Linked Data Query Processing
Tutorial at the 22nd International World Wide Web Conference (WWW 2013)
May 14, 2013

http://db.uwaterloo.ca/LDQQTut2013/

Olaf Hartig
University of Waterloo
Tutorial Outline

(1) Introduction

(2) Theoretical Foundations

(3) Source Selection Strategies

(4) Execution Process

(5) Query Planning and Optimization
1. Introduction

Olaf Hartig
University of Waterloo
Outline

- The Linked Data Principles
- Paradigms for Querying Linked Data
- Characteristics of the “Database System”
The Traditional, Hypertext Web

Data exposed to the Web via HTML

CIA World Factbook

IMDb

The World Factbook

Albania

Location: Southeastern Europe, bordering the Adriatic Sea and Ionian Sea, between Greece in the south and Montenegro and Kosovo to the north

Geographic coordinates: 41 00 N, 20 00 E

Map references: Europe

Area: total: 28,748 sq km
land: 27,396 sq km
water: 1,350 sq km

Area - comparative: slightly smaller than Maryland

Land boundaries: total: 717 km
border countries: Greece 282 km, Macedonia 151 km, Montenegro 172 km, Kosovo 112 km

Coastline: 362 km

Maritime claims: territorial sea: 12 nm
continental shelf: 200 m depth or to the depth of exploitation

Overview
User Rating: [Rating] (Investigating 5 votes)

Director: Michael Davis
Release Date: 12 July 1999 (USA) more>
Genre: Documentary | War more>
Tagline: Young people affected by the war in Kosovo
Plot Keywords: Spoiler alert! more

Additional Details
Parents Guide: Add content advisory for parents
Runtime: USA: 26 min
Country: USA
Language: English
Towards a Web of Linked Data

War Child, release date, 12 July 1999

War Child, filming location, Albania

Michael Davie, directed, War Child

CIA World Factbook

Data model: RDF

MovieDB

CIA World Factbook

(Albania, unemployment rate, 13.2%)
Towards a Web of Linked Data

Data model: **RDF**
Global identifier: **URI**

MovieDB

CIA World Factbook

( http://...imdb.../WarChild , release date , 12 July 1999 )
( http://...imdb.../WarChild , filming location , http://cia.../Albania )
( http://...imdb.../MichaelDavie , directed , http://...imdb.../WarChild )

( http://cia.../Albania , unemployment rate , 13.2% )
Towards a Web of Linked Data

**Data model:** RDF

**Global identifier:** URI

**Access mechanism:** HTTP

((http://...imdb.../WarChild, release date, 12 July 1999)
((http://...imdb.../WarChild, filming location, http://cia.../Albania)
((http://...imdb.../MichaelDavie, directed, http://...imdb.../WarChild)

((http://cia.../Albania, unemployment rate, 13.2%)

**MovieDB**

**CIA World Factbook**
Towards a Web of Linked Data

Data model: RDF
Global identifier: URI
Access mechanism: HTTP
Connection: data links

MovieDB

CIA World Factbook

( http://...imdb.../WarChild , release date , 12 July 1999 )
( http://...imdb.../WarChild , filming location , http://cia.../Albania )
( http://...imdb.../MichaelDavie , directed , http://...imdb.../WarChild )

( http://cia.../Albania , unemployment rate , 13.2% )
Supplementary Access Methods

- **RDF dump**: the whole dataset provided as a big file
- **SPARQL endpoint**: Web service that allows for executing SPARQL queries over the dataset

**Caveat**: these access methods cannot be assumed to be available for all datasets

- Creating dumps is not feasible if data changes very frequently
- Dumps or endpoints may not be feasible if Linked Data interface is simply a wrapper for some other data source
- Providing and maintaining a **reliable** SPARQL endpoint is a significant additional effort
Outline

- The Linked Data Principles ✓
- Paradigms for Querying Linked Data
- Characteristics of the “Database System”
Traditional Paradigm 1: Warehousing

- **Copy** data into a centralized repository
- Query this repository

+ Almost instant results
- Misses unknown or new sources
- Collection possibly out of date
Traditional Paradigm 2: Federation

- Distribute query execution over a federation of SPARQL endpoints

+ Current data
- Misses sources without SPARQL endpoint
Linked Data Query Processing

Principle 1: Rely on the Linked Data principles only

Principle 2: On-line execution
 Principle 1: Rely on the Linked Data principles only

 Principle 2: On-line execution

 Consequence: Obtain data for executing a given query by looking up URIs during the query execution process itself
“Ingredients” for LD Query Execution
“Ingredients” for LD Query Execution

- Data retrieval approach
- Data source selection
- Data source ranking (optional, for optimization)
“Ingredients” for LD Query Execution

- Data retrieval approach
- Data source selection
- Data source ranking (optional, for optimization)

Result construction approach
- i.e., query-local data processing

Query-local data

GET http://.../movie2449

|------------|--------------------|-------------------|
“Ingredients” for LD Query Execution

- Data retrieval approach
- Data source selection
- Data source ranking (optional, for optimization)

- Combining data retrieval and result construction

- Result construction approach
  - i.e., query-local data processing

<table>
<thead>
<tr>
<th>?actor</th>
<th>?loc</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://mdb.../Ric">http://mdb.../Ric</a></td>
<td><a href="http://geo.../Rome">http://geo.../Rome</a></td>
</tr>
</tbody>
</table>
Properties of LD Query Processing

+ Current data
+ May make use of any Linked Data available on the Web
  – Least efficient due to data shipping

Use cases: live querying where freshness and discovery of results is more important than an almost instant answer
Combination with other Paradigms

- Linked Data query processing with a query-local dataset
  - Query-local dataset contains additional data [LT11]
  - Query-local dataset for caching [Har11b, HH11]
- Linked Data query processing with a SPARQL endpoint
  - SPARQL endpoint exposes a cache of Linked Data [UKH+12]
Our Topic Today …

… pure Linked Data query processing

**Linked Data query:** a query that ranges over data made available using the Linked Data principles

**Web of Linked Data:** network of data that evolves by publishing data according to the Linked Data principles
Outline

- The Linked Data Principles ✓
- Paradigms for Querying Linked Data ✓
- Characteristics of the “Database System”
An Analogy ...
Traditional, Central Database Systems
Distributed Database Systems
The Web of Linked Data
The Web of Linked Data
The Web of Linked Data

- Number of potential data sources infinite
The Web of Linked Data

- Number of potential data sources infinite
- No (a priori) information
The Web of Linked Data

- Number of potential data sources infinite
- No (a priori) information
The Web of Linked Data

- Number of potential data sources infinite
- No (a priori) information
The Web of Linked Data

- Number of potential data sources infinite
- No (a priori) information
The Web of Linked Data

- Number of potential data sources infinite
- No (a priori) information
The Web of Linked Data

- Number of potential data sources infinite
- No (a priori) information
- Number of actual data sources infinite
Issues due to the Openness

- Data quality issues
  - Accuracy
  - Freshness / timeliness
  - Believability / trustworthiness

- Data source quality issues
  - Availability
  - Reliability

- Data integration issues
  - Coreferences: Publishers may use different URIs for denoting the same entity
  - Schema heterogeneity: Publishers may use different vocabularies for their data

For the purpose of discussing execution of queries in this tutorial, we largely ignore these issues.
Outline

- The Linked Data Principles
- Paradigms for Querying Linked Data
- Characteristics of the “Database System”

Next part: 2. Theoretical Foundations...
These slides have been created by **Olaf Hartig**
for the
WWW 2013 tutorial on
Link Data Query Processing

Tutorial Website: http://db.uwaterloo.ca/LDQTut2013/

This work is licensed under a
**Creative Commons Attribution-Share Alike 3.0 License**
(http://creativecommons.org/licenses/by-sa/3.0/)