LinkedDataLens: Linked Data as a Network of Networks

Paul Groth
VU University Amsterdam
De Boelelaan 1081a
Amsterdam, 1081 HV
The Netherlands
p.t.groth@vu.nl

Yolanda Gil
Information Sciences Institute
University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292, USA
gil@isi.edu

ABSTRACT

With billions of assertions and counting, the Web of Data represents the largest multi-contributor interlinked knowledge base that ever existed. We present a novel framework for analyzing and using the Web of Data based on extracting and analyzing thematic subsets of it. We view the Web of Data as a "network of networks" from which to extract meaningful subsets that can be converted them into self-contained networks to be further analyzed and reused. These extracted networks can then be analyzed through network analysis and discovery algorithms, and the results of these analyses can be published back on the Web of Data. We describe LinkedDataLens, an implementation of this framework that uses the Wings workflow system to represent multi-step network extraction and analysis processes.

Categories and Subject Descriptors

I.2.11 Distributed Artificial Intelligence; I.2.8 Problem Solving, Control Methods, and Search; H.4 Information Systems Applications; I.2.4 Knowledge Representation Formalisms and Methods.

General Terms

Algorithms, Languages.

Keywords

Knowledge capture, network analysis, linked open data.

INTRODUCTION

There is an increasing amount of open interlinked data sets on the Web collectively known as the Web of Data (WoD) [2]. As of September 2010, it contained more than 203 data sets totaling over 25 billion RDF triples, which are interlinked by around 395 million RDF links and released under open licenses. Techniques for analyzing and understanding the current WoD as a complex artifact are of great value to the community [5].

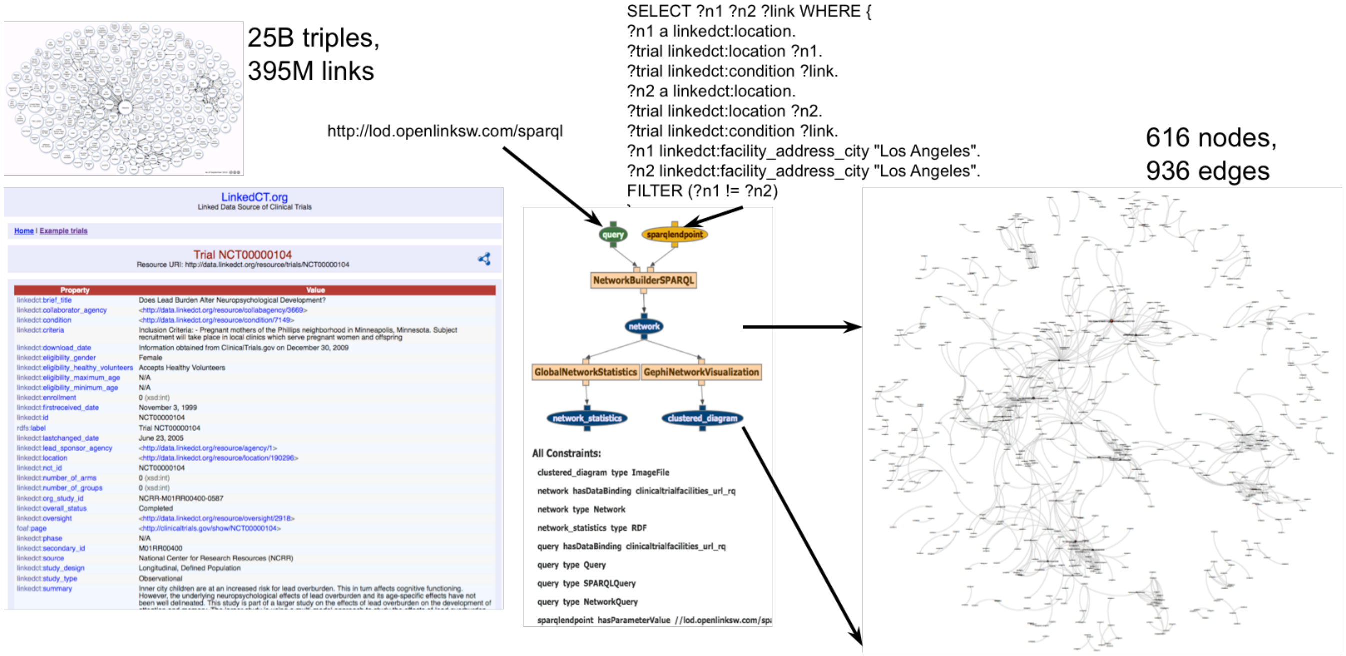
This paper presents LinkedDataLens, a framework for understanding WoD by extracting meaningful portions of it and characterizing them. We view the WoD as a "network of networks." Diverse datasets, such as Geonames and DBpedia, are interlinked into a massive network. Within this WoD network, one could identify smaller self-contained subsets represented in turn as networks. These extracted networks could span more than one dataset. For example, one could construct a temporal network of events containing all rock concerts in a geographical region, which would integrate information from event and geospatial sources. Each of these extracted networks represents a meaningful aspect of some phenomenon, and can be studied and characterized in its own right. Using network analysis algorithms, we can derive useful summary statistics, detect clusters, and infer new links. The resulting analyses can be seen as *metadata* of the extracted networks. This metadata can be used to formulate queries to search for networks or entities of interest with particular characteristics. For example, finding whether social networks have parallel network properties to the content networks they are associated with.

LINKEDDATALENS

LinkedDataLens addresses a number of challenges to extracting and analyzing the WoD as a network of networks. First, the networks to be analyzed may not be directly accessible within the WoD network. For example, WoD resources may be connected by multi-hop paths rather than being directly connected in a network by a single relation. Similarly, WoD links may be represented by resources rather than by edges in a network. Secondly, most network algorithms do not directly ingest RDF data. Finally, comprehensive metadata and provenance about the extracted networks need to be maintained in order to facilitate search.

Figure 1 gives an overview of how LinkedDataLens works. It uses a workflow system [3,4] to represent network analysis processes. The inputs to the workflow are typically a query to the WoD and a location to access it. When workflows are executed, networks of interest are extracted and analyzed. Finally, the networks extracted and their derived characteristics are published back to the WoD.

K-CAP'11, June 26–29, 2011, Banff, Alberta, Canada.
ACM 978-1-4503-0396-5/11/06.



```
SELECT ?n1 ?n2 ?link WHERE {
?n1 a linkedct:location.
?trial linkedct:location ?n1.
?trial linkedct:condition ?link.
?n2 a linkedct:location.
?trial linkedct:location ?n2.
?trial linkedct:condition ?link.
?n1 linkedct:facility_address_city "Los Angeles".
?n2 linkedct:facility_address_city "Los Angeles".
FILTER (?n1 != ?n2)
}
```

LinkedCT.org
Linked Data Source of Clinical Trials

Home | Example trials

Trial NCT00000104
Resource URI: http://data.linkedct.org/resource/trials/NCT00000104

Property	Value
linkedct:brief_title	Does Lead Burden Alter Neuropsychological Development?
linkedct:collaborator_agency	<http://data.linkedct.org/resource/collabagency/3669>
linkedct:condition	<http://data.linkedct.org/resource/condition/7149>
linkedct:criteria	Inclusion Criteria: - Pregnant mothers of the Phillips neighborhood in Minneapolis, Minnesota. Subject recruitment will take place in local clinics which serve pregnant women and offspring
linkedct:download_date	Information obtained from ClinicalTrials.gov on December 30, 2009
linkedct:eligibility_gender	Female
linkedct:eligibility_healthy_volunteers	Accepts Healthy Volunteers
linkedct:eligibility_maximum_age	N/A
linkedct:eligibility_minimum_age	N/A
linkedct:enrollment	0 (xsd:int)
linkedct:firstreceived_date	November 3, 1999
linkedct:id	NCT00000104
rdfs:label	Trial NCT00000104
linkedct:lastchanged_date	June 23, 2005
linkedct:lead_sponsor_agency	<http://data.linkedct.org/resource/agency/1>
linkedct:location	<http://data.linkedct.org/resource/location/190296>
linkedct:nct_id	NCT00000104
linkedct:number_of_arms	0 (xsd:int)
linkedct:number_of_groups	0 (xsd:int)
linkedct:org_study_id	NCRR-M01RR00400-0587
linkedct:overall_status	Completed
linkedct:oversight	<http://data.linkedct.org/resource/oversight/2918>
foaf:page	<http://clinicaltrials.gov/show/NCT00000104>
linkedct:phase	N/A
linkedct:secondary_id	M01RR00400
linkedct:source	National Center for Research Resources (NCRR)
linkedct:study_design	Longitudinal, Defined Population
linkedct:study_type	Observational
linkedct:summary	Inner city children are at an increased risk for lead overburden. This in turn affects cognitive functioning. However, the underlying neuropsychological effects of lead overburden and its age-specific effects have not been well delineated. This study is part of a larger study on the effects of lead overburden on the development of

All Constraints:

```
clustered_diagram type ImageFile
network hasDataBinding clinicaltrialfacilities_url_rq
network type Network
network_statistics type RDF
query hasDataBinding clinicaltrialfacilities_url_rq
query type Query
query type SPARQLQuery
query type NetworkQuery
sparqlendpoint hasParameterValue //lod.openlinksw.com/sp
```

```
<96941b6b3c52aec66bfc14238d46a7cb> <nd#numberOfNodes> "616"^^<http://www.w3.org/2001/XMLSchema#integer>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#numberOfEdges> "936"^^<http://www.w3.org/2001/XMLSchema#integer>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#density> "0.00494140006335"^^<http://www.w3.org/2001/XMLSchema#decimal>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#averageClusteringCoefficient> "0.386224819698"^^<http://www.w3.org/2001/XMLSchema#decimal>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#isConnected> "False"^^<http://www.w3.org/2001/XMLSchema#boolean>.
<96941b6b3c52aec66bfc14238d46a7cb> <nd#numberOfConnectedComponents> "118"^^<http://www.w3.org/2001/XMLSchema#integer>.
```

Figure 1. An overview of how LinkedDataLens works.

Our framework consists of the following three main steps:

1. *Pattern-based network extraction from the WoD.* Our workflows typically start off with a generic component that is given a patterned query and a SPARQL Endpoint and extracts a network.
2. *Characterization of the extracted networks with statistics*. The workflow includes steps to analyze and visualize the network using multiple network analysis algorithms. To facilitate interoperability between components, we adopt the PAJEK format as as a standard serialization to communicate networks among components [1].
3. *Publication of networks back to WoD with associated statistics and provenance metadata.* LinkedDataLens takes advantage of the capabilities offered by workflow systems to record the provenance of the network and its characterization. Using the provenance we can navigate from the characterizations of the network to the network itself, as well as from the network to its characterization.

An example of a network is shown on the right side of Figure 1. It is a network of facilities within Los Angeles that have investigated the same condition in a clinical trial. The network represents 616 facilities with 936 connections between them. It makes apparent which universities are involved in many clinical trials, and which pharmaceutical companies are running most clinical trials in the area.

Through LinkedDataLens, we generate three kinds of useful artifacts: 1) the extracted networks themselves, 2) their derived characteristics as metadata, and 3) the analytic processes used to derive them.

ACKNOWLEDGMENTS

This research was funded in part by the National Science Foundation under grant number IIS-0948429.

REFERENCES

[1] Batagelj, V. and Mrvar, A. “Pajek: Analysis and visualization of large networks.” In M. Junger and P. Mutzel (Eds), Graph Drawing Software. Springer, 2003.

[2] Bizer, C.; Heath, T.; and Berners-Lee, T. “Linked Data - The Story So Far.” International Journal on Semantic Web and Information Systems, 5(3), 2009.

[3] Gil, Y.; Ratnakar, V.; Kim, J.; Gonzalez-Calero, P. A.; Groth, P.; Moody, J.; and Deelman, E. “Wings: Intelligent Workflow-Based Design of Computational Experiments.” IEEE Intelligent Systems, 26(1), 2011.

[4] Gil, Y.; Gonzalez-Calero, P. A.; Kim, J.; Moody, J.; and Ratnakar, V. “A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs.” To appear in the Journal of Experimental and Theoretical Artificial Intelligence, 2011.

[5] Guéret, C.; Groth, P.; Harmelen, F.V.; and Schlobach, S. “Finding the Achilles Heel of the Web of Data: using network analysis for link-recommendation,” 9th International Semantic Web Conference (ISWC), 2010.