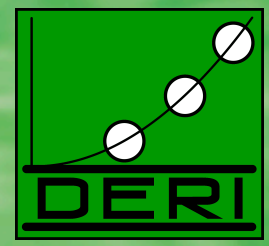


Scalable Authoritative OWL Reasoner

= *"free" (Irish)*

Aidan Hogan, Andreas Harth, **Axel Polleres**

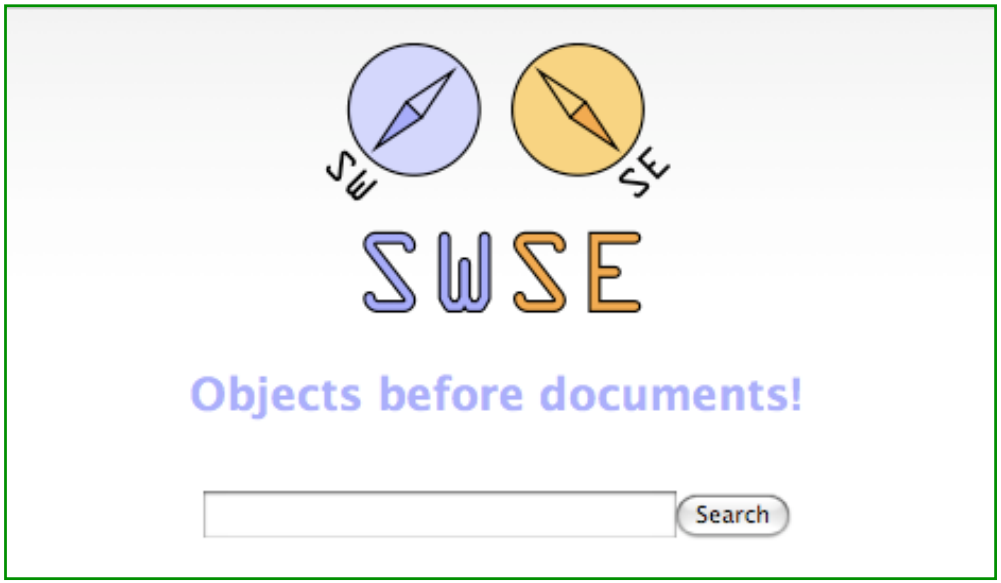
Digital Enterprise Research Institute
National University of Ireland, Galway

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

sfi
science foundation ireland
fondúireacht eolaíochta éireann

SAOR - Reasoning for SWSE

http://swse.deri.org/



- We want the challenge data **plus** OWL inferred data in the search results!
- Our approach:

  SAOR – **S**calable **A**uthoritative **O**WL **R**easoning

Idea

- Apply a subset of OWL reasoning to the billion triple challenge dataset
- Forward-chaining rule based approach, e.g.[ter Horst, 2005]

- Reduced output statements for the SWSE use case…
  - Must be *scalable*, must be *reasonable*
- … incomplete w.r.t. OWL **BY DESIGN!**
  - **SCALABLE:** Tailored ruleset
    - file-scan processing
    - avoid joins
  - **AUTHORITATIVE:** Avoid Non-Authoritative inference ("hijacking", "non-standard vocabulary use")

Scalable Reasoning

- **Scan 1:**

  Scan all data (1.1b statements), separate T-Box statements, load T-Box statements (8.5m) into memory, perform authoritative analysis.

- **Scan 2:**

  Scan all data and join all statements with in-memory T-Box .

  - Only works for inference rules with 0-1 A-Box patterns
  - No T-Box expansion by inference

  → Needs “tailored” ruleset

Rules Applied:
Tailored version of [ter Horst, 2005]

#	DL Syntax	Rule	# Inferred
		$\mathcal{G}0$: NO A-BOX PATTERNS IN ANTECEDENT	
00	$\{o_i....o_n\}$	?C :oneOf (?o$_1$ ... ?o$_n$) . ⇒ ?o$_1$ ... ?o$_n$ a ?C .	35,161
		$\mathcal{G}1$: ONE A-BOX PATTERN IN ANTECEDENT	
01	$C \sqsubseteq D$	?C rdfs:subClassOf ?D . ?s a ?C . ⇒ ?s a ?D .	1,124,758,631
02$_a$	$C \equiv D$	?C :equivalentClass ?D . ?s a ?C . ⇒ ?s a ?D .	8,137,162
02$_b$		?C :equivalentClass ?D . ?s a ?D . ⇒ ?s a ?C .	90,372
03	$P \sqsubseteq Q$	?P rdfs:subPropertyOf ?Q . ?s ?P ?o . ⇒ ?s ?Q ?o .	156,462,399
04$_a$	$P \equiv Q$	?P :equivalentProperty ?Q . ?s ?P ?o . ⇒ ?s ?Q ?o .	5,667,464
04$_b$		?P :equivalentProperty ?Q . ?s ?Q ?o . ⇒ ?s ?P ?o .	6,642
05$_a$	$P \equiv P_0^-$	?P :inverseOf ?Q . ?s ?P ?o . ⇒ ?o ?Q ?s .	230,945,040
05$_b$		?P :inverseOf ?Q . ?s ?Q ?o . ⇒ ?o ?P ?s .	230,941,648
06	$\top \sqsubseteq \forall P^-.C$	?P rdfs:domain ?C . ?s ?P ?o . ⇒ ?s a ?C .	588,530,865
07	$\top \sqsubseteq \forall P.C$	?P rdfs:range ?C . ?s ?P ?o . ⇒ ?o a ?C .	528,995,909
08	$P \equiv P^-$	?P a :SymmetricProperty . ?s ?P ?o . ⇒ ?o ?P ?s .	560,460
09$_a$	$\exists P.x$	?C :hasValue ?x; :onProperty ?P . ?y ?P ?x . ⇒ ?y a ?C .	98,601
09$_b$		?C :hasValue ?x; :onProperty ?P . ?y a ?C . ⇒ ?y ?P ?x .	104,780
10	$C_1 \sqcup ... \sqcup C_n$	?C :unionOf (?C$_1$...?C$_i$...?C$_n$) . ?x a ?C$_i$. ⇒ ?x a ?C .	81,736,234
11	$(\geq 1P)$	?C :minCardinality 1; :onProperty ?P . ?x ?P ?y . ⇒ ?x a ?C .	65,283,322
12$_a$	$C_1 \sqcap ... \sqcap C_n$	?C :intersectionOf (?C$_1$ ... ?C$_n$) . ?y a ?C . ⇒ ?y a ?C$_1$, ..., ?C$_n$.	115,383
12$_b$	$C_1 \sqcap ... \sqcap C_n$	?C :intersectionOf (?C$_1$) . ?y a ?C$_1$. ⇒ ?y a ?C .	42

Good "excuses" to avoid G2 rules

- The obvious:
  - *G2* rules would need joins, i.e. to trigger restart of file-scan
- The interesting one:
  - Take for instance IFP rule:

$\top \sqsubseteq \forall \leq 1 P^{-}$ **?P** a :InverseFunctionalProperty . ?x ?P ?o . ?y ?P ?o . ⇒ ?x :sameAs ?y .

  - Maybe not such a good idea on real Web data

  - More experiments including *G2*, *G3* rules in [Hogan, Harth, Polleres, ASWC2008]

Authoritative Reasoning

- Document **D** authoritative for concept **C** iff:
  - **C** not identified by URI
    - OR
  - De-referenced URI of **C** coincides with or redirects to **D**
  - FOAF spec authoritative for `foaf:Person` ✓
  - MY spec not authoritative for `foaf:Person` ✘

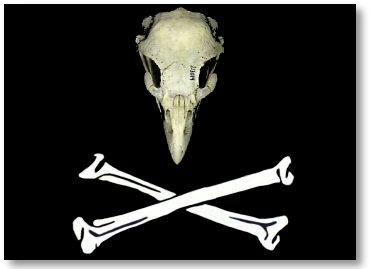
- Only allow extension in authoritative documents
  - `my:Person rdfs:subClassOf foaf:Person .` (MY spec) ✓
- BUT: Reduce obscure memberships
  - `foaf:Person rdfs:subClassOf my:Person .` (MY spec) ✘
- Similarly for other T-Box statements.
- In-memory T-Box stores authoritative values for rule execution

Rules Applied

#	DL Syntax	Rule	# Inferred
		$\mathcal{G}0$: NO A-BOX PATTERNS IN ANTECEDENT	
00	$\{o_i....o_n\}$	?C :oneOf (?o$_1$ ... ?o$_n$) . ⇒ ?o$_1$ ... ?o$_n$ a ?C .	35,161
		$\mathcal{G}1$: ONE A-BOX PATTERN IN ANTECEDENT	
01	$C \sqsubseteq D$	?C rdfs:subClassOf ?D . ?s a ?C . ⇒ ?s a ?D .	1,124,758,631
02$_a$	$C \equiv D$	?C :equivalentClass ?D . ?s a ?C . ⇒ ?s a ?D .	8,137,162
02$_b$		?C :equivalentClass ?D . ?s a ?D . ⇒ ?s a ?C .	90,372
03	$P \sqsubseteq Q$	?P rdfs:subPropertyOf ?Q . ?s ?P ?o . ⇒ ?s ?Q ?o .	156,462,399
04$_a$	$P \equiv Q$	?P :equivalentProperty ?Q . ?s ?P ?o . ⇒ ?s ?Q ?o .	5,667,464
04$_b$		?P :equivalentProperty ?Q . ?s ?Q ?o . ⇒ ?s ?P ?o .	6,642
05$_a$	$P \equiv P_0^-$	?P :inverseOf ?Q . ?s ?P ?o . ⇒ ?o ?Q ?s .	230,945,040
05$_b$		?P :inverseOf ?Q . ?s ?Q ?o . ⇒ ?o ?P ?s .	230,941,648
06	$\top \sqsubseteq \forall P^-.C$	?P rdfs:domain ?C . ?s ?P ?o . ⇒ ?s a ?C .	588,530,865
07	$\top \sqsubseteq \forall P.C$	?P rdfs:range ?C . ?s ?P ?o . ⇒ ?o a ?C .	528,995,909
08	$P \equiv P^-$	?P a :SymmetricProperty . ?s ?P ?o . ⇒ ?o ?P ?s .	560,460
09$_a$	$\exists P.x$	?C :hasValue ?x; :onProperty ?P . ?y ?P ?x . ⇒ ?y a ?C .	98,601
09$_b$		?C :hasValue ?x; :onProperty ?P . ?y a ?C . ⇒ ?y ?P ?x .	104,780
10	$C_1 \sqcup ... \sqcup C_n$	?C :unionOf (?C$_1$...?C$_i$...?C$_n$) . ?x a ?C$_i$. ⇒ ?x a ?C .	81,736,234
11	$(\geq 1P)$	?C :minCardinality 1; :onProperty ?P . ?x ?P ?y . ⇒ ?x a ?C .	65,283,322
12$_a$	$C_1 \sqcap ... \sqcap C_n$	?C :intersectionOf (?C$_1$ ... ?C$_n$) . ?y a ?C . ⇒ ?y a ?C$_1$, ..., ?C$_n$.	115,383
12$_b$	$C_1 \sqcap ... \sqcap C_n$	?C :intersectionOf (?C$_1$) . ?y a ?C$_1$. ⇒ ?y a ?C .	42

*The 17 rules applied including statements considered to be T-Box, elements which must be **authoritatively** spoken for (including for bnode OWL abstract syntax), and output count*

Authoritative Resoning covers rdfs: owl: vocabulary misuse



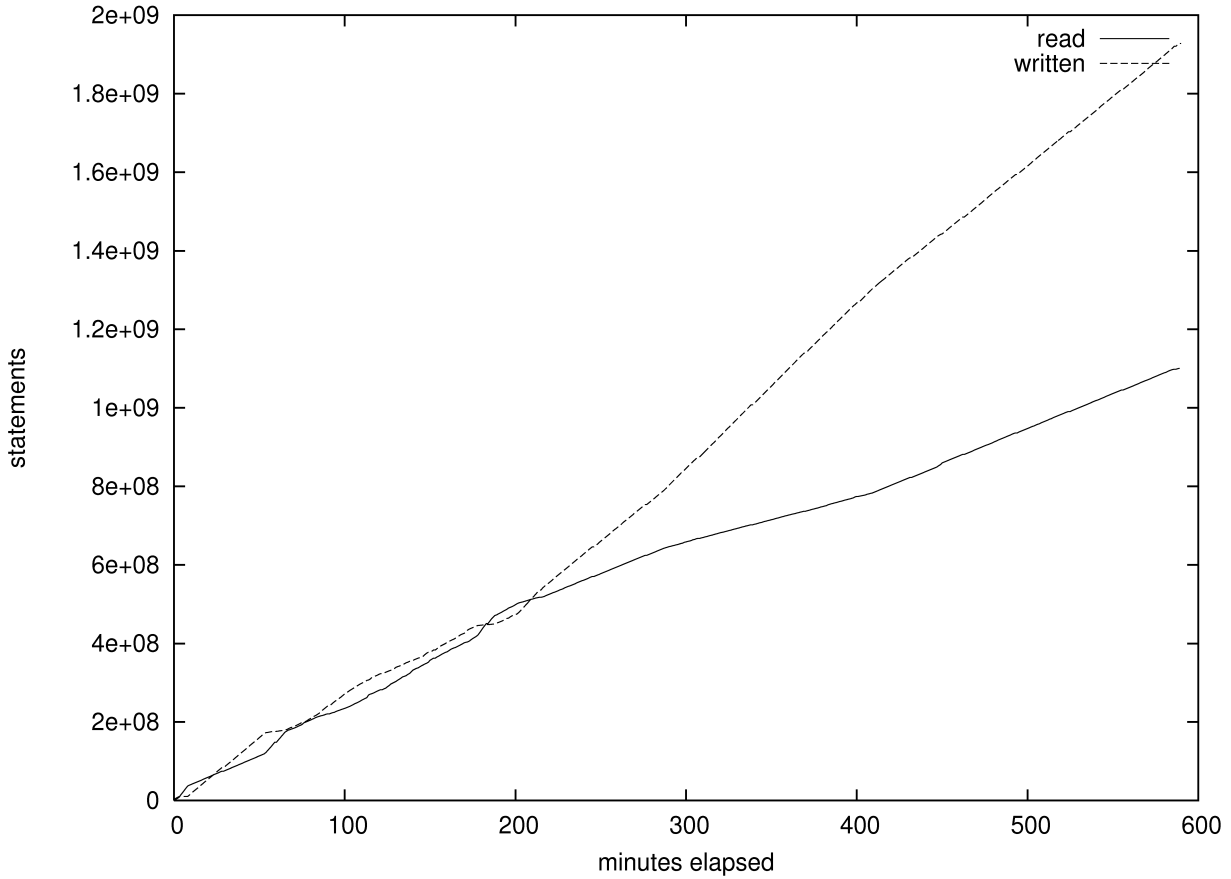
:rdfs :owl Hijacking

- http://www.polleres.net/nasty.rdf:

```
rdfs:subClassOf rdfs:subPropertyOf rdfs:Resource.
rdfs:subClassOf rdfs:subPropertyOf rdfs:subPropertyOf.
rdf:type rdfs:subPropertyOf rdfs:subClassOf.
rdfs:subClassOf rdf:type owl:SymmetricProperty.
```

- Naïve rules application would infer $O(n^3)$ triples
- By use of authoritative reasoning SAOR/SWSE doesn't stumble over these ☺

Performance



Graph showing SAOR's rate of input/output statements per minute for reasoning on 1.1b statements: reduced input rate correlates with increased output rate and vice-versa

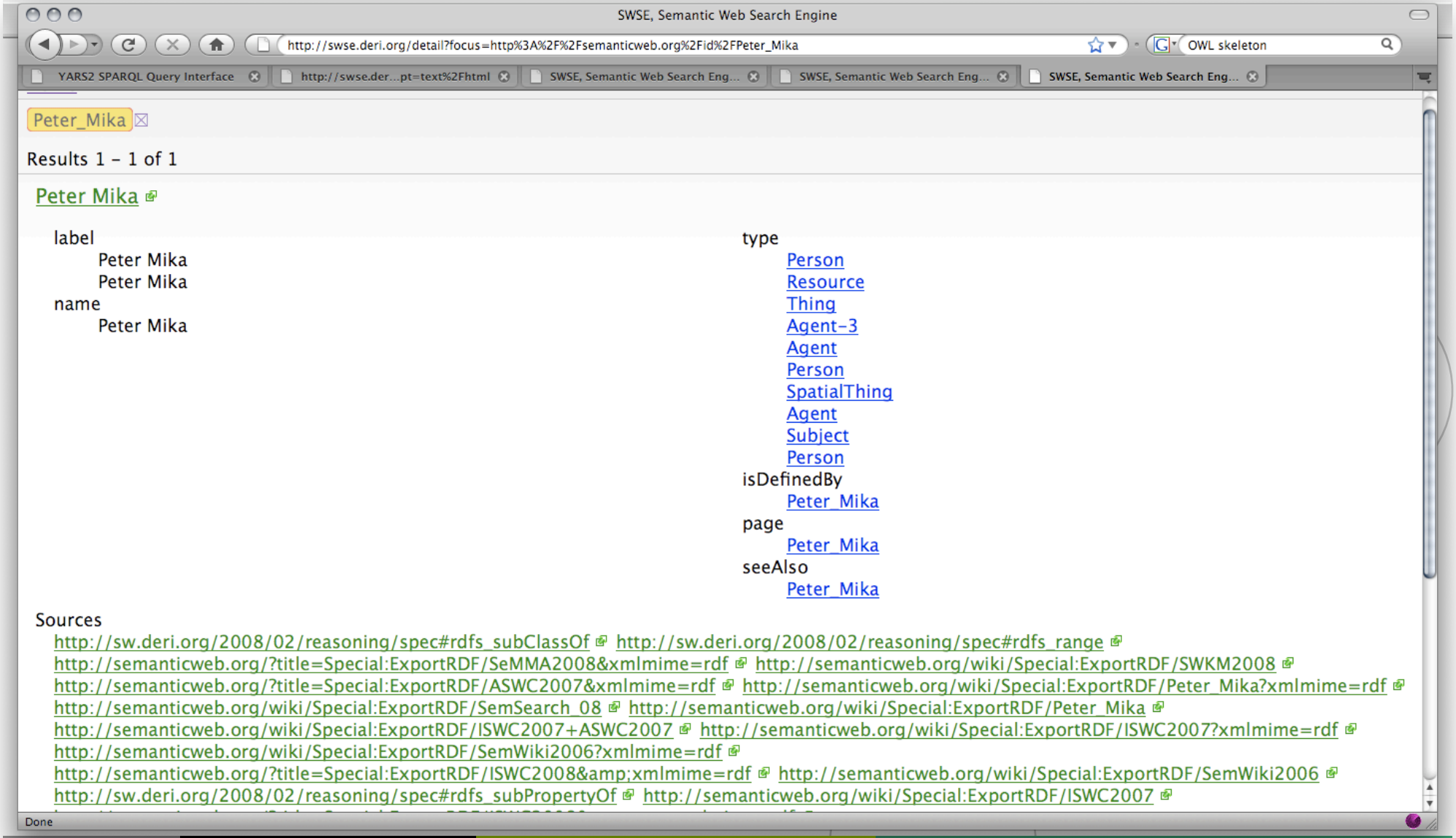
Results

- **SCAN 1: 6.47 hrs**
  - In-mem T-Box creation, authoritative analysis:
- **SCAN 2: 9.82 hrs**
  - Scan reasoning – join A-Box with in-mem authoritative T-Box:
- **1.925b new statements inferred in 16.29 hrs**

1.1b + 1.9b inferred = 3 billion triples in SWSE

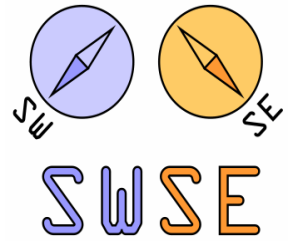
- **On our agenda:**
  - **More valuable insights on our experiences from Web data**
  - **G2 and G3 rules?**
  - **Detailed comparison to OWL RL**

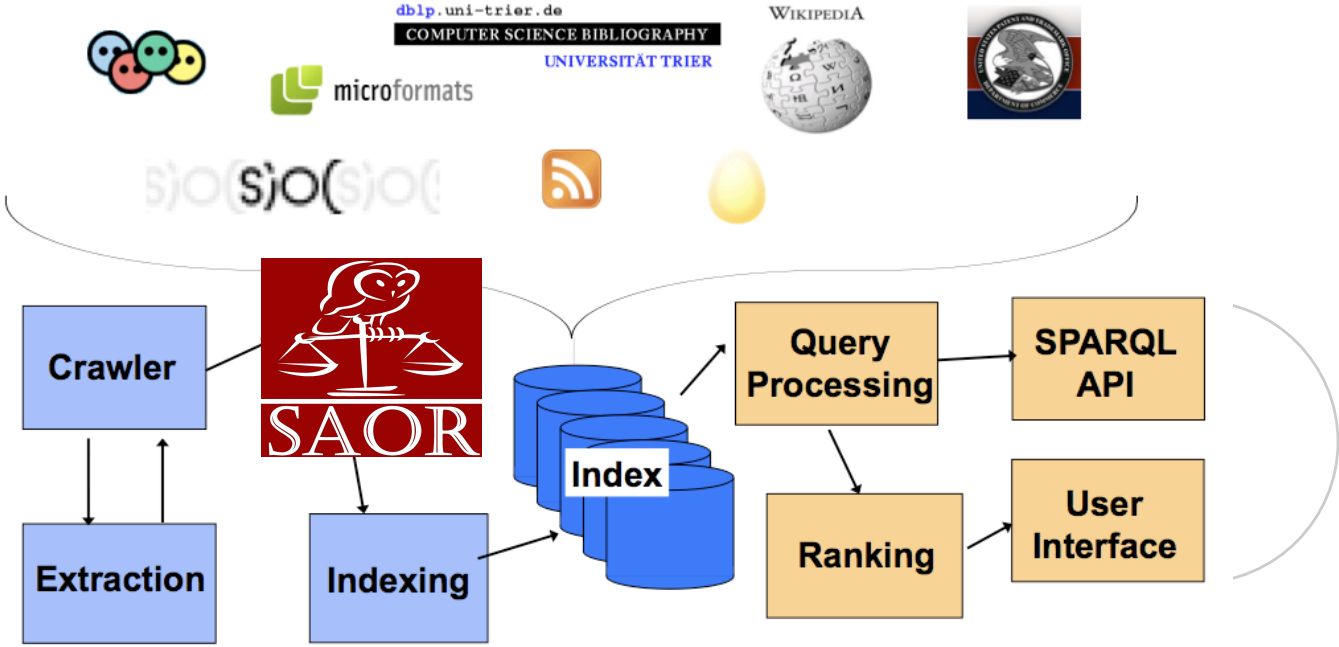
Search result example:



Le Fin…

Enjoy the data…





GUI: http://swse.deri.org/

SPARQL interface: http://swse.deri.org/yars2/

Contact us for feedback!