Linked Open Data Publication Strategies: Application in Networking Performance Measurement Data

Renan F. Souza[1], Les Cottrell[2] , Bebo White[2], Maria L. Campos[1], Marta Mattoso[1]

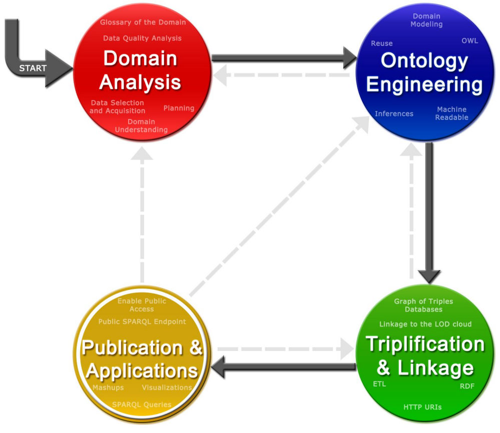
[1]Federal University of Rio de Janeiro, Brazil, [2]SLAC National Accelerator Laboratory

ABSTRACT

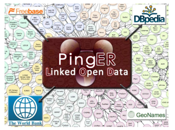
- Most of the data published on the web is unstructured or does not follow a standard.
- It makes the data harder to be retrieved and interchanged between different data sources
- Linked Open Data (LOD) technologies are applied in a scenario that deals with a large amount of computer network measurement data.
- As a result, we generated more structured data, hence easier to be retrieved, analyzed, and more interoperable.
- The challenges of processing large amount of data to: transform it into a standard format (RDF); link it to other data sources; and analyze and visualize the transformed data are discussed.
- An ontology that aims to minimize the number of triples is proposed and a discussion on how ontologies may impact query performance is presented.
- We emphasize the advantages of having the data in RDF format and show use cases on the scenario of the project.

RESEARCH DESIGN AND METHODOLOGY

We proposed the following methodology for Linked Open Data publication:



In the end, we want to have structured, retrievable, and publicly accessible PingER data directly linked to DBpedia [8], Geonames [9], and Freebase [10]. Also, indirectly, to any other data source on the LOD cloud [12].



www.pingerlod.slac.stanford.edu

SYSTEM MODELING AND RESULTS

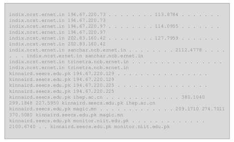
Domain Analysis

PingER Ping end-to-end reporting

Understanding PingER Project's domain

- It envolves data about network performance measurement
- 80 monitoring nodes
- 800 monitored nodes
- 8000 pairs of nodes (monitor-monitored)
- 160 countries, several cities within each country
- 16 network metrics (e.g. TCP throughput, packet loss, average RTT)
- Hourly data, since 1998
- Data can be applied to many different situations such as economical, geographical, and seasonal events.

Data stored in multiple flat CSV files



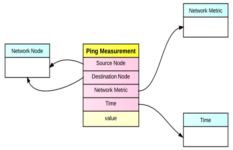
```
indix.ncst.ernet.in 194.67.220.73 . . . . . . . . 113.8784 . . . . . . .
indix.ncst.ernet.in 194.67.220.73
indix.ncst.ernet.in 194.67.220.97 . . . . . . . . 114.0955 . . . . . . .
indix.ncst.ernet.in 194.67.220.97
indix.ncst.ernet.in 202.83.160.42 . . . . . . . . 127.7959 . . . . . . .
indix.ncst.ernet.in 202.83.160.42
indix.ncst.ernet.in sanchar.ncb.ernet.in . . . . . . . . 2112.4778 . . . .
. . . indix.ncst.ernet.in sanchar.ncb.ernet.in
indix.ncst.ernet.in trinetra.ncb.ernet.in . . . . . . . . . . . . . . .
indix.ncst.ernet.in trinetra.ncb.ernet.in
kinnaird.seecs.edu.pk 194.67.220.129 . . . . . . . . . . . . . . . .
kinnaird.seecs.edu.pk 194.67.220.129
kinnaird.seecs.edu.pk 194.67.220.225 . . . . . . . . . . . . . . . .
kinnaird.seecs.edu.pk 194.67.220.225
kinnaird.seecs.edu.pk ihep.ac.cn . . . . . . . . . . . . . 380.1040
299.1848 227.5950 kinnaird.seecs.edu.pk ihep.ac.cn
kinnaird.seecs.edu.pk magic.mn . . . . . . . . . . . . . . 209.1710 274.7011
370.5080 kinnaird.seecs.edu.pk magic.mn
kinnaird.seecs.edu.pk monitor.niit.edu.pk . . . . . . . . . . . . .
2100.6740 . . kinnaird.seecs.edu.pk monitor.niit.edu.pk
```

Problem and Strategies

- Hard to query the CSV files to retrieve specific data, comparing to traditional DBMS
- Hard to produce informative graphs, reports, and dashboards
- Data not interoperable with other data sources

- Data could be published in an open standard format to enable wider consumption
- **Semantic Web** and **Linked Open Data** strategies can be applied to publish PingER structured data in an open standard web format, enabling complex queries to the data and interoperability with other external data sources.

Simple Domain Model



How a ping measurement can be modeled

Ontology Engineering

Ontology Reuse

- An Ontology is needed to model the domain following W3C recommendations.
- Reusing existing ontologies supports the idea of standardization and interoperability within LOD community.

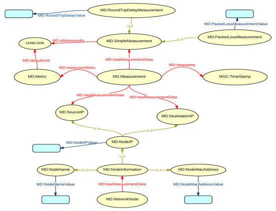
Reuse Evaluation

- Semantic expressivity
- Completeness in relation to the domain
- Impacts on query performance

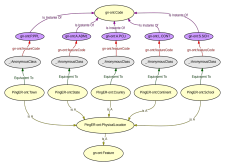
Ontologies being reused

- Geonames [3]
- W3C Time Ontology [4]
- MOMENT [5][6]

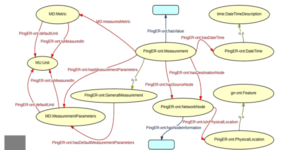
MOMENT Ontology



Geonames Ontology

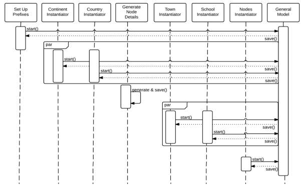


PingER LOD Ontology



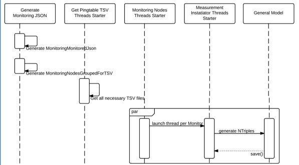
Triplification & Linkage

ETL Process for General Data



- Parallel and Distributed approach to triplify multiple CSV files
- ETL - Extract data from the CSV files, Transform it into Triples format, and Load it into the RDF DBMS
- While the data is being transformed into triples, it is also being linked to external data sources in the LOD cloud.
- Each process is independent, hence can be simultaneously executed in different machines.

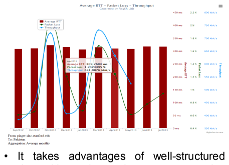
ETL Process for Measurement Data


Generate Monitoring JSON
Get Pingtable TSV Threads Starter
Monitoring Nodes Threads Starter
Measurement Instatiator Threads Starter
General Model
Generate MonitoringMonitoredJson
Generate MonitoringNodesGroupedForTSV
Get all necessary TSV files
par
launch thread per Monitor
generate NTriples
save()

- Each ETL process for Measurement data is responsible for a single network metric and a single time aggregation
- 11 network metrics (throughput, packet loss, etc) and 3 time aggregations (daily, monthly, and yearly)
- 33 processes that can run in distributed machines
- Each process is further parallelized

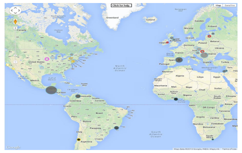
Publication & Applications

Mutiple Network Metrics



- It takes advantages of well-structured data with a schema, in a very expressive format (RDF).
- It explores complex SPARQL queries to capture precisely what is being searched.
- Any possible combination of paramaters is able to be retrieved.

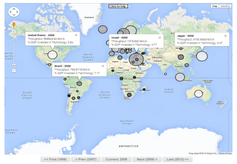
Network Metrics vs. University Metrics



- Illustration of a mashup of PingER data with Dbpedia [9] data about universities (information about number of students, endowment, etc).
- Using this graph, one could visually verify that well-funded universities have better network connectivity.

Network Metrics vs. % of GDP Invested in Research and Development

- PingER data mashed up with World Bank [11] Data.
- It is possible to verify how the countries have invested in Research and Development throughout the years
- And how it has affected network connectivity.



CONCLUSIONS

This work followed the methodology proposed to publish Linked Open Data applied in a real scenario that deals with big datasets about internet measurement. This methodology is based on:

- Domain analysis: understanding the domain and selecting which should be triplified.
- Ontology engineering: reuse evaluation and number of triples minimization
- Triplification project based on a parallel and distributed approach, linking to other data sources in the LOD cloud
- Publication: Enabling public access to both the data and the ontology in a standard, open, structured, and interoperable format, utilizing Semantic Web and LOD technologies.

Results: SPARQL Endpoint is available to query and to interoperate the data; RDF dump of the database is available; and the Ontology is public in OWL format.

FUTURE WORK

- Utilizing complex SPARQL queries (those that are common in database with OLAP characteristics) on the PingER LOD database is still taking undesirable amount of time.
- Thus, in terms of query performance, more research is needed to provide an efficient way of querying very large Triple Stores with OLAP characteristics.

REFERENCES

1. PingER Project. (2014) PingER - Ping end-to-end reporting. [online]. http://www-iepm.slac.stanford.edu/pinger/
2. World Wide Web Consortium. (2013) Semantic Web. [Online]. http://www.w3.org/standards/semanticweb/
3. Geonames. (2013) GeoNames Ontology. [Online]. http://www.geonames.org/ontology/documentation.html
4. World Wide Web Consortium. (2006) Time Ontology. [Online]. http://www.w3.org/TR/owl-time/
5. Sathya Rao, "Monitoring and measurement in the next generation technologies," 2010.
6. European Telecommunications Standards Institute, "Measurement Ontology for IP traffic (MOI); Requirements for IP traffic measurement ontologies development," 2012.
7. Giancarlo Guizzard, "Uma abordagem metodológica de desenvolvimento para e com reúso, baseada em ontologias formais de domínio," 2000.
8. DBpedia. (2014) DBpedia. [Online]. http://dbpedia.org/About
9. Geonames. (2014) About GeoNames. [Online]. http://www.geonames.org/about.html
10. Freebase. (2014) Freebase. [Online]. http://www.freebase.com/
11. World Bank. (2014) The World Bank Data. [online]. http://data.worldbank.org/
12. Richard Cyganiak and Anja Jentzsch. (2012) Linking Open Data Cloud Diagram. [Online]. http://lod-cloud.net/

ACKNOWLEDGEMENT

- This work was supported in part by the Department of Energy contract DE-AC02-76SF0051.
- This work was supported in part by the Rio de Janeiro State Science Foundation (FAPERJ).