

Squirrel

Crawling RDF Knowledge Graphs on the Web

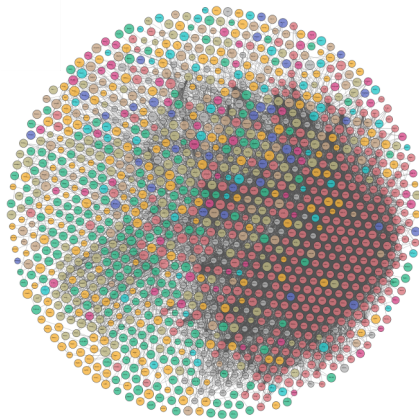
Michael Röder, Geraldo de Souza Jr.,
Axel-Cyrille Ngonga Ngomo
DICE Research Group
Paderborn University, Germany



ISWC 2020

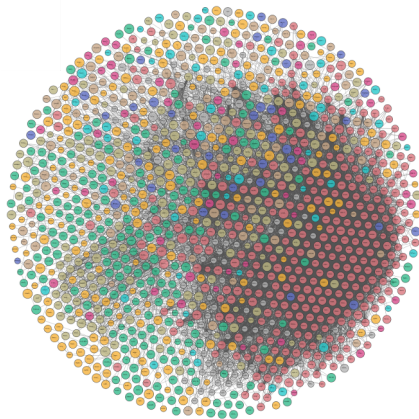
- Search
- Question answering
- Intelligent assistants
- Machine learning
- . . .

- Search
- Question answering
- Intelligent assistants
- Machine learning
- . . .



<https://www.lod-cloud.net/>

- Search
- Question answering
- Intelligent assistants
- Machine learning
- . . .



→ We need a crawler.

<https://www.lod-cloud.net/>

R1: Distributed and scalable

R1: Distributed and scalable

R2: “Respectful” behaviour

- R1:** Distributed and scalable
- R2:** “Respectful” behaviour
- R3:** Semi-structured data

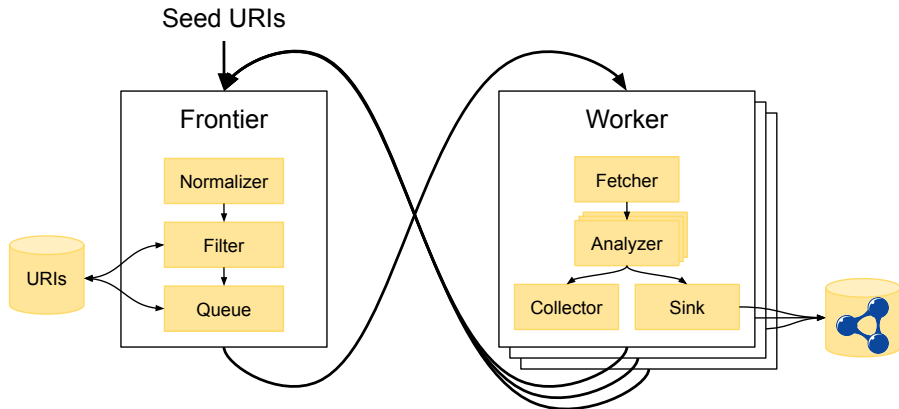
R1: Distributed and scalable

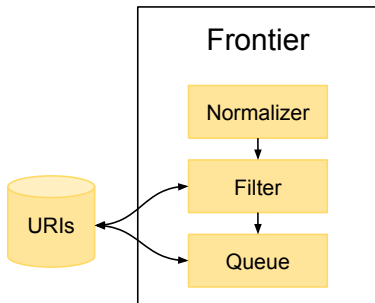
R2: “Respectful” behaviour

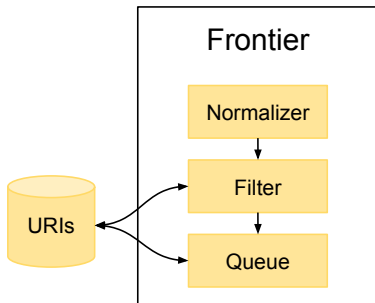
R3: Semi-structured data

R4: Fully extensible architecture

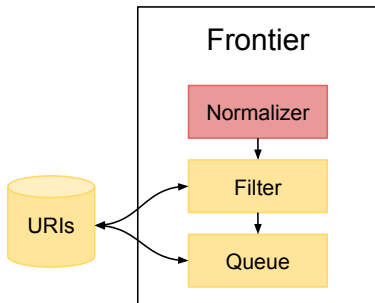
- R1:** Distributed and scalable
- R2:** “Respectful” behaviour
- R3:** Semi-structured data
- R4:** Fully extensible architecture
- R5:** Metadata about the crawling process



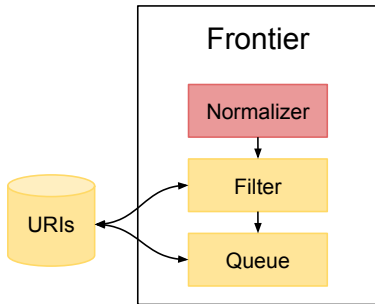




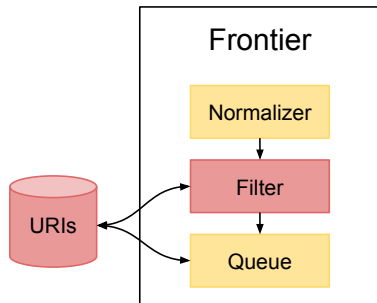
`http://ex.org:80/data/entity1`
`http://ex.org/data/entity%50`
`http://ex.com/ontology#property`
`http://ex.com/ontology#Class`
`http://dbpedia.org/resource/./Rome`



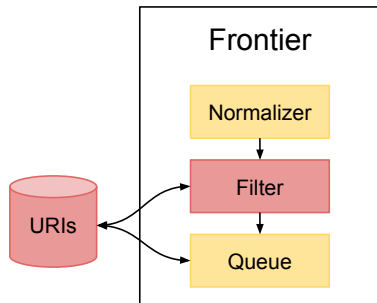
`http://ex.org:80/data/entity1`
`http://ex.org/data/entity%50`
`http://ex.com/ontology#property`
`http://ex.com/ontology#Class`
`http://dbpedia.org/resource/./Rome`



`http://ex.org:80/data/entity1`
`http://ex.org/data/entity%50`
`http://ex.com/ontology#property`
`http://ex.com/ontology#Class`
`http://dbpedia.org/resource/. /Rome`



`http://ex.org/data/entity1`
`http://ex.org/data/entity2`
`http://ex.com/ontology`
`http://dbpedia.org/resource/Rome`

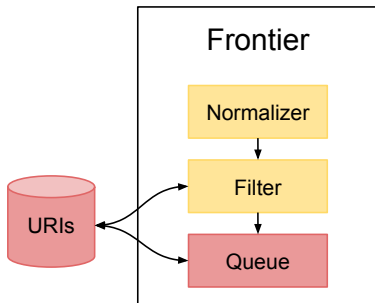


`http://ex.org/data/entity1`

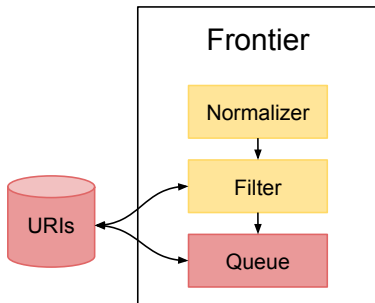
`http://ex.org/data/entity2`

`http://ex.com/ontology`

`http://dbpedia.org/resource/Rome`



`http://ex.org/data/entity1`
`http://ex.org/data/entity2`
`http://ex.com/ontology`



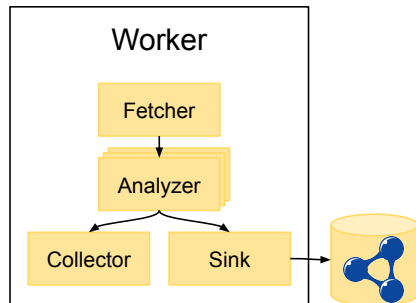
ex.org

<http://ex.org/data/entity2>

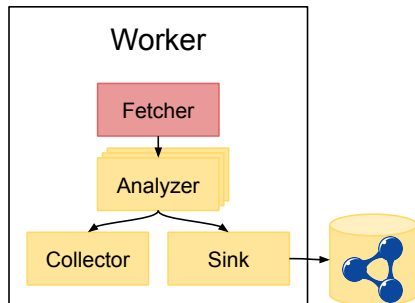
<http://ex.org/data/entity1>

ex.com

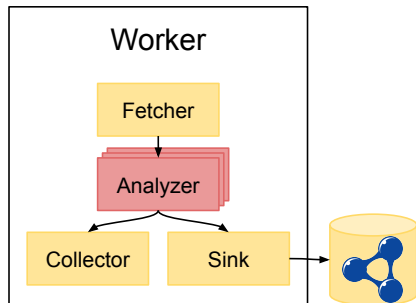
<http://ex.com/ontology>



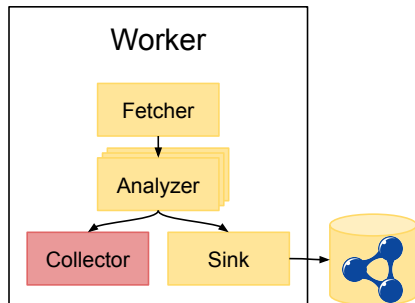
- HTTP
- FTP
- SPARQL (via HTTP)
- CKAN API



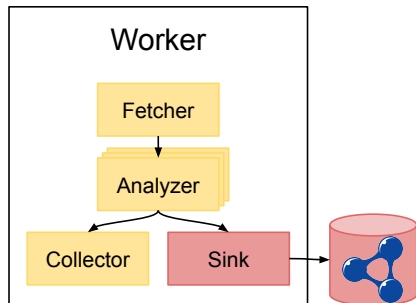
- Apache Jena
- Apache Any23
- Semargl
- HTML scraping
- ...

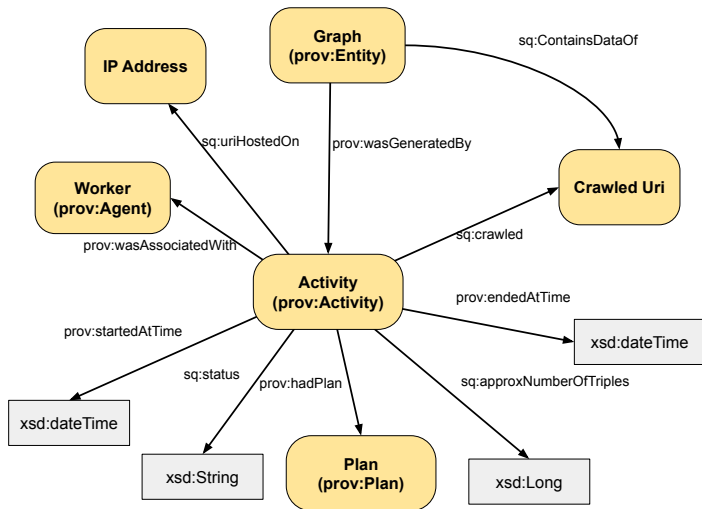


- Set of new URIs



- SPARQL
- Files



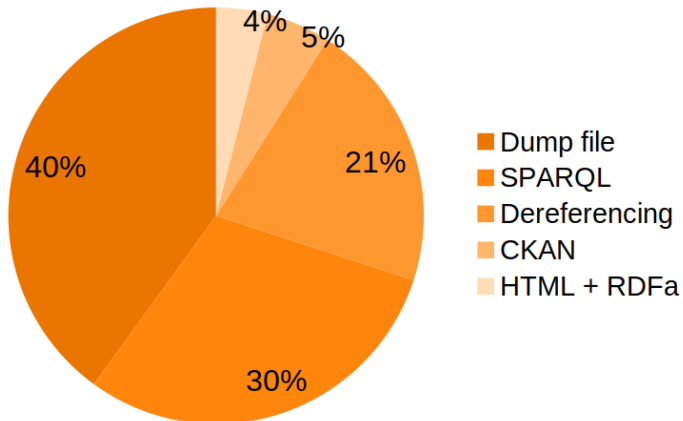


- Apache Nutch
 - Web crawler
 - RDF plugin
 - Plugin is not maintained anymore

- Apache Nutch
 - Web crawler
 - RDF plugin
 - Plugin is not maintained anymore
- LDSpider
 - Breadth first strategy
 - Load-balancing strategy (LBS)
 - 1, 8, 16 or 32 threads

	RDF Serialisations										Comp.					HTML					
	RDF/XML	RDF/JSON	Turtle	N-Triples	N-Quads	Notation 3	JSON-LD	TriG	TriX	HDT	ZIP	Gzip	bzip2	7zip	tar	RDFa	Microdata	Microformat	HTML (scraping)	SPARQL	CKAN
Apache Nutch	✓	–	✓	✓	–	✓	–	–	–	–	✓	✓	–	–	–	✓	✓	✓	–	–	–
LDSpider	✓	–	✓	✓	✓	✓	✓	–	–	–	–	–	–	–	–	✓	✓	✓	–	–	–
Squirrel	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

100 servers



Crawler	Micro Recall	Run time (in s)
LDSpider (T1)	0.31	1798
LDSpider (T8)	0.30	1792
LDSpider (T16)	0.31	1858
LDSpider (T32)	0.31	1847
LDSpider (T32,LBS)	0.03	66
Squirrel (W1)	0.98	6 663
Squirrel (W3)	0.98	2 686
Squirrel (W9)	0.98	1 412
Squirrel (W18)	0.97	1 551

- 200 servers
- Only dereferencing servers



Crawler	Micro Recall	Run time (in s)	RAM (in GB)
LDSpider (T1)	1.00	2 031	1.2
LDSpider (T8)	1.00	2 295	2.8
LDSpider (T16)	1.00	1 945	1.6
LDSpider (T32)	1.00	2 635	2.6
LDSpider (T32,LBS)	0.54	765	7.5
Squirrel (W1)	1.00	11 821	3.9
Squirrel (W3)	1.00	4 100	8.6
Squirrel (W9)	1.00	1 591	18.1
Squirrel (W18)	1.00	1 091	22.1

- Squirrel achieved a
 - Higher effectiveness
 - Lower run time

SUMMARY



- Squirrel achieved a
 - Higher effectiveness
 - Lower run time
- LDSpider showed a better RAM efficiency
 - Several JVMs for Squirrel vs. 1 for LDSpider
 - Part of future work in the Squirrel project

SUMMARY

Squirrel is a

- distributed
- scalable
- extensible
- open-source

crawler for the Web of Data.



<https://github.com/dice-group/Squirrel>

This work has been supported by the BMVI (Bundesministerium für Verkehr und digitale Infrastruktur) projects LIMBO (GA no. 19F2029C) and OPAL (GA no. 19F2028A).



Bundesministerium
für Verkehr und
digitale Infrastruktur

