Linked Open Data Aggregation: Conflict Resolution and Aggregate Quality

Tomáš Knap, Jan Michelfeit, Martin Nečaský
{tomas.knap@mff.cuni.cz, jan.michelfeit@seznam.cz}
XML and Web Engineering Research Group (XRG)
Department of Software Engineering
Charles University, Prague, Czech Republic

METHOD, 16.7.2012

Outline

Linked Data Framework and Project ODCleanStore

Data Aggregation Algorithm

Computation of the Aggregate Quality

Experiments & Conclusion

Motivation

- Journalist: “Give me suppliers of public contracts for the Ministry of Finance from the region Prague with just one offer; for each public contract show me the list of payments, links to budget and the person responsible for that contract. Show me the results in the iPhone application”
- Questions:
  - Where to get the data (more sources)
  - How to get the data (different formats, retrieval methods)
  - How to merge and link the data together
  - How to show the data in the iPhone application
- To address the needs of (not just) the journalist: an OpenData.cz initiative with the goals to:
  - Open governmental data in Czech Republic
  - Clean and connect the data
  - Enable exploration of the data

Linked Data

Set of best practises for publishing structured data on the Web, Tim Berners-Lee presented four principles:

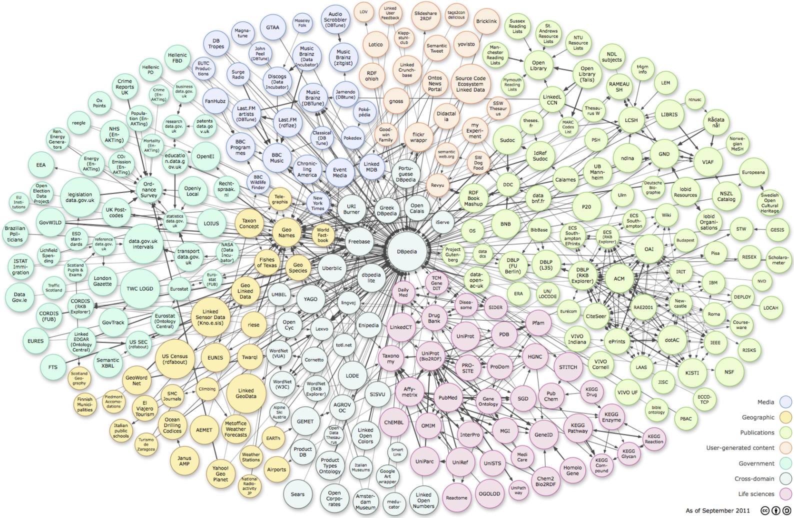
- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF)
- Include links to other URIs. so that they can discover more things.

See: `http://www.w3.org/DesignIssues/LinkedData.html`

RDF

- Sample RDF statement (triple):
  - (http://dbpedia.org/resource/Izmir http://dbpedia.org/ontology/populationTotal ”3450889”)
  - (http://dbpedia.org/resource/Izmir http://www.w3.org/2002/07/owl#sameAs http://rdf.freebase.com/ns/en.izmir)
- RDF data are represented as typed statements – *triples* $(s, p, o) \in U^3$ – consisting of a *subject s*, a *predicate (property) p* and an *object (value) o*.
  - U = all possible nodes, URI resources or literals (optionally typed)
- A triple may be part of a *named graph* – a set of triples identified by an URI
  - Triples can be then extended to *quads* $(s, p, o, g) \in Q$ where $g \in G$ is the named graph (its URI) to which the data belongs.
- The RDF data model can be viewed as a directed graph where edges, labeled with a predicate, lead from a subject to an object.

Linked Data Cloud

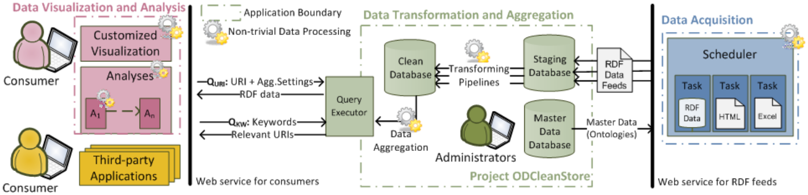


Obrázek: Linked Data Cloud, `http://linkeddata.org/`

Linked Data Framework

- Is built as part of OpenData.cz initiative and LOD2 project

- Data acquisition
- Data transformation and aggregation = ODCleanStore project
- Data visualization and analysis



Obrázek: Linked Data Framework

Outline

Linked Data Framework and Project ODCleanStore

Data Aggregation Algorithm

Computation of the Aggregate Quality

Experiments & Conclusion

Motivational Scenario

- Suppose we have in the clean database data about the city Izmir coming from multiple sources – DBpedia, GeoNames, and Freebase
  - `http://dbpedia.org/resource/Izmir`
  - `http://sws.geonames.org/311046/`
  - `http://rdf.freebase.com/ns/en.izmir`

  .
- Consumer would like to get data about the resource `http://dbpedia.org/resource/Izmir`
- Tasks:
  - Discover and follow `owl:sameAs` links between resources representing the same concepts
  - Discover that meaning of the predicates `geo:lat` and `fb:location.geocode.latitude` is the same
  - Compute average value for the values of the properties `geo:long` and `geo:lat`
  - Select the best value (with the highest aggregate quality) for `rdfs:label`
  - Select the maximum (latest) value from the values of the property `dbpedia:populationTotal`

Data Aggregation - Basics

- Schema mapping
  - Enabled by proper mappings between ontologies in the master data database
- Duplicate detection
  - Enabled by proper linker
- Data fusion
  - Instance level conflicts (data conflicts)

J. Bleiholder and F. Naumann. Data fusion. ACM Comput. Surv., 2009.

Data Fusion Algorithm - Inputs/Outputs

- Inputs:
  - A collection of quads from the clean database to be fused – e.g. the quads (x,*,*,*),(*,*,x,*), where *x* is the requested URI in a URI query
  - Data fusion settings (e.g. a selected conflict resolution policies – global or per property)
  - `owl:sameAs` links between URI resources occurring in the quads
    - result of deduplication and schema mapping
  - Quality scores for named graphs of the quads.
- Outputs:
  - Collection of aggregated triples enriched with the aggregate quality and source named graphs for each quad.

Phase 1 of Data Fusion Algorithm – an Overview

Step 1.1) Replace URIs of resources representing the same entity (i.e. connected by the `owl:sameAs` links) with a single URI. Prefer URI in the consumer's query.

Step 1.2) Remove duplicate quads.

Step 1.3) Group quads to sets of o-conflicting quads.

- Suppose $g_1, g_2 \in G$; quads (s, p, o_1, g_1) and (s, p, o_2, g_2) are called *o-conflicting quads* if $o_1 \neq o_2$

Phase 2 of Data Fusion Algorithm – an Overview

Step 2.1) Choose and apply a conflict resolution policy

Step 2.2) Compute aggregate quality for the conflict resolved quads

- Note: Phase 2 of the algorithm is applied to each set of *o-conflicting quads*

Conflict Resolution Policies

- Deciding - selects one or more values
  - ANY,MIN,MAX,SHORTEST,LONGEST – an arbitrary value, minimum, maximum, shortest, or longest is selected from the conflicting values V
  - BEST – the value with the highest aggregate quality is selected
  - LATEST – the value with the newest time is selected
- Mediating - computes new values
  - AVG, MEDIAN, CONCAT – computes the average, median, or concatenation of conflicting values
- Ignoring
  - ALL – ignores conflicts, fuses equal triples
  - NONE – ignores conflicts, no fusion

Outline

Linked Data Framework and Project ODCleanStore

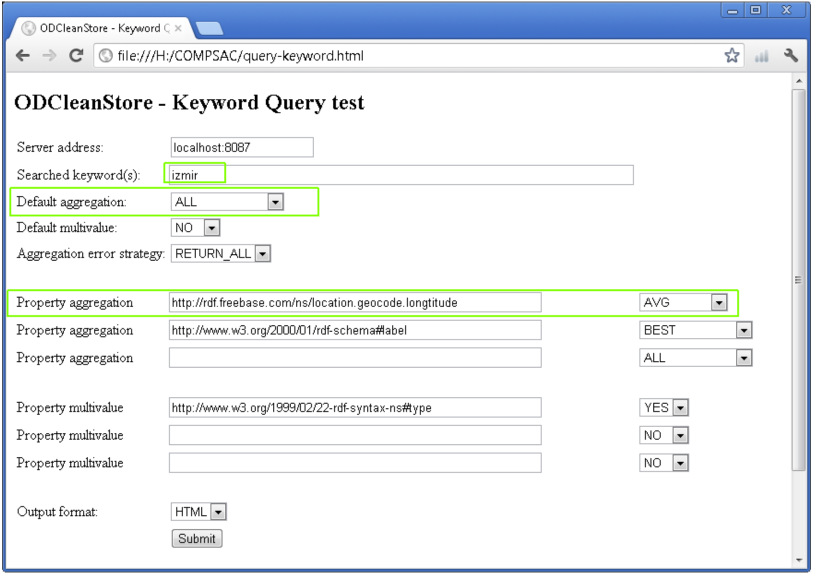
Data Aggregation Algorithm

Computation of the Aggregate Quality

Experiments & Conclusion

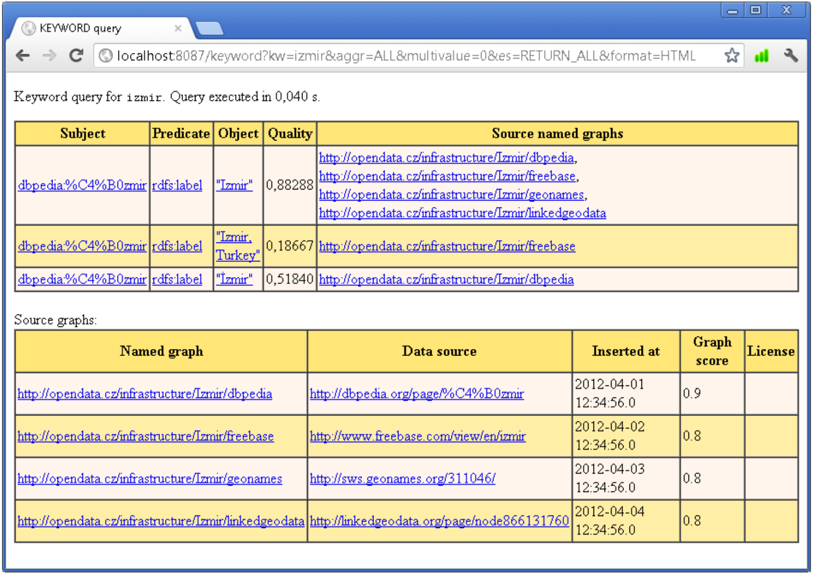
Computation of the Aggregate Quality - Overview

- Several factors based on real-world examples
- Let’s show it on Izmir!



ODCleanStore - Keyword Query test

Server address:	localhost:8087	
Searched keyword(s):	izmir	
Default aggregation:	ALL	
Default multivalue:	NO	
Aggregation error strategy:	RETURN_ALL	
Property aggregation	http://rdf.freebase.com/ns/location.geocode.longtitude	AVG
Property aggregation	http://www.w3.org/2000/01/rdf-schema#label	BEST
Property aggregation		ALL
Property multivalue	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	YES
Property multivalue		NO
Property multivalue		NO
Output format:	HTML	


KEYWORD query
localhost:8087/keyword?kw=izmir&aggr=ALL&multivalue=0&es=RETURN_ALL&format=HTML

Keyword query for `izmir`. Query executed in 0,040 s.

Subject	Predicate	Object	Quality	Source named graphs
dbpedia:%C4%B0zmir	rdfs:label	"Izmir"	0,88288	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/geonames, http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	rdfs:label	"Izmir, Turkey"	0,18667	http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdfs:label	"İzmir"	0,51840	http://opendata.cz/infrastructure/Izmir/dbpedia

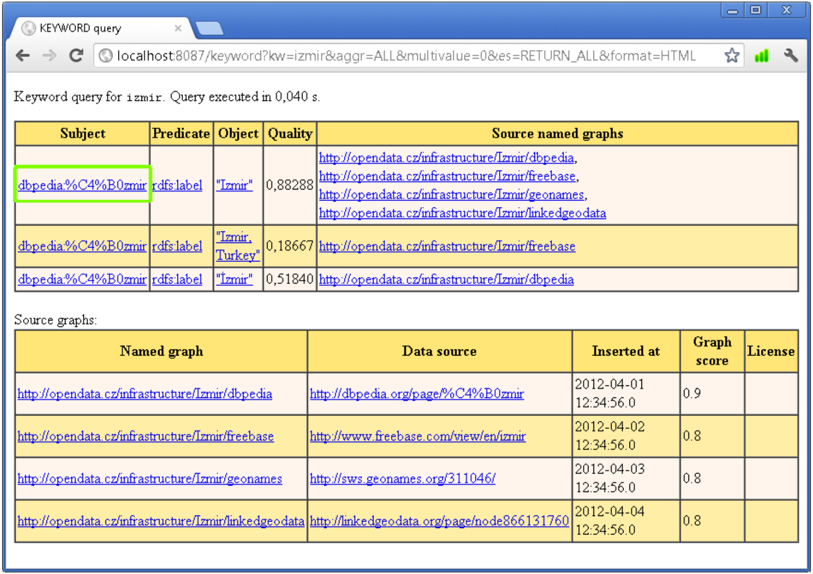
Source graphs:

Named graph	Data source	Inserted at	Graph score	License
http://opendata.cz/infrastructure/Izmir/dbpedia	http://dbpedia.org/page/%C4%B0zmir	2012-04-01 12:34:56.0	0.9	
http://opendata.cz/infrastructure/Izmir/freebase	http://www.freebase.com/view/en/izmir	2012-04-02 12:34:56.0	0.8	
http://opendata.cz/infrastructure/Izmir/geonames	http://sws.geonames.org/311046/	2012-04-03 12:34:56.0	0.8	
http://opendata.cz/infrastructure/Izmir/linkedgeodata	http://linkedgeodata.org/page/node866131760	2012-04-04 12:34:56.0	0.8	

First Quality Factor - Scores of Source Named Graphs

- A value $v \in A$ may
  - (a) be calculated from all the sources (in case of conflict resolution policies AVG, MEDIAN, CONCAT)
  - (b) come from named graphs containing a quad (s, p, v, g_i) (in case of other conflict resolution policies)

$$q_1(v) = \begin{cases} \mathrm{avg}\,\{s(g) \mid g \in \{g_1, \ldots, g_n\}\} & \text{(a)} \\ \max\,\{s(g) \mid g \in agree(v)\} & \text{(b)} \end{cases}$$

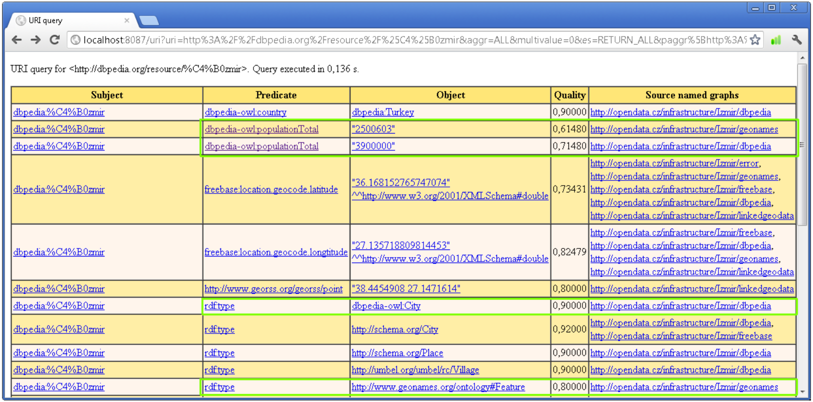

KEYWORD query
localhost:8087/keyword?kw=izmir&aggr=ALL&multivalue=0&es=RETURN_ALL&format=HTML

Keyword query for `izmir`. Query executed in 0,040 s.

Subject	Predicate	Object	Quality	Source named graphs
dbpedia:%C4%B0zmir	rdfs:label	"Izmir"	0,88288	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/geonames, http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	rdfs:label	"Izmir, Turkey"	0,18667	http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdfs:label	"İzmir"	0,51840	http://opendata.cz/infrastructure/Izmir/dbpedia

Source graphs:

Named graph	Data source	Inserted at	Graph score	License
http://opendata.cz/infrastructure/Izmir/dbpedia	http://dbpedia.org/page/%C4%B0zmir	2012-04-01 12:34:56.0	0.9	
http://opendata.cz/infrastructure/Izmir/freebase	http://www.freebase.com/view/en/izmir	2012-04-02 12:34:56.0	0.8	
http://opendata.cz/infrastructure/Izmir/geonames	http://sws.geonames.org/311046/	2012-04-03 12:34:56.0	0.8	
http://opendata.cz/infrastructure/Izmir/linkedgeodata	http://linkedgeodata.org/page/node866131760	2012-04-04 12:34:56.0	0.8	



localhost:8087/uri?uri=http%3A%2F%2Fdbpedia.org%2Fresource%2F%25C4%25B0zmir&aggr=ALL&multivalue=0&es=RETURN_ALL&paggr%5Bhttp%3A%

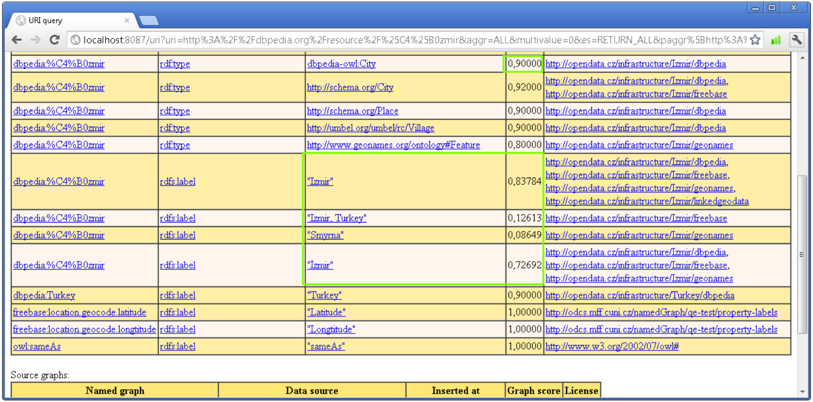
URI query for <http://dbpedia.org/resource/%C4%B0zmir>. Query executed in 0,136 s.

Subject	Predicate	Object	Quality	Source named graphs
dbpedia:%C4%B0zmir	dbpedia-owl:country	dbpedia:Turkey	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	dbpedia-owl:populationTotal	"2500603"	0,61480	http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:%C4%B0zmir	dbpedia-owl:populationTotal	"3900000"	0,71480	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	freebase:location.geocode.latitude	"36.168152765747074"^^http://www.w3.org/2001/XMLSchema#double	0,73431	http://opendata.cz/infrastructure/Izmir/error, http://opendata.cz/infrastructure/Izmir/geonames, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	freebase:location.geocode.longtitude	"27.135718809814453"^^http://www.w3.org/2001/XMLSchema#double	0,82479	http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/geonames, http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	http://www.georss.org/georss/point	"38.4454908 27.1471614"	0,80000	http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	rdf:type	dbpedia-owl:City	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/City	0,92000	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/Place	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://umbel.org/umbel/rc/Village	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://www.geonames.org/ontology#Feature	0,80000	http://opendata.cz/infrastructure/Izmir/geonames

Second Quality Factor - Differences between Conflicting Values

- We use a metric $d : U \times U \to [0, 1]$ for each type of values (numbers, strings, dates, ...).
- Different values reduce score increasingly with their distance and their scores (weighted average)).
- Can be turned off by the *multivalue* parameter.

$$q_2(v) = q_1(v) \cdot \left(1 - \frac{\sum_{i=1}^{n} s(g_i)d(v, v_i)}{\sum_{i=1}^{n} s(g_i)}\right)$$



localhost:8087/uri?uri=http%3A%2F%2Fdbpedia.org%2Fresource%2F%25C4%25B0zmir&aggr=ALL&multivalue=0&es=RETURN_ALL&paggr%5Bhttp%3A9

dbpedia:%C4%B0zmir	rdf:type	dbpedia-owl:City	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/City	0,92000	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/Place	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://umbel.org/umbel/rc/Village	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://www.geonames.org/ontology#Feature	0,80000	http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:%C4%B0zmir	rdfs:label	"Izmir"	0,83784	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/geonames, http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	rdfs:label	"Izmir, Turkey"	0,12613	http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdfs:label	"Smyrna"	0,08649	http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:%C4%B0zmir	rdfs:label	"Izmir"	0,72692	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:Turkey	rdfs:label	"Turkey"	0,90000	http://opendata.cz/infrastructure/Turkey/dbpedia
freebase:location.geocode.latitude	rdfs:label	"Latitude"	1,00000	http://odcs.mff.cuni.cz/namedGraph/qe-test/property-labels
freebase:location.geocode.longtitude	rdfs:label	"Longtitude"	1,00000	http://odcs.mff.cuni.cz/namedGraph/qe-test/property-labels
owl:sameAs	rdfs:label	"sameAs"	1,00000	http://www.w3.org/2002/07/owl#

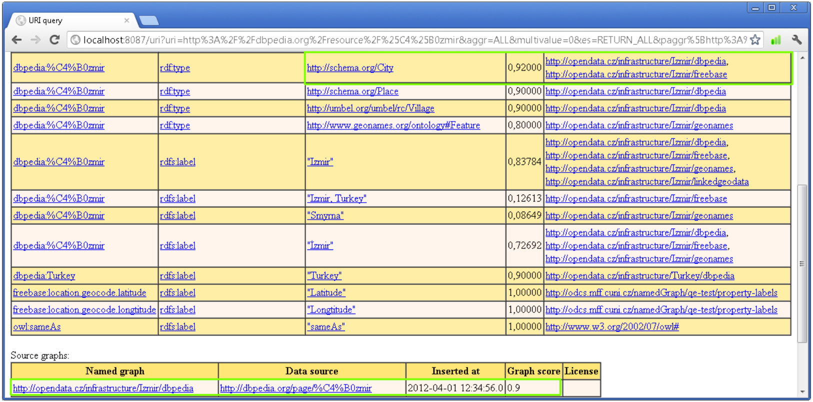
Source graphs:

Named graph	Data source	Inserted at	Graph score	License

Third Quality Factor - Confirmation by Multiple Sources

- Agreement on a single value by multiple sources increases its value.
- Weighted by scores of the sources.

$$q_3(v) = q_2(v) + \\ + (1 - q_2(v)) \cdot \min\left(\frac{-q_1(v) + \sum_{g \in agree(v)} s(g)}{C}, 1\right)$$



localhost:8087/uri?uri=http%3A%2F%2Fdbpedia.org%2Fresource%2F%25C4%25B0zmir&aggr=ALL&multivalue=0&es=RETURN_ALL&paggr%5Bhttp%3A9

dbpedia:%C4%B0zmir	rdf:type	http://schema.org/City	0,92000	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/Place	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://umbel.org/umbel/rc/Village	0,90000	http://opendata.cz/infrastructure/Izmir/dbpedia
dbpedia:%C4%B0zmir	rdf:type	http://www.geonames.org/ontology#Feature	0,80000	http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:%C4%B0zmir	rdfs:label	"Izmir"	0,83784	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/geonames, http://opendata.cz/infrastructure/Izmir/linkedgeodata
dbpedia:%C4%B0zmir	rdfs:label	"Izmir, Turkey"	0,12613	http://opendata.cz/infrastructure/Izmir/freebase
dbpedia:%C4%B0zmir	rdfs:label	"Smyrna"	0,08649	http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:%C4%B0zmir	rdfs:label	"İzmir"	0,72692	http://opendata.cz/infrastructure/Izmir/dbpedia, http://opendata.cz/infrastructure/Izmir/freebase, http://opendata.cz/infrastructure/Izmir/geonames
dbpedia:Turkey	rdfs:label	"Turkey"	0,90000	http://opendata.cz/infrastructure/Turkey/dbpedia
freebase:location.geocode.latitude	rdfs:label	"Latitude"	1,00000	http://odcs.mff.cuni.cz/namedGraph/qe-test/property-labels
freebase:location.geocode.longtitude	rdfs:label	"Longtitude"	1,00000	http://odcs.mff.cuni.cz/namedGraph/qe-test/property-labels
owl:sameAs	rdfs:label	"sameAs"	1,00000	http://www.w3.org/2002/07/owl#

Source graphs:

Named graph	Data source	Inserted at	Graph score	License
http://opendata.cz/infrastructure/Izmir/dbpedia	http://dbpedia.org/page/%C4%B0zmir	2012-04-01 12:34:56.0	0.9	

Computation of the Aggregate Quality - Summary

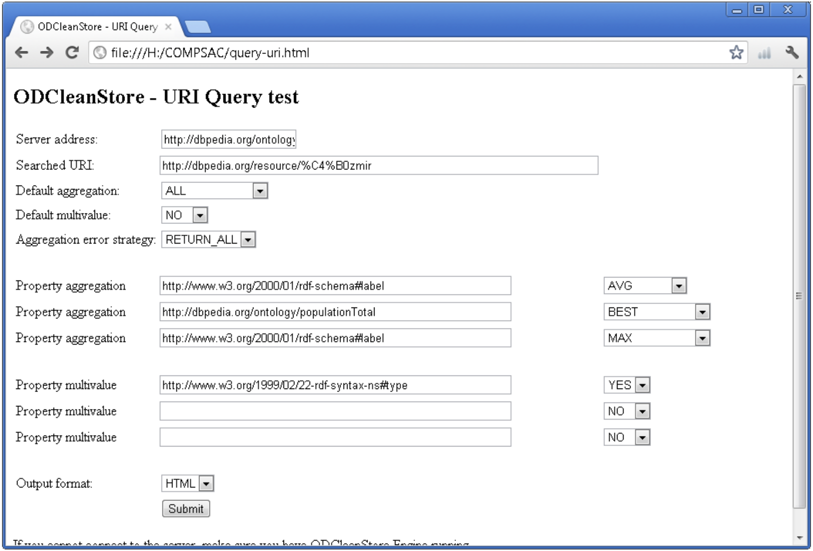
- The result $q(v) = q_3(v)$ is the aggregate quality.
- The second or the third step of the quality computation may be omitted when its use doesn’t make sense (e.g. CONCAT).

The quality satisfies the following constraints:

- If there is a named graph g asserting a non-conflicting value v, the aggregate quality (based just on the value v) should be at least $s(g)$.
- $q(v)$ is increasing with quality scores of source named graphs v was selected from or calculated from.
- $q(v)$ is decreasing with difference of other values $v_i \in V$, taking their quality scores $s(g_i)$ into consideration.
- If multiple sources agree on the same value, the aggregate quality is increased.

Other Interesting Features of the Data Aggregation Algorithm

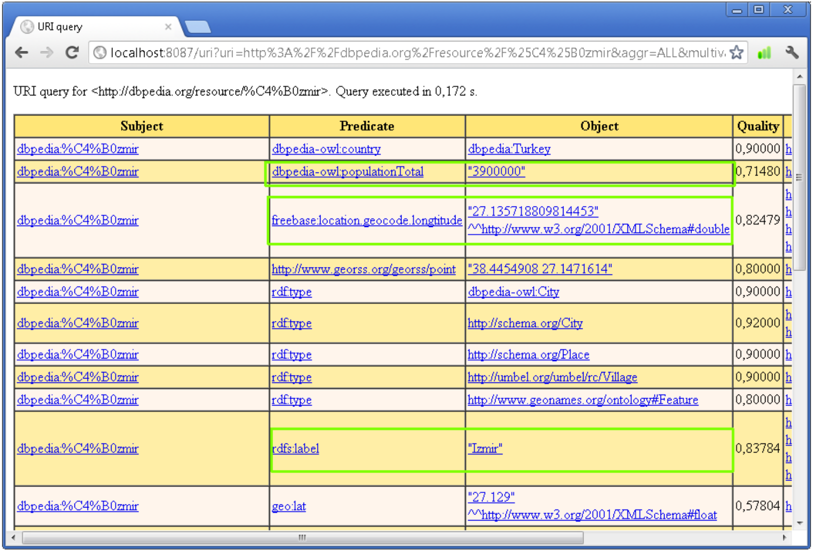
- Automatic translation of URIs:
  - `http://dbpedia.org/resource/Izmir` vs. `http://rdf.freebase.com/ns/en.izmir`
  - `http://www.w3.org/2003/01/geo/wgs84_pos#long` vs. `http://rdf.freebase.com/ns/location.geocode.longtitude`
  - preference given implicitly
- Various aggregation methods - BEST, AVG, conflict tolerating ALL
  - again URI translation



ODCleanStore - URI Query test

Server address:	http://dbpedia.org/ontology	
Searched URI:	http://dbpedia.org/resource/%C4%B0zmir	
Default aggregation:	ALL	
Default multivalue:	NO	
Aggregation error strategy:	RETURN_ALL	
Property aggregation	http://www.w3.org/2000/01/rdf-schema#label	AVG
Property aggregation	http://dbpedia.org/ontology/populationTotal	BEST
Property aggregation	http://www.w3.org/2000/01/rdf-schema#label	MAX
Property multivalue	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	YES
Property multivalue		NO
Property multivalue		NO
Output format:	HTML	

If you cannot connect to the server, make sure you have ODCleanStore Engine running



URI query
localhost:8087/uri?uri=http%3A%2F%2Fdbpedia.org%2Fresource%2F%25C4%25B0zmir&aggr=ALL&multiv

URI query for <http://dbpedia.org/resource/%C4%B0zmir>. Query executed in 0,172 s.

Subject	Predicate	Object	Quality	
dbpedia:%C4%B0zmir	dbpedia-owl:country	dbpedia:Turkey	0,90000	h
dbpedia:%C4%B0zmir	dbpedia-owl:populationTotal	"3900000"	0,71480	h
dbpedia:%C4%B0zmir	freebase:location.geocode.longtitude	"27.135718809814453"^^http://www.w3.org/2001/XMLSchema#double	0,82479	h h h h
dbpedia:%C4%B0zmir	http://www.georss.org/georss/point	"38.4454908 27.1471614"	0,80000	h
dbpedia:%C4%B0zmir	rdf:type	dbpedia-owl:City	0,90000	h
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/City	0,92000	h h
dbpedia:%C4%B0zmir	rdf:type	http://schema.org/Place	0,90000	h
dbpedia:%C4%B0zmir	rdf:type	http://umbel.org/umbel/rc/Village	0,90000	h
dbpedia:%C4%B0zmir	rdf:type	http://www.geonames.org/ontology#Feature	0,80000	h
dbpedia:%C4%B0zmir	rdfs:label	"Izmir"	0,83784	h h h h
dbpedia:%C4%B0zmir	geo:lat	"27.129"^^http://www.w3.org/2001/XMLSchema#float	0,57804	h

Outline

Linked Data Framework and Project ODCleanStore

Data Aggregation Algorithm

Computation of the Aggregate Quality

Experiments & Conclusion

Experiments

- Data fusion execution times for various conflict resolution policies

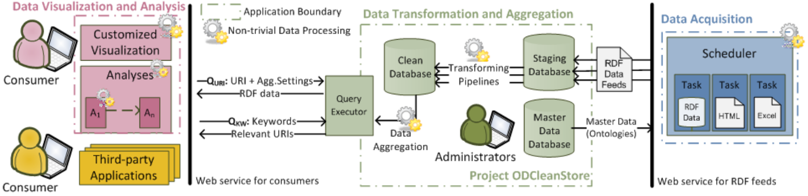
Tabulka: DBpedia evaluation – Execution times

Triples	Conflict resolution	Multivalue	Time
100,000	ALL	no	1.75 s
100,000	ANY	no	1.02 s
100,000	ALL	yes	1.01 s
100,000	CONCAT	yes	0.96 s
100,000	ANY	yes	0.83 s

- Plus time for RDF store query
- Current prototype queries under 0.5 s even on larger dataset

Conclusions

- Linked Data Framework
  - Data Aggregation - Data Fusion



Obrázek: Linked Data Framework

Thank You!