

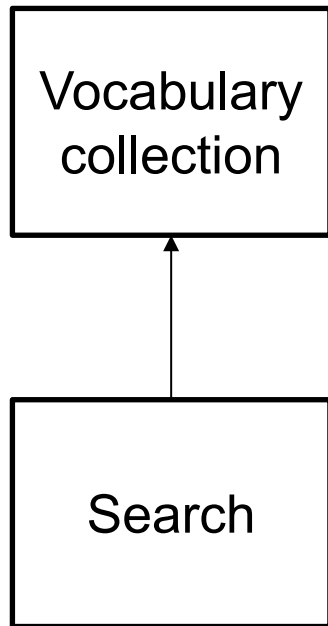
Vocabulary search

08/10/2019

Niklas Kolbe, PhD student

SnT – Interdisciplinary Centre for Security, Reliability and Trust

Ontology repositories



<i>Ontokhoj</i>	2003	[105]	■	✗
<i>OntoSelect</i>	2004	[19]	■	✗
<i>Swoogle</i>	2004	[36]	■	✓
<i>Ontosearch</i>	2005	[170]	■	✗
<i>SWSE + ReConRank</i>	2007	[55]	■	✗
<i>Sindice</i>	2007	[155]	■	✗
<i>Watson</i>	2007	[32]	■	✓
<i>Falcons Concept & Entity Search</i>	2009	[28, 117]	■	✗
<i>VisiNav</i>	2009	[53]	■	✗
<i>WebOWL</i>	2012	[12]	■	✗
<i>LODstats</i>	2012	[8]	■	✓
<i>vocab.cc</i>	2013	[140]	■	✓
<i>OUSAF</i>	2015	[5]	■	✗
<i>Supekar et al.</i>	2004	[147]	■	✗
<i>OntoMetric</i>	2004	[83]	■	✗
<i>Ontology Auditor</i>	2005	[20]	■	✗
<i>OntoQA</i>	2005	[151]	■	✗
<i>Knowledge Zone + TS-ORS</i>	2006	[79, 148]	■	✗

<i>Open Metadata Registry</i>	2006	[60]	■	✓
<i>Ontosearch2</i>	2006	[104]	■	✗
<i>Oyster</i>	2006	[103]	■	✓
<i>OBO Foundry</i>	2007	[139]	■	✓
<i>BioPortal</i>	2009	[100]	■	✓
<i>Cupboard</i>	2009	[33]	■	✗
<i>MMI</i>	2009	[124]	■	✓
<i>Ontobee</i>	2011	[165]	■	✓
<i>BiOSS</i>	2010	[89]	■	✗
<i>Manchester OWL Repository</i>	2014	[90]	■	✓
<i>smartcity</i>	2014	[112]	■	✓
<i>.linkeddata.es</i>	2014	[158]	■	✓
<i>LOV</i>	2014	[158]	■	✓
<i>Ontology Lookup Service</i>	2015	[68]	■	✓
<i>Ontohub</i>	2017	[29]	■	✓
<i>(Web)CORE</i>	2006	[25, 38]	■	✗
<i>DWRank</i>	2014	[22, 23]	■	✗
<i>TermPicker</i>	2016	[132]	■	✗
<i>NCBO 2.0</i>	2017	[88]	■	✓
<i>AKTiveRank</i>	2006	[3]	■	✗
<i>(combi)SQORE</i>	2007	[156, 157]	■	✗
<i>LOVR</i>	2015	[142]	■	✓
<i>RecoOn</i>	2016	[24]	■	✓

LOV

- <https://lov.linkeddata.es/dataset/lov/>
- 680 ontologies, 43 domains
- 280+ queries / day (UI)
- Ontology + term search
- Curation, metadata, evolution

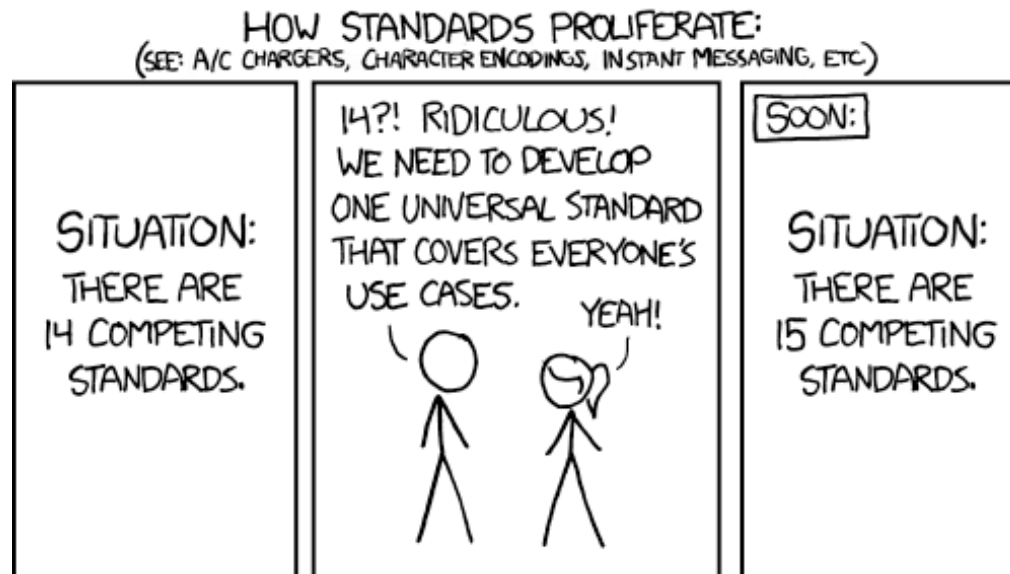
The screenshot shows the LOV website interface. The browser address bar displays 'lov.linkeddata.es'. The navigation menu includes 'VOCABS', 'TERMS', 'AGENTS', and 'SPARQL/DUMP'. A search bar labeled 'TERMS' contains the query 'LOV is all :)'.

The main content area displays search results for the query. On the left, a summary indicates '63K results'. The results are listed in a table with columns for the property name, its occurrences in LOD datasets, and a count. On the right, there are two sidebars: 'Type' and 'Tag'. The 'Type' sidebar shows a hierarchy: 'vocabulary >' (selected), 'property/class', 'property (35919)', and 'class (27541)'. The 'Tag' sidebar lists various categories such as 'Health (8988)', 'General & Upper (6670)', 'Biology (6394)', 'Catalogs (3490)', 'Geography (2885)', 'Services (2870)', 'FRBR (2734)', 'Metadata (2560)', 'Society (2517)', and 'Vocabularies (2111)'. A 'show more...' link is visible below the tag list. At the bottom, a 'Vocabulary' sidebar lists 'dicom (8625)' and 'owl (3678)'. A 'Display a menu' button is located at the bottom left of the results area.

Property	Occurrences in LOD datasets	Count
rdf:type (rdf) 60,189,781 occurrences in 500 LOD datasets http://www.w3.org/1999/02/22-rdf-syntax-ns#type	60,189,781	1.000
rdfs:label (rdfs) 44,393,909 occurrences in 410 LOD datasets http://www.w3.org/2000/01/rdf-schema#label	44,393,909	0.888
owl:sameAs (owl) 17,236,787 occurrences in 147 LOD datasets http://www.w3.org/2002/07/owl#sameAs	17,236,787	0.540
dcterms:title (dcterms) 13,420,023 occurrences in 132 LOD datasets http://purl.org/dc/terms/title	13,420,023	0.498
dcterms:subject (dcterms) 20,568,980 occurrences in 60 LOD datasets http://purl.org/dc/terms/subject	20,568,980	0.436
dcterms:identifier (dcterms) 19,016,180 occurrences in 58 LOD datasets http://purl.org/dc/terms/identifier	19,016,180	0.424
dcterms:creator (dcterms) 4,781,374 occurrences in 129 LOD datasets http://purl.org/dc/terms/creator	4,781,374	0.423
dce:title (dce) 7,626,720 occurrences in 107 LOD datasets http://purl.org/dc/elements/1.1/title	7,626,720	0.423
skos:prefLabel (skos) 8,542,993 occurrences in 97 LOD datasets http://www.w3.org/2004/02/skos/core#prefLabel	8,542,993	0.417
foaf:primaryTopic (foaf) 6,852,815 occurrences in 102 LOD datasets http://xmlns.com/foaf/0.1/primaryTopic	6,852,815	0.409

Motivation

- To help users finding most relevant ontologies and terms through keyword search, e.g., when they are not familiar with a domain
- A good ranking will make it easier to find a suitable ontology for reuse



Relevance

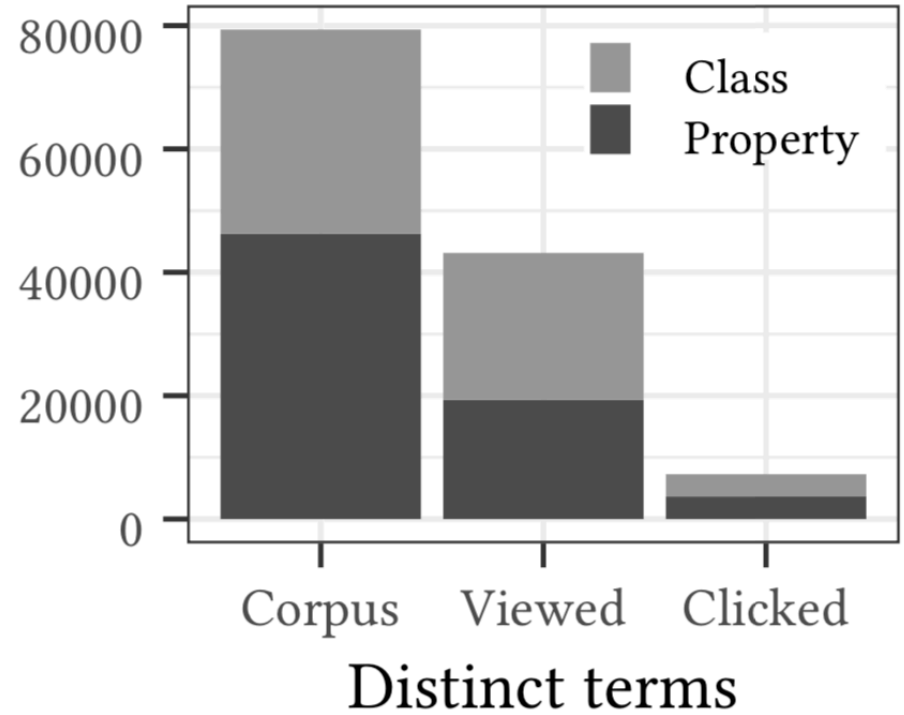
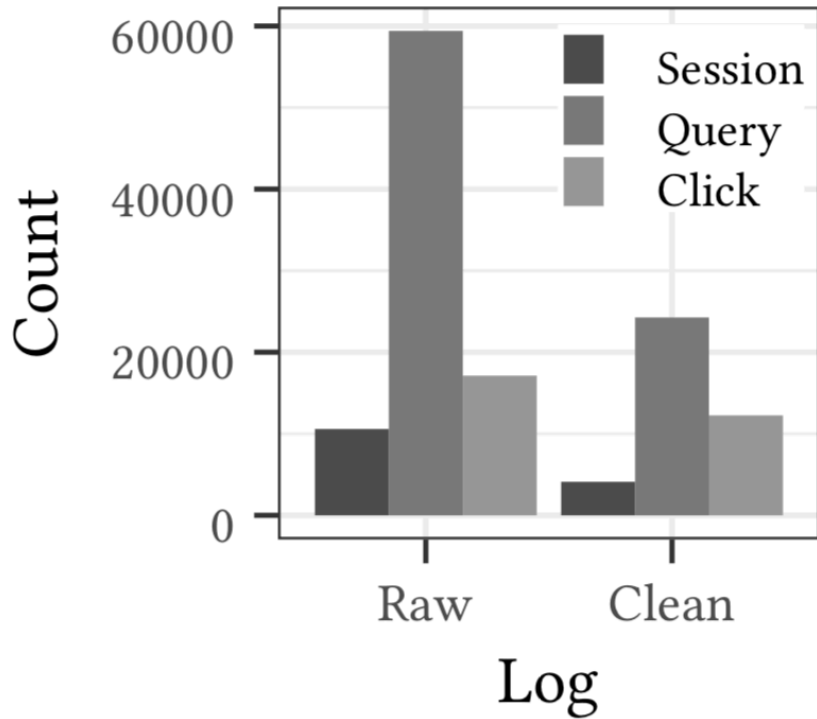
- Expert judgments
 - Costly to scale
 - May contain bias
 - E.g., CBRBench (10 queries)
- Implicit User feedback
 - In the form of observed queries and clicks
 - Contains bias
 - Easy to scale, continuous
- A large dataset with many relevance judgments allows to learn a ranking model with the optimal combination of several ranking criteria

Collecting user feedback in LOV

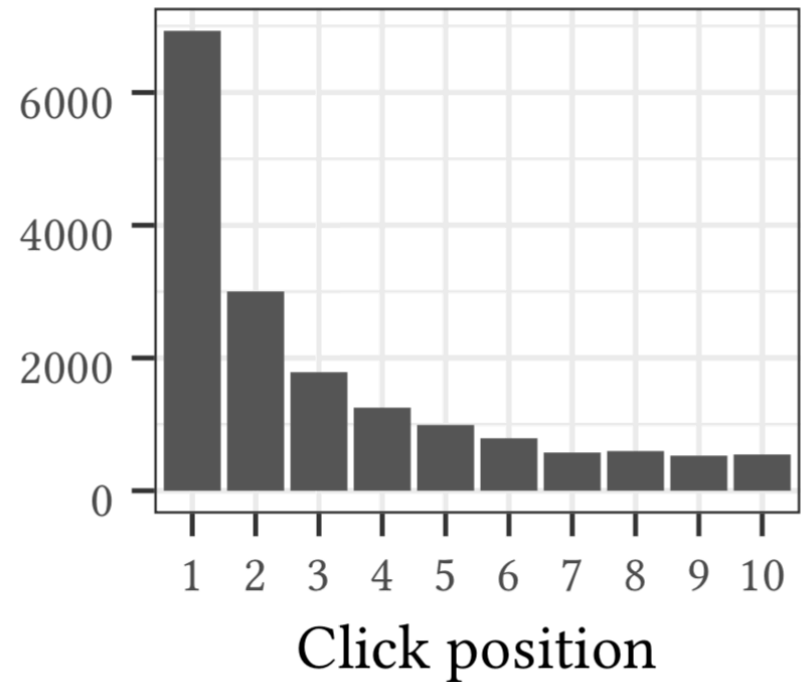
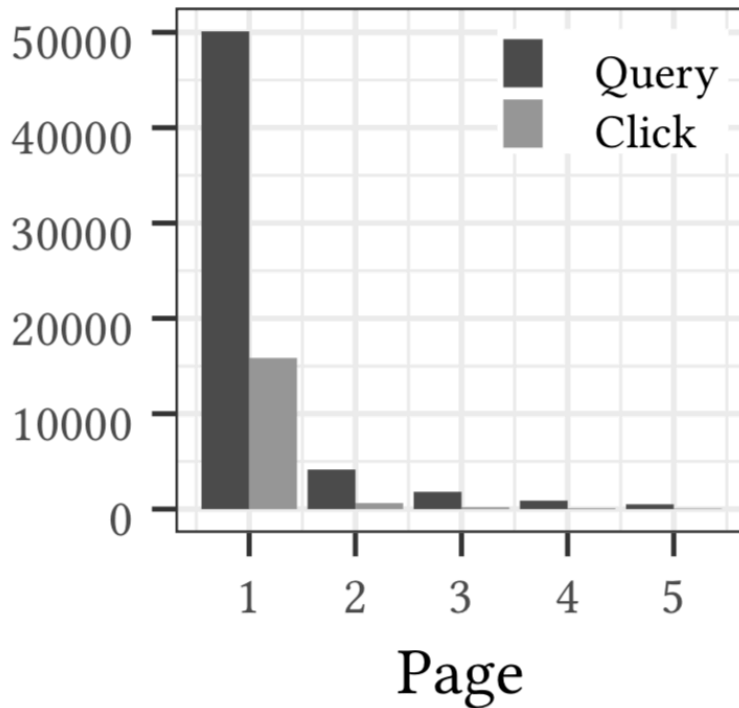
The screenshot shows a web browser window at lov.linkeddata.es. The interface has a navigation bar with 'VOCABS', 'TERMS', 'AGENTS', and 'SPARQL/DUMP'. A search bar is highlighted with a red border, containing the text 'TERMS' and 'person'. Below the search bar, the results are displayed in a table-like format. The first result is for 'foaf:Person' (foaf) with a score of 0.627. It includes the URI <http://xmlns.com/foaf/0.1/Person> and properties: `rdfs:comment` 'A person.', `rdfs:label` 'Person', and `localName` 'Person'. The second result is for 'npg:Person' (npg) with a score of 0.556, with a note 'n/a (use in LOD)' and URI <http://ns.nature.com/terms/Person>. A 'skos:definition' is provided: 'The :Person class represents a single person entity. @en'. To the right of the results is a 'Type' sidebar with a tree view: 'vocabulary >', 'property/class', 'property (1750)', 'class (512)', and 'agent >'.

Count	URI	Score
2262 results	http://xmlns.com/foaf/0.1/Person	0.627
	http://ns.nature.com/terms/Person	0.556

LOV user logs

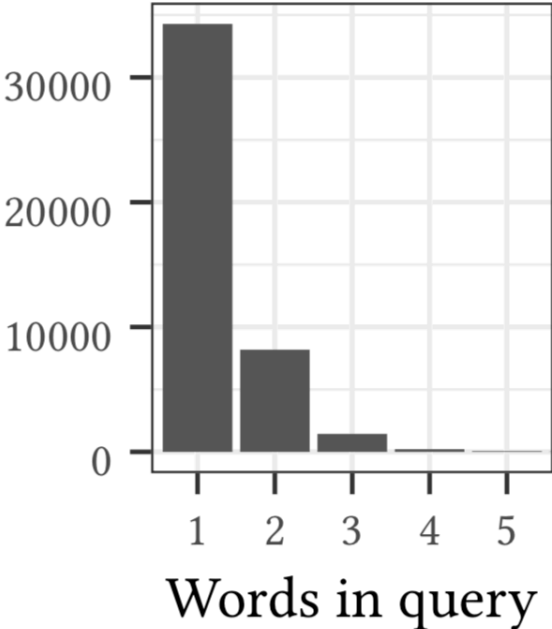


LOV user logs



“The best place to hide a dead body is page 2 of Google”

LOV user logs

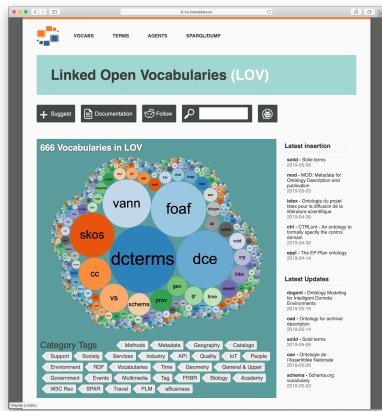
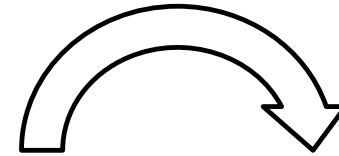
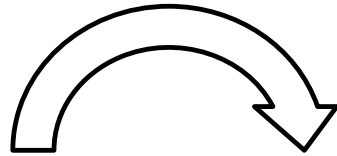
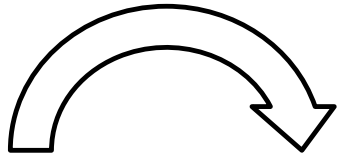


Benchmark (ground truth)

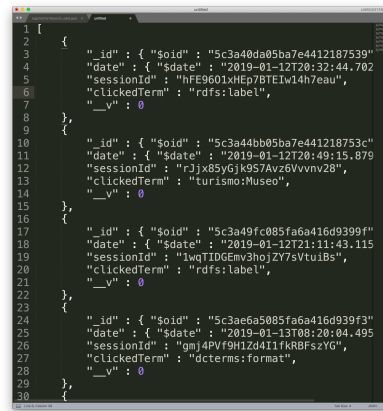
Server-side logging

Data cleaning (R script)

Learning click model
(Python, using PyClick)



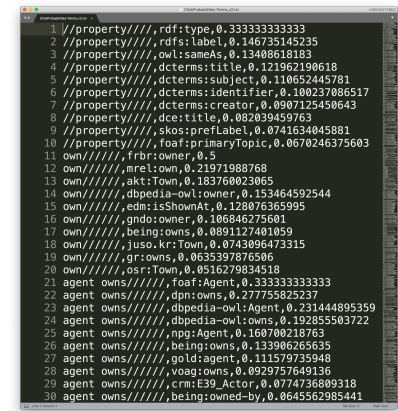
LOV



Query and click logs



Cleaned and merged logs



Click probabilities

Benchmark (ground truth)

- Query – Term – Relevance Score

```
"person", "http://xmlns.com/foaf/0.1/Person", 4
"person", "http://ns.nature.com/terms/Person", 0
"person", "http://www.bbc.co.uk/ontologies/sport/Person", 0
"person", "http://schema.org/Person", 0
"person", "http://purl.org/dc/terms/creator", 0
"person", "http://purl.org/ontology/af/person", 0
"person", "http://vocab.getty.edu/ontology#ulan1105_apprentice_of", 0
"person", "http://d-nb.info/standards/elementset/gnd#dateOfBirth", 0
"person", "http://dati.san.beniculturali.it/SAN/persona", 0
"person", "http://purl.org/saws/ontology#marginaliaAddedBy", 0
"person", "http://d-nb.info/standards/elementset/gnd#NameOfThePerson", 0
"person", "http://data.vlaanderen.be/ns/persoon#heeftPersoonsrelatie", 0
"person", "http://d-nb.info/standards/elementset/gnd#firstArtist", 0
"person", "http://purl.org/dc/elements/1.1/contributor", 0
"person", "http://vocab.getty.edu/ontology#ulan1310_advised_by", 0
"person", "http://purl.org/vocab/bio/0.1/Birth", 0
"person", "http://opendata.aragon.es/def/ei2a#personGender", 0
"person", "http://vocab.getty.edu/ontology#ulan2841_performer_was", 0
"person", "http://xmlns.com/foaf/0.1/PersonalProfileDocument", 0
"person", "http://sparql.cwrc.ca/ontologies/cwrc#FictionalPerson", 0
"person", "http://rdvocab.info/ElementsGr2/identifierForThePerson", 0
"person", "http://d-nb.info/standards/elementset/gnd#founder", 0
"person", "http://www.w3.org/2000/10/swap/pim/contact#preferredURI", 0
"person", "http://d-nb.info/standards/elementset/gnd#firstComposer", 0
"person", "http://www.bbc.co.uk/ontologies/coreconcepts/Person", -2
"person", "http://www.data-knowledge.org/dk/Person", -2
"person", "http://www.aktors.org/ontology/portal#Person", -2
"person", "http://purl.org/dc/terms/publisher", -2
```

- 7000+ queries
- 180000+ relevance judgments

Ontology ranking

- Query match
 - How well does the description match the keyword query?
 - E.g., Lucene search of rdfs:label and rdfs:comment
- Importance
 - What is the standing of the ontology/term in the repository?
 - E.g., how often has it been imported by others?
- Quality
 - what are the characteristics and does it apply to best practices?
 - E.g., existence of labels, consistencies, availability
- Metadata
 - “External” information about the ontologies
 - E.g., how often has an ontology already been used to model data?

Challenges

- Combination of ontologies
- Personalized search
- Better expression of user information need