

Sindice.com: Weaving the open linked data

Giovanni Tummarello Renaud Delbru **Eyal Oren**

Digital Enterprise Research Institute
National University of Ireland, Galway

November 14, 2007

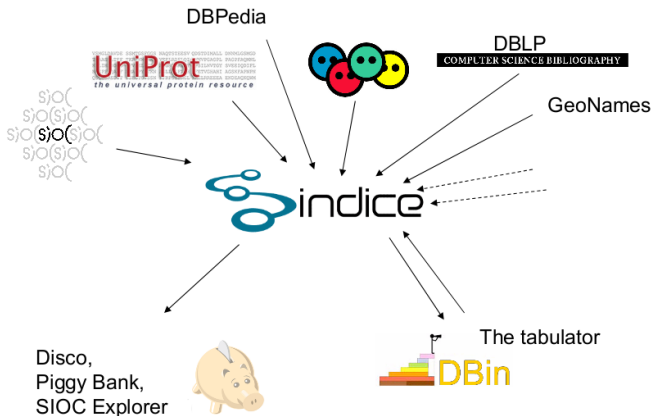
Scenario (1)

- ▶ Tom surfs to `http://dbpedia.org/resource/Pusan`
- ▶ 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 Pusan 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 Pusan
- ▶ 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



Sindice: discover Semantic Web resources

- ▶ A lookup index over Semantic Web resources
- ▶ Retrieval through URIs, IFPs and keyword search
- ▶ Focus on machine clients: API, linked data
- ▶ Only resource lookups: no (join) queries
- ▶ 11m documents, 46m URIs, 2.7m IFPs, 2.1b triples

Problem description: discovering data sources

- ▶ Clients can browse/download/display RDF data
- ▶ Challenge: **where to find data sources**
- ▶ Problem: linked data finds only authoritative data
- ▶ Challenge: **identify common resources**
- ▶ Problem: URI reuse is low, IFP reasoning expensive

Requires: lightweight lookup index

Indexing approach

- ▶ IR viewpoint: SW is bunch of documents (inverted indices, term search)
- ▶ DB viewpoint: SW is bunch of triples (compound indices, conjunctive queries)
- ▶ We take IR viewpoint: we index all identifiers and provide simple lookups

Clients must process RDF documents themselves

Sindice overview



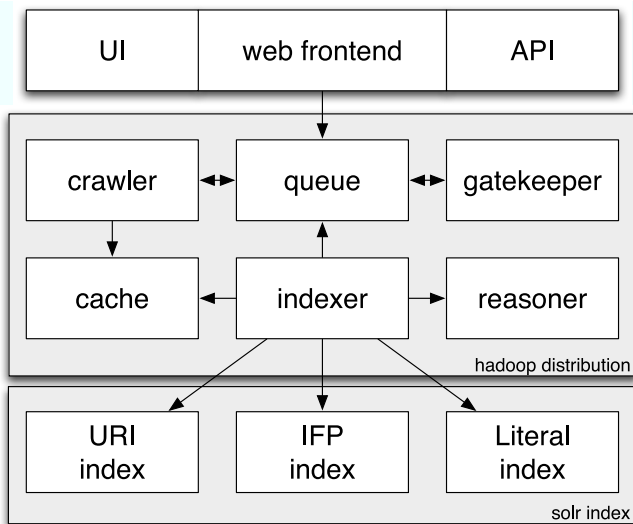
The screenshot shows a web browser window titled "Sindice semantic index". The address bar contains the URL "http://sindice.com/query/keyword". The main content area features the Sindice logo, which consists of a stylized blue 'S' made of three overlapping circles followed by the word "indice" in a black, lowercase, sans-serif font. Below the logo is a search interface with three tabs: "keyword" (highlighted in dark teal), "uri", and "ifp". Under the "keyword" tab is a text input field and a "Search" button. Below the search area are three links: "Submit your RDF", "About Sindice", and "Sindice Blog". At the bottom, there is a copyright notice: "© 2007 Digital Enterprise Research Institute (β version: indexing around 11.0 million RDF documents) (around 46.2 million URIs, around 2.1 billion triples)".

Sindice functionality (operators)

- ▶ $index : url \rightarrow \emptyset$
- ▶ $lookup : uri \rightarrow \{url\}$
- ▶ $lookup : ifp \times value \rightarrow \{url\}$
- ▶ $lookup : text \rightarrow \{url\}$
- ▶ Available through Web interface and REST API

Natural data structure: inverted index over documents

Sindice architecture



Sindice components

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

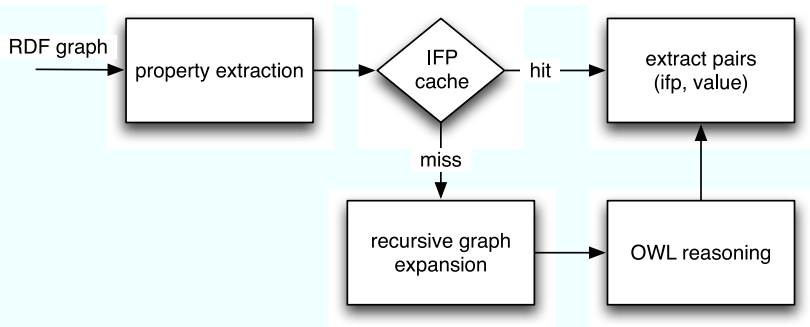
Graph processing

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
- ▶ Desirable: reasoning cache to reuse computation
- ▶ Undesirable: 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
- ▶ $\{\text{foaf:name, dc:title, foaf:homepage, foaf:mbox}\} \rightarrow \{\text{foaf:mbox}\}$
- ▶ For any document that uses same properties you know the set of possible IFPs

Tools for data providers

- ▶ You want **all** your data to be found
- ▶ You want your data to be indexed **correctly**
- ▶ HTML: Sitemap to describe your website
- ▶ HTML: Validator and link checker

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:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
  http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd"
  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.org
    </sc:datasetLabel>

    <sc:dataDumpLocation>
      http://example.org/cataloguedump.rdf
    </sc:dataDumpLocation>

    <sc:linkedDataPrefix>
      http://example.org/products/
    </sc:linkedDataPrefix>
    <changefreq>monthly</changefreq>
  </sc:dataset>
</urlset>
```

Linked data validator



Validated <http://www.eyaloren.org/foaf.rdf>

Found 7 errors:

- URI processing failed for "http://purl.org/rss/1.0/modules/syndication/updateFrequency".
Cannot retrieve data from http://web.resource.org/rss/1.0/modules/syndication/updateFrequency. Received 404 response code.
- URI processing failed for "http://www.w3.org/2001/XMLSchema#nonNegativeInteger".
The document located at "http://www.w3.org/2001/XMLSchema" does not contain an RDF graph.
- URI processing failed for "http://www.w3.org/2001/XMLSchema#string".
The document located at "http://www.w3.org/2001/XMLSchema" does not contain an RDF graph.
- URI processing failed for "http://xmins.com/wordnet/1.6/Agent".
Cannot retrieve data from http://xmins.com/wordnet/1.6/Agent. Received 404 response code.
- URI processing failed for "http://xmins.com/wordnet/1.6/Document".
Cannot retrieve data from http://xmins.com/wordnet/1.6/Document. Received 404 response code.
- URI processing failed for "http://xmins.com/wordnet/1.6/Organization".
Cannot retrieve data from http://xmins.com/wordnet/1.6/Organization. Received 404 response code.
- URI processing failed for "http://xmins.com/wordnet/1.6/Person".
Cannot retrieve data from http://xmins.com/wordnet/1.6/Person. Received 404 response code.

Found 7 inverse functional properties:

- http://xmins.com/foaf/0.1/isPrimaryTopicOf - http://www.eyaloren.org/
- http://xmins.com/foaf/0.1/isPrimaryTopicOf - http://www.activerdf.org/
- http://xmins.com/foaf/0.1/isPrimaryTopicOf - http://www.browserdf.org/
- http://xmins.com/foaf/0.1/homepage - http://www.browserdf.org/
- http://xmins.com/foaf/0.1/homepage - http://www.activerdf.org/
- http://xmins.com/foaf/0.1/homepage - http://www.eyaloren.org/
- http://xmins.com/foaf/0.1/mbox - mailto:eyal.oren@deri.org

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: all ideas and plans
- ▶ Michele Catasta: architecture and parallelisation
- ▶ Renaud Delbru: indexing, reasoning and validator
- ▶ Holger Stenzhorn, Adam Westerski, Richard Cyganiak
- ▶ Openlink: technical support
- ▶ SFI: financial support

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
- ▶ 11m documents, 46m URIs, 2.7m IFPs, 2.1b triples