

# Towards Question Answering on Statistical Linked Data

Konrad Höffner

Universität Leipzig, AKSW/MOLE, PhD Student

2014-9-5

- 1 Motivation
  - Use Cases
  - Standard Approaches for Linked Data Consumption
- 2 Introduction
  - Question Answering
  - Data Cubes
- 3 Approach
  - Research Problems
  - Goals
  - Corpus
- 4 Future Work

- increasing amount of statistical linked data
- highly relevant for decision making

# Financial Data—Where Does My Money Go?



## WHERE DOES MY MONEY GO?

*Showing you where your taxes get spent*

[The Daily Bread](#) [Country & Regional Analysis](#) [Departmental Spending](#) [About](#)

## The Daily Bread Costs for the British Taxpayer per Day

**SALARY**

**£22,000**

**SELECT YOUR SALARY**



**YOUR TAX**

**£8,774**

Running  
Government



£3.26

Defence



£1.69

Health



£5.86

Helping Others



£8.60

Culture



£0.38

Education



£1.54

Running The  
Country, Social  
Systems



£1.42

Order & Safety



£0.78

Our Streets



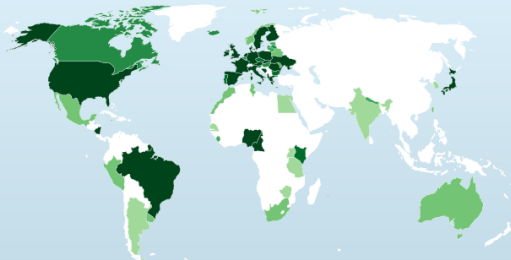
£0.28

The Environment



£0.24

## Financial Data—OpenSpending

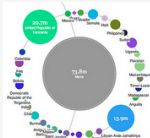


## Mapping the money.

Our aim is to track every government financial transaction across the world and present it in useful and engaging forms for everyone from a school-child to a data geek.

Search 20,214,833 government transactions in 395 datasets...

Search



## Upload and visualize data

Upload any kind of financial data to OpenSpending and explore it with our built-in interactive visualizations. Users publish **budgets**, **procurements**, **spending data** and even **public employee salaries**.

Use our [widgets](#) to embed your visualization on your own website.

## Upload a dataset

## GETTING STARTED

What can I do here?

## FAQ

[Browse datasets](#)

## THE PROJECT

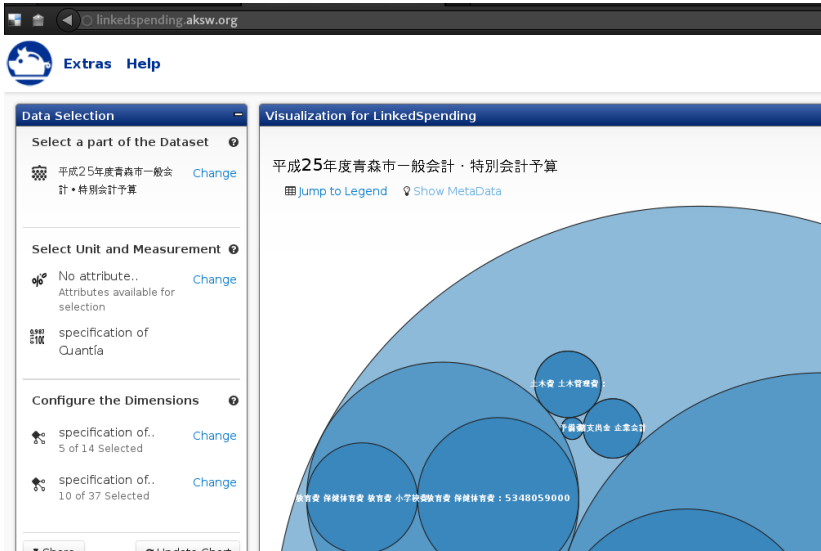
Spending Blog

## Projects Portfolio

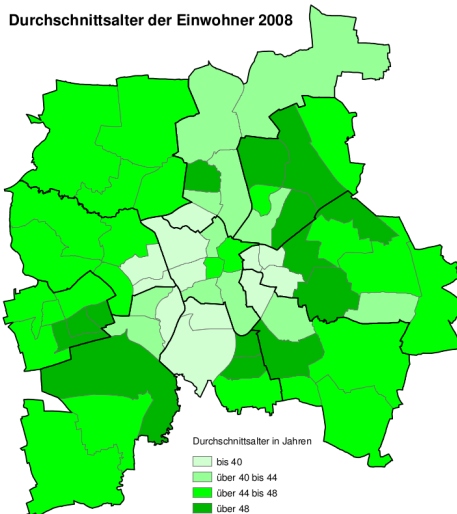
[Mailing List](#)

Contribute

# Financial Data—LinkedSpending



# Demographic Data—Average Age in Leipzig by District



Karte und Datenquelle: Amt für Statistik und Wahlen Leipzig

# SPARQL Endpoint

## Virtuoso SPARQL Query Editor

[About](#) | [Namespace Prefixes](#) | [Inference rules](#)


Default Data Set Name (Graph IRI)

Query Text

```
select distinct ?d ?c ?r
{
  ?o qb:dataset ?d.
  ?o dbo:currency ?c.
  ?c dbp:inflationRate ?r.
  filter(?r > 10)
}
```



# Faceted Browsing



**Data Portal**  
The Open Data Hub of the European Union **BETA**

[Legal notice](#) [Contact](#) [Search](#)

European Commission > Open Data Portal > Home

**Settings**

- undefined
  - type 1/31
  - label 1475
  - funding 14773
    - type 1
    - label 14754
    - partner 4718
      - type 1
      - label 4747
      - address 4718
        - city 1914
          - label 1910
          - someAs 1526
            - country 96
              - organisation class 4
              - NUTS region 733
              - amount 12793
                - partner role 3
                - cell 30
                - year 5
                - strategic objective 128
                - instrument 3
                - id 1494

prc

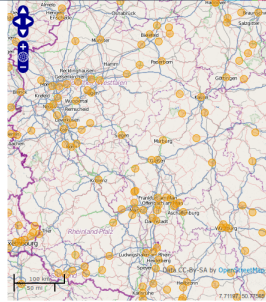
  - ☐ Property
  - ☐ OntologyProperty
  - ☐ AnnotationProperty
  - ☒ Project

lod2

s	fv_172	fv_173
LOD2	OPENLINK GROUP LIMITED	870380
LOD2	Korea Advanced Institute of Science and Technology	0
LOD2	INSTITUT MIHAJLO PUPIN	193500
LOD2	WOLTERS KLUWER DEUTSCHLAND GMBH	371815
LOD2	OPEN KNOWLEDGE FOUNDATION LIMITED LBG	362304
LOD2	SEMANTIC WEB COMPANY GMBH	518526
LOD2	TENFORCE BVBA*	706280
LOD2	UNIVERSITAET LEIPZIG	1333901
LOD2	STICHTING CENTRUM VOOR WISKUNDE EN INFORMATICA	505275
LOD2	EXALEAD	470220

< 1 2 >

[Export CSV](#) [Export RDF](#)



Center on user location

# Question Answering

autosparql-tbtl-dl-leader.org

Oxford - Real estate

houses with more than 2 bedrooms

Run

Found answer for "houses whose number of bedrooms is greater than 2".

Wrong!

Sort by price(highest first)

**Park Lane, Stanford in the Vale, Faringdon, Oxfordshire**

£3,950,000.00

Available for the first time to the market in approximately fifty years, a rare opportunity to acquire a fine Grade II listed substantial seven bedroom Georgian farmhouse circa 1800's with later additions, with delightful views towards the south downs, together with a comprehensive range of modern and traditional farm buildings and stabling, all in need of extensive renovation/restoration, situated centrally within 223 acres (90.279 ha) of mainly pasture land

**Street:** [Park Lane, Stanford in the Vale, Faringdon, Oxfordshire](#)  
**Locality:** [Stanford in the Vale, Faringdon, Oxfordshire](#)  
**#bedrooms:** 7

**Horse Shoe Lane, Wootton, Woodstock, Oxfordshire, OX20**

£2,700,000.00

An impressive and extensive conversion of three barns into one very spacious and individual c.6,500 sq ft house with a private garden and paddocks of approximately 9.7 acres and frontage onto the River Glyme. The two storey accommodation has a flexible layout enabling it to accommodate a variety of lifestyles with a total of five reception rooms, six bedrooms, four bathrooms and an office. Outside to the front there is ample parking and a double garage, while on the south side lies a private garden which opens onto extensive meadow and paddocks. The property is approached discreetly via a short 'No Through' lane tucked away in the favoured and unspoilt south side of the village. Upon entering the gated

**Street:** [Horse Shoe Lane, Wootton, Woodstock, Oxfordshire](#)  
**Locality:** [Wootton, Woodstock, Oxfordshire](#)  
**#bedrooms:** 6

**Harberton Mead, Headington**

- 1 Motivation
  - Use Cases
  - Standard Approaches for Linked Data Consumption
- 2 Introduction
  - Question Answering
  - Data Cubes
- 3 Approach
  - Research Problems
  - Goals
  - Corpus
- 4 Future Work

# Definition of QA

- users ask questions in natural language (NL)
- using their own terminology
- receive concise answer

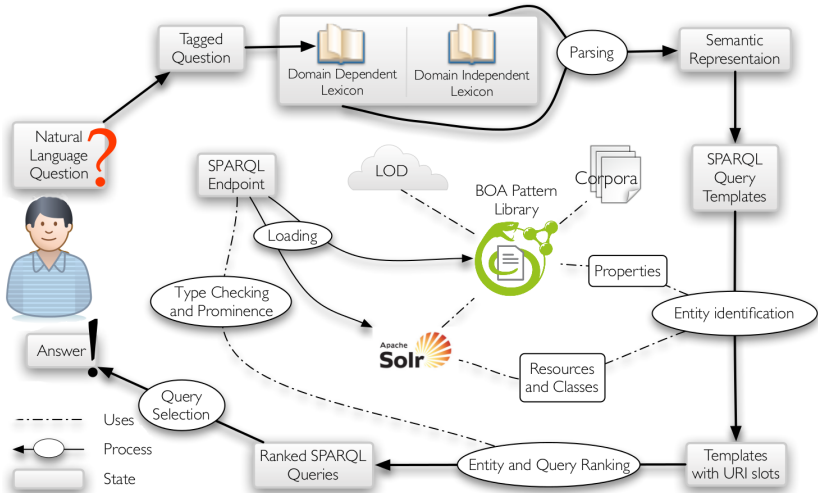
# Properties of QA

- intuitive for users
- expressive
- some effort to set up

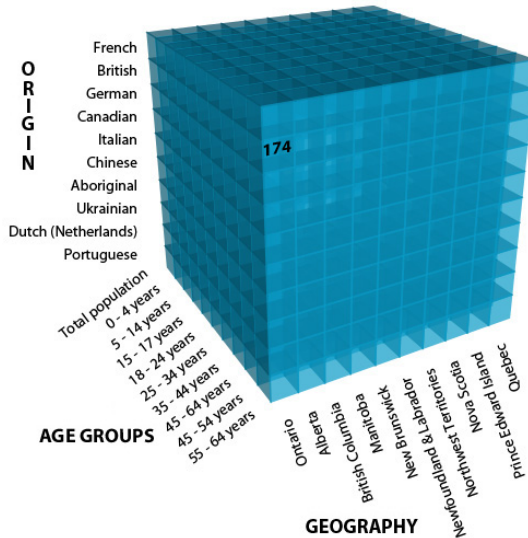
# Dimensions of QA

dimension	values
input type	keywords, <b>factual queries (who, what, how many, ...)</b> , affirmation/negation, causality (why, how)
source scope	structured, semistructured, textual or <b>semantic</b> domain dependent, <b>domain independent</b>

# AutoSPARQL TBSL

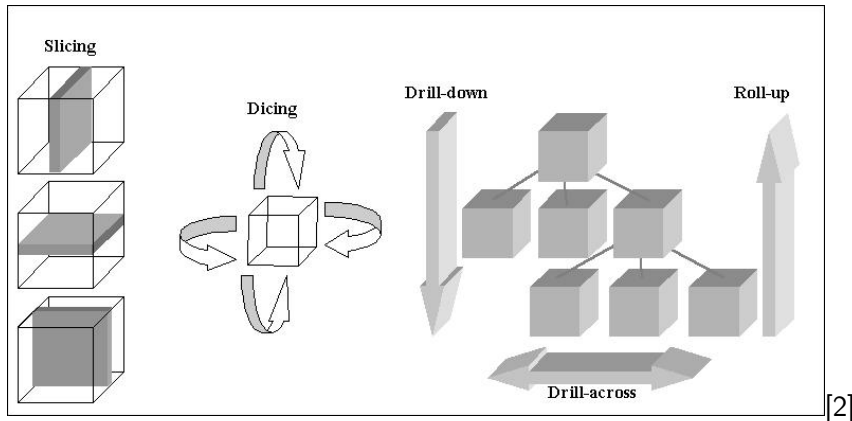


# Multidimensional Dataset: Data Cube

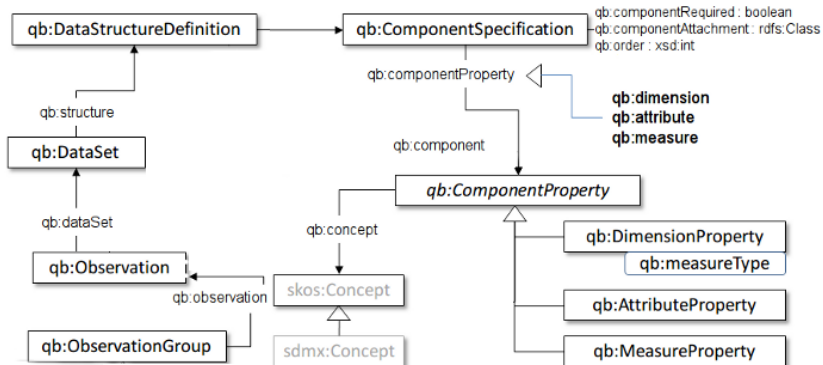




# Operations on Data Cubes



# RDF Data Cube (RDC) Vocabulary



- 1 Motivation
  - Use Cases
  - Standard Approaches for Linked Data Consumption
- 2 Introduction
  - Question Answering
  - Data Cubes
- 3 **Approach**
  - Research Problems
  - Goals
  - Corpus
- 4 Future Work

# Research Problems

- “What steps are necessary to adapt an existing semantic question answering system to statistical linked data in the form of RDF Data Cubes?”
- “How can research on question answering on statistical linked data be stimulated and evaluated?”

# Problem—Different Questions

- “What is the name of the wife of Barack Obama?”
- “What was the average Leipzig district budget in 2012?”

# Problem—Different Structure

- RDF Data Cube Vocabulary (RDC) is a meta model
- RDF as underlying structure only provides the base

# Goals

- ① create statistical question corpus to identify required functionality
- ② using (1)—create benchmark on LinkedSpending datasets
- ③ using (1)—adapt TBSL to RDCs
- ④ using (2)—evaluate and optimize new system

# Statistical Question Corpus

- collection of 50 representative questions from volunteers

## Excerpt of Questions

How much money, does Leipzig and Dresden spend on child care in relation to the birth rate in comparison to the average in Saxony.

What is the average monthly income of a German citizen?

How much money was invested to fight bicycle thefts in Leipzig?

How many citizens live in a certain area?

How much does Germany spend on research a year?



# Challenges

- aggregation mapping
- implied aggregation
- dataset selection
- expected presentation type
- expected answer type

# Expected Answer Types

question word	expected answer type	<i>f</i>
how much	quantity (uncountable)	19
what	any	12
how many	quantity (countable)	11
which	equivalent to “what”	3
where	location or purpose	2
how is	any	1
relate	comparison or visualization	1
none (statement)	any	1
total		50

- 1 Motivation
  - Use Cases
  - Standard Approaches for Linked Data Consumption
- 2 Introduction
  - Question Answering
  - Data Cubes
- 3 Approach
  - Research Problems
  - Goals
  - Corpus
- 4 Future Work

# Future Work

- still very much work in progress
- extend the corpus, more participants from a more varied background
- refine corpus to benchmark
- add RDC functionality to TBSL
- continuously evaluate and optimize the new system using the benchmark

## References



Konrad Höffner, Michael Martin, and Jens Lehmann.  
LinkedSpending: OpenSpending becomes Linked Open Data.  
*Semantic Web Journal*, 2015.



Harry Mucksch and Wolfgang Behme.  
Das data warehouse-konzept.  
2000.



Claus Stadler, Michael Martin, and Sören Auer.  
Exploring the Web of Spatial Data with Facete.  
In *Companion proceedings of 23rd International World Wide Web Conference (WWW)*, pages 175–178, 2014.



Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano.  
Template-based question answering over RDF data.  
In *Proceedings of the 21st international conference on World*