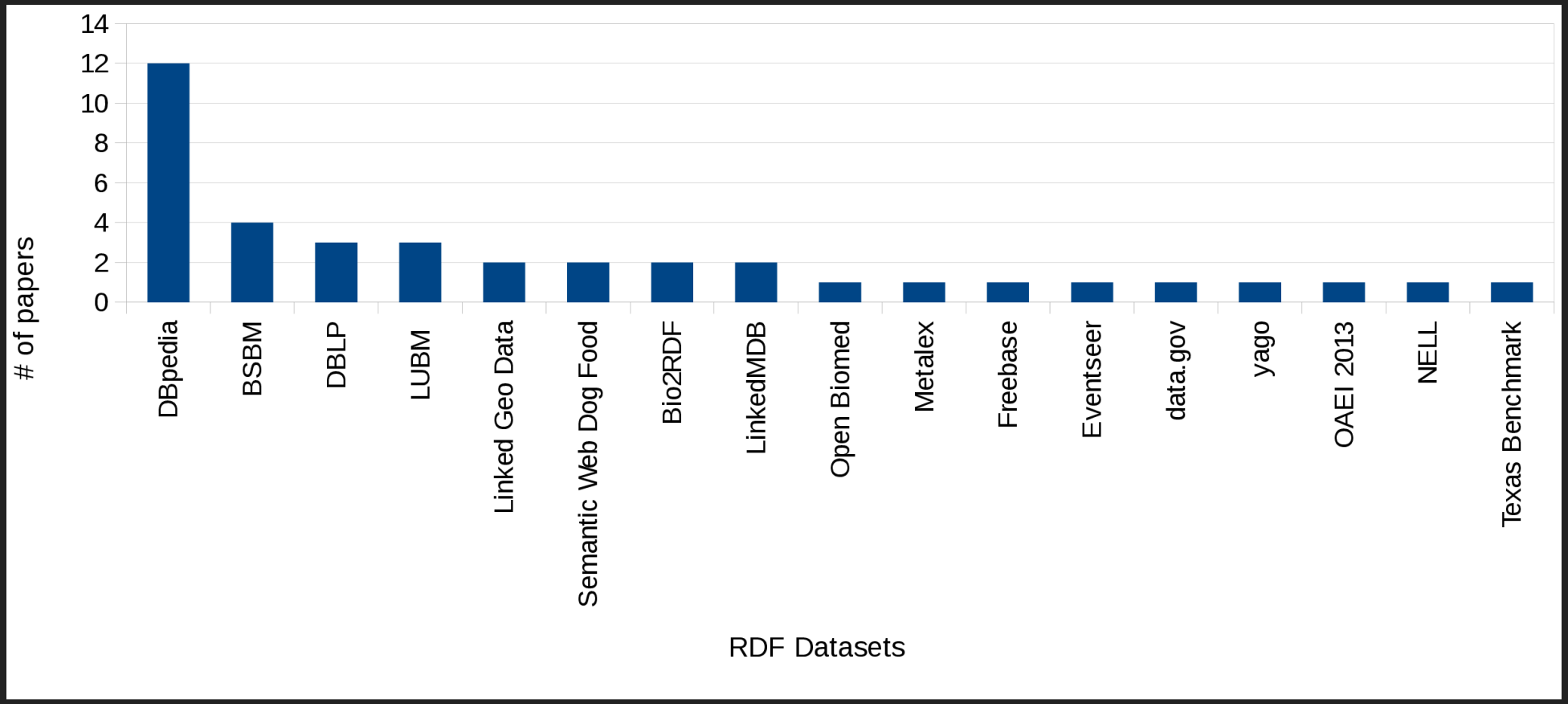
LOD Lab

Experiments at LOD Scale

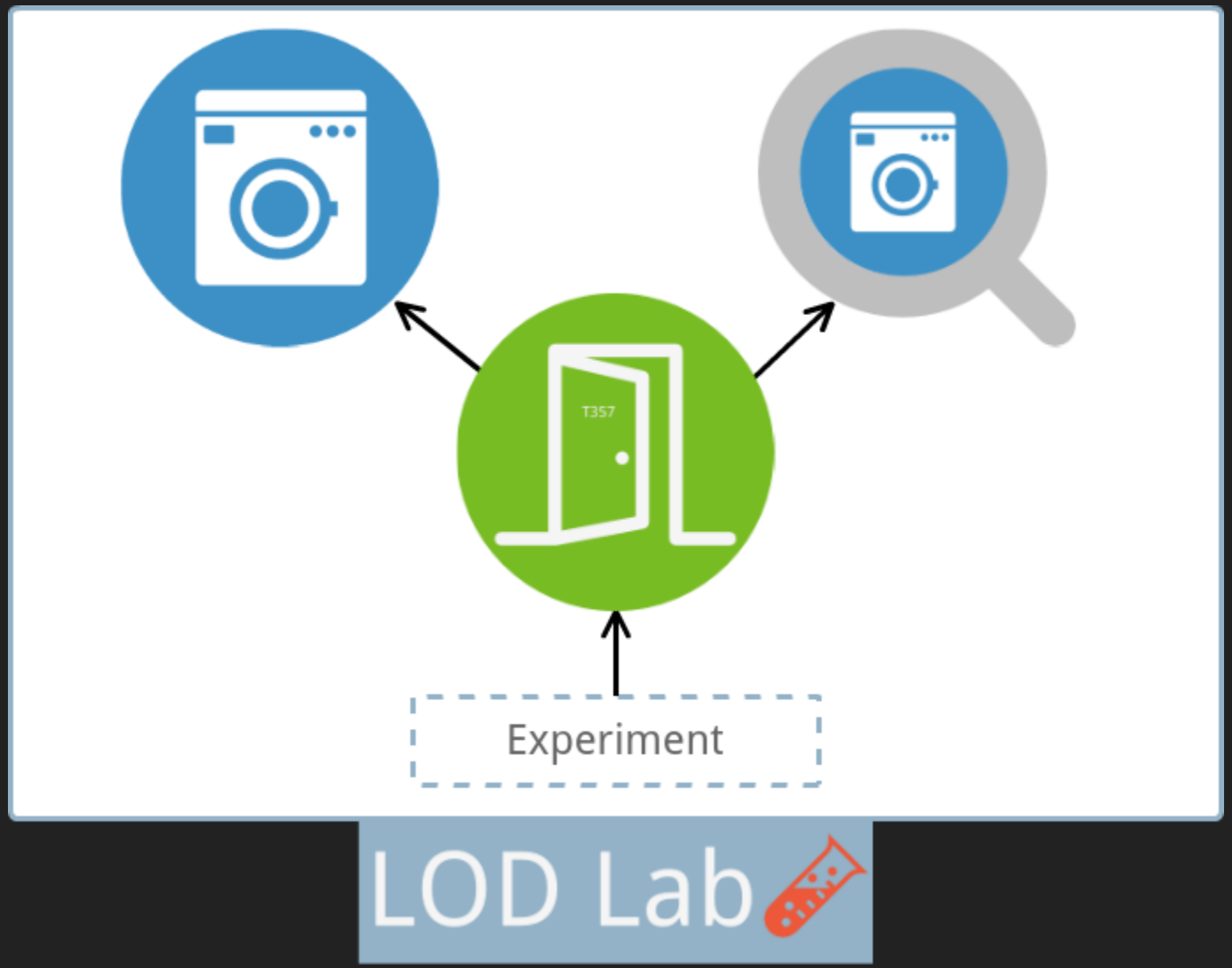
LAURENS RIETVELD, WOUTER BEEK AND STEFAN SCHLOBACH

http://presentations.laurensrietveld.nl/iswc2015/


14
12
10
8
6
4
2
0
# of papers
DBpedia
BSBM
DBLP
LUBM
Linked Geo Data
Semantic Web Dog Food
Bio2RDF
LinkedMDB
Open Biomed
Metalex
Freebase
Eventseer
data.gov
yago
OAEI 2013
NELL
Texas Benchmark
RDF Datasets

WHY IS WEB SCALE LINKED DATA EVALUATION DIFFICULT?

1. 'Messy' datasets
2. Datasets are hard to find
3. Inaccessible via a uniform interface



lodlaundromat.org/services

PROBLEM 1: MESSY DATASETS

http://lodlaundromat.org

LOD LAUNDROMAT

- 650k documents, 38 billion cleaned triples
- Gzipped N-Triples/N-quads files
- Triple Pattern Fragment APIs

PROBLEM 2: DATASETS ARE HARD TO FIND

lodlaundromat.org/sparql
index.lodlaundromat.org

STRUCTURE DESCRIPTIONS

lodlaundromat.org/sparql

- Aggregate Descriptions
- Syntactic Descriptions
- Network Properties

CONTENT DESCRIPTIONS

index.lodlaundromat.org

- Resource → Document
- Namespace → Document

PROBLEM 3: INACCESSIBILITY

https://github.com/LODLaundry/Frank

FRANK

- Glue between LOD Laundromat service, Meta-Data and Indexes
- Programming Language Independent: Bash Pipes

FRANK

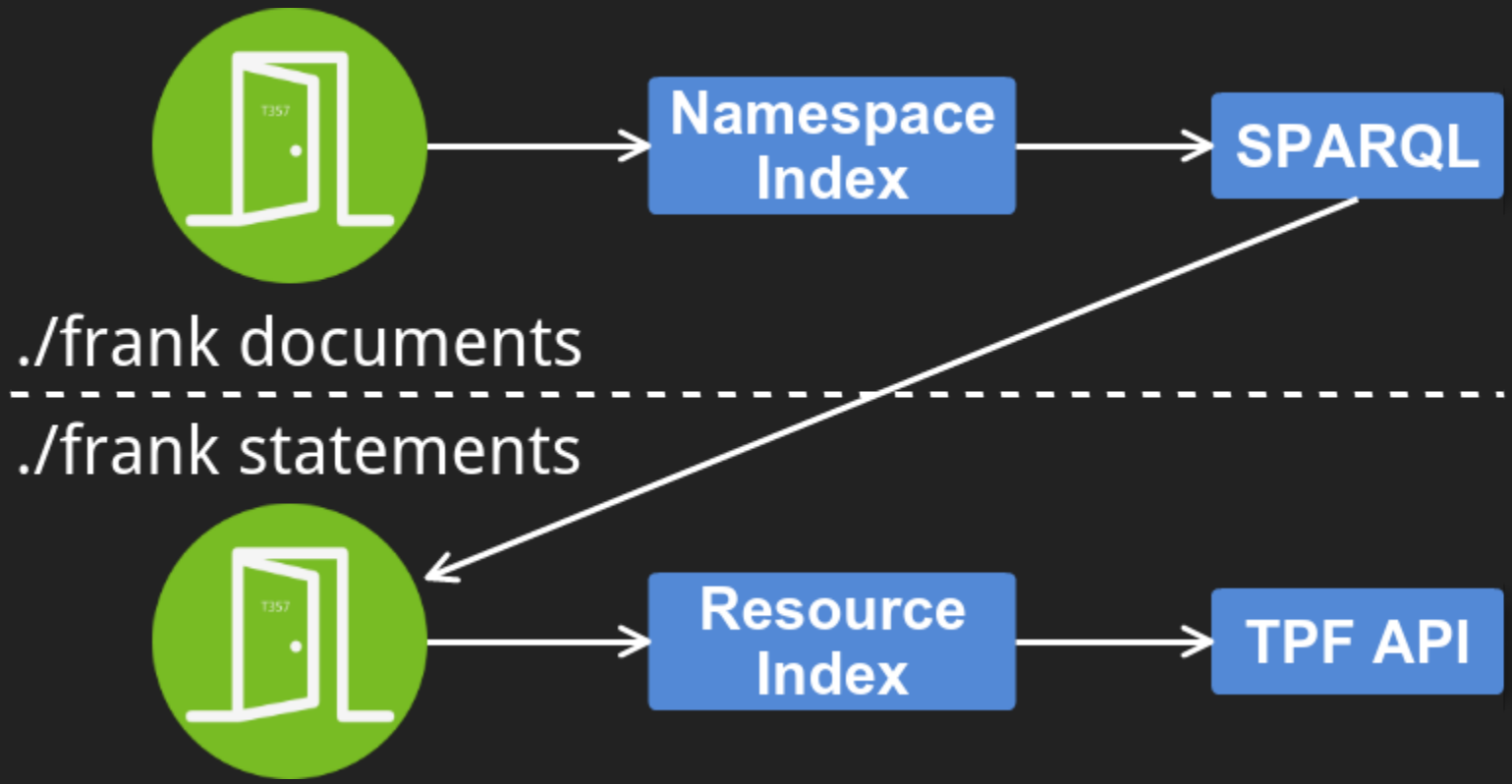
```
$ ./frank statements --predicate foaf:name | head -n 5

eurostat:void.rdf#Eurostat foaf:name "Eurostat".
author:5ff33...1c4 foaf:name "Dong-Mei Shi".
author:d873s...19b foaf:name "Feng-Xia Ma".
author:fbbcf...54c foaf:name "Ya-Guang Chen".
author:1ec76...f4b foaf:name; "Jian Yu".
```

(3/3) ACCESSIBILITY

```
$ ./frank documents --namespace void --minTriples 1000 \
| ./frank statements --predicate foaf:name \
| head -n 5;

europa:Eurostat foaf:name "Eurostat".
tw:ReviewCommission foaf:name "Review Commission"^^xsd:string.
sw:gianluca-demartini foaf:name "Gianluca Demartini".
sw:mohammad-mannan foaf:name "Mohammad Mannan".
sw:tom-minka foaf:name "Tom Minka".
```


Namespace Index
SPARQL
./frank documents
./frank statements
Resource Index
TPF API

LOD LAB DEMONSTRATION

1. RDF Vault

   Bazoobandi, Hamid R., et al. "A Compact In-Memory Dictionary for RDF Data." The Semantic Web. Latest Advances and New Domains. Springer International Publishing, 2015. 205-220.

2. RDF Header Dictionary Triples (HDT)

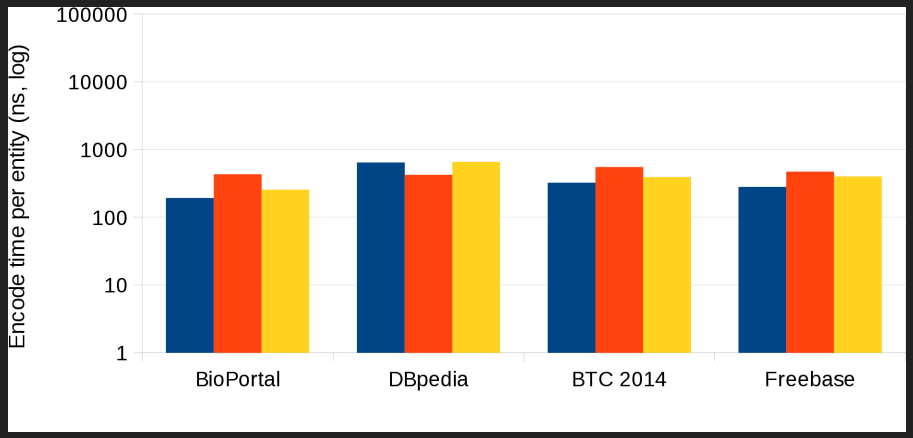
   Fernández, Javier D., et al. "Binary RDF representation for publication and exchange (HDT)." Web Semantics: Science, Services and Agents on the World Wide Web 19, 2013. 22-41.

3. Linked Data Best Practices

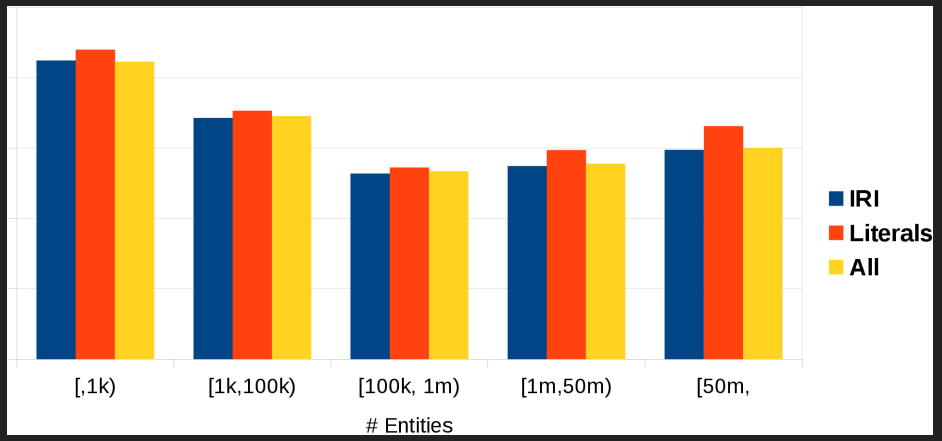
   Schmachtenberg, M., et al. "Adoption of the linked data best practices in different topical domains." The Semantic Web–ISWC 2014. Springer International Publishing, 2014. 245-260.

(1/3) RDF VAULT

RDF Vault



LOD Lab



```
$ ./frank documents --downloadUri \
--minTriples 1000 --maxTriples 100000 \
| ./runVaultExperimentForFile
```

(2/3) RDF HDT

Triples (millions)	RDF HDT (Uniprot)		LOD Lab	
	# Docs	Compression Ratio	# Docs	Compression Ratio
1	1	3.73%	179	11.23%
5	1	3.48%	74	4.99%
10	1	3.27%	50	5.43%
20	1	3.31%	17	4.15%
30	1	3.27%	15	5.09%
40	1	3.26%	8	7.25%

(2/3) RDF HDT

Avg. Degree	# Docs	Compression Ratio
1-5	92	21.68%
5-10	80	6.67%
10-∞	99	4.85%

```
$ ./frank documents \
--minAvgDegree 5 --maxAvgDegree 10 \
| ./hdtCompressDocument
```

(3/3) LINKED DATA BEST PRACTICES

Original			LOD Lab		
Prefix	**#datasets**	**%datasets**	**Prefix**	**#documents**	**%documents**
rdf	996	98.22%	rdf	639,575	98.40%
rdfs	736	72.58%	time	443,222	68.19%
foaf	701	69.13%	cube	155,460	23.92%
dcterm	568	56.01%	sdmxdim	154,940	23.84%
owl	370	36.49%	worldbank	147,362	22.67%

```
$ ./frank documents --downloadUri
| ./countNamespacesForDocument
```

CONCLUSION

- Toolkit to scale evaluations to web size
- Showcased on three recent SW Publications
- Relating experiment results to structural properties of datasets
- Our goal: Improving Linked Data Evaluation Best Practices