Sindice.com: A Semantic Web Search Engine

Giovanni Tummarello Renaud Delbru
Eyal Oren **Richard Cyganiak**

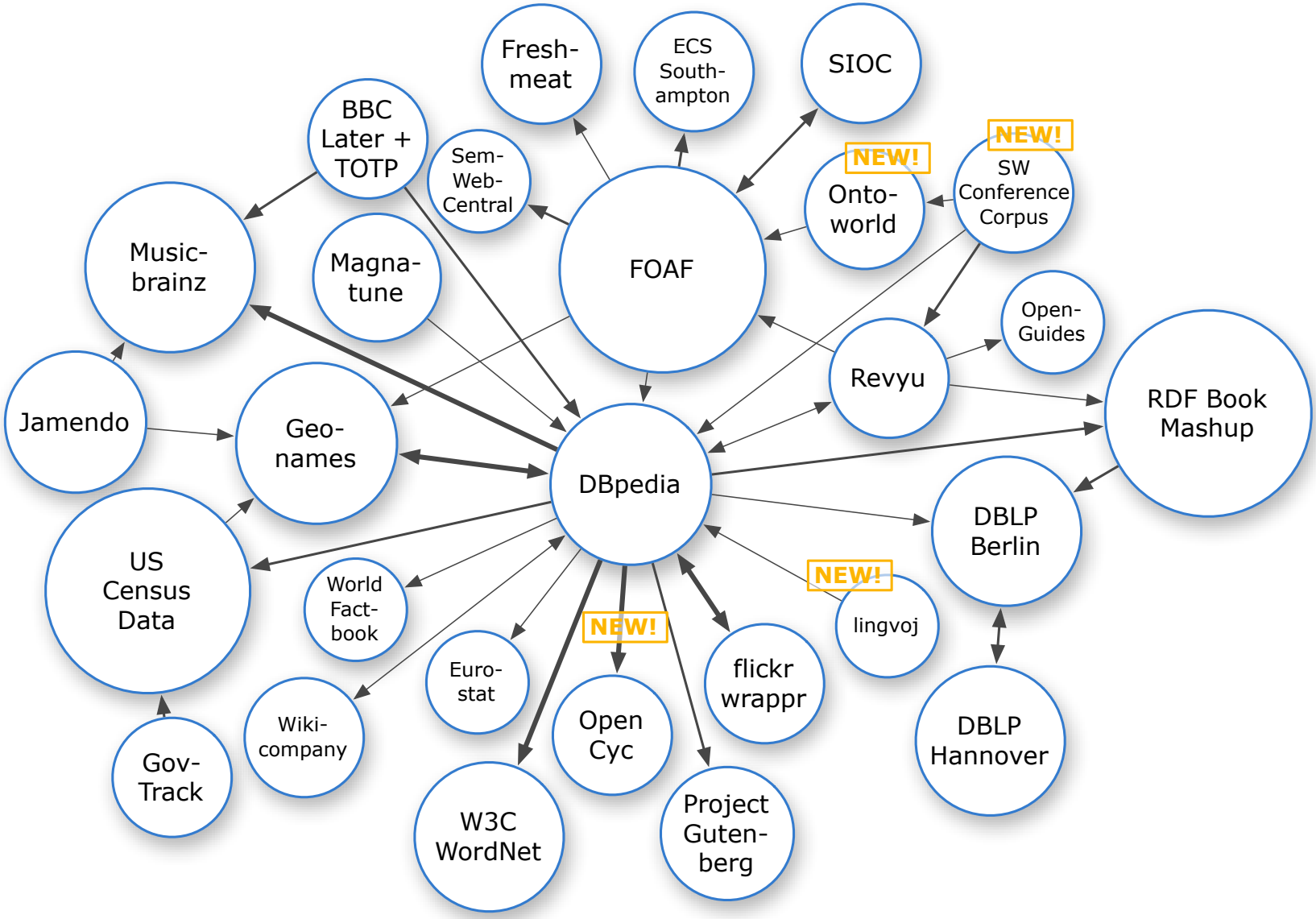
Digital Enterprise Research Institute
National University of Ireland, Galway

November 23, 2007

The Semantic Web is a reality

- Many Gigs of RDF dumps
- 30+ public SPARQL endpoints
- Linked Data, 5+ different browsers
- RDFa

The Semantic Web is a reality



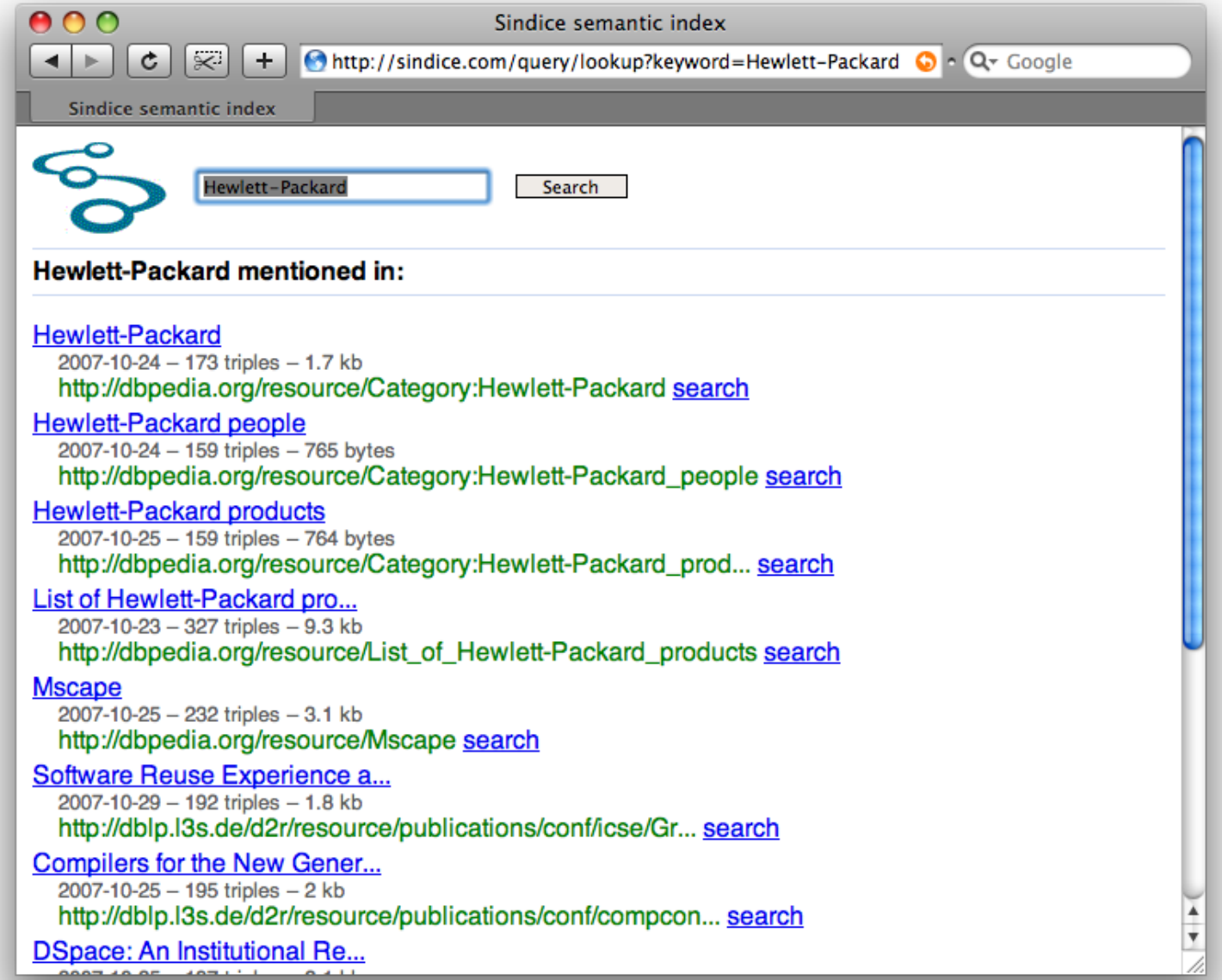
The Semantic Web is a reality

We don't worry about running out of data

Sindice.com


Sindice semantic index
http://sindice.com/query/keyword
Google
Sindice semantic index
sindice
keyword
uri
ifp
Hewlett-Packard
Search
Submit your RDF
About Sindice
Sindice Blog
© 2007 Digital Enterprise Research Institute
(β version: indexing around 21.9 million RDF documents)

Sindice.com


Sindice semantic index
http://sindice.com/query/lookup?keyword=Hewlett-Packard
Google
Sindice semantic index
Hewlett-Packard
Search
Hewlett-Packard mentioned in:
Hewlett-Packard
2007-10-24 – 173 triples – 1.7 kb
http://dbpedia.org/resource/Category:Hewlett-Packard search
Hewlett-Packard people
2007-10-24 – 159 triples – 765 bytes
http://dbpedia.org/resource/Category:Hewlett-Packard_people search
Hewlett-Packard products
2007-10-25 – 159 triples – 764 bytes
http://dbpedia.org/resource/Category:Hewlett-Packard_prod... search
List of Hewlett-Packard pro...
2007-10-23 – 327 triples – 9.3 kb
http://dbpedia.org/resource/List_of_Hewlett-Packard_products search
Mscape
2007-10-25 – 232 triples – 3.1 kb
http://dbpedia.org/resource/Mscape search
Software Reuse Experience a...
2007-10-29 – 192 triples – 1.8 kb
http://dblp.l3s.de/d2r/resource/publications/conf/icse/Gr... search
Compilers for the New Gener...
2007-10-25 – 195 triples – 2 kb
http://dblp.l3s.de/d2r/resource/publications/conf/compcon... search
DSpace: An Institutional Re...

Sindice API

- `http://sindice.com/query/lookup?uri=...`
- `http://sindice.com/query/lookup?keyword=...`
- `http://sindice.com/query/lookup?property=...&object=...`

- Ask for HTML, plain text, RDF/XML or JSON via content negotiation

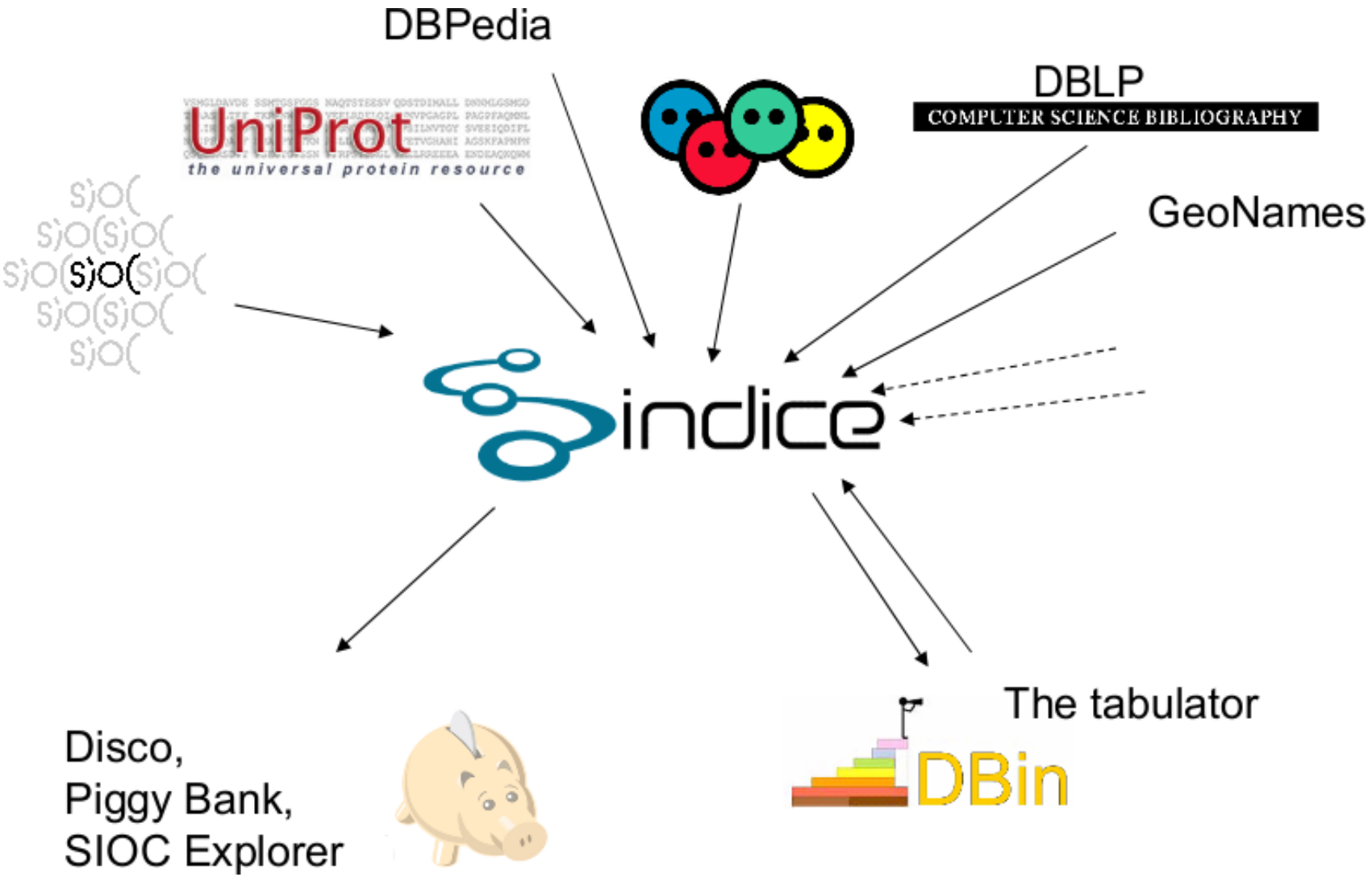
Scenario (1)

- Tom surfs to `http://dbpedia.org/resource/Busan`
- Tom wants more than just DBpedia's information
- Tom's Tabulator has a Sindice plugin
- Tom presses 'lookup on Sindice'
- Tom gets a top-ten list of Busan sources
- Tom selects his two trustworthy sources
- Tom's Tabulator downloads this data
- Tom continues his happy data-surfing

Scenario (2)

- Tom goes eating in Busan
- Tom likes the food and reviews the restaurant
- Tom's review site pings Sindice with the update
- Within an hour, others can find this info
- Tom continues his happy fish-eating

Sindice: discover Semantic Web resources



Indexing approach

- IR viewpoint: SW is bunch of documents
- DB viewpoint: SW is bunch of triples
- We take IR viewpoint: we index all identifiers and provide simple lookups; no RDF queries

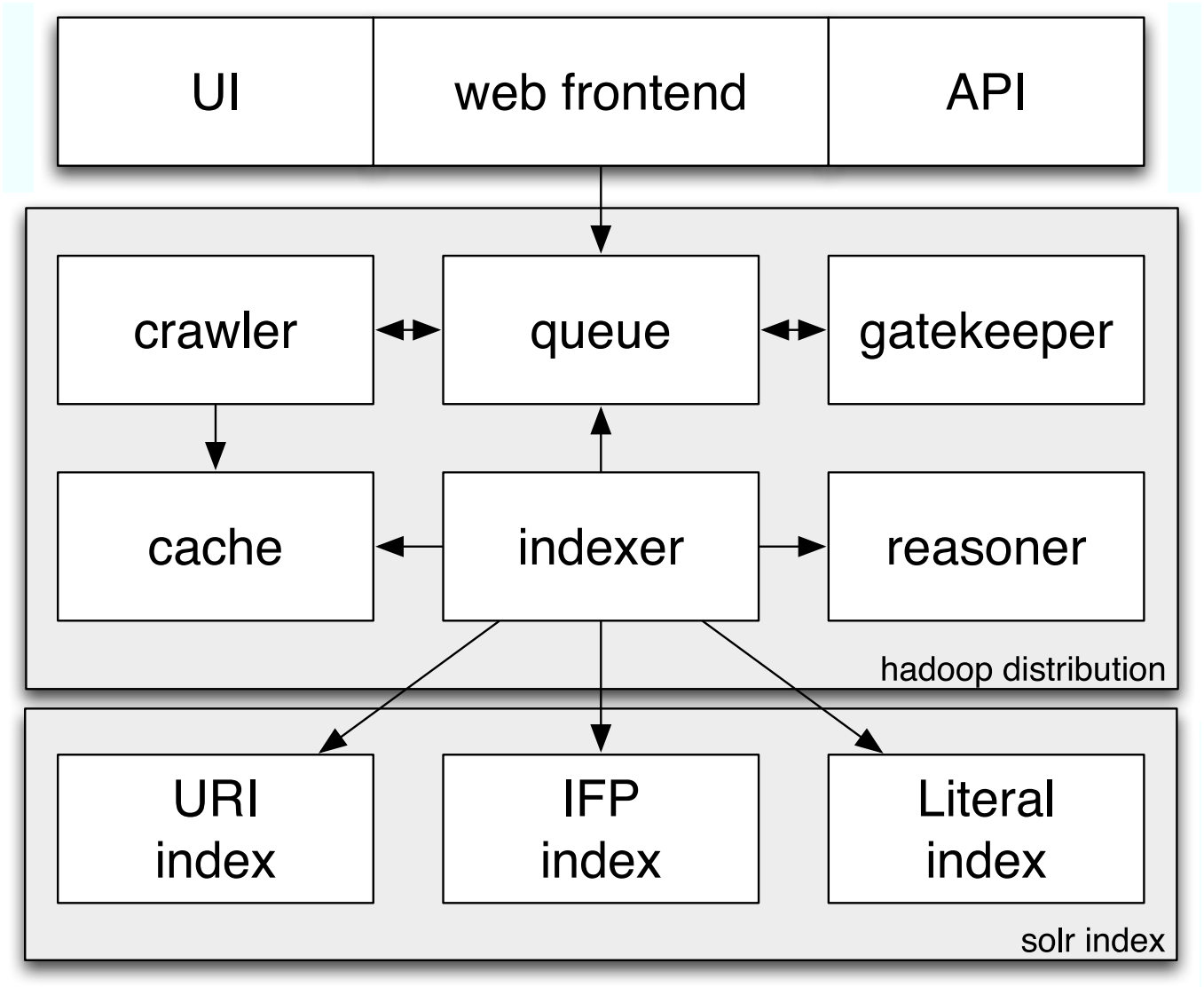
- Clients can browse/download/display RDF data themselves; we tell them where to find it

Sindice functionality (operators)

- $index : url \longrightarrow \emptyset$
- $lookup : uri \longrightarrow \{url\}$
- $lookup : text \longrightarrow \{url\}$
- $lookup : ifp \times value \longrightarrow \{url\}$

Natural data structure: inverted index over documents

Sindice architecture



Index lookup

- Index retrieval
- Ranking phase
- Result generation

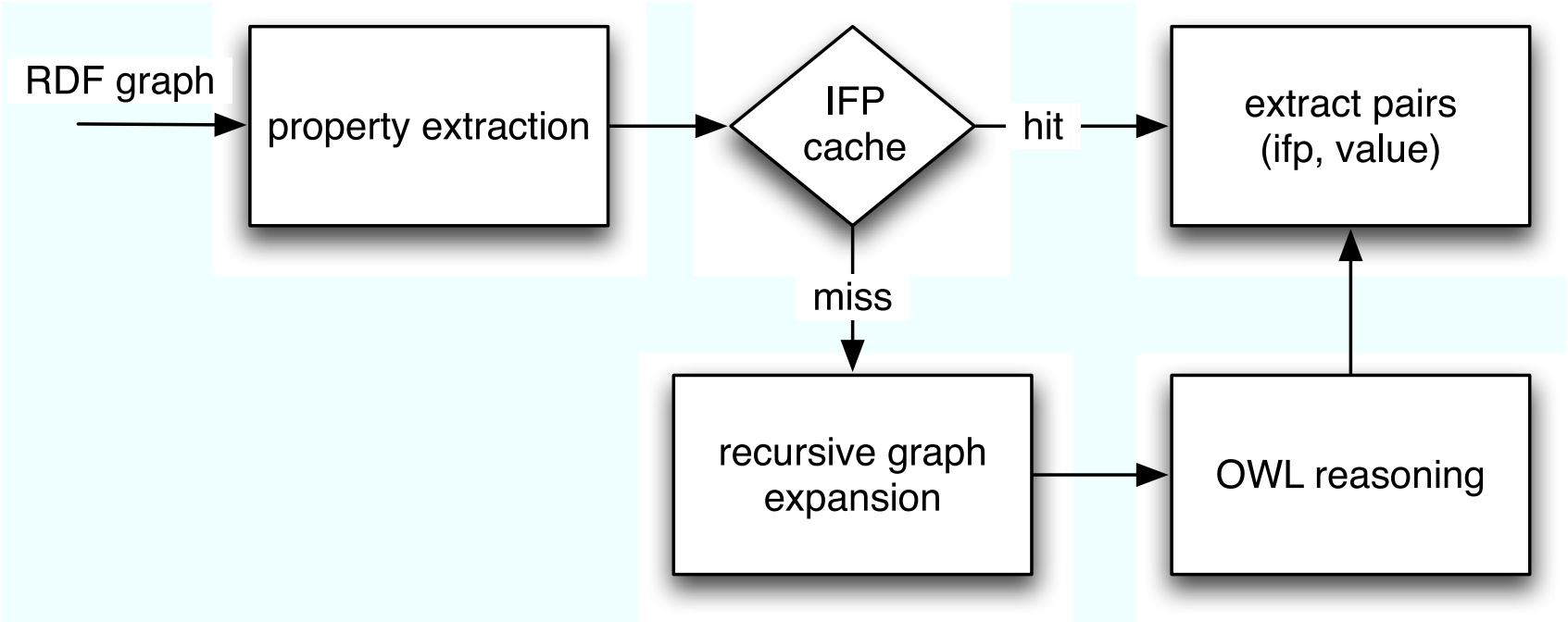
Graph processing

1. Fetch RDF data
2. Extract and index full-text literals
3. Extract and index mentioned URIs
4. Extract graph metadata (size and length)
5. Graph expansion and inferencing
6. Extract labels
7. Extract and index mentioned IFP pairs

Graph processing

1. Fetch RDF data
2. Extract and index full-text literals
3. Extract and index mentioned URIs
4. Extract graph metadata (size and length)
5. Graph expansion and inferencing
6. Extract labels
7. Extract and index mentioned IFP pairs

IFP processing



Graph processing: IFP extraction

- OWL reasoning needed to find IFPs, but computationally expensive
- Desireable: reasoning cache to reuse computation
- Undesireable: global trust in all statements

Solution: quarantained reasoning cache

- Recursively fetch all mentioned schemas
- Compute closure of schemas union
- Query and store all properties that are an IFP
- {foaf:name, dc:title, foaf:homepage, foaf:mbox} $\rightarrow$ {foaf:mbox}
- For any document that uses same properties you know the set of possible IFPs

Sindice components

- Hadoop (parallel processing)
- HTable (document cache)
- Solr (document index)
- Sesame & OWLIM (reasoning)
- Ruby on Rails (frontend)
- `pingthesemanticweb.com`

Tool for data providers: Semantic Sitemap

- Sitemap protocol exposes "deep web" to crawlers
- Semantic sitemap adds Semantic Web data
- `http://sw.deri.org/2007/07/sitemapextension/`
- Used by: Geonames, DBLP, Uniprot, DBpedia, data.semanticweb.org

Semantic Sitemap: example

```
<urlset xmlns=http://www.sitemaps.org/schemas/sitemap/0.9
        xmlns:sc=http://sw.deri.org/2007/07/sitemapextension/scschema.xsd>
  <sc:dataset>

    <sc:datasetLabel>Product Catalog for Example.com</sc:datasetLabel>

    <sc:linkedDataPrefix>http://example.com/products/</sc:linkedDataPrefix>
    <sc:sparqlEndpoint>http://example.com/sparql</sc:sparqlEndpoint>
    <sc:dataDumpLocation>http://example.com/all.rdf</sc:dataDumpLocation>

    <changefreq>weekly</changefreq>

  </sc:dataset>
</urlset>
```

What about other search engines?

- We do not answer queries but refer to data sources
- We have IFP lookup using OWL reasoning
- We have semantic sitemap for data-dumps
- We support linked data (input and output)
- We have fully open client APIs
- We have Hadoop infrastructure
- We have live, continuous, updates
- Simplicity, efficiency, scalability

Credits

- Giovanni Tummarello
- Eyal Oren
- Michele Catasta
- Renaud Delbru
- Holger Stenzhorn
- Adam Westerski
- OpenLink Software

Upcoming as we speak . . .

- Validator API
- Trust assessment API
- SW Pipes and widgets platform
- Entity-based API (Okkam)
- Growing hardware cluster (possibly 100 nodes)

Summary

- Sindice: lookup service for Semantic Web resources
- Lookup: resource by URIs, IFPs, keyword
- Architecture: Based on Hadoop, Solr and OWLIM
- Data: DBLP, DBpedia, Uniprot, Geonames and more
- 20M+ documents, 80M+ URIs, 4M+ IFPs, 2B+ triples