PingER Linked Open Data (PingERLOD) overview

Introduction

The proposal of this project is to publish PingER data in Linked Open Data[1] Semantic Web[2] standards. Part of PingER data has been published in Resource Description Framework[3] (RDF) format using an OWL ontology[4] for network measurements and linked to other existing databases on the web. The data is also retrievable through a SPARQL[5] Endpoint, using structured SPARQL queries. Use cases were added as visualization examples to show how the data in this standard format can be useful.

---

[1] Linked Data, see http://en.wikipedia.org/wiki/Linked_data
[2] Semantic Web, see http://en.wikipedia.org/wiki/Semantic_Web
[3] Resource Description Framework, see http://en.wikipedia.org/wiki/Resource_Description_Framework
[4] Web Ontology language, see http://en.wikipedia.org/wiki/Web_Ontology_Language
[5] Search RDF data with SPARQL, see Resource Description Framework RDF

PingER Linked Open Data

Since 1998 PingER stores network measurement data, hourly and daily, measuring more than 10 different metrics between more than 8000 pairs of nodes in more than 160 countries. Thus, there is a huge amount of data to be dealt with. The data is stored in millions of flat CSV files, which are organized using meaningful file names, and the easiest way to get access to it is using the Pingtable[1] application.

However, despite this files structure organization that makes it possible to access the right data, it is far from a standard database system which has well known features and benefits[2]. These features would significantly improve the retrieval and management of the data hence making it much easier to build more complex structured queries in order to retrieve a very specific data to support more informative graphs, reports, and dashboards.

Further, to interoperate and interchange PingER data with other existing data sources is not a very simple task. Since the data could be highly useful and applied to many different situations (economical, geographical, seasonal events, etc.), it would be convenient to make the data easily interoperable to any other kind of data source. If PingER data were available in a standard and widely used format, there could be more joint exploration of the many possibilities that the data could offer and combine it with other kinds of sources, bringing more diversity to its usage and analysis. Besides, utilizing a standard common and open format, more people would consume the data, enabling a possible specific use that has not been anticipated.

Therefore, the target problem of PingER Linked Open Data (LOD) project can be stated as: PingER data is not stored in a conveniently accessible way, hence the ease of production of reports, smart visualizations, and dashboards could be significantly improved, especially when the data involves different data sources. Moreover, the data could be published in an open standard format to enable wider consumption. The project draws on the Semantic Web and LOD strategies and techniques to publish the data according to community and World Wide Web Consortium recommendations.

Once defined the problem, the solution approach begins with selecting part of the huge amount of PingER data. Due to the short period of time the project had to start, only a specific section of the data was considered to be converted and stored in RDF format:

- Regarding the network nodes scope: measurement between all pairs of nodes considered by PingER;
- Regarding the geographic level of detail: PingER stores not only node to node data, but also site to site, country to country, region to region, etc. PingER LOD considers only node to node data, but the geographic hierarchy of the nodes is very well defined so one could easily aggregate the data by country, continent, etc.;
- Regarding the time level of detail: the smallest time grain considered by PingER is *hour*, but the smallest grain considered by PingER LOD is *day*. It was analyzed that to process the entire data, which includes the 24 hours of all days, of all months, of all years, since 1998, would take an impractical amount of time. Additionally, the last 60 days data is continuously being inserted into the PingER LOD triples store, since August 2013.
- Regarding the types of network metrics: Pingtable contemplates 16 metrics, but PingER LOD considers only the 11 more important ones: Mean Opinion Score, Directivity, Average Round Trip Time, Conditional Loss Probability, Duplicate Packets, Inter Packet Delay Variation, Minimum Round Trip Delay, Packet Loss, TCP Throughput, Unreachability, and Zero Packet Loss Frequency.
- Regarding the network packet size: Pingtable contemplates the packet sizes 100 and 1000 bytes. It was analyzed that considering both sizes would take approximately double of the time and only one of them would be enough to satisfy the initial goals of the project. Thus, PingER LOD contemplates only packet sizes of 100 bytes.

After selecting the section of the data that will be published in RDF, a simple conceptual model was designed based on the PingER data characteristics: each measurement is basically defined by a ping sent from a source (or monitor) node to a destination (or monitored) node, sent in a determined time, and related to a specific network metric. Additionally, the data is accumulated in a historical base within the years and can be used to support to decision making. Data with these characteristics can be modeled using a well-known and studied data model: star schem[3]a which the designed conceptual was based on.

This model was used to support the construction of the OWL ontology for the PingER conceptual domain. The PingER LOD ontology is useful for organizing, structuring, and formalizing (part of) the PingER knowledge so it can easily be retrieved and processed by a machine. Most of the ideas behind the PingER LOD ontology were reused based on an existing network measurement domain ontology proposed by the project “Monitoring and Measurement in the Next Generation Technologies”[4] (MOMENT), which produced a Measurement Ontology for IP (MOI) traffic, that is an European Telecommunications Standards Institute (ETSI) Group Specification[5]. In addition to the network measurement part, the Geonames ontology[6] covered the geographic concepts used by PingER and the W3C Time Ontology[7] covered the time concepts. The entire OWL PingER LOD ontology in [CL1] and a further documentation are available[8].

Having the ontology, the next phase in the process of publishing LOD is the triplification. The triplification consists of Extracting PingER raw data, Transforming it into RDF triples format, and Loading onto the RDF database; this process is known as Extract, Transform and Load (ETL). Currently, the RDF database management system used on the project is Open RDF Sesame Native 2.7.2[9], but more complex queries run very slowly given the big number of triples stored for measurements (more than 50 million). However, experiments are being performed using other alternatives such as OWLIM[10] and Open Link Virtuoso[11] technologies, but to upload the data takes a very long time so more tests are needed to give more precise results. The ETL process for general concepts (like time, geography, and universities) is not complex and takes a considerably short time (the longest part takes around 4 hours to complete). General concepts data is extracted from DBPedia[12], Freebase[13], and Geonames, transformed into the right format to be adequate to the ontology, and loaded onto the RDF database. The ETL process for network measurement data is very complex since it deals with the big data part. Network measurement data is extracted using Pingtable mechanism, transformed into triples accordingly to the ontology, and loaded onto the RDF database. This is one the most complex tasks within the whole project; parallel and distributed programming techniques were widely used in order to make it possible.

Finally, after populating the RDF database with PingER data, a SPARQL Endpoint was made available[14] so people could access the RDF data using structured SPARQL queries.

Additionally, easy interfaces were developed to help people to build SPARQL Queries and to plot graphs in order to show some of the advantages of using LOD techniques. Specifically, three cases were developed (they are available on the website[15], Visualizations tab):

- Multiple network metrics analysis: This case utilizes PingER data only. It exemplifies how LOD aids even when not using mashups (crossing PingER data with other data sources). It highlights the advantage of having well-structured data with a schema, in a very expressive format: RDF triples. It also explores the use of complex SPARQL queries which are able to capture precisely what is being searched. A single query can retrieve network measurements using any possible combination of parameters. After running the query, a graph is plotted showing multiple network metrics simultaneously for the specified parameters. Before the project, the task of combining multiple metrics in a single data sheet to build a graph was not simple.
- Crossing network metrics with university metrics: Since most of the nodes considered by PingER are educational institutions, it was convenient to verify the relation between network quality and university quality. This case exemplifies a mashup, crossing PingER data with DBPedia data about universities. To measure university quality, some metrics were retrieved: number of students, number of undergrad and grad students, faculty size, and endowment. After running the mashed query, a map is drawn plotting circles on the universities PingER monitors. Size of the circles represents the value of the university metric (e.g., the bigger the circle, the greater the number of students), the filling color of the circles represents the value of the network metric (e.g., the whiter the color, the greater the value of throughput from a PingER monitor to that university), and the stroke color represents the type of universities (e.g., gold color represents private universities). Thus, using this graph one could visually verify that better universities have better network connection.
- Crossing network metrics with percentage of GDP that countries invest in research and development: this exemplifies another mashup with PingER data with another RDF data source, which is, in this case, World Bank Data. PingER data gives network measurements to many countries on the globe and World Bank gives many different interesting indicators[16], including countries' Research and Development (R&D) expenditure (% of GDP). Similarly, a map is drawn and circles are plotted on countries. The bigger the circle, the more it is invested in R&D by the country and the whiter the color, the greater the value of the network metric is. Moreover, an evolution within the years (since 1998) of the countries' investment in R&D as well as their network quality can be easily visualized on the map.

In addition to those listed cases, since the data is well-structured in a standard and interoperable format, anyone can now use PingER LOD to retrieve data and make more interesting analysis, especially combining network measurement data with data of any other different characteristic.

For the future, the most urgent need is to determine which RDF triplestore to use. As stated previously, experiments are being currently made using Open Link Virtuoso and OWLIM repositories. Furthermore, other points can be listed:

- The updating system is not very efficient: If, for example, a specific node changes its full name, the system will store both old and new names, with no distinction between the values. Another worse case happens if, for some reason, two different measurement values are captured for the exact same query parameters: both of them will be stored with no distinction. This needs to be understood and fixed.
- Many universities are not being captured using the current system, even though they are monitored by PingER. More research is needed to provide better results.
- More documentation needs to be provided.
- More tests are needed to verify the consistency of the data.

A better analysis of the ontology occurred recently and changes have happened to provide better semantic and organization to the terms. The data should be re-uploaded accordingly to the new ontology.

---

[1] Pingtable, see http://www-wanmon.slac.stanford.edu/cgi-wrap/pingtable.pl
[2]. Fundamentals of Database Systems. 6.ed, NAVATHE, Shamkant; ELMASRI, Ramez, Addison-Wesley, 2010.
[3] Star schema, see http://www.learndatamodeling.com/star.php#.UogBUsTUDIF
[4] MOMENT: Monitoring and Measurement in the Next Generation Technologies, see http://www.salzburgresearch.at/en/projekt/moment_en/
[5] ETSI Industry Specification Groups, see http://www.etsi.org/about/how-we-work/industry-specification-groups
[6] http://www.geonames.org/ontology/documentation.html
[7] http://www.w3.org/TR/owl-time/
[8] https://confluence.slac.stanford.edu/display/IEPM/Ontology
[9] http://openrdf.callimachus.net/sesame/2.7/docs/users.docbook?view
[10] Ontotext, see http://en.wikipedia.org/wiki/Ontotext
[11] Virtuoso, see http://virtuoso.openlinksw.com/
[12] DBpedia, see http://en.wikipedia.org/wiki/DBpedia
[13] Freebase, see http://en.wikipedia.org/wiki/Freebase
[14] http://pingerlod.slac.stanford.edu/sparql
[15] http://pingerlod.slac.stanford.edu
[16] http://data.worldbank.org/indicator

[CL1]in what, this does not make grammatical sense tehre is something missing.