LINKED DATA AS-A-SERVICE
THE SEMANTIC WEB REDEPLOYED
LAURENS RIETVELD, RUBEN VERGORGH, WOUTER BEEK, MIEL VANDER SANDE

Slides at http://wouterbeek.github.io
PROBLEM 1

The LOD Cloud cannot be uniformly queried today.
Most SW datasets are not available online
Those that are online are often data dumps
Many of those are not standards-compliant
Many datasets that can be queried live have a custom API
Most custom APIs are not self-describing
Many live-queryable datasets pose restrictions
Others have low availability
Different SPARQL endpoints pose different restrictions
PROBLEM 2

Existing deployment techniques are unable to close the gap between downloadable data dumps and live queryable data.
Even though a technology stack for publishing Semantic Web data exists today, there is currently no simplified Web Service that does the same thing on a Web-scale.
PROBLEM 4: FEDERATION

There are no LOD Cloud-wide guarantees as to whether, and if so how, sub-queries will be evaluated by different endpoints.
RELATED WORK

Large Linked Datasets: BTC, Freebase
Large Linked Data Idexes: Sindice, LODCache, DyLDO
Cloud-based triple store: Dydra
SOLUTION 1: MACHINE READABILITY

Clean & republish all data documents.
LOD LAUNDROMAT

HTTPS://LODЛАУНДРОМАТ.ORG

Open source (of course)
SOLUTION 2: AVAILABILITY

Strike a balance between server- and client-side processing.
SERVER-SIDE

SPARQL server

client

client

client

client

low availability
CLEANT- AND SERVERS-SIDE (LDF)
SOLUTION 3: SERVICABILITY

- Integration with popular services (Dropbox)
- Self-descriptive Web Services (Hydra)
- Command-line tool *Frank*
- Libraries for popular programming languages
SCALABLE

- 38B ground statements
- 650K data documents (265GB HDT, 193GB raw)
USAGE NUMBERS (APRIL STATS)

- 2,150 users
- 2,119,218 downloads
- 8,586,193 queries
Fig. 3. All FedBench queries complete slowly, but successfully, with high average recall (shown on top of each bar) when ran on the deployed LDaaS.
QUESTIONS?