Linked Data in the Clouds: a Sindice.com perspective

Giovanni Tummarello, FBK - DERI
Some definitions

- **Linked Open Data: Official Definition**

  - The data that is available on the web published according to the “Linked Data Principles”
    1. Entities must be given resolvable names (URLs/URIs) minted in a namespace that belongs to the entity publishing the data.
    2. Resolving such URLs/URIs returns a “useful” resource description, possibly URLs/URIs of other related resources.
    3. Among these descriptions, there should be references to URIs/URLs of others datasets, notably “SameAs” links when the entities are known to be described somewhere else in the LOD cloud.
Web Data: a underestimate

- LOD Cloud: approx 300 datasets
  - Relatively clean, with some mutual integration
  - Not very updated
- 0.1..% of the databases on the "Web of Data"
  - 4000000+ sites with some form of data
  - 200000+ datasets (sites with semantic content on many pages)
  - Very updated!
  - ... Likely 10x this actually. Lots of data! And very important sources
- #SEO ppl agree: marking up your data is a must
Sindice challenges

Questions

- How to find sources?
- How to clean and transform them for one’s specific needs?
- Very high performance querying (production systems)?
Sindice.com: Portal for Sindice Services

**LaTeX Code for the Page:**

```
\documentclass{article}
\usepackage{graphicx}
\usepackage{hyperref}

\begin{document}
\section*{Sindice.com: Portal for Sindice Services}

\begin{center}
\includegraphics[width=\textwidth]{sindice_portal.png}
\end{center}

\section*{Sindice - Data Web Services}

Billions of pieces of reusable information can already be found across
hundreds of millions of web pages which embed RDF and Microformats. Start
consuming this data today with Sindice Data Web services.

\subsection*{Latest Data}

- 11:03:13 (sigma) 11\ triples: http://semantic.eurobau.com...09466505p525.rdf
- 11:03:13 (sigma) 57\ triples: http://www.freezebox.net...9_11_01_archive.html
- 11:03:12 (site_manager) 2\ triples: http://www.freebase.com...sic/album/genre
- 11:03:12 (site_manager) 2\ triples: http://www.freebase.com..._episode/writer

\section*{Sindice Tweet}

Tue Sep 07 09:49:30 2010 Sindice participates in the new eu LOD2 project, 4 years long, just started. http://lod2.eu Truly great competences in the consortium

\section*{Sindice Blog}

- Sindice now supports Efficient Data discovery and Sync
  - So far semantic web search engines and semantic aggregation services have been inserting datasets by hand or have been based on “random... (More →)
  - Sindice planned downtime this weekend
    - Due to an expansion of one of our datacentres (and the electrical work that this implies), Sindice and related services such as sigma w... (More →)

\end{document}
```
A fuzzier definition

- **Enterprise Linked Data**
  - Data that might or might not be published on the web. But has a "webby updated feeling"
  - But at some point is in RDF format
  - Should be querable in an integrated way
    - .. "across" datasets
    - ..if links are there (e.g. sameAs properties) this is easy
    - Otherwise we hope Sparql is expressive enough to still do the join

- **So in practice**
  - Live Semantic Data Warehousing
Sindice Tools to address Enterprise Linked Data

- A lot of Sindice cluster technology immediately maps to Enterprise Linked Data
  - Hadoop workflows
  - Semantic Information Retrieval (Siren)
  - Cluster RDF Databases
  - Assisted Interaction tools
    - Dataset searching, data inspection and browsing, query writing etc.

- It’s Web Scale already.. But we need more
  - An extra dimension: to effectively tackle diversity
Introducing Sindice cloud dataspaces
Zooming in on one Semantic Web Data

A typical implementation template
Clouds are behind the scenes

<dataspace id="iphonedataspace">
  <dependencies>
    http://ecommerce01.dataspace.sindice.net/
    http://price01.dataspace.sindice.net/
  </dependencies>
  <resources>
    <mysql name="sql">
      <hbase size="10g">
        <siren name="index">
          <triplestore name="sparql" kind="virtuoso" />
        </siren>
      </hbase>
    </mysql>
  </resources>
  <retention> (see later)
  <update-rate>1D</update-rate>
  <timeout>1D</timeout>
</dataspace>
But enable the magic

- **Key feature: hotswap**
  1. Cloud creates a whole new set of resources
  2. These are loaded up from 0 possibly
  3. When ready these are swapped with the old ones
  4. Old ones are retained as long as they keep serving existing long standing requests
  5. Old one is discarded, resources are reclaimed
Considerations

- RDF is not a magic bullet SPARQL on the fly integration capabilities have clear and well known limits
- Data transformation capabilities at very large scale is served via cluster technologies
- We believe **cloud technology is fundamental** to deliver on the promise of Semantic Web integration systems.
References:


Semantic Information Retrieval

- **Sindice first work: Semantic Data Search**
  - Applying IR technologies for the retrieval of semantic data
  - Searching through billions of complex entities instead of textual documents
- **Return “data answers”**
  - To textual and structured queries
SIREn: Semantic IR Engine

- Extension to Enterprise Search Engine Solr
- Semantic, full-text, incremental updates, distributed search

![Graph showing performance comparison between Semantic Databases and SIREn]
Limitations of Apache Solr

- Not efficient with highly heterogeneous structured data sources
  - Limitation on the number of attributes:
    - Dictionary size explosion
# Dictionary Size Explosion

## Record 1

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>Renaud Delbru</td>
</tr>
<tr>
<td>name</td>
<td>Renaud Delbru</td>
</tr>
</tbody>
</table>

Digital Enterprise Research Institute

www.deri.ie
Dictionary Size Explosion

Dictionary construction

- Concatenation of attribute name and term
- \(N \times M\) complexity (worst case)

2 attributes * 2 terms = 4 dictionary entries

100K attributes * 1B terms = 100B entries

Record 1

<table>
<thead>
<tr>
<th>label</th>
<th>Renaud Delbru</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>Renaud Delbru</td>
</tr>
</tbody>
</table>

Dictionary

| label:renaud |
| label:delbru |
| name:renaud |
| name:delbru |
Limitations of Apache Solr

- Not efficient with highly heterogeneous structured data sources
  - Limitation on the number of attributes:
    - Dictionary size explosion
    - Query clause explosion when searching across all attributes
Limitations of Apache Solr

- Not efficient with highly heterogeneous structured data sources
  - Limitation on the number of attributes:
    - Dictionary size explosion
    - Query clause explosion when searching across all attributes
- Limited support for structured query
  - Multi-valued attributes
Multi-valued attributes

- No support in Solr for "all words must match in the same value of a multi-valued field".
- A field value is a bag of words
  - No distinction between multiple values

<table>
<thead>
<tr>
<th>Record 1</th>
<th>Record 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>label</td>
</tr>
<tr>
<td><strong>man's best friend</strong></td>
<td><strong>man's worst enemy</strong></td>
</tr>
<tr>
<td>pooch</td>
<td>friend to no one</td>
</tr>
</tbody>
</table>
Multi-valued attributes

- No support in Solr for "all words must match in the same value of a multi-valued field".
- A field value is a bag of words
  - No distinction between multiple values
- Query example
  - label: man’s friend
  - Solr returns Record 1 & 2 as results

<table>
<thead>
<tr>
<th>Record 1</th>
<th>Record 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>label</td>
<td>label</td>
</tr>
<tr>
<td><strong>man's best friend</strong> pooch</td>
<td><strong>man's worst enemy friend</strong> to no one</td>
</tr>
</tbody>
</table>
Limitations of Apache Solr

- Not efficient with highly heterogeneous structured data sources
  - Limitation on the number of attributes:
    - Dictionary size explosion
    - Query clause explosion when searching across all attributes
- Limited support for structured query
  - Multi-valued attributes
  - No full-text search on attribute names
Full-text search on attribute names

- No support in Solr for “keyword search in attribute names”.
- Query example
  - (name OR label) = “Renaud Delbru”
  - Solr is unable to find the records without the exact attribute name

<table>
<thead>
<tr>
<th>Record 1</th>
<th>Record 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:label</td>
<td>Renaud Delbru</td>
</tr>
<tr>
<td>foaf:name</td>
<td>Renaud Delbru</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Record 3</th>
<th>Record 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>sioc:name</td>
<td>Renaud Delbru</td>
</tr>
<tr>
<td>full_name</td>
<td>Renaud Delbru</td>
</tr>
</tbody>
</table>
Limitations of Apache Solr

- **Not efficient with highly heterogeneous structured data sources**
  - Limitation on the number of attributes:
    - Dictionary size explosion
    - Query clause explosion when searching across all attributes

- **Limited support for structured query**
  - Multi-valued attributes
  - No full-text search on attribute names
  - No 1:N relationship materialisation
What is SIREn?

- Record-centric indexing model
  - Based on Tree focused IR technique
  - Built on top of SOLr

- Built for dealing with highly heterogeneous structured data sources
  - No attribute limit
  - Efficient query across all attributes
  - Extended support for structured query
Some numbers: Siren on Sindice

**Data Collection**
- 260M web data documents (RDF, RDFa, Microformat, etc.)
- 100K datasets
- 22B triples

**Settings**
- Cluster of 4 nodes
  - 2 nodes for indexing
  - 2 nodes for querying
- Replication

**Indexing Performance**
- Full index construction takes 14 hours
- 436K triples / second

**Services**
- Keyword and structured queries
- Dataset search
- >> 99% uptime
Query time ranking

Query: Ireland Galway City

<table>
<thead>
<tr>
<th><strong>Dbpedia:Galway</strong></th>
<th><strong>Dbpedia:James Galway</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Label</strong></td>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>Galway City</td>
<td>James Galway</td>
</tr>
<tr>
<td><strong>Comment</strong></td>
<td><strong>Comment</strong></td>
</tr>
<tr>
<td>Galway the third largest</td>
<td>James Galway is a Northern Ireland born</td>
</tr>
<tr>
<td>and the fastest-growing</td>
<td>virtuoso flutist</td>
</tr>
<tr>
<td>city in Ireland</td>
<td>from Belfast city</td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td><strong>Band</strong></td>
</tr>
<tr>
<td>72729</td>
<td>Royal Philharmonic Orchestra</td>
</tr>
<tr>
<td><strong>Type</strong></td>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>City</td>
<td>Person</td>
</tr>
</tbody>
</table>
Semantic IR – Data Ranking

- Static Ranking: Ding Model
SIREn In Action

- Developed for and in production
- Read the use case
- Open Source
  - We are aware of about dozen enterprise deployments
  - E.g. see presentations like
    “Improving Your Semantic Web Apps Performance with Lucene, SIREn and RDF”
Building on Sindice: Sig.MA

- Browser
- Mashup Generator
- API

...For the Web of Data
Building on Sindice: Sig.ma

Giovanni Tummarello

given name: Giovanni [3,4,8,9,10,11,12,18,19,21,24,28]
family name: Tummarello [3,4,8,9,10,11,12,18,19,21,24,25,28]
is creator of:
- A Node Indexing Scheme for Web Entity Retrieval [3,4]
- Hierarchical Link Analysis for Ranking Web Data [3,4]
- Rapid Prototyping of Semantic Mash-Ups through Semantic Web Pipes [3,4]
- Context Dependent Reasoning for Semantic Documents in Sindice [3,4]
- An Entity Name System for Linking Semantic Web Data [3,4]
- Sindice.com: Weaving the Open Linked Data [3,4,5,6,7,13,14,15,16]
- RDFSync: efficient remote synchronization of RDF models [3,4,5,6,7,13,14,15,16]
- Exposing Large Datasets with Semantic Sitemaps [5,6,7,13,14,15,16]
- Enabling Semantic Web communities with DBin: an overview [5,6,7,13,14,15,16]

show 100 more values →
Building on Sindice (2): Site Services

“Got an I Like It button(*)? You are 1 click away to leverage the entire Web of Data (for your needs specific visitor needs).”

Sindice Site Services - for website owners

Sindice Widgets

Get Web Data Powered widgets to deliver highly valuable extra information and power smart recommendations across your pages.
World most unfortunate DVD store

Menu just demo purpose

The Thirteenth Floor 2012 10,000 B.C. The Day After Tomorrow Trade The Secret Life of Bees Alien Resurrection Aliens Alien 3 AVP - Alien Vs. Predator

10,000 B.C.
Same site, but "I like it" more

Menu just demo purpose

The Thirteenth Floor 2012 10,000 B.C. The Day After Tomorrow Trade The Secret Life of Bees Alien Resurrection Aliens Alien 3 AVP - Alien Vs. Predator

10,000 B.C.

Like 125 people like this. Be the first of your friends.
Just a small code addition! (?)

```html
<meta property="og:url" content="http://demo.sindice.net/..." />
<meta property="og:title" content="10000 BC" />
<meta property="og:type" content="movie" />
<meta property="og:image" content="http://demo.sindice.net/..."/>
```
Just make it known to Sindice

Share your Web Data presence with Sindice

For individual semantically enabled pages (RDF, RDFa, Microformats), use the PING form to the right (also see the ping API). Pings are processed with high priority and they're likely indexed within minutes.

For effective indexing of large datasets, please follow the guidelines on How to Publish Web Data for Effective Discovery and Synchronization and then ping any of your pages and or use the Submit Sitemap form to the right.

Any questions related to sindice data acquisition should be post to:
http://groups.google.com/group/sindice-dev

Ping Form (one URL per line)

http://demo.sindice.net/sindice-widget/og-data/movie/200

Submit Pages

Submit Sitemap form (read me)

Email (optional, used for feedback):

Sitemap URL:

Submit Sitemap
.. Which will happily acquire your pages

.. And keep itself in sync
Pick a widget from our “Store”
Copypaste Tag into your HTML

```html
<html>
<head>
<script src="http://localhost:8080/sindice-widget/ja/jquery/js/jquery-1.4.3.min.js" type="text/javascript"></script>
<script type="text/javascript">
$(document).ready(function()
{
  var myWidget = new SindiceWidget({
    widgetName: 'demo_get_more_info_about_movie',
    initial:
      {attributes:
        header: 'My header',
        footer: 'Powered by Sindice',
        width: 200,
        type: 'more_info'},
    params:
       content: 'The Day After Tomorrow',
       role: 'image'},
    body:
      {div id: 'my-widget'>"div>
</head>
</html>
```
<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Starring</th>
<th>Writer</th>
<th>Director</th>
<th>Budget</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000 BC</td>
<td></td>
<td>Steven Strait</td>
<td>Harald Klocher</td>
<td>Roland Emmerich</td>
<td>105000000</td>
<td>109 min</td>
</tr>
<tr>
<td>The Day After Tomorrow</td>
<td></td>
<td>Jake Gyllenhaal</td>
<td>Roland Emmerich</td>
<td>Roland Emmerich</td>
<td>125000000</td>
<td>124 mins</td>
</tr>
<tr>
<td>Alien 3</td>
<td></td>
<td>Sigourney Weaver</td>
<td>Dan O'Bannon</td>
<td>David Fincher</td>
<td>50000000</td>
<td></td>
</tr>
<tr>
<td>Alien Resurrection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Thirteenth Floor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aliens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sorted by: **labels**; then by... **✓** grouped as sorted.
### 10,000 B.C.

**More info:**

**Title:**
- 10,000 BC

**Starring:**
- Steven Strait
- Cliff Curtis
- Camilla Belle

**Writer:**
- Harald Kloser
- Roland Emmerich

**Director:**
- Roland Emmerich

**Budget:**
- 105000000

**Runtime:**
- 109 min.

powered by Sindice