Trust-aware Curation of Linked Open Data Logs

Dihia Lanasri[1] Selma Khouri[1] Ladjel Bellatreche[2]

[1]Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria
[2]LIAS/ISAE-ENSMA, Futuroscope, France

ad_lanasri@esi.dz, s_khouri@esi.dz, bellatreche@ensma.fr

@ER 2020
November 3-6, 2020

Trustworthiness concept

Reference definition

Trust is "the subjective probability with which an agent expects that another agent or group of agents will perform a particular action on which its welfare depends" **[Gambetta, 2000]**

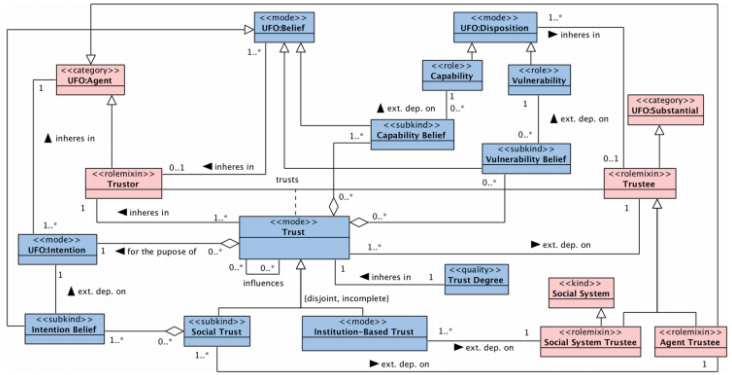
Trust is used in different IT fields

- *Information systems [Valacich et al , HICSS'04]*
- *Social networks [Gong et al., Information Sciences'20]*
- *Requirement engineering [Giorgini et al, Int. J. Inf.'06]*
- *Big data (Veracity 4th V) [Laure Berti-Équille et al. , CIKM'15]*
- *Knowledge bases [Xin Luna Dong et al., Talk VLDB'15]*

Linked Open Data (LOD)

Trust

Ontologies of trust



[Amaral et al, 2019]

Trust in LOD

- *Ownership* ***[Hallo et al, 2016]***
- *Quality* ***[Hallo et al, 2016]***
- *Usefulness* ***[Hitzler et al ,2013]***
- *Correctness* ***[Petrucciani et al, 2015]***
- *Certainty* ***[Behkamal, et al, 2015] …***

Trust in LOD

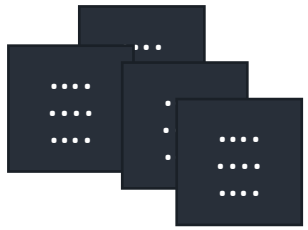
LOD dataset

- A priori : Trust value representation **(tRDF)** & querying **(tSparql)** [Hartig et al, 2009]
- A posteriori: ETL, Curations tools, Provenance annotation [Nath et al, 2020][Abedjan et al, 2014][Behkamal, et al, 2015] [Anam et al., 2015]

What about LOD query-logs ?

Multiple users

Sparql query logs



```
PREFIX swc: http/data.semanticweb.org/ns/swc/ontology
SELECT DISTINCT conf_uri, conf_name, conf_acronym
WHERE {conf_uri a swc:ConferenceEvent. conf_uri rdfs: label conf_name. conf_uri swc:hasAcronym conf_acronym.}
```

Exploitation

- Statistical Analysis [Bonifati et al, 2020]
- Source Selection [Tian et al, 2012]
- Multidimensional Exploration [Khouri et al, 2019]

Trust issues:

- **Veracity** : unknown users, unknown provenance
- **Quality**: users with different expertise level
- **Representation?**

Two goals :
- Need to **define the concept of Trust** for LOD query-logs
- Need to define an **approach for curing logs**

Agenda

1- Ontology-based metamodel of Trust in LOD logs

2- Trust-based curation approach of LOD logs

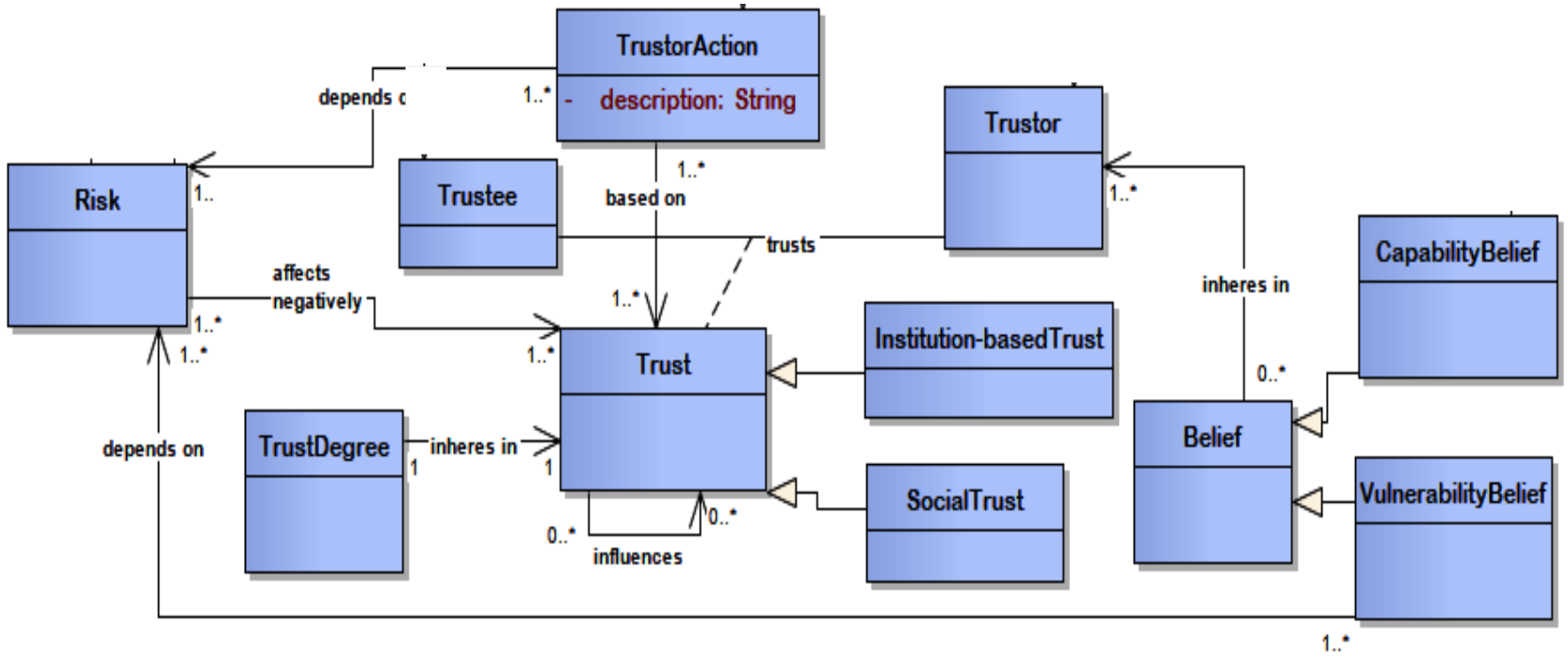
- Log profiling
- ETL-like operators

3- Experimentations & Results

4- Summarization

Meta-Model of Trust in LOD Logs

Reference ontology of Trust **[Amaral et al, 2019]**

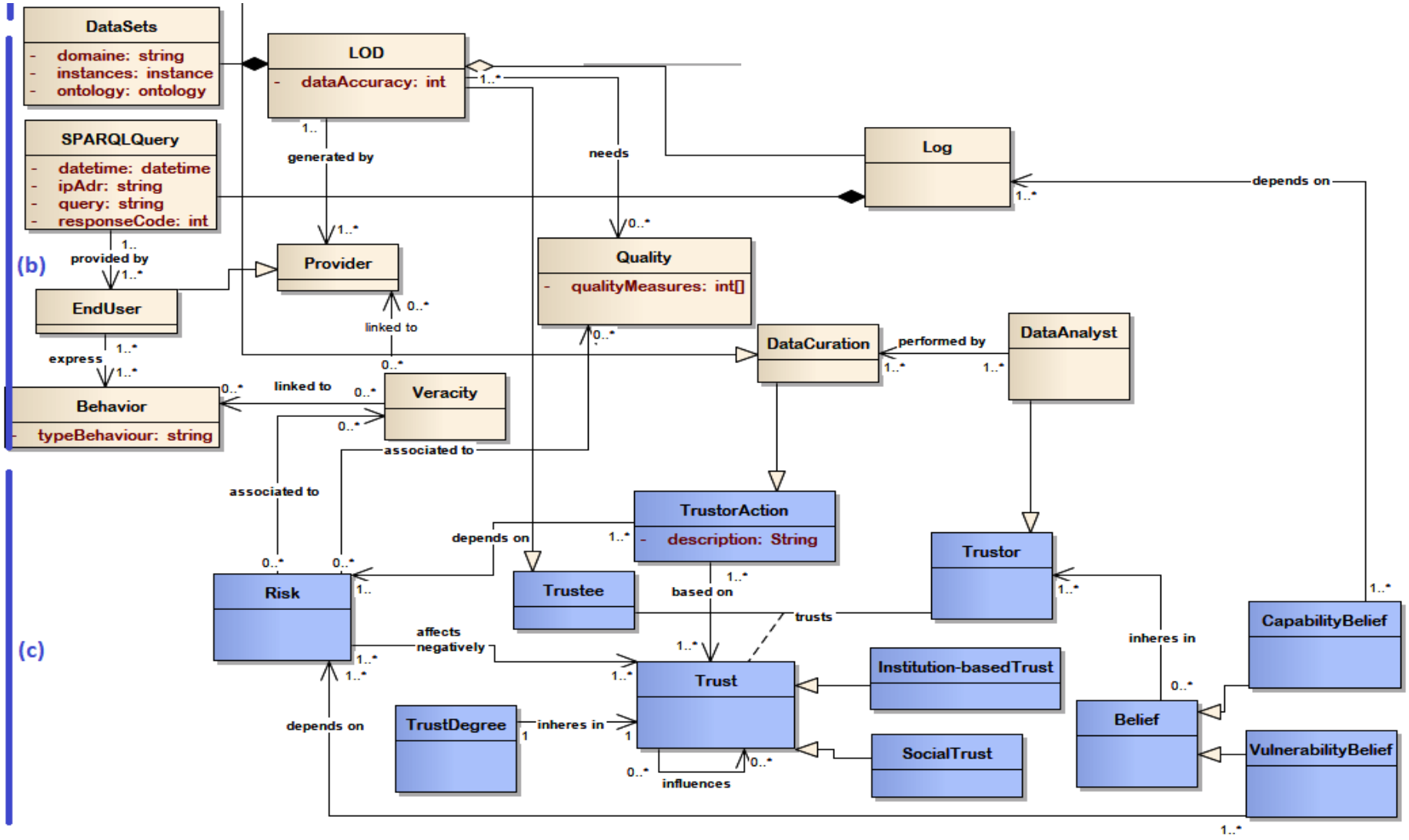


- Ontology-based metamodel of trust for LOD logs:
  - Fragment of Reference Ontology of Trust [Amaral et al, 2019] as a foundation for defining our metamodel
  - Trust is linked to Trustor and Trustee.
  - Trust is composed of a set of Capability & Vulnerability Beliefs to perform the desired action.
  - Vulnerability is manifested by risk events.

Meta-Model of Trust in LOD Logs

In our context of LOD logs:

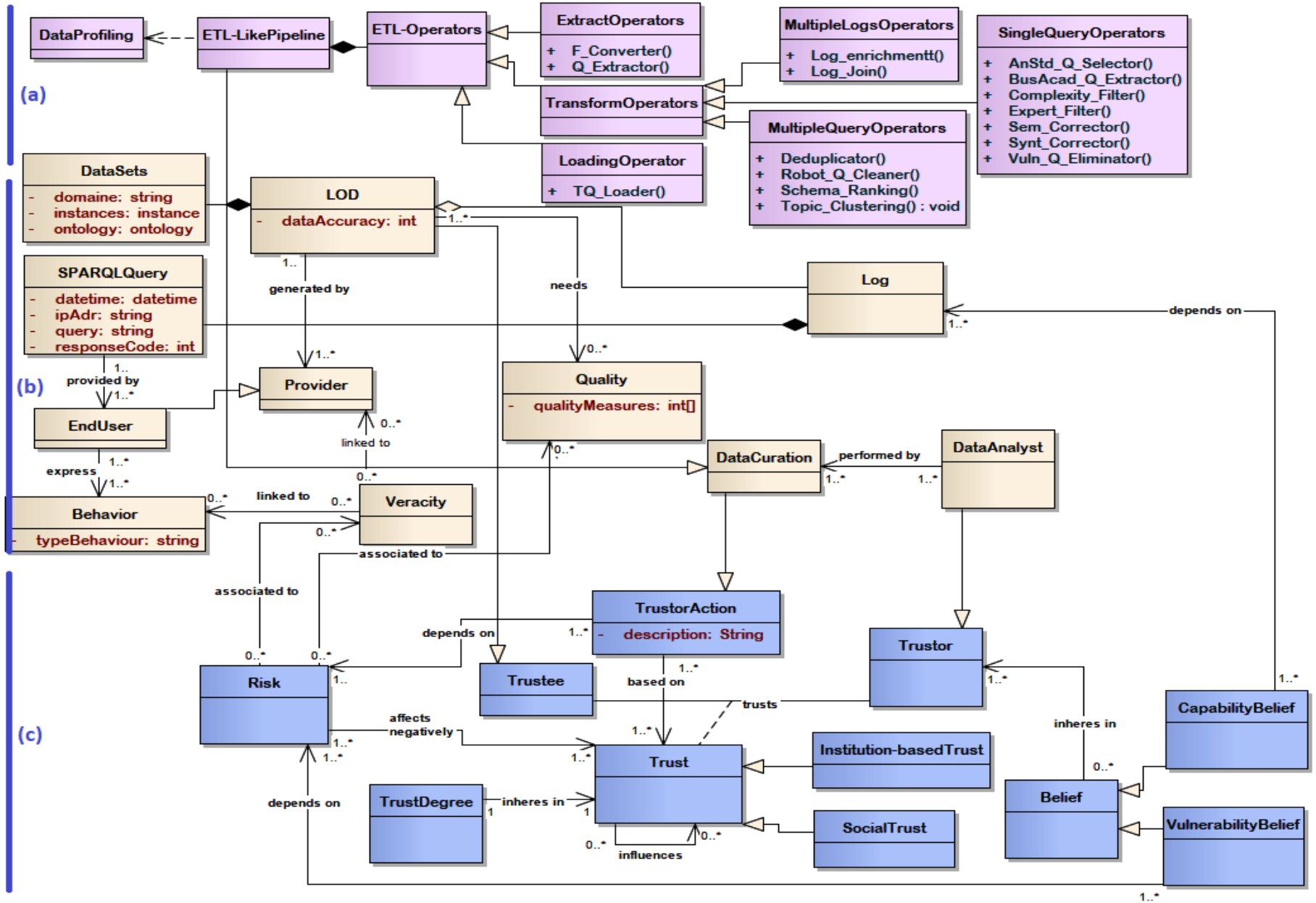
- Trustor : the data analyst
- Trustee : LOD logs
- Capability Belief : set of queries generated in the logs,
- Vulnerability Belief: two risks dimensions : Veracity & Quality



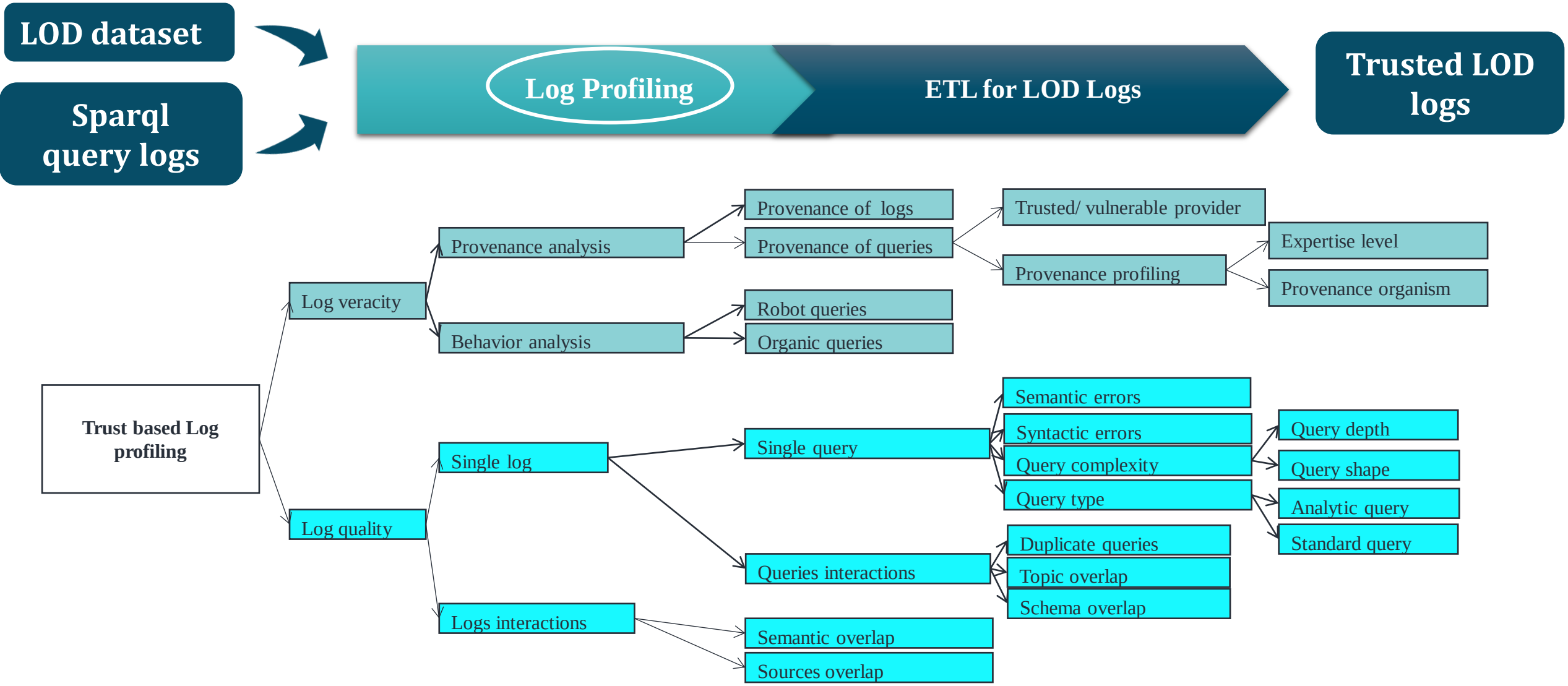
Meta-Model of Trust in LOD Logs

ETL-like operators for LOD logs curation

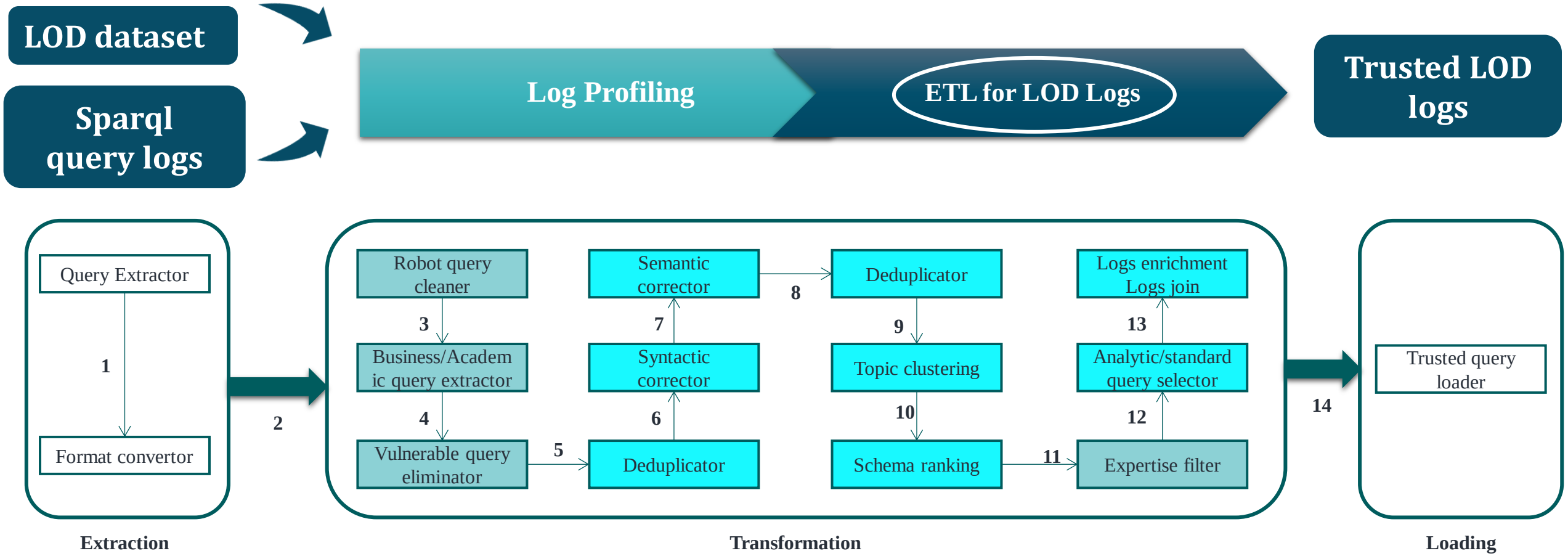
→ ETL operators defined as a curation action for each risk



Our Approach –LOD log profiling



Our Approach–Trust ETL for LOD Logs



- Extract-Transform and Load operators that curate queries.
- ETL operators adapted from traditional ETL and new ETL operators defined
- ETL operators orchestrated to form an ETL-pipeline to be used as a service by analysts.

Our Approach –*Query extractor, parser*

140.203.154.206 - - [*16/May/2014:03:22:51* +0100] "GET /sparql?query=PREFIX+swc%3A+%3Chttp%3A%2F%2Fdata.semanticweb.org%2Fns%2Fswc%2Fontology%23%3E+PREFIX+rdfs%3A+%3Chttp%3A%2F%2Fwww.w3.org%2F2000%2F01%2Frdf-schema%23%3E++++++++++++SELECT+DISTINCT+%3Fconf_uri+%3Fconf_name++%3Fconf_acronym+WHERE+%7B+%3Fconf_uri+a+swc%3AConferenceEvent+.+%3Fconf_uri+rdfs%3Alabel+%3Fconf_name+.%3Fconf_uri+swc%3AhasAcronym+%3Fconf_acronym+.%7D+ORDER+BY+%3Fconf_acronym HTTP/1.0" *200* 15287 "-" "-"

Discard non-relevant information
Extract query
Parsing using HTML parser
Extract Metadata (IP address, date-time, etc)

```
PREFIX  swc: http/data.semanticweb.org/ns/swc/ontology
PREFIX rdfs: http://www.w3.org/rdf-schema
SELECT DISTINCT conf_uri, conf_name, conf_acronym
WHERE {    conf_uri a swc:ConferenceEvent.
           conf_uri  rdfs: label conf_name.
           conf_uri swc:hasAcronym conf_acronym.}
ORDER BY conf_acronym
```

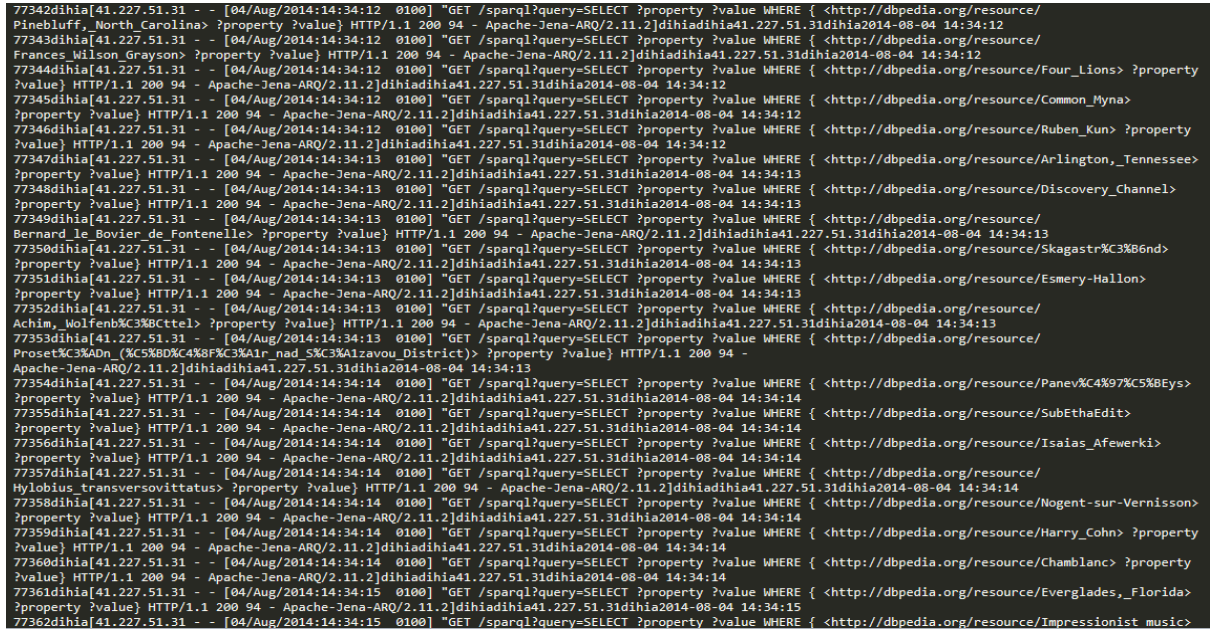
IP: *140.203.154.206*

DateTime: *16/May/2014:03:22:51*

Response code: *200*

Our Approach –*Robot query cleaner*

[**41.227.51.31** - - **04/Aug/2014:14:34:12** 0100] "GET /sparql?query= *SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?value}* HTTP/1.1 **200** 94 - Apache-Jena-ARQ/2.11.2]



More than 1000 queries in around 2 min

Same IP: 41.227.51.31
Same type of query

[**41.227.51.31** - - **04/Aug/2014:14:36:46** 0100] "GET /*sparql?query=SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Eupomatia**> ?property ?value*} HTTP/1.1 **200** 94 - Apache-Jena-ARQ/2.11.2]

Our Approach –Vulnerable query eliminator

193.157.226.245 *SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?value}*

41.227.51.31 *SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?value}*

Database of Blacklisted IPs

Correct

193.157.226.245 *SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?value}*

False

41.227.51.31 *SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?value}*

https://github.com/stamparm/ipsum

Our Approach–Business/academic query extractor

193.157.226.245 *SELECT ?property ?value WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?value}*

Whois ip API

Org Name: National Higher school of computer science

OrgId: ESI

Adress: Oued Smar,

City: Algiers

Country: Algeria

……

Academic Query

Our Approach –Syntactic/Semantic Correctors

SELECT property ***?value*** *WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property ?val*

Missing ?

Undeclared var

Missing ()

Missing }

SELECT ***(?property)*** ***(?value)*** *WHERE { <http://dbpedia.org/resource/**Pinebluff,_North_Carolina**> ?property* ***?value}***

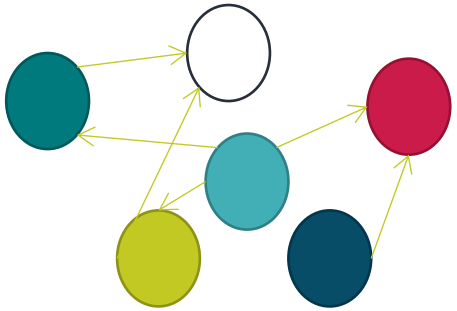
Semantic correction is based on **[Jiménez et al, 2017]**

Our Approach –Topic clustering

```
SELECT DISTINCT ?property ?hasValue

WHERE {<http://data.semanticweb.org/conference/dc/2010/proceedings> ?property ?hasValue .}

ORDER BY ?hasValue
```



LOD ontology

SELECT ?class WHERE {<http://data.semanticweb.org/conference/dc/2010/proceedings> rdfs:subClassOf ?class }

Topic= Document

Our Approach –Schema Ranking

Q1

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .}
ORDER BY ?hasValue
```

Depth = 1

Q2

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .}
```

Depth = 1

Q3

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .
?author  a   foaf:person}
```

Depth = 2

Query ranking based on Query Depth in descendent order

Our Approach –Schema Ranking

Q3

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .
?author  a   foaf:person}
```

Q1

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .}
ORDER BY ?hasValue
```

Q2

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .}
```

Overlap Q3 and Q1, Q2

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .
?author  a   foaf:person}
```

Keep

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .}
ORDER BY ?hasValue
```

delete

```
SELECT DISTINCT  ?author
WHERE {swc:proceedings :hasAuthor ?author .}
```

delete

Our Approach –Complexity Filter

```
SELECT  *    WHERE        { ?SUBJECT <http://purl.org/dc/terms/modified> ?OBJECT}
```

Simple 1

beginner

```
select distinct ?person ?place1 ?place2
where {?person <http://xmlns.com/foaf/0.1/based_near> ?place1.
          ?person <http://xmlns.com/foaf/0.1/based_near> ?place2.
          ?person <http://xmlns.com/foaf/0.1/based_near> ?place3.
          ?person <http://xmlns.com/foaf/0.1/based_near> ?place4
}
```

Star 4

Intermediate

```
SELECT DISTINCT  ?s ?a ?first ?last ?sha              WHERE        { ?s swc:isPartOf
<http://data.semanticweb.org/conference/iswc-aswc/2013/proceedings> .              ?s
swrc:abstract ?a .              ?s foaf:maker ?au .          ?au foaf:firstName ?first .              ?au
foaf:mbox_sha1sum ?sha .                  ?au foaf:lastName ?last                  }
```

Tree 6

Expert

Based on **[Bonifati et al, 2018]**

Experimental Study

Scholarly Data

- Domain: research conferences in SW domain
- Log file size: *5.499.797* queries
- SPARQL queries: 139.932 queries

Dbpedia

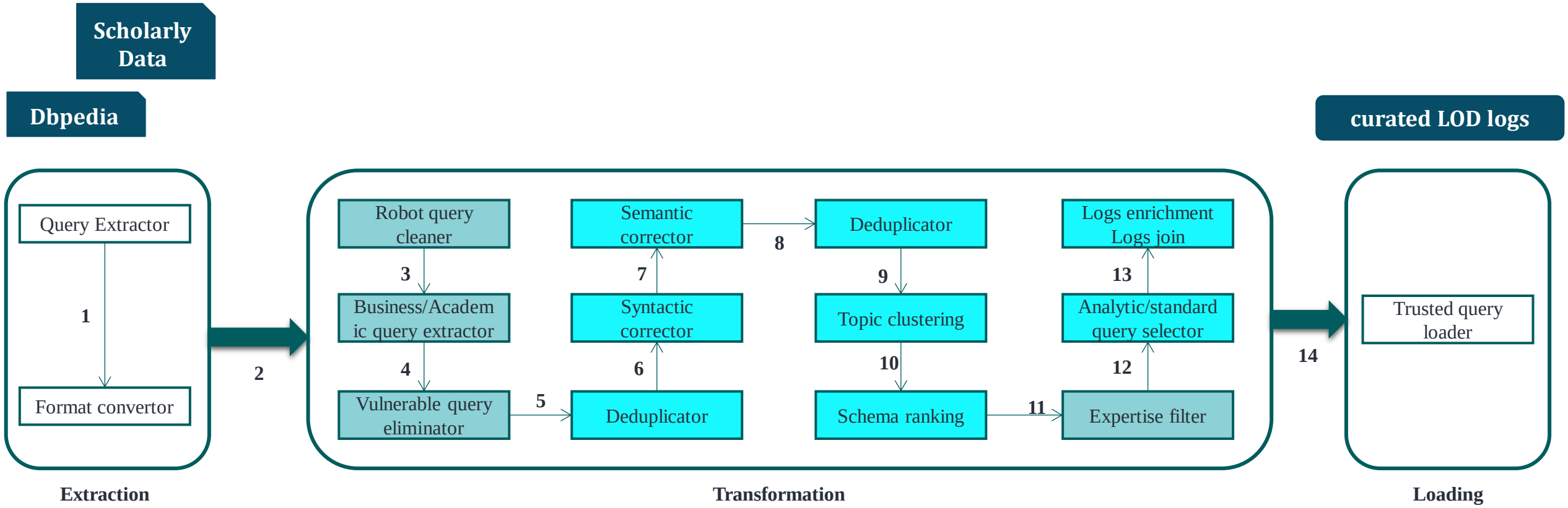
DBpedia-3.5.1

- Domain: Generalist KB
- Log file size: *3.193.672* queries
- Research topic queries: *6.680*

Our experiments are performed on the scholarly data logs and DBpedia logs.

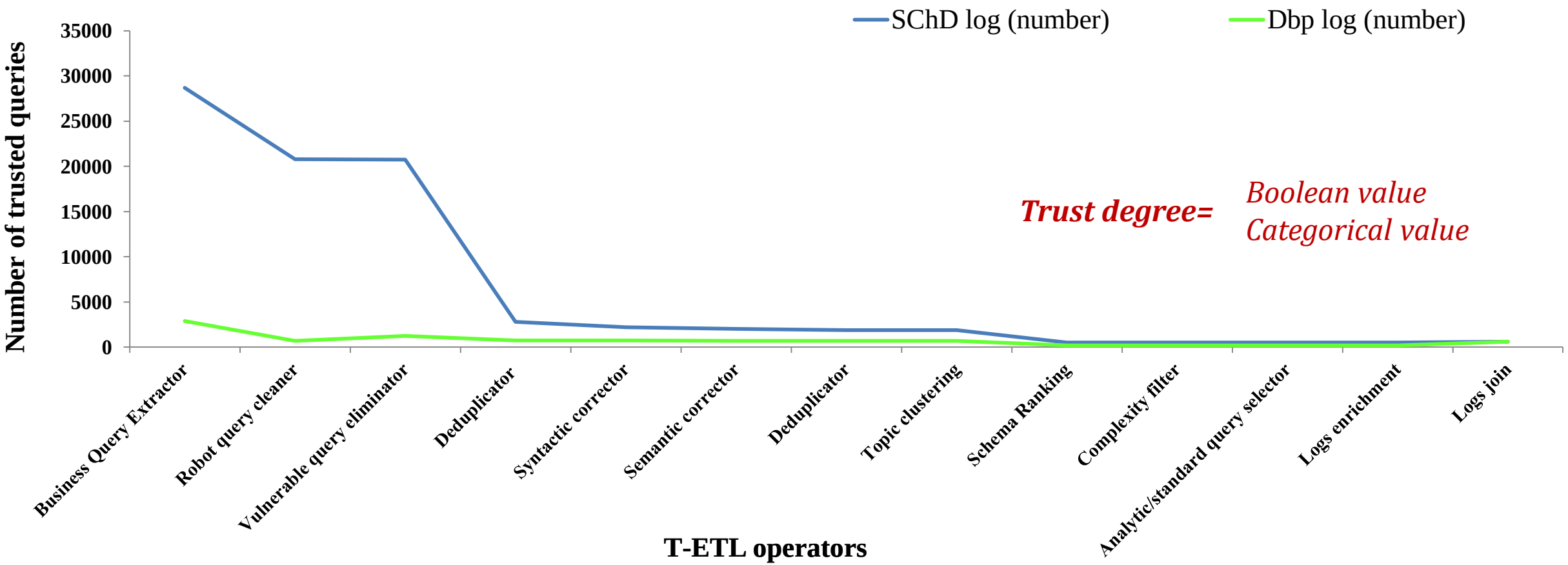
https://aksw.github.io/LSQ/

Experimental Study



→ Trust-aware ETL pipeline that orchestrates the ETL operators

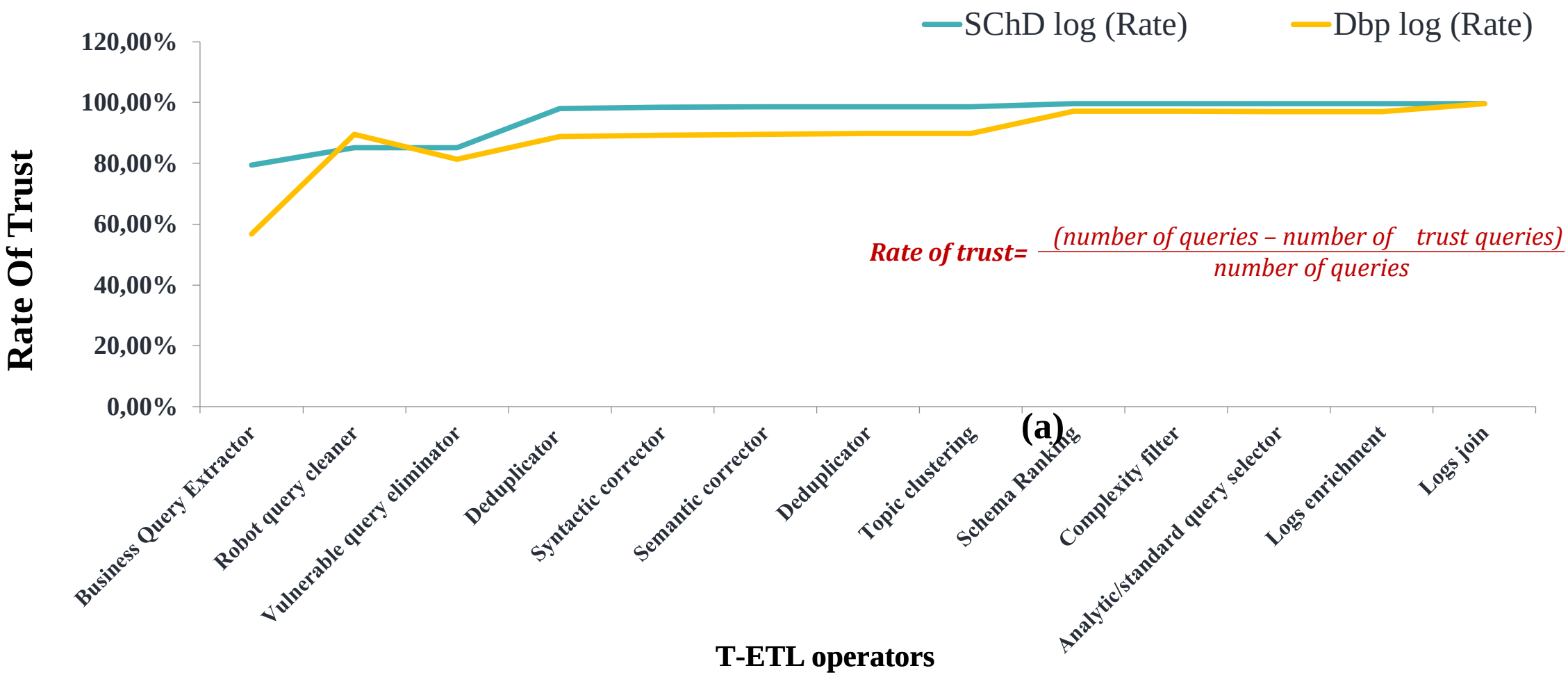
Experimental Study



Two metrics:

- Metric 1: number of trusted queries which are the curated queries parsed by trust-aware ETL operators and annotated by a TrustDegree value

Experimental Study



- Metric 2: Rate of trust
- This metric is improved from 79% to achieve 99% for scholarly data log and from 56% to 97% for Dbpedia log

Conclusion & Perspectives

- LOD logs are rich sources to be exploited for data analysis, but their trust is questionable
- Ontology-based modeling of the concept of trust for LOD query-logs
- Trust-based curation approach for logs: based on Log profiling, and ETL-like mechanism
- Experimentations & tool

As perspectives:

- Consider the trust of LOD datasets
- Trust ponderation
- Impact of trusted logs on decision support systems

References

- Li, X., Valacich, J. S., & Hess, T. J. (2004, January). Predicting user trust in information systems: A comparison of competing trust model In *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the* (pp. 10-pp). IEEE.
- Gong, Z., Wang, H., Guo, W., Gong, Z., & Wei, G. (2020). Measuring trust in social networks based on linear uncertainty theory. *Informatic Sciences*, *508*, 154-172.
- Giorgini, P., Massacci, F., Mylopoulos, J., & Zannone, N. (2006). Requirements engineering for trust management: model, methodology, an reasoning. *International Journal of Information Security*, *5*(4), 257-274.
- Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., ... & Zhang, W. (2015). Knowledge-based trust: Estimating th trustworthiness of web sources. *arXiv preprint arXiv:1502.03519*.
- Gambetta, D. (2000). Can we trust trust. *Trust: Making and breaking cooperative relations*, *13*, 213-237.
- Hallo, M., Luján-Mora, S., Maté, A., & Trujillo, J. (2016). Current state of Linked Data in digital libraries. *Journal of Informatic Science*, *42*(2), 117-127.
- Hitzler, P., & Janowicz, K. (2013). Linked Data, Big Data, and the 4th Paradigm. *Semantic Web*, *4*(3), 233-235.
- Petrucciani, A. (2015). Quality of library catalogs and value of (good) catalogs. *Cataloging & Classification Quarterly*, *53*(3-4), 303-313.
- Abedjan, Z., Grütze, T., Jentzsch, A., & Naumann, F. (2014, March). Profiling and mining RDF data with ProLOD++. In *2014 IEEE 30t International Conference on Data Engineering* (pp. 1198-1201). IEEE.
- Behkamal, B., Kahani, M., & Bagheri, E. (2015, September). Quality metrics for linked open data. In *Database and Expert Systen Applications* (pp. 144-152). Springer, Cham.
- Anam, S., Kang, B. H., Kim, Y. S., & Liu, Q. (2015). Linked data provenance: State of the art and challenges. In *3rd Australasian we conference (AWC 2015)* (Vol. 166, pp. 19-28).
- Deb Nath, R. P., Hose, K., Pedersen, T. B., Romero, O., & Bhattacharjee, A. (2020, April). SETLBI: An Integrated Platform for Semant Business Intelligence. In *Companion Proceedings of the Web Conference 2020* (pp. 167-171).
- Laure Berti-Equille and Javier Borge-Holthoefer. Veracity of big data: From truth discovery computation algorithms to models misinformation dynamics (tutorial). In Proceedings of the 24th ACM International on Conference on Information and Knowledg Management, (CIKM 2015), 2015

References

- Hartig, O. (2009, May). Querying trust in rdf data with tsparql. In *European Semantic Web Conference* (pp. 5-20). Springer, Berlin, Heidelberg.
- Bonifati, A., Martens, W., & Timm, T. (2020). An analytical study of large SPARQL query logs. *The VLDB Journal*, *29*(2), 655-679.
- Tian, Y., Umbrich, J., & Yu, Y. (2011, December). Enhancing source selection for live queries over linked data via query log mining. In *Joint International Semantic Technology Conference* (pp. 176-191). Springer, Berlin, Heidelberg.
- Khouri, S., Lanasri, D., Saidoune, R., Boudoukha, K., & Bellatreche, L. (2019, August). LogLInc: LoG queries of linked open data investigator for cube design. In *International Conference on Database and Expert Systems Applications* (pp. 352-367). Springer, Cham.
- Almendros Jiménez, J. M., Becerra Terón, A., & Cuzzocrea, A. M. (2017). Detecting and Diagnosing Syntactic and Semantic Errors in SPARQL Queries. In *7th ACM International Workshop on Linked Web Data Management*. CEUR-WS.
- Amaral, G., Sales, T. P., Guizzardi, G., & Porello, D. (2019, October). Towards a reference ontology of trust. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 3-21). Springer, Cham.
- Bonifati, A., Martens, W., & Timm, T. (2018, April). Darql: Deep analysis of sparql queries. In *Companion Proceedings of the The Web Conference 2018* (pp. 187-190).

THANK YOU