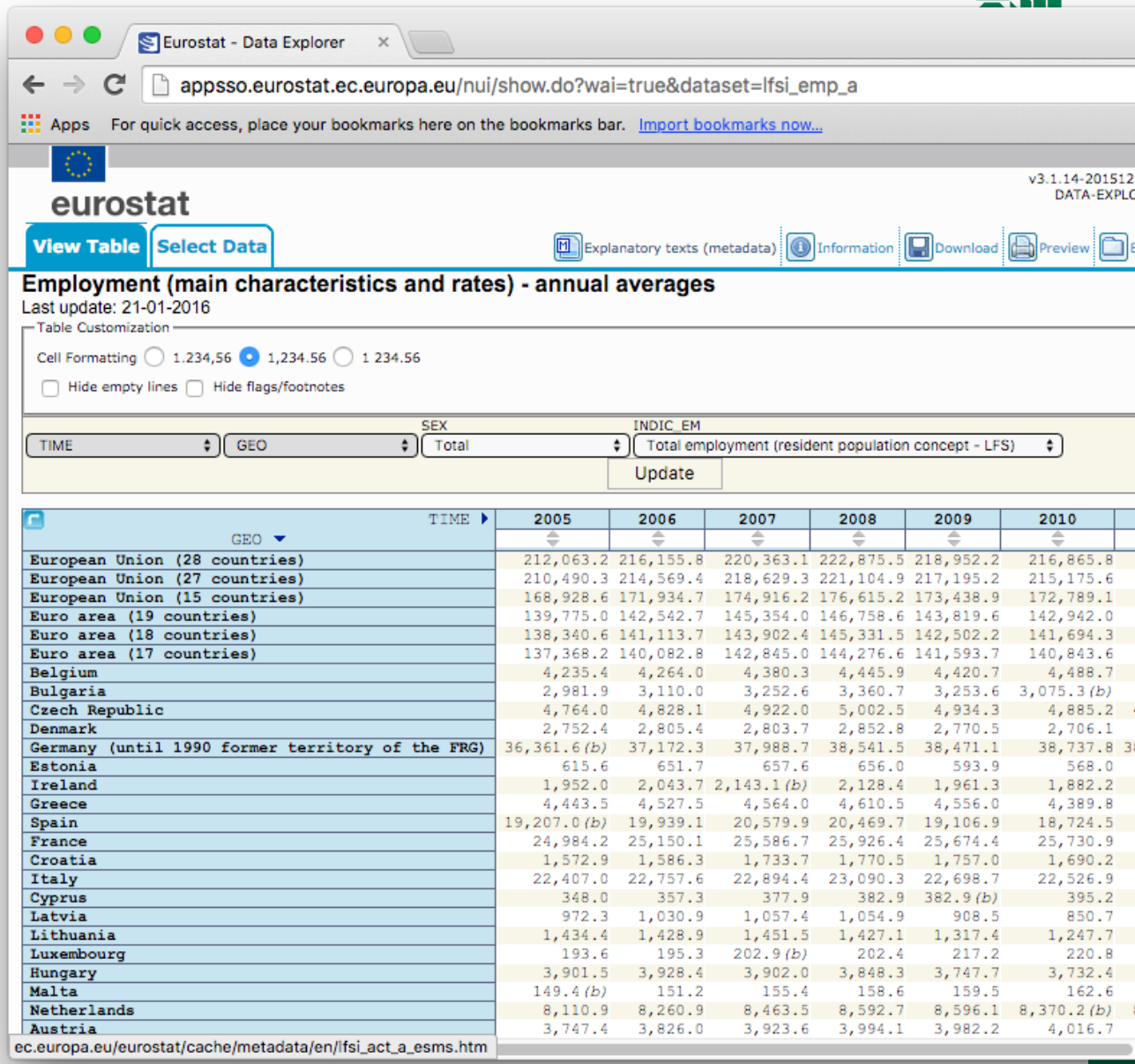
SEO4OLAP – Search Engine Optimized Presentation of Statistical Linked Data

Daniel Breucker, Benedikt Kämpgen, Andreas Harth

INFORMATIK, BDSDST 2016, 2016-09-30

Motivation

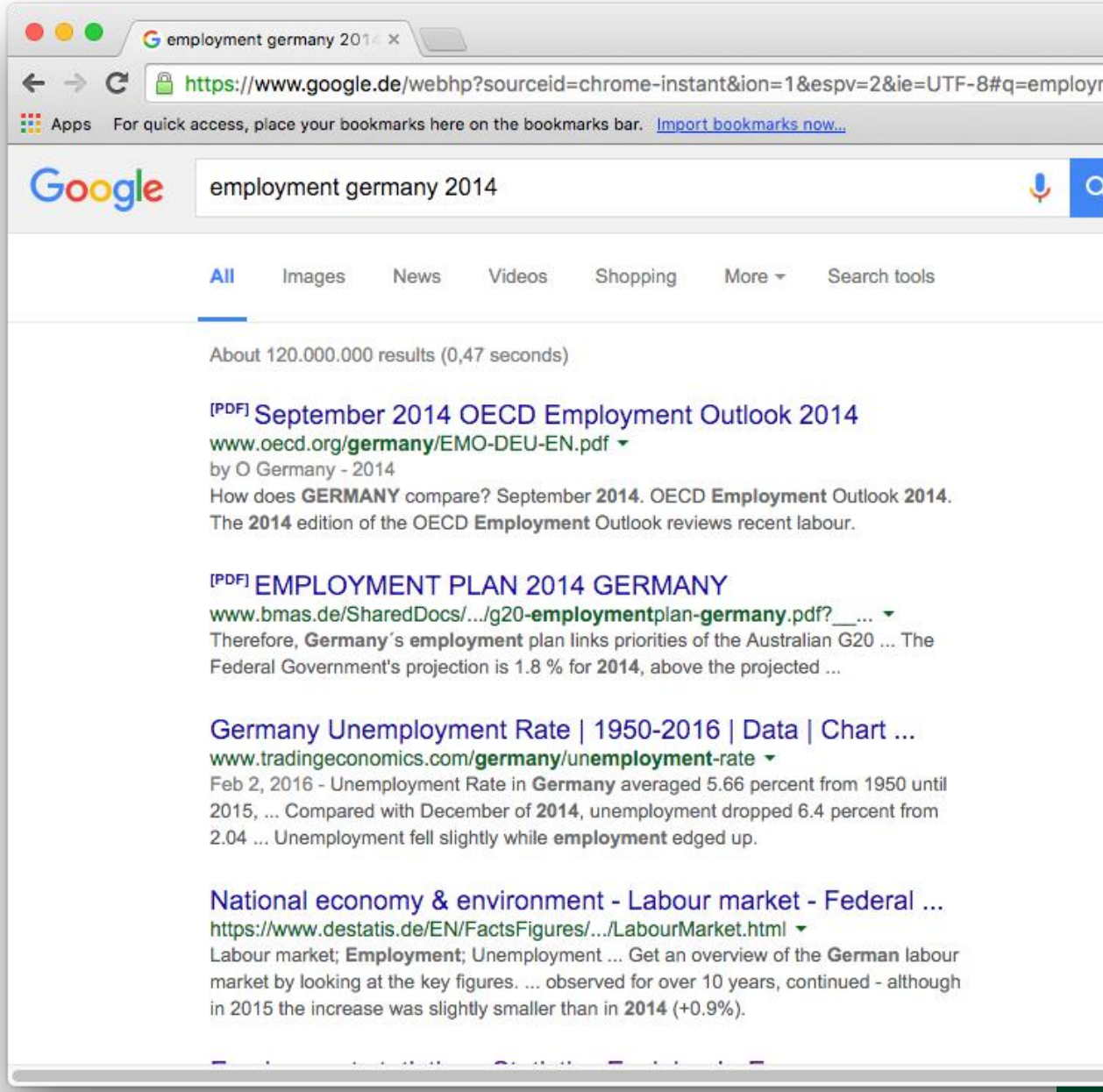
- G8 Open Data Charter pushes open statistics
- Large amounts of numeric data already available
- In theory: Important information at your fingertips
  - Computations
  - Comparisons
  - …



© FZI Forschungszentrum Informatik

Problem

- How to find the right number for a specific question?
- Standard way = search engines
- Numeric data often not properly indexed
- Reasons
  - Deep web
  - CSV, SDMX…
  - JavaScript
  - …

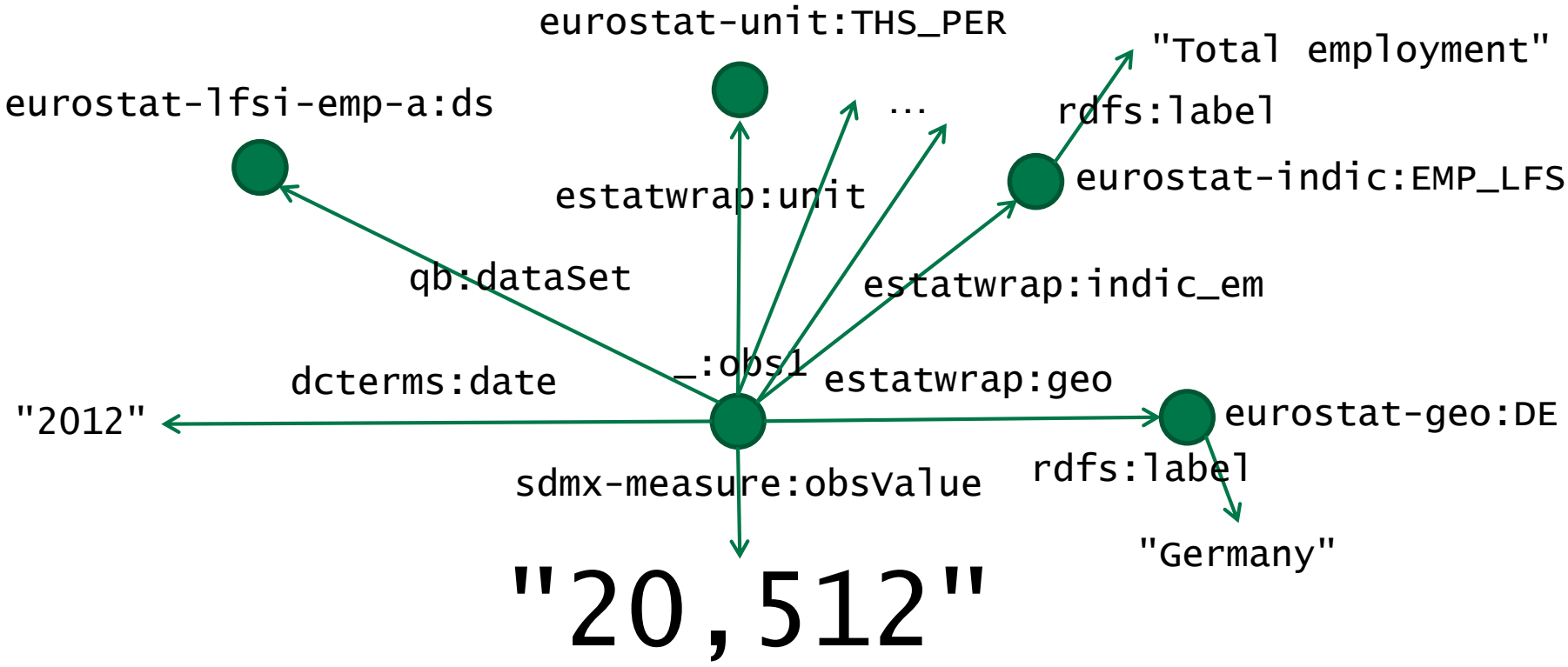


At the same time… Statistical Linked Data

How to publish numeric data on the Web?

"20,512"

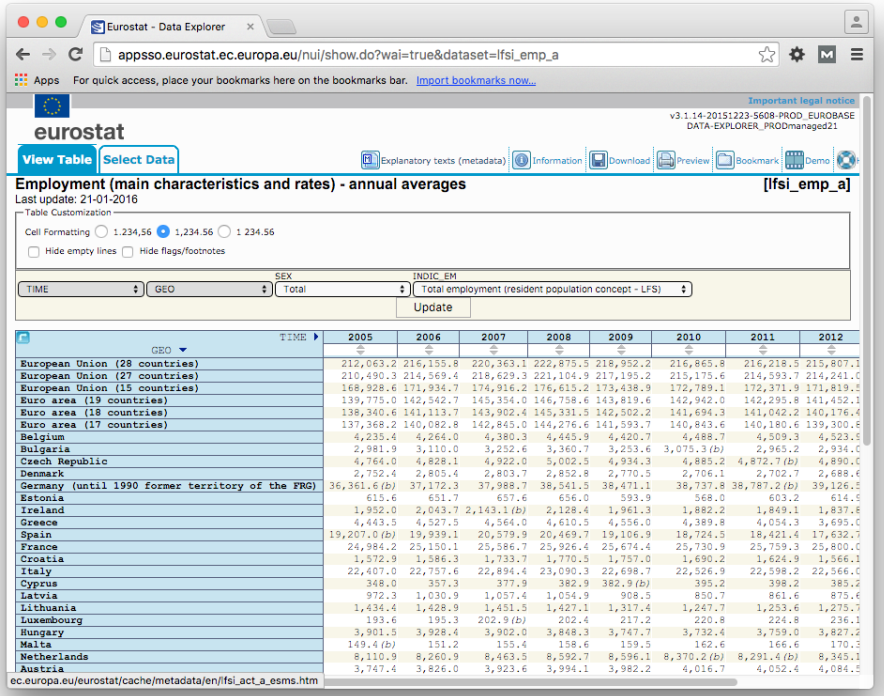
At the same time… Statistical Linked Data (2)

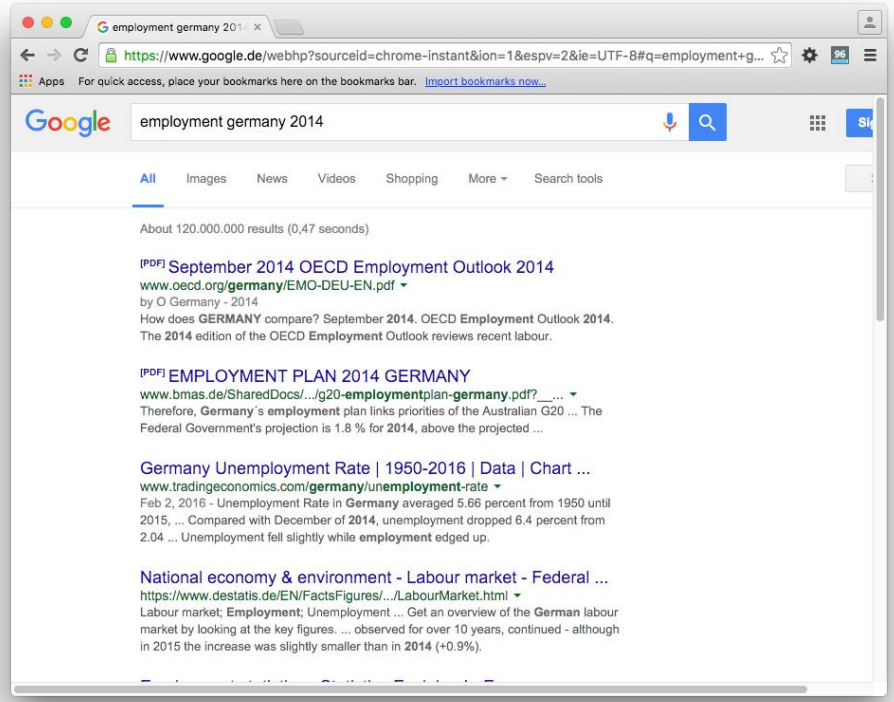


Dataset URI: http://estatwrap.ontologycentral.com/page/lfsi_emp_a
RDF Data Cube Vocabulary: *https://www.w3.org/TR/vocab-data-cube/*

© FZI Forschungszentrum Informatik

Problem





Research questions

- Can we semi-automatically built a version of arbitrary statistics on the web that is indexed by search engines?
- Is it true that the more specific a query, the better does our version in comparison to existing pages?

Outline

- Motivation/Problem
- **Related Work**
- Approach
- Evaluation
- Discussions
- Conclusions

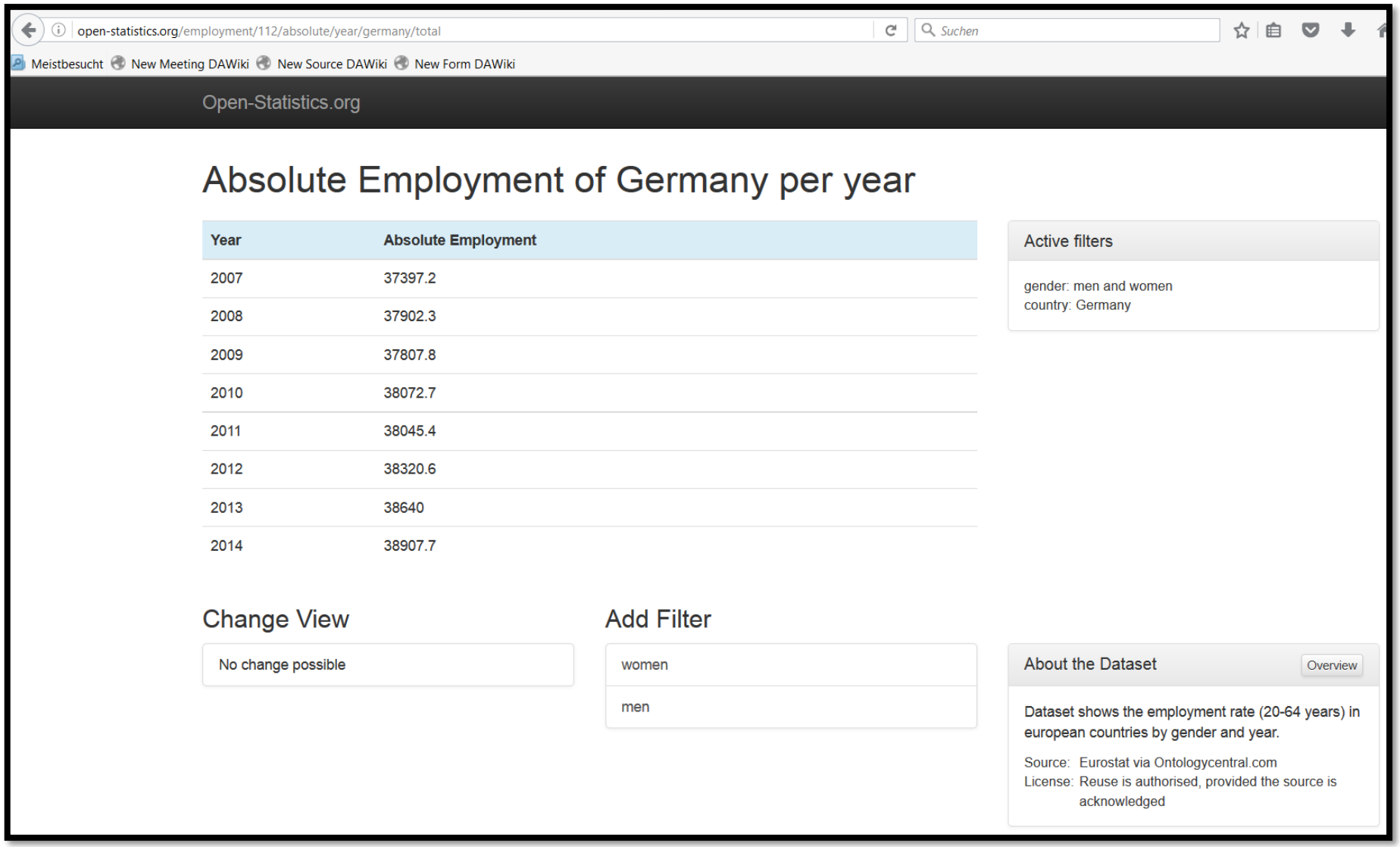
Related Work

- We generate **search engine optimized pages**
  - Statistical Linked Data visualisation: CODE Query Wizard [Hoe13], CubeViz [SMM+12]
- We use **semantic annotations**
  - Automatic semantic annotation of documents: (automatic) annotation of websites with schema.org [VE13,Mik15]
- We do not do **off-page SEO** (e.g., promoting incoming links)
  - SEO: Web 2.0 techniques [Mal09], empirical evaluations of SEO techniques [SCC13]
- We make **transparent the sources** of numeric data
  - Commercial websites with numeric information: Wolfram Alpha, Tradingeconomics, Statista...

Outline

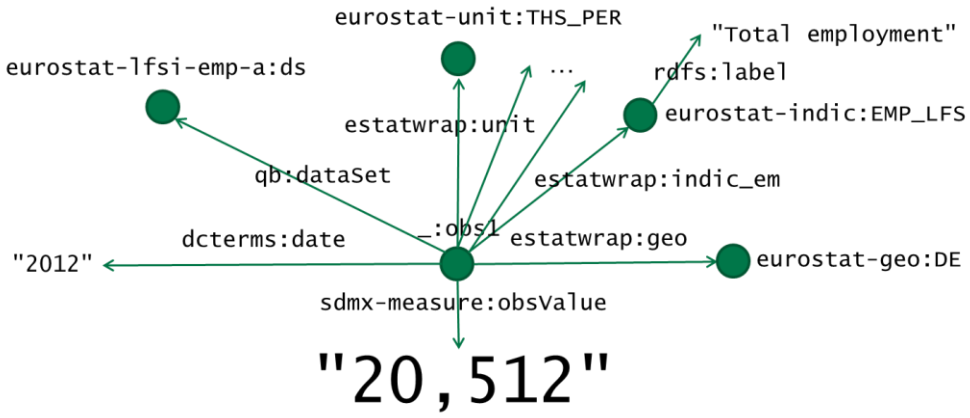
- Motivation/Problem
- Related Work
- **Approach**
- Evaluation
- Discussions
- Conclusions

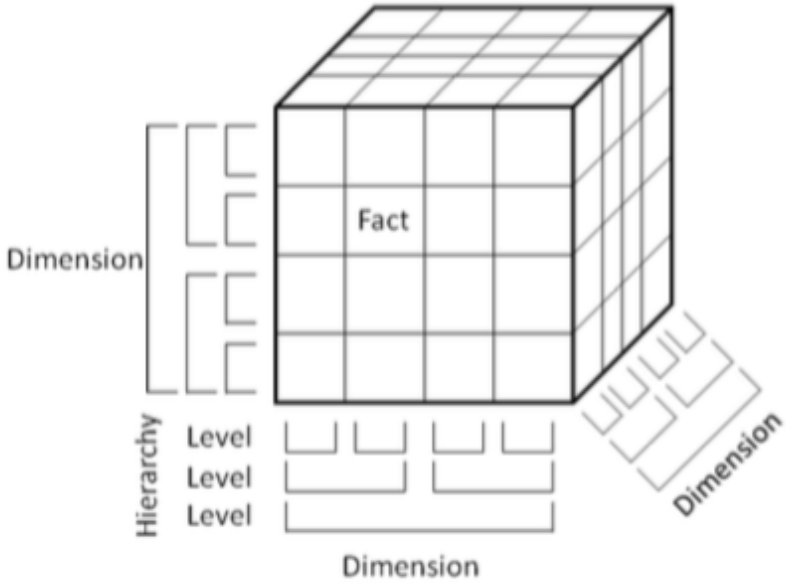
Approach: For every possible "QB view" a search engine optimized landing page



QB View?

QB view = OLAP subcube query [KOH12] = Diced X Sliced X Freed





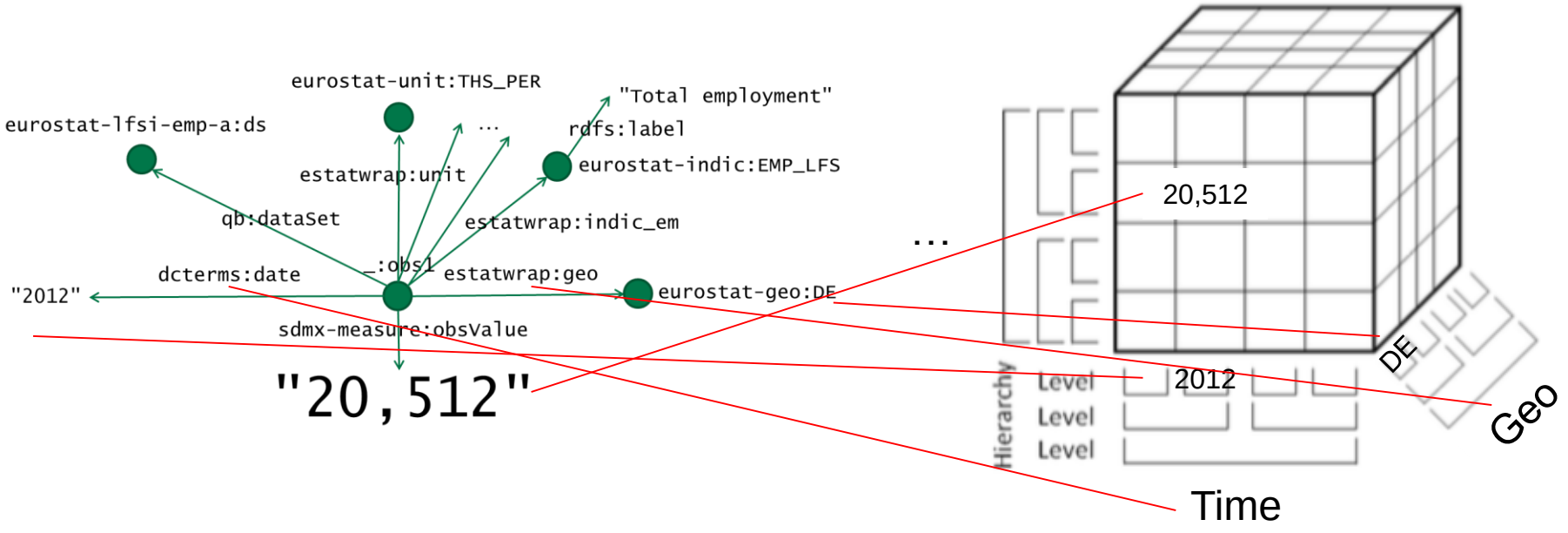
Statistical Linked Data

Data Cube

© FZI Forschungszentrum Informatik

QB View?

QB view = OLAP subcube query [KOH12] = Diced X Sliced X Freed



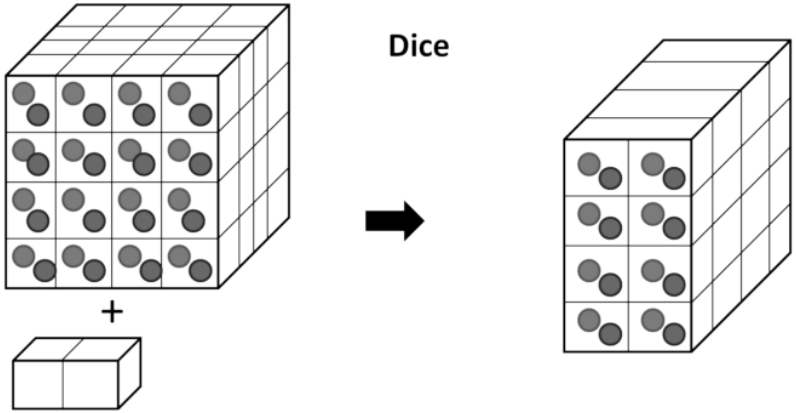
Statistical Linked Data

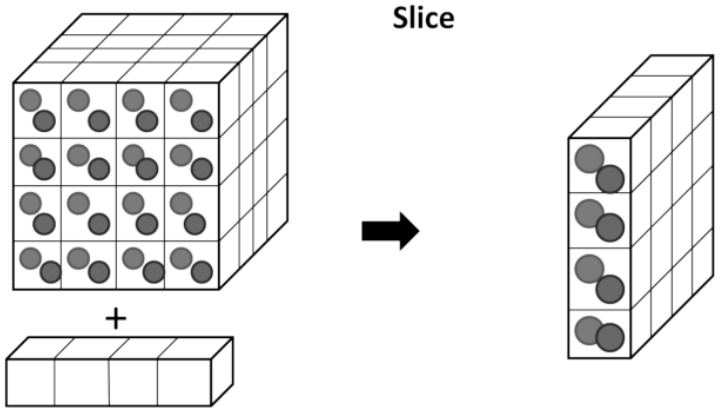
Data Cube

© FZI Forschungszentrum Informatik

QB View? (2)

Dice	Slice
Filter for one (or more) dimension values	Remove and aggregate over one dimension





© FZI Forschungszentrum Informatik

QB View? (3)

- OLAP subcube query [KOH12]: `Diced X Sliced X Freed`
- Example: `({(Unit,THS_PER),(Indicator,Total employment),…}, {Sex,Geo}, {Time})`

Dice	Slice	Free
Unit = Thousand persons Indicator = Total employment Age = 15-64 years Geo = DE	Sex = ALL	Time

Time	Value
2007	37,397
2008	37,902
…	…

Numeric data on the landing page

© FZI Forschungszentrum Informatik

For every possible "QB view"?

- # QB view = $\prod_{i=1}^{n}(d_i + 2)$
- with
  - n : number of dimensions
  - d_i : number of values of dimension i
  - 2 : additional values per dimension for "Sliced" and "Freed"
- Example:
  - 6 dimensions (time, geo, age, unit, sex, indicator)
  - 10 values for time, 42 (geo), 2 (age), 2 (unit), 2 (sex), 2 (indicator)
  - 12 x 44 x 4 x 4 x 4 x 4 = 135,168

© FZI Forschungszentrum Informatik

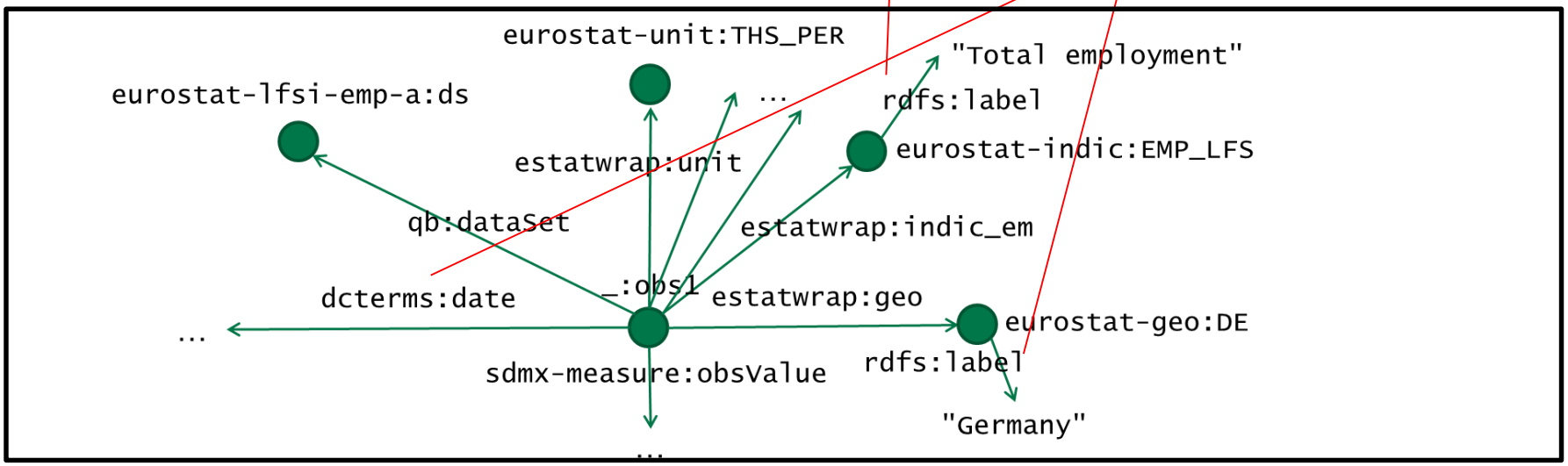
Search engine optimized landing page?

Verbalisation (using templates)

<title> … </title>
<h1>{{diced}}{{freed}}</h1>
…

<title> … </title>
<h1>Employment of Germany per year</h1>
…

HTML



Statistical Linked Data

Search engine optimized landing page? (2)

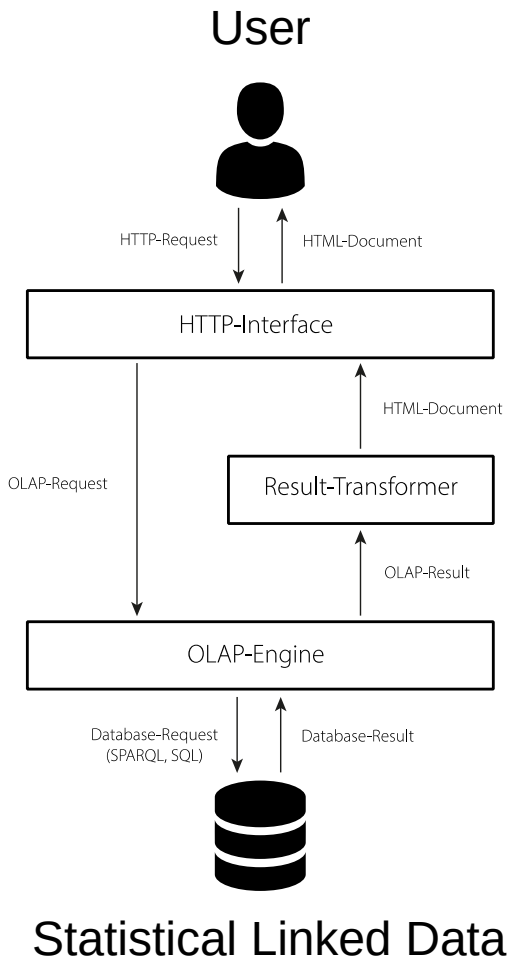
- URI schema
  - `({(Unit,THS_PER),(Indicator,Total employment),…},{Sex,Geo},{Time})`
  - HTTP Request
    http://baseUri/endpoint?dataset=id1&freed=id2&diced=id3...
  - URL Scheme
    http://baseUri/id1/pattern/id1/id2/id3/... (pattern defines how many freed dimensions)
  - Bsp:
    http://example.org/lfsi-emp-a/1/date/total_employment/germany
- No off-page optimisation (e.g., promoting incoming links)

Outline

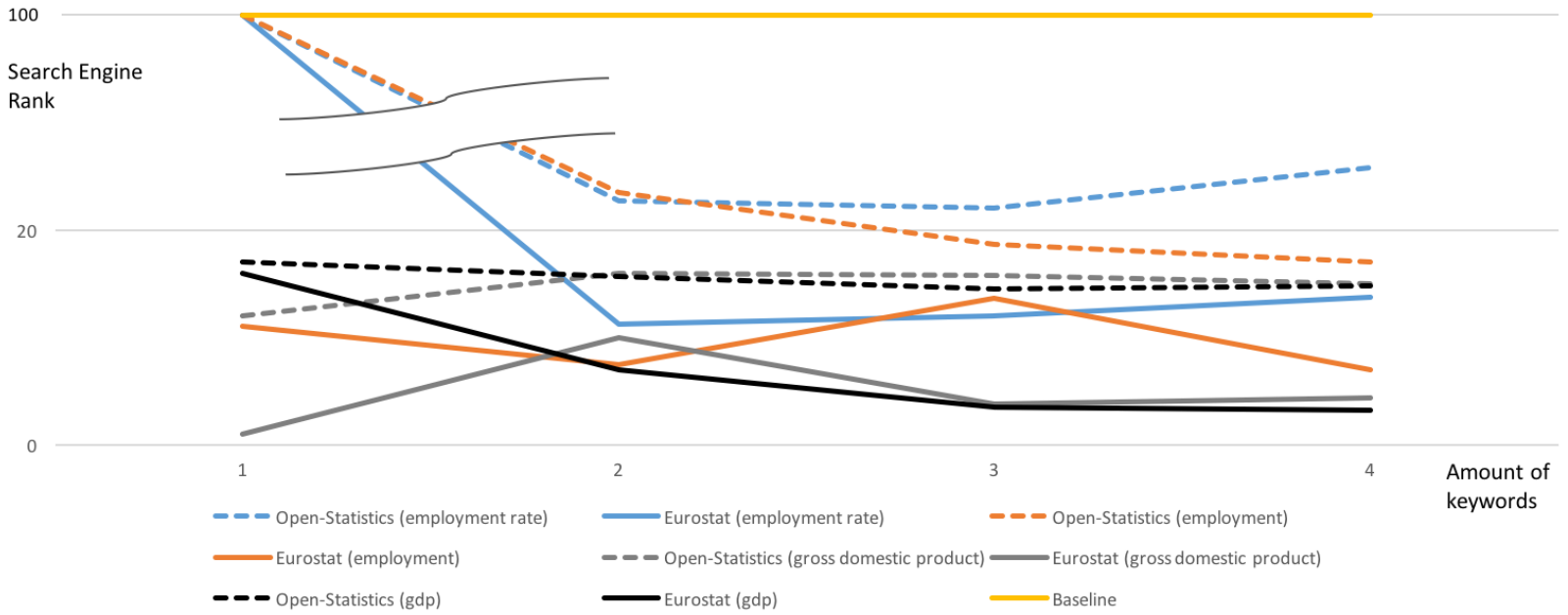
- Motivation/Problem
- Related Work
- Approach
- **Evaluation**
- Discussions
- Conclusions

Evaluation

- Hypotheses
  - We can built a version of arbitrary Statistical Linked Data that is indexed by search engines
  - The more specific a query, the better does our version in comparison to existing pages
- Implementation
  - Website: http://open-statistics.org/
  - HTTP-Interface+Result-Transformer: https://github.com/dbreucker/seo4olap-evaluation/
  - Statistical Linked Data: Estatwrap: http://estatwrap.ontologycentral.com/
  - OLAP-Engine: OLAP4LD [KH14]
- Evaluation
  - 2 datasets: Employment in the EU - annual averages, Gross Domestic Product (GDP)
  - Restricted number of pages: max. 2x free dimensions, max. 2x diced dimensions
  - Time: 614 pages generated in 30min



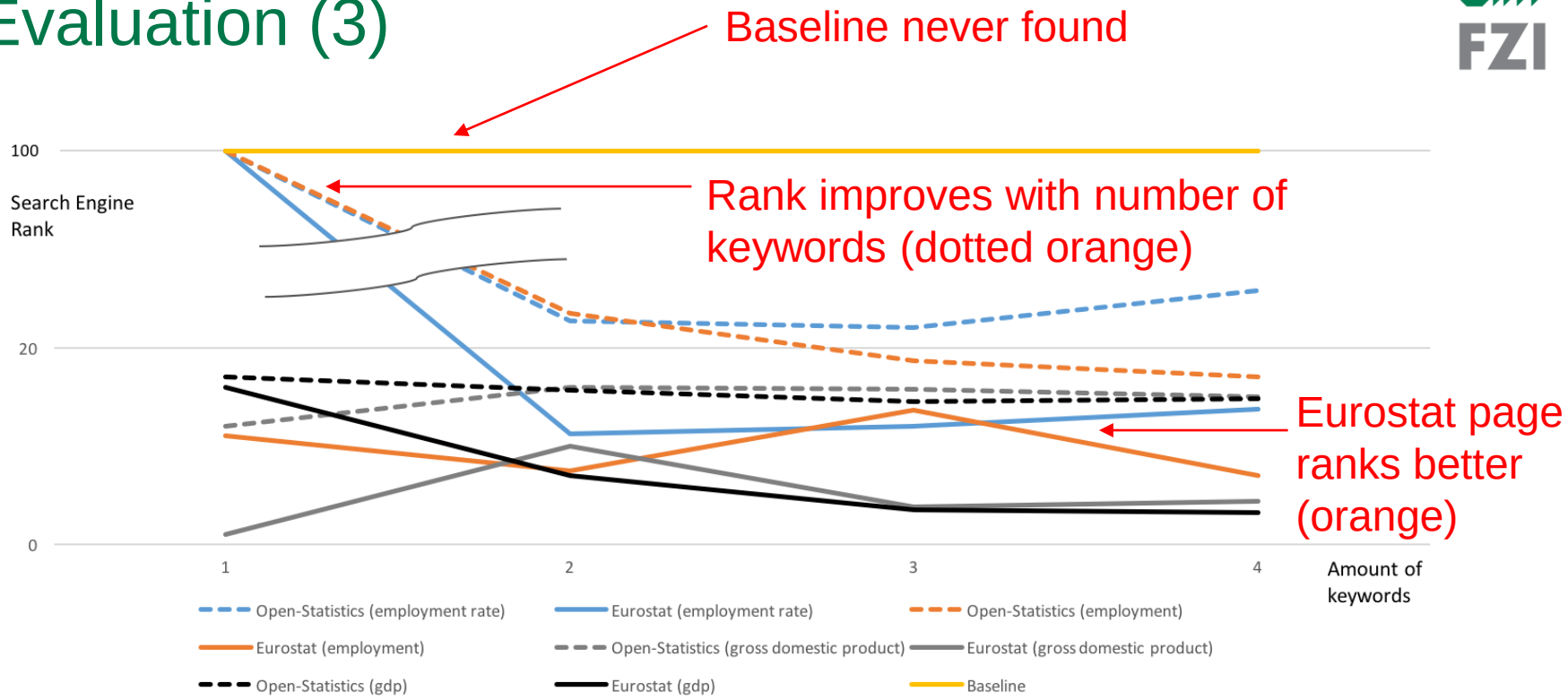
Evaluation (3)



- Comparing Google rank of 3x page types using CuteRank tool
  - Baseline = Eurostat Pivot Pages
  - Eurostat = Landing pages of Eurostat (without actual numbers)
  - Open-Statistics = Our landing pages

© FZI Forschungszentrum Informatik

Evaluation (3)



- Comparing Google rank of 3x page types using CuteRank tool
  - Baseline = Eurostat Pivot Pages
  - Eurostat = Landing pages of Eurostat (without actual numbers)
  - Open-Statistics = Our landing pages

© FZI Forschungszentrum Informatik

Outline

- Motivation/Problem
- Related Work
- Approach
- Evaluation
- **Discussions**
- Conclusions

© FZI Forschungszentrum Informatik

Discussions

- About generation of pages…
  - **Measure dimension:** some dimensions refer to the indicator and should not be sliced since aggregation would be meaningless
  - **Slice member:** some dimensions contain special type members "total" that should not be included in aggregations
  - **Alternative labels** based on simple rules ("female" => "women")
  - **Unit** not always available in a machine readable format (e.g., ambiguous "ratio or thousand" simply in the textual documentation)
  - Some datasets only mention their **indicator** neither as a specific measure nor in a dimension but only in the title
  - **Schema.org** so far does not provide a way to annotate arbitrary numeric facts for search engines
- About the evaluation…
  - We can built a version of arbitrary Statistical Linked Data that is indexed by search engines – **Confirmed.**
  - The more specific a query, the better does our version in comparison to existing pages – **Valid for some pages but cannot be generally shown.**

© FZI Forschungszentrum Informatik

Outline

- Motivation/Problem
- Related Work
- Approach
- Evaluation
- Discussions
- **Conclusions**

© FZI Forschungszentrum Informatik

Conclusions

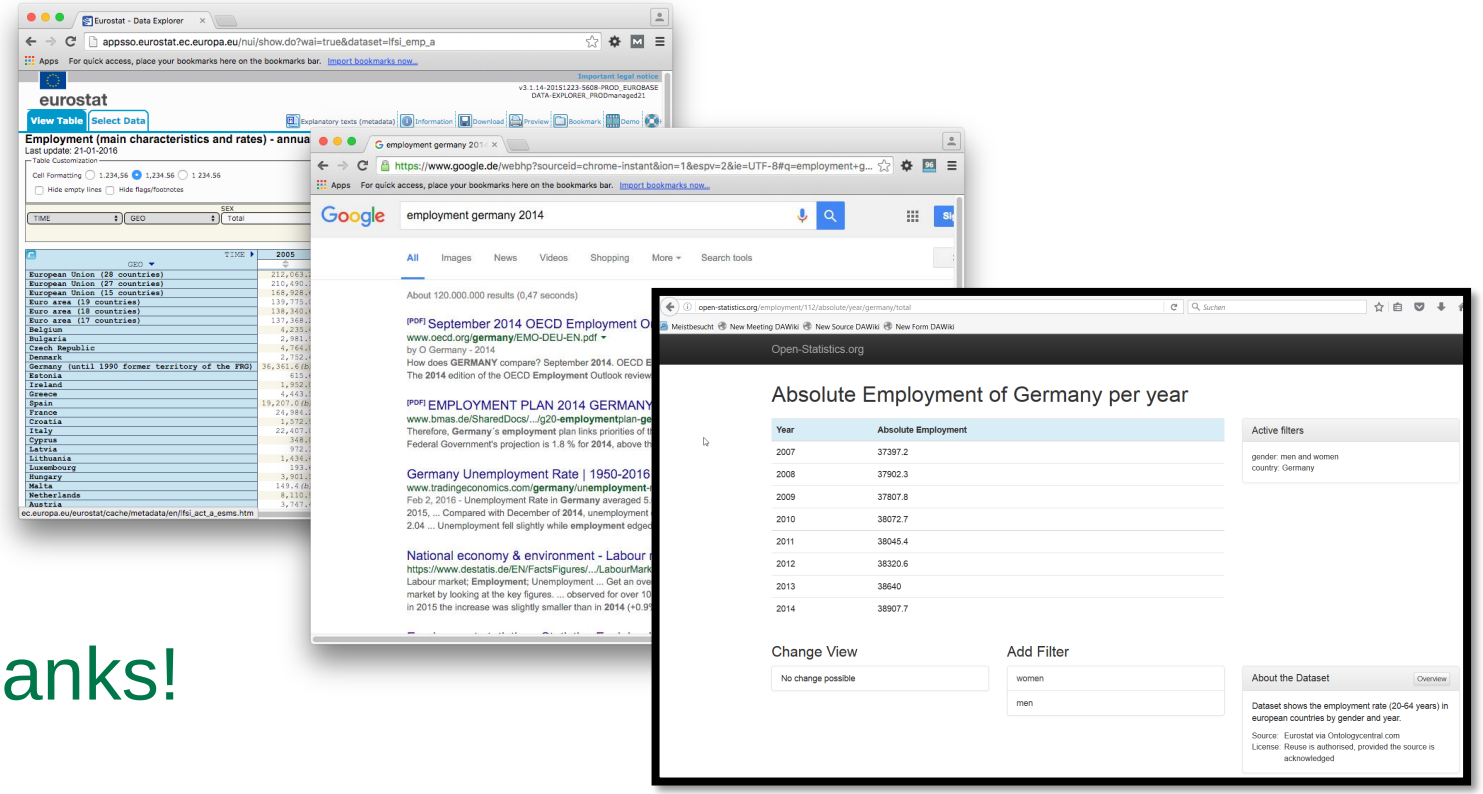
Summary

- We can automatically generate SEO landing pages for arbitrary statistics on the web using Statistical Linked Data
- Although we cannot compete with more prominent off-page optimized websites, some of our pages may eventually prove to be a useful source of specific numeric information for search engines and users

Open work

- So far only two datasets published
  - What to do if we want to publish all 5,000 datasets of Eurostat and of other available Statistical Data sources?
- Is it possible to automatically provide additional information besides the specific numbers such as
  - definitions of indicators from DBpedia/Wikipedia
  - news articles mentioning the numbers
  - highlighting anomalies
  - showing related numbers (for comparisons)
- Is it possible to allow arbitrary combinations (e.g., derived number "GDP per capita")…
- …and comparisons of numbers from the "Global Cube" of all available Statistical Linked Data on the web?

© FZI Forschungszentrum Informatik



Thanks!

Kontakt:

Benedikt Kämpgen: kaempgen@fzi.de

Andreas Harth: harth@kit.edu

FZI FORSCHUNGSZENTRUM INFORMATIK

References

- [Hoe13] P. Hoefler, “Linked data interfaces for non-expert users,” in The Semantic Web: Semantics and Big Data. Springer, 2013, pp. 702–706.
- [KH14] Benedikt Kämpgen, Andreas Harth: OLAP4LD - A Framework for Building Analysis Applications Over Governmental Statistics. ESWC (Satellite Events) 2014: 389-394.
- [KOH12] Benedikt Kämpgen, Seán O'Riain, Andreas Harth: Interacting with Statistical Linked Data via OLAP Operations. ESWC (Satellite Events) 2012: 87-101.
- [Mal09] R. Malaga, “Web 2.0 Techniques for search engine optimization: Two case studies,” Review of Business Research, vol. 9, no. 1, pp. 132–139, 2009.
- [Mik15] P. Mika, “On Schema.org and Why It Matters for the Web,” Internet Computing, IEEE, vol. 19, no. 4, pp. 52–55, 2015.
- [SCC13] B.-Y. Shih, C.-Y. Chen, and Z.-S. Chen, “An empirical study of an internet marketing strategy for search engine optimization,” Human Factors and Ergonomics in Manufacturing & Service Industries, vol. 23, no. 6, pp. 528–540, 2013.
- [SMM+12] P. E. R. Salas, M. Martin, F. M. D. Mota, S. Auer, K. Breitman, M. Casanova et al., “Publishing statistical data on the web,” in Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on. IEEE, 2012, pp. 285–292.
- [VE13] C. Veres and E. Elseth, “Schema.org for the Semantic Web with MaDaME.” in I-SEMANTICS (Posters & Demos). Citeseer, 2013, pp. 11–15.

© FZI Forschungszentrum Informatik