2
Availability	+
Ground Truth	Reference Alignment
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	-

Data Integration (OAEI 2011) [EHH+11]

- **Datasets (No information about classes and instances)**
 - New York Times
 - DBPedia
 - Freebase
 - Geonames
- **Tests cases**
 - DBPedia locations
 - DBPedia organizations
 - DBPedia people
 - Freebase locations
 - Freebase organizations
 - Freebase people
 - Geonames

New York Times Subject headings
- **Reference Alignments**
 - Based on the links present in the datasets
 - Provided matches are accurate but may not be complete

Data Integration – New York Times

	People	Organizations	Locations
# NYT resources	9958	6088	3840
# Links to Freebase	4979	3044	1920
# Links to DBPedia	4977	1949	1920
# Links to Geonames	0	0	1789

DI Results

	AgreementMaker			SERIMI			Zhishi.links		
Dataset	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.	Prec.	FMeas.	Rec.
DI-nyt-dbpedia-loc.	.79	.69	.61	.69	.68	.67	.92	.92	.91
DI-nyt-dbpedia-org.	.84	.74	.67	.89	.88	.87	.90	.91	.93
DI-nyt-dbpedia-peo.	.98	.88	.80	.94	.94	.94	.97	.97	.97
DI-nyt-freebase-loc.	.88	.85	.81	.92	.91	.90	.90	.88	.86
DI-nyt-freebase-org.	.87	.80	.74	.92	.91	.89	.89	.87	.85
DI-nyt-freebase-peo.	.97	.96	.95	.93	.92	.91	.93	.93	.92
DI-nyt-geonames.	.90	.85	.80	.79	.80	.81	.94	.91	.88
H-mean.	.92	.85	.80	.89	.89	.88	.93	.92	.92

1. Good results from all the systems
2. Well known domain and datasets
3. No logical variations

*Source OAEI 2010 <http://oei.ontologymatching.org/2010/vlcr/index.html>

Overview DI 2011

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	+
Volume	?
Dissemination	3
Availability	-
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	-

Comparison of Real Benchmarks

	ARS	VLCR 2009	DI 2010	DI 2011
Systematic Procedure	+	+	+	+
Quality	+	+	+	+
Equity	+	+	+	+
Availability	+	+	+	—
Volume	~1M	~2M	~6000	?
Dissemination	5	2	2	3
Ground Truth	R.A.	R.A.	R.A.	+
Value variations	+	+	+	+
Structural variations	+	+	+	+
Logical variations	—	—	—	—
Multilinguality	—	+	—	—
Blind Evaluations	—	+	—	—

- Introduction into Linked Data
- Instance Matching
- Benchmarks for linked Data
 - Why Benchmarks?
 - Benchmarks Characteristics
 - Benchmarks Dimensions
- **Benchmarks in the literature**
 - Synthetic Benchmarks
 - Real Benchmarks
 - Isolated Benchmarks
- Outcomes & Conclusions

Isolated Benchmarks

ONTOBI

OpenPhacts

ONTOlogy matching Benchmark with many Instances (ONTOBI) [Z10]

Synthetic Benchmark

- **Datasets**
 - RDF/OWL benchmark created by extracting data from DBPedia v. 3.4
 - 205 classes, 1144 object properties and 1024 data types properties
 - 13.704 instances
- **Divided into 16 Test cases**
- **Variations**
 - *Value variations*
 - *Structural variations*
 - *Combination of the above*
- **Ground Truth**
 - Automatically created Gold Standard

ONTOBI Variations

Simple Variations

Spelling mistakes

(Value Variations)

Change format

(Value Variation)

Suppressed
Comments

(Structural Variation)

Delete data types

(Structural Variation)

ONTOBI Variations

Complex Variations	Flatten/Expand Structure	(Structural Variation)
	Language modification	(Value Variation)
	Random names	(Value Variation)
	Synonyms	(Value Variation)
	Disjunct Dataset	(Value Variation)

ONTOBI Predefined Variations

Simple
tests
cases

OS1: spelling mistakes

OS2: suppressed comments

OS3: disjunct dataset

OS4: another language

OS5: random names

OS6: synonyms

OS7: expanded structure

OS8: flatten structure

ONTOBI Predefined Variations

Complex tests (2 mods)

OC1: spelling mistakes, suppressed comments

OC2: random names, no datatype

OC3: synonyms, overlapping datasets

OC4: flatten structure, overlapping datasets

Complex tests (>3 mods)

OCC1: spelling mistakes, suppressed comments, no datatype, disjunct datasets

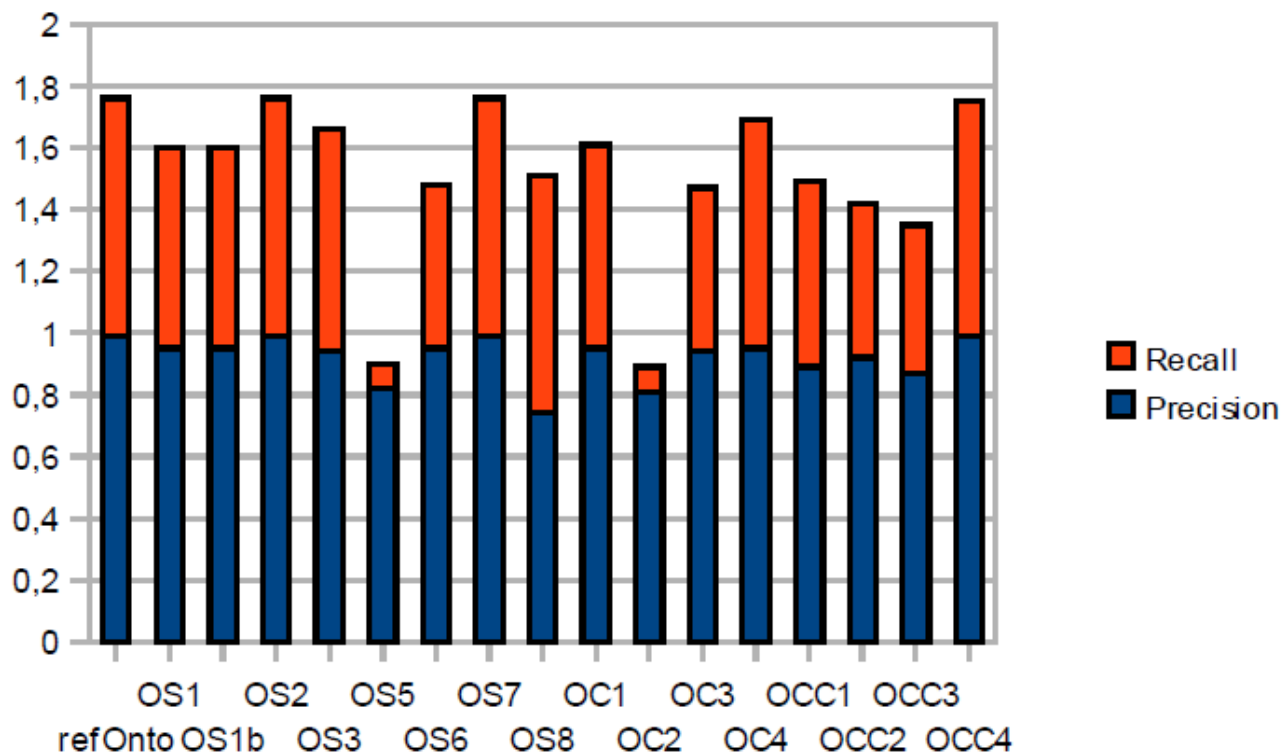
OCC2: spelling mistakes, synonyms, no data types

OCC3: synonyms, expanded structure, disjunct data sets,

OCC4: suppressed comments, changed format, overlapping datasets

ONTOBI Systems & Results

MICU system



*Source K. Zaiß: Instance-Based Ontology Matching and the Evaluation of Matching Systems , 2011, Dissertation

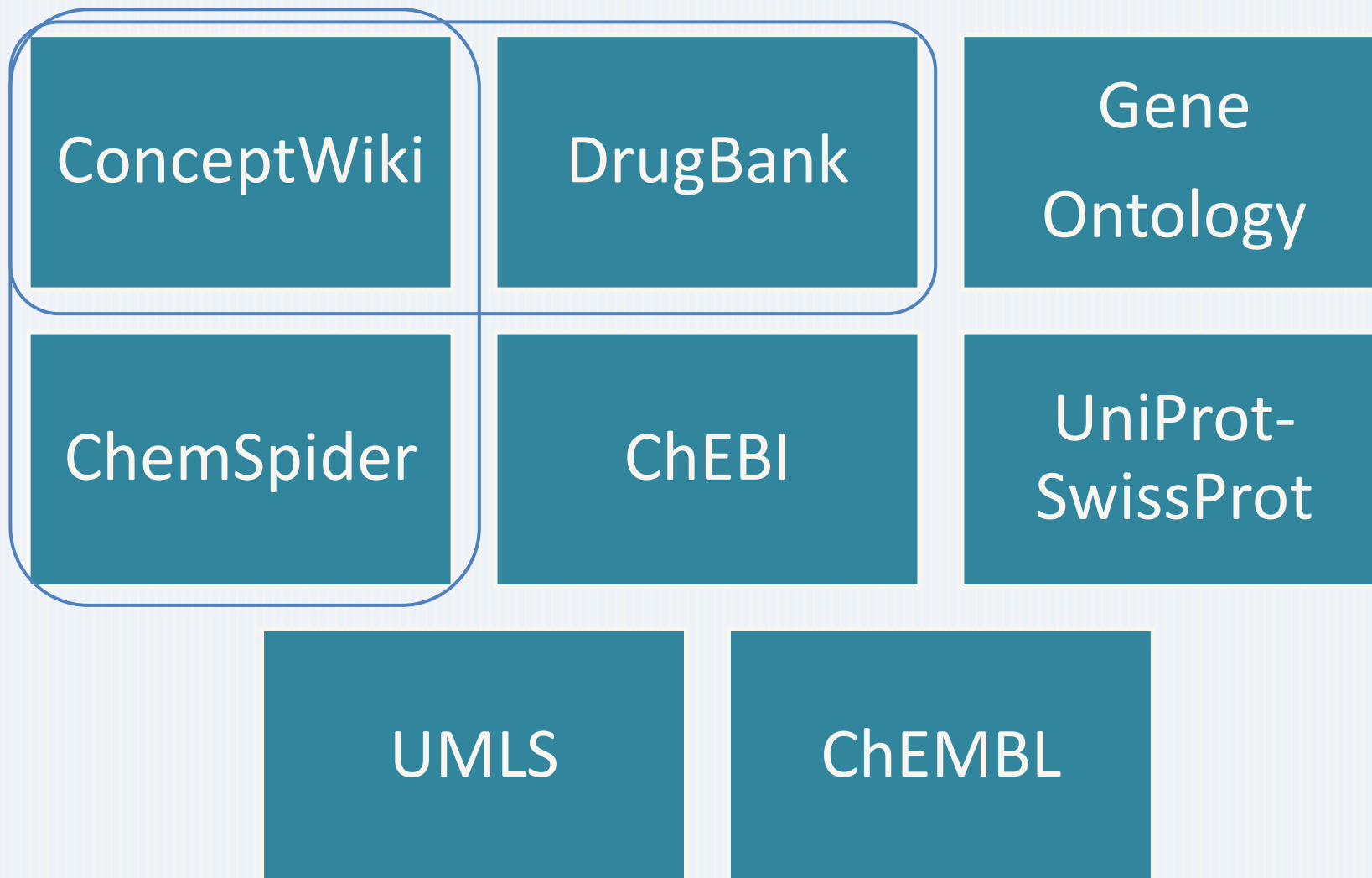
Overview ONTOBI 2010

Characteristics

Variations

Systematic Procedure	+
Quality	+
Equity	?
Volume	~13700
Dissemination	1
Availability	-
Ground Truth	+
Value Variations	+
Structural Variations	+
Logical Variations	-
Multilinguality	+

Open Pharmacological Space (Open PHACTS) [GGL+12]



Open PHACTS Reference Alignment

- Creation of sophisticated SPARQL queries for the Identity Mapping Service (IMS)
- Semi-automatic creation of reference alignments, with the curation of domain experts
- Links of <skos:exactMatch>

<<http://www.conceptwiki.org/concept/4918acc2-23e4-4bea-886b-b167d56f5a72>>
skos:exactMatch <<http://www4.wiwiss.fu-berlin.de/drugbank/resource/targets/6511>>.

<<http://www.conceptwiki.org/concept/09a60eb9-90f3-4938-92d8-b12133e27716>>
skos:exactMatch <<http://www4.wiwiss.fu-berlin.de/drugbank/resource/targets/2686>>.

<<http://www.conceptwiki.org/concept/8c847e1b-bf16-45b1-b899-f7403aa70e12>>
skos:exactMatch <<http://www4.wiwiss.fu-berlin.de/drugbank/resource/targets/3417>>.

Systems and Results

TC1 : ConceptWiki – DrugBank Targets

TC2 : ConceptWiki – ChempSpider

Results in terms of F-measure

TEST CASE	DISTANCE METHOD	LIMES		SILK	
		0.9	0.8	0.1	0.2
TC1	<i>jaccard</i>	0.0	0.0	0.0	0.0
	<i>exactMatch/equality</i>	0.009	0.065	0.0	0.0
TC2	<i>jaccard</i>	0.08	0.29	0.68	0.68
	<i>exactMatch/equality</i>	0.26	0.77	0.68	0.68

1. Bad results of the systems was not due to a problem of systems
2. Matching methods did only take into consideration string matching
3. Pharmacology domain is very difficult , because of the gene/drug labels
4. Needed more sophisticated methods to match the datasets

*Source <http://ldbc.eu/sites/default/files/D4.4.1-final.pdf>

- Introduction into Linked Data
- Instance Matching
- Benchmarks for linked Data
 - Why Benchmarks?
 - Benchmarks Characteristics
 - Benchmarks Dimensions
- Benchmarks in the literature
 - Synthetic Benchmarks
 - Real Benchmarks
 - Isolated Benchmarks
- **Summary and Conclusions**

Wrapping up: Benchmarks

Which benchmarks included multilingual datasets?

OAEI RDFT
2013 (French-
English)

VLCR (Dutch-
English)

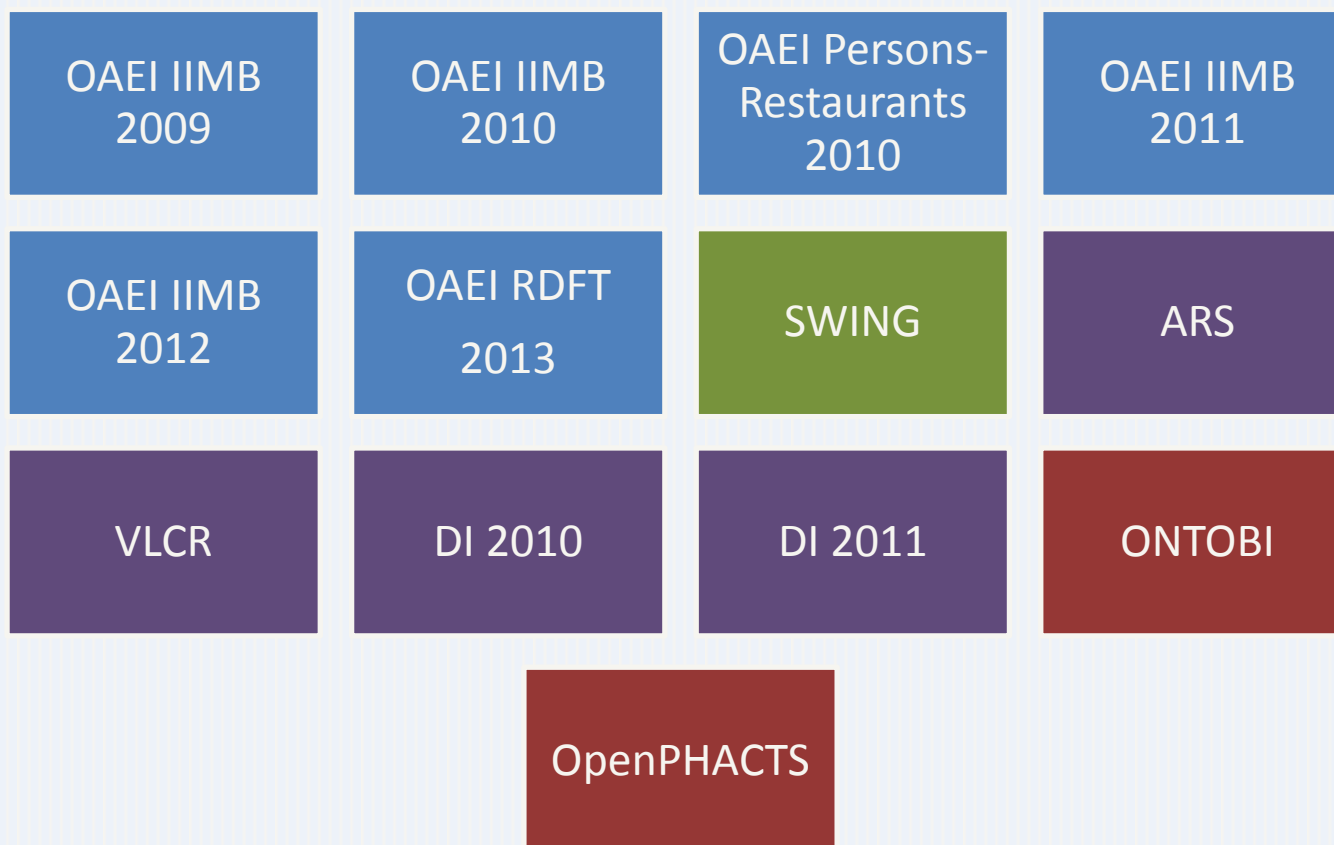
Wrapping up: Benchmarks

Which benchmarks included value variations into the test cases?

OAEI IIMB 2009	OAEI IIMB 2010	OAEI Persons-Restaurants 2010	OAEI IIMB 2011
Sandbox	OAEI IIMB 2012	OAEI RDFT 2013	SWING
ARS	VLCR	DI 2010	DI 2011
	ONTOBI	OpenPHACTS	

Wrapping up: Benchmarks

Which benchmarks included structural variations into the test cases?



Wrapping up: Benchmarks

Which benchmarks included logical variations into the test cases?

OAEI IIMB
2009

OAEI IIMB
2010

OAEI IIMB
2011

OAEI IIMB
2012

SWING

Wrapping up: Benchmarks

Which benchmarks included combination of the variations into the test cases?

OAEI IIMB
2009

OAEI IIMB
2010

OAEI IIMB
2011

OAEI IIMB
2012

SWING

Wrapping up: Benchmarks

Which benchmarks are more voluminous?

ARS

VLCR

DI 2011

OpenPHACTS

Wrapping up: Benchmarks

Which benchmarks included both combination of the variations and was voluminous at the same time?

None

Issue 1:

No IM benchmark tackles both, **combination** of variations and **scalability** issues

Issue 2 :

No IM benchmark using the **full expressiveness** of RDF/OWL language

- Complex class definitions (union, intersection)
- Cardinality constraints (functional property)
- Disjointness (properties)

Systems for Benchmarks

Outcomes as far as systems are concerned:

- Systems can handle the value variations, the structural variation, and the simple logical variations **separately**.
- Systems can cope with **multilingual datasets**
- More work needed for **complex variations** (combination of value, structural, and logical)
- Enhancement of systems to cope with the clustering of the mappings (1-n mappings)

Conclusion

- Need for benchmarks that will “**show the way to the future**” to the systems.
- **Standard Organization for IM Benchmarks** , in the line of TPC.
 - OAEI not yet an Organizations
 - The **Linked Data Benchmark Council (LDBC)** is established as an independent authority responsible for specifying benchmarks, benchmarking procedures and verifying/publishing results for software systems designed to manage graph and RDF data.
(<http://ldbcouncil.org/>)

Thank you!

Questions? Comments?

References (1)

#	Reference	Abbreviation
1	J. L. Aguirre, K. Eckert, A. F. J. Euzenat, W. R. van Hage, L. Hollink, C. Meilicke, A. N. D. Ritze, F. Scharffe, P. Shvaiko, O. Svab-Zamazal, C. Trojahn, E. Jimenez-Ruiz, B. C. Grau, and B. Zapolko. Results of the ontology alignment evaluation initiative 2012. In OM, 2012.	[AEE+12]
2	I. Bhattacharya and L. Getoor. Entity resolution in graphs. Mining Graph Data. Wiley and Sons, 2006.	[BG06]
3	J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. Joslyn, V. Malaise, C. Meilicken, A. Nikolov, J. Pane, M. Sabou, F. Scharffe, P. Shvaiko, V. S. H., Stuckenschmidt, O. Svab-Zamazal, V. Svatek, , C. Trojahn, G. Vouros, and S. Wang. Results of the Ontology Alignment Evaluation Initiative 2009. In OM, 2009.	[EFH+09]
4	J. Euzenat, A. Ferrara, C. Meilicke, J. Pane, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, V. Svatek, and C. Trojahn. Results of the Ontology Alignment Evaluation Initiative 2010. In OM, 2010.	[EFM+10]
5	A. F. J. Euzenat, W. R. van Hage, L. Hollink, C. Meilicke, A. N. D. Ritze, F. Scharffe, P. Shvaiko, H. Stuckenschmidt, O. Svab-Zamazal, and C. Trojahn. Results of the Ontology Alignment Evaluation Initiative 2011. In OM, 2011.	[EHH+11]
6	A. K. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering, 19(1), 2007.	[EIV07]
7	J. Euzenat and P. Shvaiko, editors. Ontology Matching. Springer-Verlag, 2007.	[ES07]
8	A. Ferrara, D. Lorusso, S. Montanelli, and G. Varese. Towards a Benchmark for Instance Matching. In OM, 2008.	[FLM08]
9	A. Ferrara, S. Montanelli, J. Noessner, and H. Stuckenschmidt. Benchmarking Matching Applications on the Semantic Web. In ESWC, 2011.	[FMN+11]
10	J. Gray, editor. The Benchmark Handbook for Database and Transaction Systems. Morgan Kaufmann, 1993.	[G93]

References (2)

#	Reference	Abbreviation
11	B. C. Grau, Z. Dragisic, K. Eckert, A. F. J. Euzenat, R. Granada, V. Ivanova, E. Jimenez-Ruiz, A. O. Kempf, P. Lambrix, A. Nikolov, H. Paulheim, D. Ritze, F. Schare, P. Shvaiko, C. Trojahn, and O. Zamazal. Results of the ontology alignment evaluation initiative 2013. In OM, 2013.	[GDE+13]
12	Gray, A.J.G., Groth, P., Loizou, A., et al.: Applying linked data approaches to pharmacology: Architectural decisions and implementation. Semantic Web. (2012).	[GGL+12]
13	P. Hayes. RDF Semantics. www.w3.org/TR/rdf-mt , February 2004.	[H04]
14	R. Isele and C. Bizer. Learning linkage rules using genetic programming. In OM, 2011.	[IB11]
15	A. Isaac, L. van der Meij, S. Schlobach, and S. Wang. An Empirical Study of Instance-Based Ontology Matching. In ISWC/ASWC, 2007.	[IMS07]
16	E. Ioannou, N. Rassadko, and Y. Velegrakis. On Generating Benchmark Data for Entity Matching. Journal of Data Semantics, 2012.	[IRV12]
17	A. Jentzsch, J. Zhao, O. Hassanzadeh, K.-H. Cheung, M. Samwald, and B. Andersson. Linking open drug data. In Linking Open Data Triplification Challenge, I-SEMANTICS, 2009.	[JZH+09]
18	C. Li, L. Jin, and S. Mehrotra. Supporting efficient record linkage for large data sets using mapping techniques. In WWW, 2006.	[LJM06]
19	D. L. McGuinness and F. van Harmelen. OWL Web Ontology Language. http://www.w3.org/TR/owl-features/ , 2004.	[MH04]
20	B. M. F. Manola, E. Miller. RDF Primer. www.w3.org/TR/rdf-primer , February 2004.	[MM04]

Reference (3)

#	Reference	Abbreviation
21	J. Noessner, M. Niepert, C. Meilicke, and H. Stuckenschmidt. Leveraging Terminological Structure for Object Reconciliation. In ESWC, 2010.	[NNM10]
22	A. Nikolov, V. Uren, E. Motta, and A. de Roeck. Refining instance coreferencing results using belief propagation. In ASWC, 2008.	[NUM+08]
23	M. Perry. TOnToGen: A Synthetic Data Set Generator for Semantic Web Applications. AIS SIGSEMIS, 2(2), 2005.	[P05]
24	E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. www.w3.org/TR/rdfsparql- query, January 2008.	[PS08]
25	S. Wang, G. Englebienne, and S. Schlobach: Learning Concept Mappingd from Instance Similarity International Semantic Web Conference 2008: 339-355	[WES08]
26	Williams, A.J., Harland, L., Groth, P., Pettifer, S., Chichester, C., Willighagen, E.L., Evelo, C.T., Blomberg, N., Ecker, G., Goble, C., Mons, B.: Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today. 17, 1188–1198 (2012).	[WHG+12]
27	K. Zaiss, S. Conrad, and S. Vater. A Benchmark for Testing Instance-Based Ontology Matching Methods. In KMIS, 2010.	[Z10]
28	Jim Gray. Benchmark Handbook: For Database and Transaction Processing Systems, ISBN:1558601597, 1992	[G92]

Acknowledgments & Contact Information

This work has been funded from the European project **LDBC (317548)** and the European project **eHealthMonitor (287509)**.

Contact Information:

Evangelia Daskalaki - eva@ics.forth.gr

Tzanina Saveta - jsaveta@ics.forth.gr

Irini Fundulaki - fundul@ics.forth.gr

Melanie Herschel - melanie.herschel@ipvs.uni-stuttgart.de