A Systematic Investigation of Explicit and Implicit Schema Information on the Linked Open Data Cloud

Thomas Gottron, Malte Knauf, Stefan Scheglmann, Ansgar Scherp

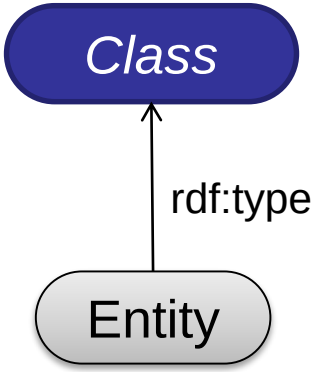
ESWC 2013, Montpellier

WeST
People and Knowledge Networks

Explicit and Implicit Schema Information on LOD

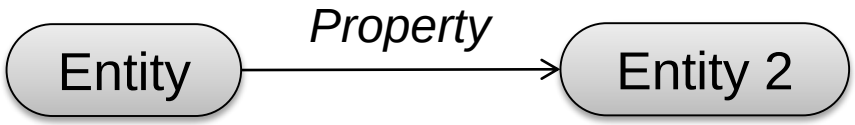
Explicit

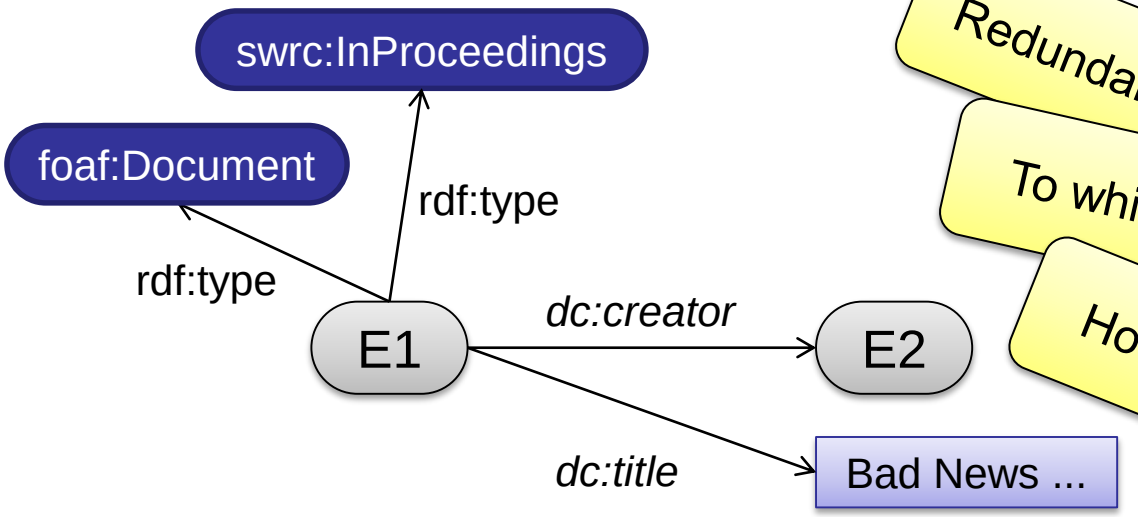
Assigning class types



Implicit

Modelling attributes





Redundancy!

To which degree?

How to measure?

Main questions

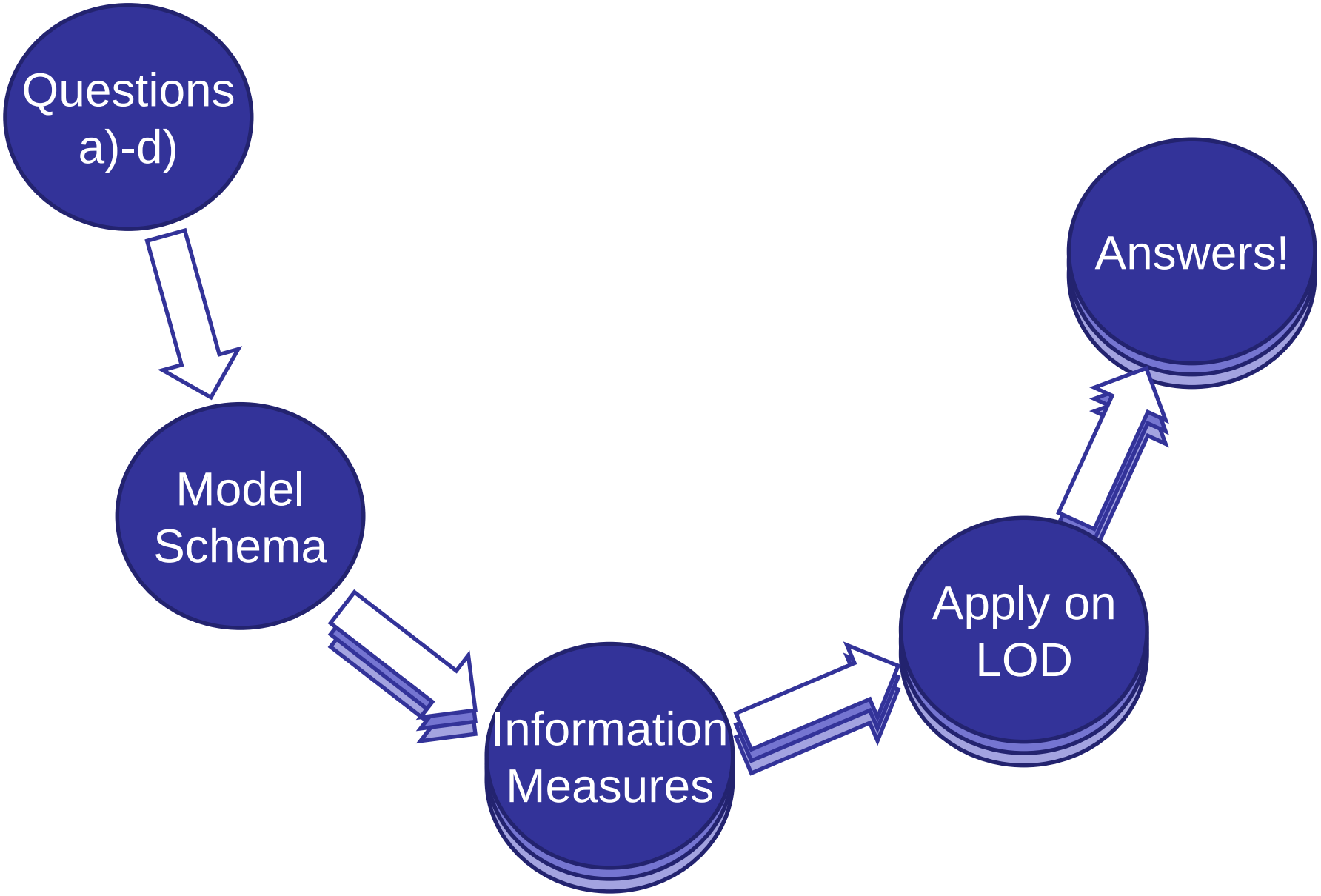
a) How much information is encoded in the type set or property set of a resource?

b) How much information is still contained in the properties, once we know the types of a resource?

c) How much information is still contained in the types, once we know the properties of a resource?

d) To which degree can one information (either properties or types) explain the respective other?

Outline


Questions a)-d)
Model Schema
Information Measures
Apply on LOD
Answers!

Probabilistic model of schema information

- Joint Distribution P(T,R) of
  - Type sets : TS
  - Property sets : PS

- $p(t,r)$: Probability of a resource having set t of class types and set r of properties.

e.g.

dc:creator →

dc:title →

P(T,R)	r_1	r_2	r_3	r_4	P(T)
t_1	14%	2%	5%	8%	29%
t_2	5%	15%	2%	3%	25%
t_3	7%	3%	30%	5%	45%
P(R)	26%	20%	37%	17%	

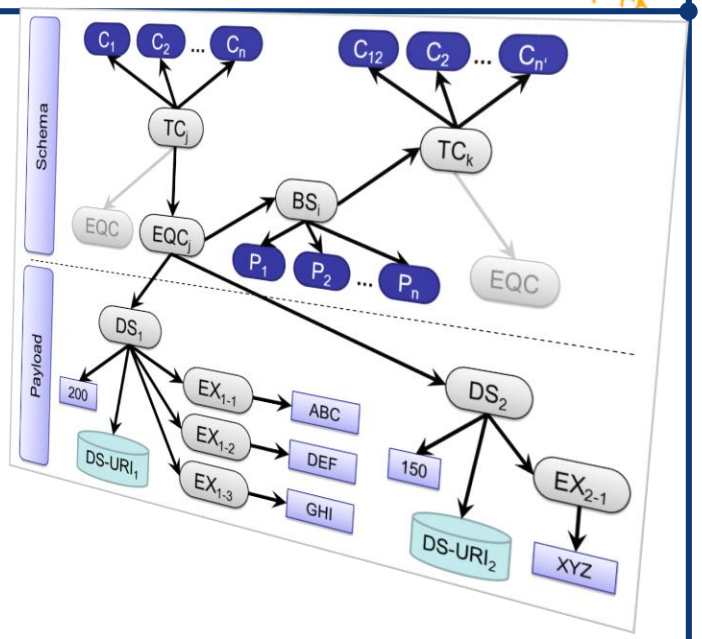
e.g.

swrc:InProceedings

foaf:Document

Marginal Distributions

Estimating probabilities



- Todo (on large-scale data sets)
  - Determine schema use
  - Aggregate
- Query a schema-level index

$$\hat{p}(t,r) = \frac{|d \in \sigma(t,r)|}{N}$$

- Data background: segments from BTC‘12

Data set	Triples	TS	PS
Rest	22.3M	793	7,522
Datahub	910.1M	28,924	14,712
Dbpedia	198.1M	1,026,272	391,170
Freebase	101.2M	69,732	162,023
Timbl	204.8M	4,139	9,619

Question a)

a) How much information is encoded in the type set or property set of a resource?

- Normalised marginal entropy

$$H_0(R) = -\sum_{r \in PS} P(R = r) \cdot \frac{\log_2(P(R = r))}{\log_2(|PS|)}$$

P(R)	26%	20%	37%	17%
P(R)	1%	97%	1%	1%
P(R)	25%	25%	25%	25%

$H_0(R) = 0.967$

$H_0(R) = 0.121$

$H_0(R) = 1.000$

a) Normalized marginal entropy

Data set	$H_0(T)$		$H_0(R)$
Rest	0.252	<	0.366
Datahub	0.263	>	0.250
Dbpedia	0.093	<	0.324
Freebase	0.127	<	0.166
Timbl	0.214	<	0.276

- Tendencies:
  - Entropy of property sets is higher
  - No very high values
  - No values close to zero

Question b)-c)

b) How much information is still contained in the properties, once we know the types of a resource?

c) How much information is still contained in the types, once we know the properties of a resource?

- „Given knowledge“ → conditional probabilities

- Expected conditional entropy

$$H(R|T) = -\sum_{t \in TS} p(T = t) \sum_{r \in PS} P(r|T = t) \cdot \log_2(P(r|T = t))$$

- Entropy we can expect, if we know the types

b)-c) Expected conditional entropy

	P(T,R)	r_1	r_2	r_3	r_4	P(T)	
1.726	t_1	14%	2%	5%	8%	29%	0.501
1.610	t_2	5%	15%	2%	3%	25%	0,403
1.441	t_3	7%	3%	30%	5%	45%	0,648
	P(R)	26%	20%	37%	17%		

$$H(R|T) = 1.567 \qquad H(T|R) = 1.170$$

b)-c) Expected conditional entropy

P(T,R)	r_1	r_2	r_3	r_4	P(T)
t_1	22%	1%	0%	1%	24%
t_2	1%	23%	1%	0%	26%
t_3	1%	0%	48%	1%	50%
P(R)	24%	24%	49%	2%	

$$H(R|T) = 0.384 \qquad H(T|R) = 0.271$$

b)-c) Expected conditional entropy

Data set	H(T|R)		H(R|T)		H(T)	H(R)
Rest	0.289	<	2.568	<	2.428	4.708
Datahub	1.319	>	0.876		3.904	3.460
Dbpedia	0.688	<	4.856		1.856	6.027
Freebase	0.286	<	1.117		2.037	2.868
Timbl	0.386	<	1.464		2.568	3.646

- Tendencies:
  - The types of a resource tell little about its properties
  - Properties tell more about types
  - Given information reduces the entropy

Question d)

d) To which degree can one information (either properties or types) explain the respective other?

- Mutual Information

$$I(T,R) = \sum_{t \in TS} \sum_{r \in PS} p(t,r) \cdot \log_2 \left(\frac{p(t,r)}{P(T=t)P(R=r)} \right)$$

- Normalised Mutual Information (Redundancy)

$$I_0(T,R) = \frac{I(T,R)}{\min(H(T), H(R))}$$

d) Normalized Mutual Information (Redundancy)

P(T,R)	r_1	r_2	r_3	r_4	P(T)
t_1	14%	2%	5%	8%	29%
t_2	5%	15%	2%	3%	25%
t_3	7%	3%	30%	5%	45%
P(R)	26%	20%	37%	17%	

$$I_0(T|R) = 0.239$$

d) Normalized Mutual Information (Redundancy)

P(T,R)	r_1	r_2	r_3	r_4	P(T)
t_1	22%	1%	0%	1%	24%
t_2	1%	23%	1%	0%	26%
t_3	1%	0%	48%	1%	50%
P(R)	24%	24%	49%	2%	

$$I_0(T|R) = 0.766$$

d) Normalized Mutual Information (Redundancy)

Data set	$H_0(T,R)$	
Rest	0.881	1
Datahub	0.747	4
Dbpedia	0.635	5
Freebase	0.860	2
Timbl	0.850	3

- Tendencies:
  - Relatively high redundancy
  - Freebase: (weakly) pre-defined schema
  - Timbl: narrow domain (FOAF profiles)
  - DBpedia: de-centralized schema

Summary

- Present a method to analyze redundancy of schema information on LOD

- Observations
  - Schema not dominated by few TS or PS combinations
  - Attributes provide more information than types
  - Attributes indicate types better than vice versa
  - High redundancy of 63 to 88% on the analyzed segments of the LOD cloud

- Future Work
  - Analysis on data provider level
  - Temporal evolution

Thanks!

WeST

People and Knowledge Networks

Contact:

Thomas Gottron

WeST – Institute for Web Science and Technologies

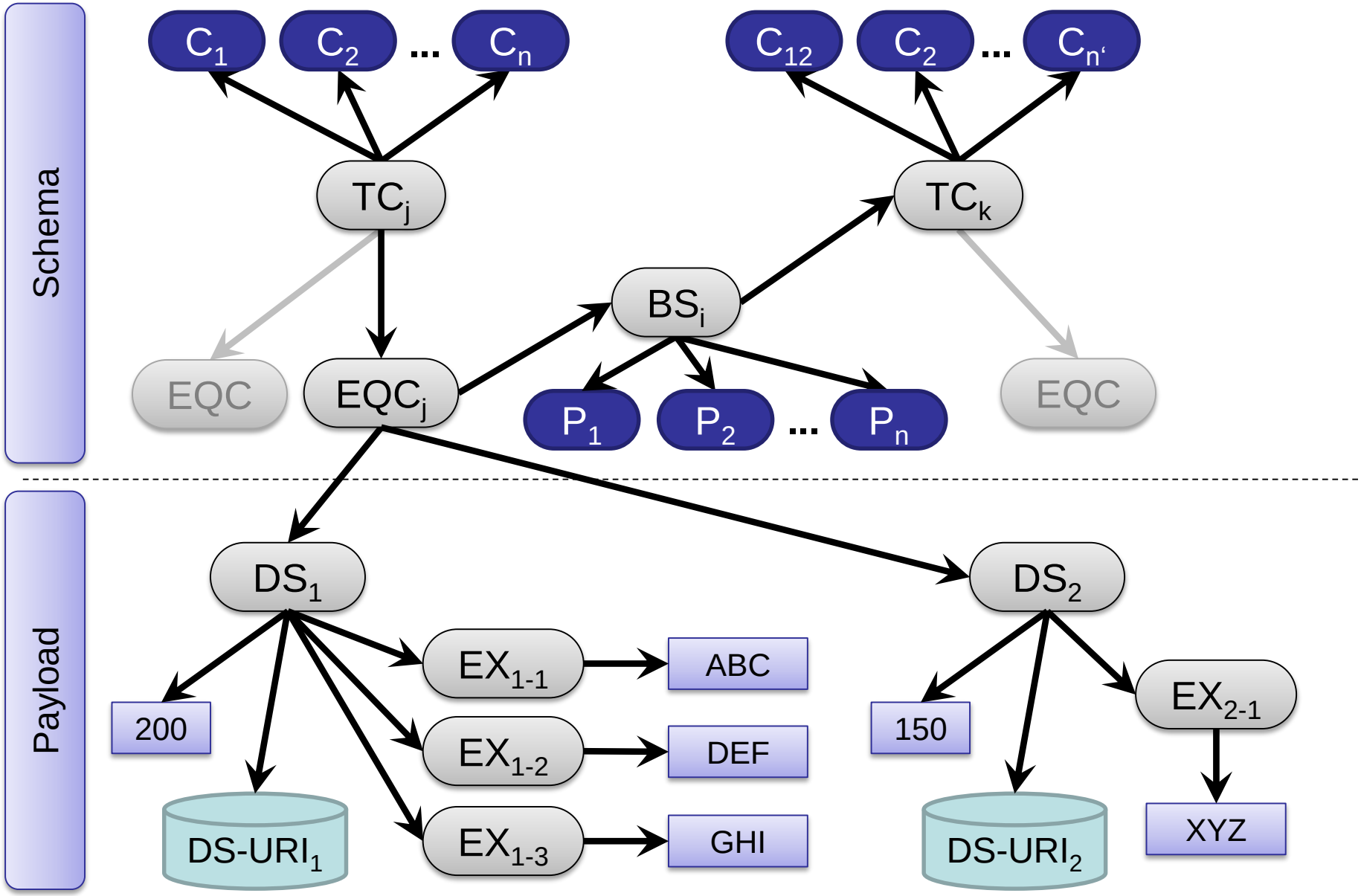
Universität Koblenz-Landau

gottron@uni-koblenz.de

References

1. M. Konrath, T. Gottron, S. Staab, and A. Scherp, “Schemex—efficient construction of a data catalogue by stream-based indexing of linked data,” Journal of Web Semantics, 2012.
2. T. Gottron and R. Pickhardt, “A detailed analysis of the quality of stream-based schema construction on linked open data,” in CSWS’12: Proceedings of the Chinese Semantic Web Symposium, 2012. to appear.
3. T. Gottron, A. Scherp, B. Krayer, and A. Peters, “LODatio: Using a Schema-Based Index to Support Users in Finding Relevant Sources of Linked Data,” in K-CAP’13: Proceedings of the Conference on Knowledge Capture, 2013.

Schema-level Index



Other metrics

Data set	P(H(T|R=r)=0)	P(H(R|T=t)=0)	H(T,R)	H(T)	H(R)
Rest	38.02%	5.31%	4.997	2.428	4.708
Datahub	11.59%	10.83%	4.779	3.904	3.46
Dbpedia	54.85%	3.73%	6.723	1.856	6.027
Freebase	80.89%	2.05%	3.154	2.037	2.868
Timbl	15.15%	1.60%	4.032	2.568	3.646

- Joint Entropy

$$H(T,R) = \sum_{t \in TS} \sum_{r \in PS} p(t,r) \cdot \log_2 p(t,r)$$

Quality of Schema Approximation

	(A) Rest		(B) Datahub (extract)		(C) Timbl (extract)	
	22.3M		20.5M		9.9M	
	lossless	efficient	lossless	efficient	lossless	efficient
$\|T\|$	791	793	3,601	3,656	1,306	1,302
$\|R\|$	8,705	7,522	4,100	4,276	3,015	3,085
$H(T)$	2.572	2.428	3.524	3.487	2.839	2.337
$H_0(T)$	0.267	0.252	0.298	0.295	0.274	0.226
$H(R)$	4.106	4.708	6.008	6.048	3.891	3.258
$H_0(R)$	0.314	0.366	0.501	0.501	0.337	0.281
$H(T\|R)$	0.295	0.289	1.158	1.131	0.670	0.512
$P(H(T\|R=r)=0)$	29.32%	38.02%	60.77%	57.79%	27.81%	21.52%
$H(R\|T)$	1.829	2.568	3.643	3.692	1.723	1.433
$P(H(R\|T=t)=0)$	6.22%	5.31%	12.01%	11.08%	6.06%	4.51%
$H(T,R)$	4.401	4.997	7.166	7.179	4.561	3.770
$I(T,R)$	2.277	2.140	2.365	2.356	2.169	1.824
$I_0(T,R)$	0.885	0.881	0.671	0.676	0.764	0.781

All Results

Data set	Rest	Datahub	DBpedia	Freebase	Timbl
Number of Triples	22.3M	910.1M	198.1M	101.2M	204.8M
$\|TS\|$	793	28,924	1,026,272	69,732	4,139
$\|PS\|$	7,522	14,712	391,170	162,023	9,619
$H(T)$	2.428	3.904	1.856	2.037	2.568
$H_0(T)$	0.252	0.263	0.093	0.127	0.214
$H(R)$	4.708	3.460	6.027	2.868	3.646
$H_0(R)$	0.366	0.250	0.324	0.166	0.276
$H(T\|R)$	0.289	1.319	0.688	0.286	0.386
$H(R\|T)$	2.568	0.876	4.856	1.117	1.464
$H(T,R)$	4.997	4.779	6.723	3.154	4.032
$I(T,R)$	2.140	2.585	1.178	1.751	2.182
$I_0(T,R)$	0.881	0.747	0.635	0.860	0.850