# Web of Data : Past Research, Current State, Clear Opportunities

Giovanni Tummarello, DERI Institute (Galway)
FBK Institute (Trento)

National University of Ireland, Galway
Ollscoil na hÉireann, Gaillimh

science foundation ireland
fondúireacht eolaíochta éireann

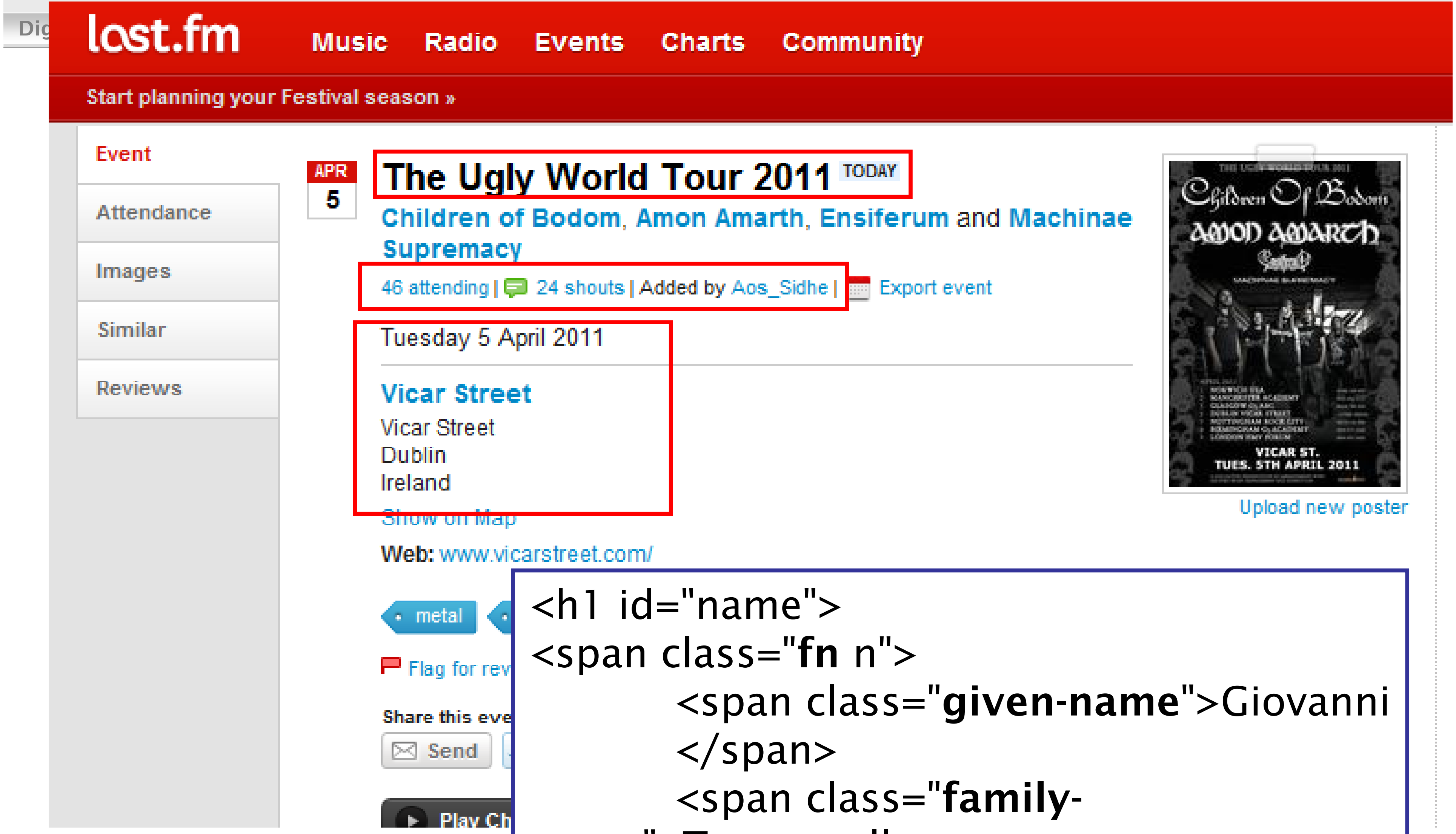
# The Web of Data

A Web where **pages** have **elements** that machines can **interpret automatically** thanks to **Metadata Standards**

Quite a significant slice of the web, already

- □ RDF, RDFa
- □ Microformats
- □ Schema.org (Microdata)

# For example:



lαst.fm Music Radio Events Charts Community

Start planning your Festival season »

Event

Attendance

Images

Similar

Reviews

APR 5

The Ugly World Tour 2011 TODAY

Children of Bodom, Amon Amarth, Ensiferum and Machinae Supremacy

46 attending | 24 shouts | Added by Aos_Sidhe | Export event

Tuesday 5 April 2011

Vicar Street

Vicar Street
Dublin
Ireland

Show on Map

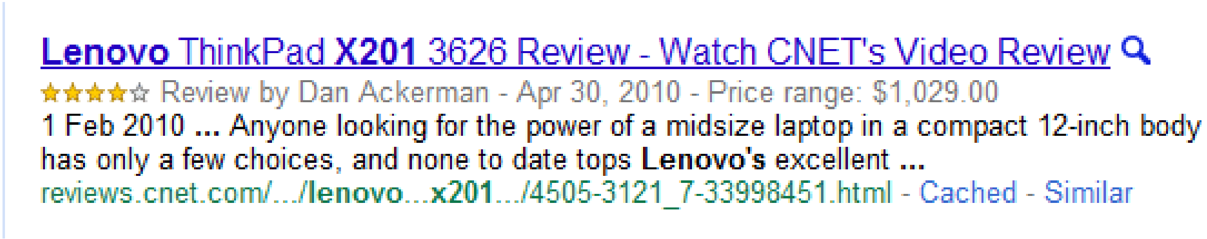
Web: www.vicarstreet.com/

Upload new poster

```
<h1 id="name">
<span class="fn n">
        <span class="given-name">Giovanni
        </span>
        <span class="family-
name">Tummarello
```

# Now becoming ubiquitous

- Google Rich Snippets Program



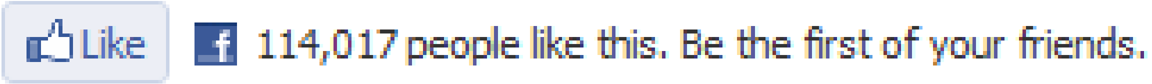
Lenovo ThinkPad **X201** 3626 Review - Watch CNET's Video Review

★★★★☆ Review by Dan Ackerman - Apr 30, 2010 - Price range: $1,029.00

1 Feb 2010 **...** Anyone looking for the power of a midsize laptop in a compact 12-inch body has only a few choices, and none to date tops **Lenovo's** excellent **...**

reviews.cnet.com/.../**lenovo**...**x201**.../4505-3121_7-33998451.html - Cached - Similar

- Facebook Opengraph Protocol

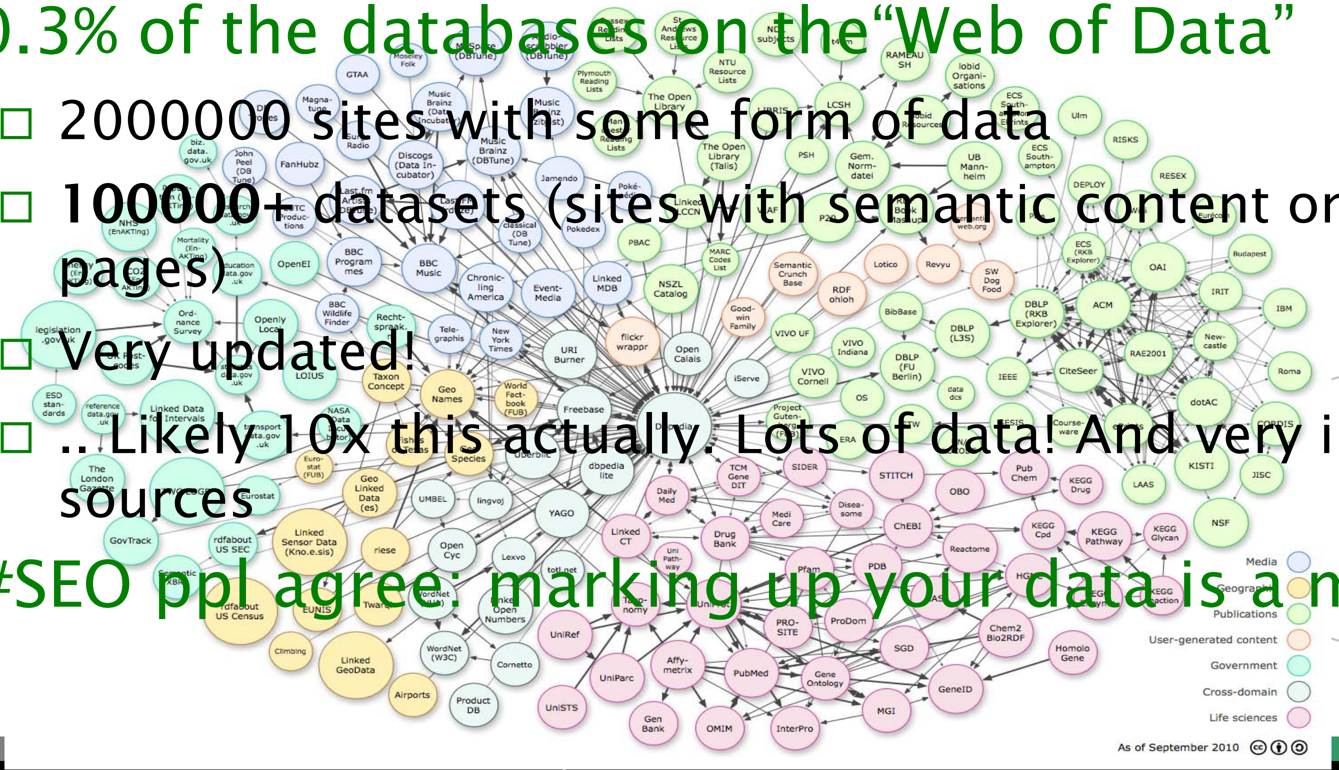


114,017 people like this. Be the first of your friends.

- Schema.org..

# Web Data: a underestimate

- LOD Cloud: approx 300 datasets
  - □ Relatively clean, with some mutual integration
  - □ Not very updated
- 0.3% of the databases on the“Web of Data”
  - □ 2000000 sites with some form of data
  - □ 100000+ 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



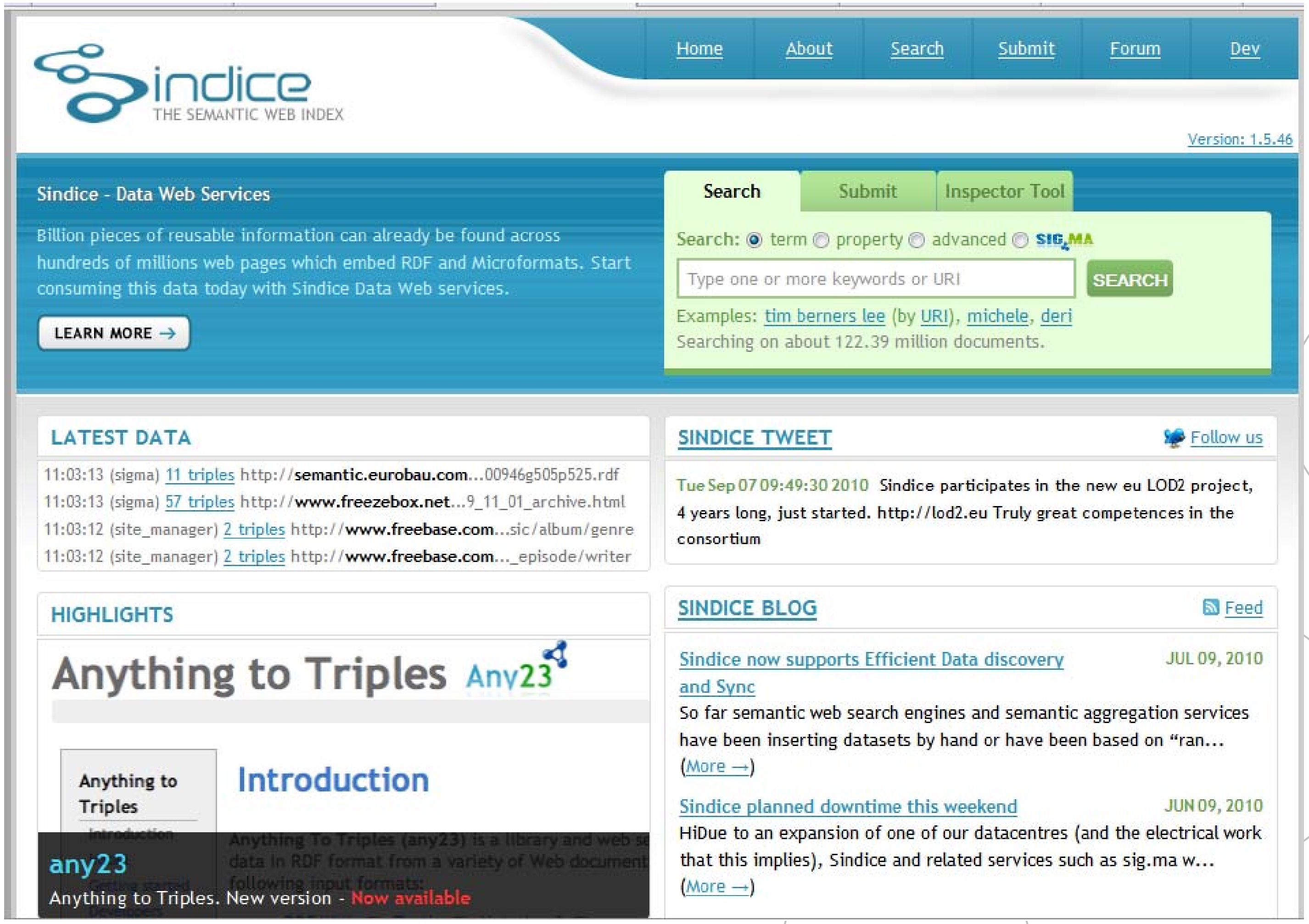
# ..but also lots of questions

- Questions
  - □ How to find sources ?
  - □ How to clean and transform them for one’s specific needs?

# The DERI Data Intensive Infrastructure lab

Sindice “WebStar” Cluster

# Sindice.com: Portal for Sindice Services



Sindice - Data Web Services

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

Search: ● term ○ property ○ advanced ○ SIG.MA

Type one or more keywords or URI SEARCH

Examples: tim berners lee (by URI), michele, deri

Searching on about 122.39 million documents.

LATEST DATA

11:03:13 (sigma) 11 triples http://semantic.eurobau.com...00946g505p525.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

HIGHLIGHTS

Anything to Triples Any23

Anything to Triples
Introduction

Anything To Triples (any23) is a library and web se... data in RDF format from a variety of Web document... following input formats:

any23
Anything to Triples. New version - Now available

SINDICE TWEET Follow us

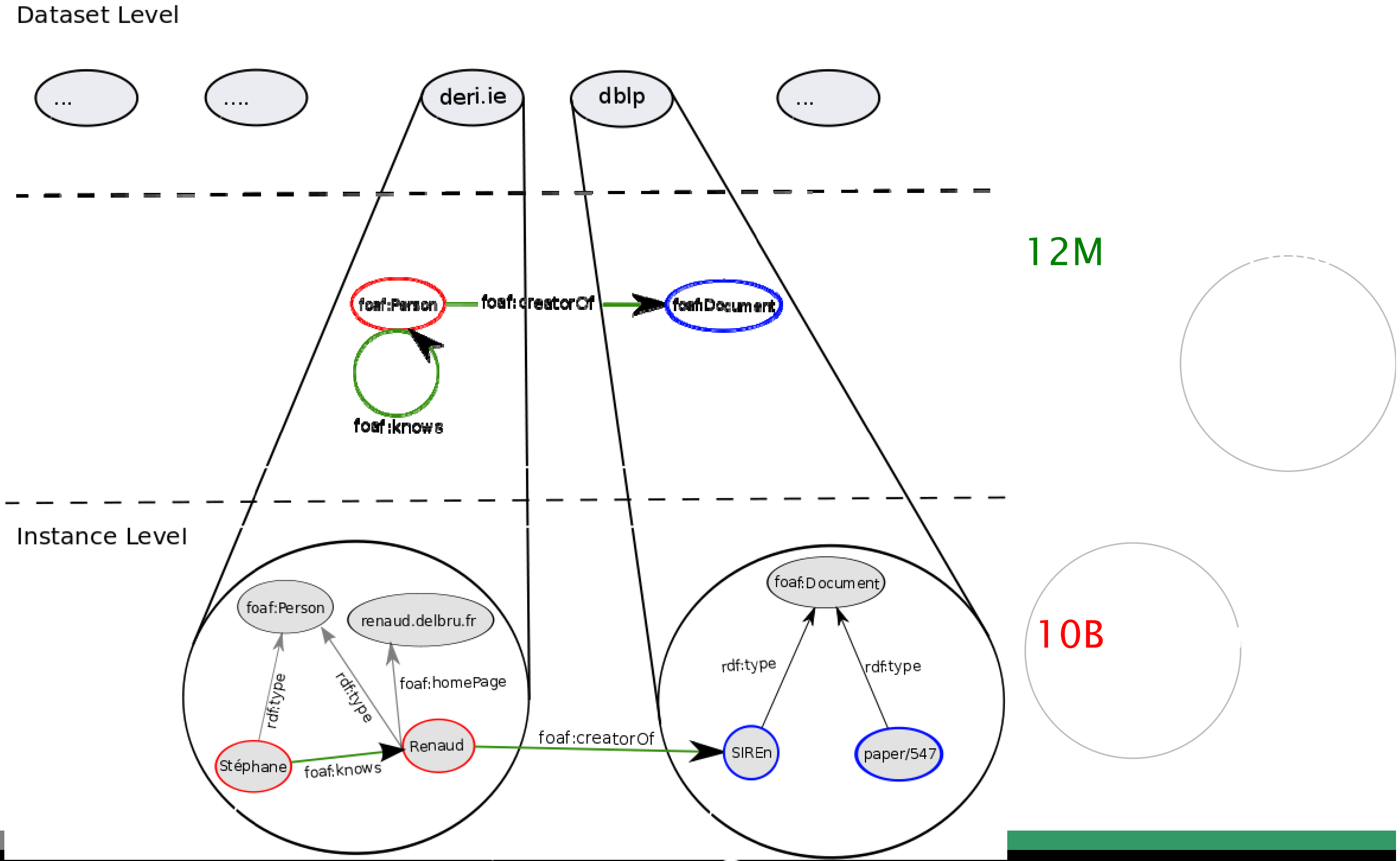
Tue Sep 07 09:49:30 2010 Sindice participates in the new eu LOD2 project, 4 years long, just started. http://lod2.eu Truly great competences in the consortium

SINDICE BLOG Feed

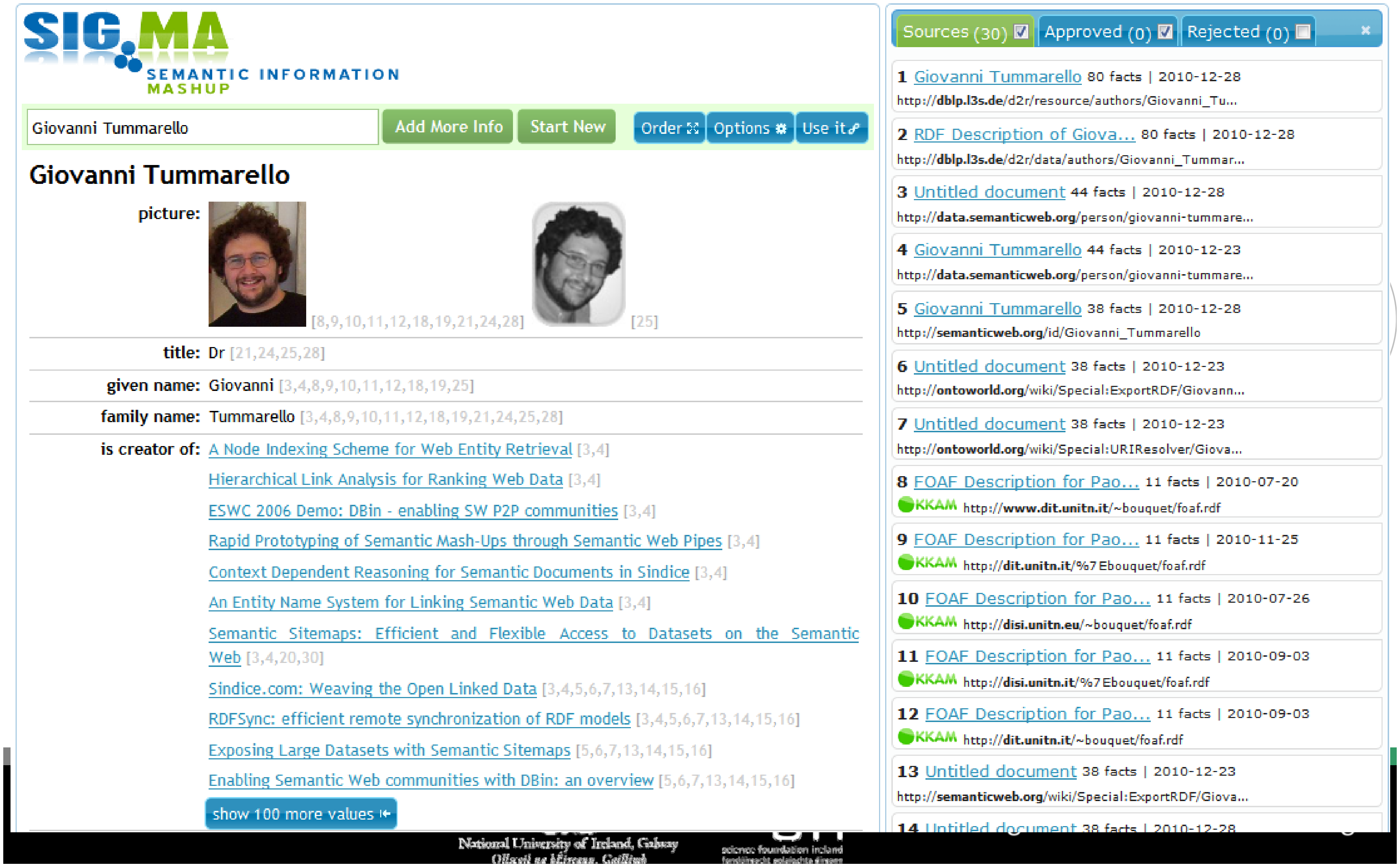
Sindice now supports Efficient Data discovery and Sync JUL 09, 2010
So far semantic web search engines and semantic aggregation services have been inserting datasets by hand or have been based on “ran... (More →)

Sindice planned downtime this weekend JUN 09, 2010
HiDue to an expansion of one of our datacentres (and the electrical work that this implies), Sindice and related services such as sig.ma w... (More →)

# Building a graph of web graphs



# Building on Sindice: Sig.ma



Giovanni Tummarello

## Giovanni Tummarello

picture: [8,9,10,11,12,18,19,21,24,28] [25]

title: Dr [21,24,25,28]

given name: Giovanni [3,4,8,9,10,11,12,18,19,25]

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]
- ESWC 2006 Demo: DBin - enabling SW P2P communities [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]
- Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web [3,4,20,30]
- 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]

Sources (30) ☑ Approved (0) ☑ Rejected (0) □ ×

1 Giovanni Tummarello 80 facts | 2010-12-28
http://dblp.l3s.de/d2r/resource/authors/Giovanni_Tu...

2 RDF Description of Giova... 80 facts | 2010-12-28
http://dblp.l3s.de/d2r/data/authors/Giovanni_Tummar...

3 Untitled document 44 facts | 2010-12-28
http://data.semanticweb.org/person/giovanni-tummare...

4 Giovanni Tummarello 44 facts | 2010-12-23
http://data.semanticweb.org/person/giovanni-tummare...

5 Giovanni Tummarello 38 facts | 2010-12-28
http://semanticweb.org/id/Giovanni_Tummarello

6 Untitled document 38 facts | 2010-12-23
http://ontoworld.org/wiki/Special:ExportRDF/Giovann...

7 Untitled document 38 facts | 2010-12-23
http://ontoworld.org/wiki/Special:URIResolver/Giova...

8 FOAF Description for Pao... 11 facts | 2010-07-20
KKAM http://www.dit.unitn.it/~bouquet/foaf.rdf

9 FOAF Description for Pao... 11 facts | 2010-11-25
KKAM http://dit.unitn.it/%7Ebouquet/foaf.rdf

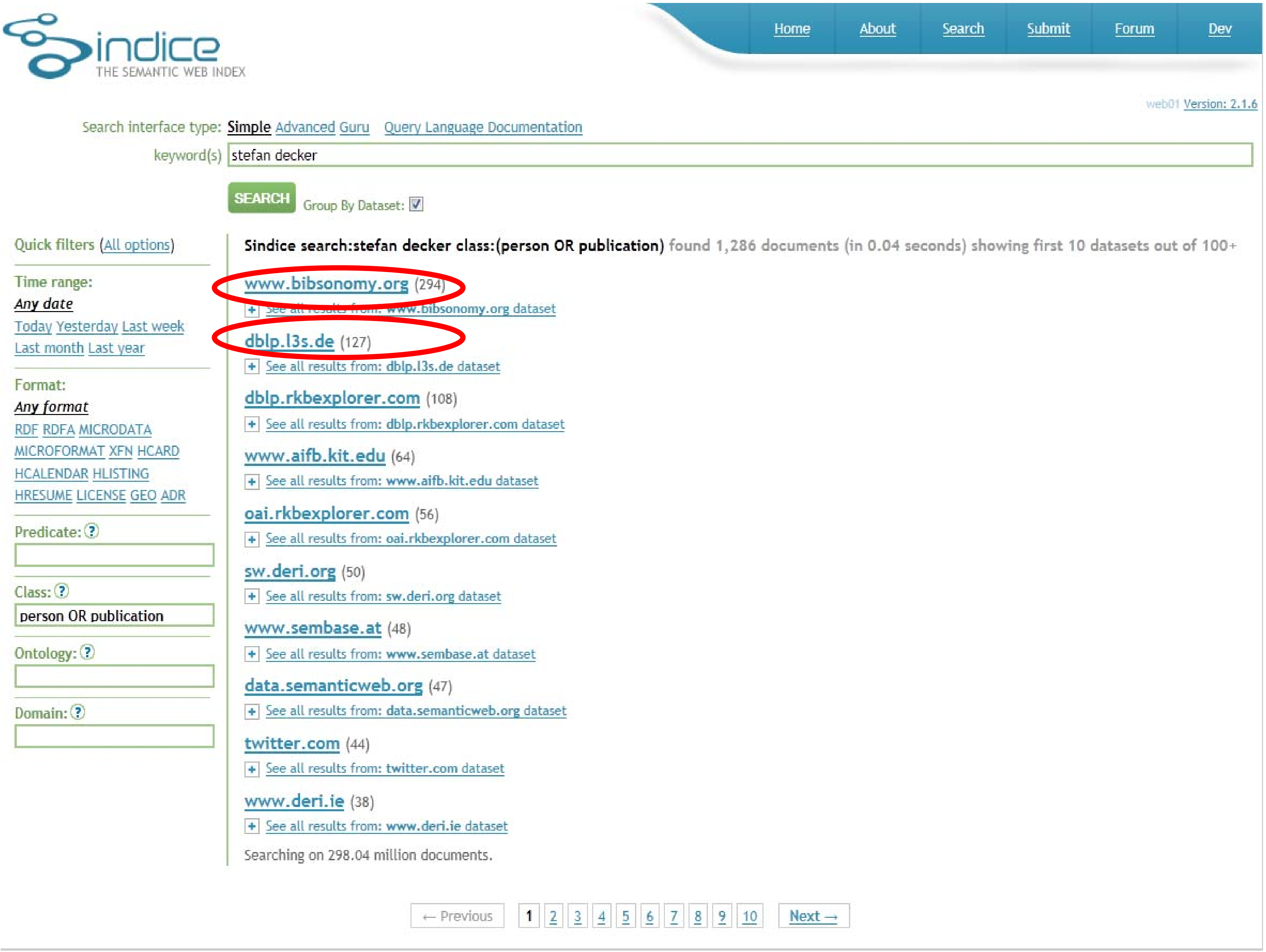
10 FOAF Description for Pao... 11 facts | 2010-07-26
KKAM http://disi.unitn.eu/~bouquet/foaf.rdf

11 FOAF Description for Pao... 11 facts | 2010-09-03
KKAM http://disi.unitn.it/%7Ebouquet/foaf.rdf

12 FOAF Description for Pao... 11 facts | 2010-09-03
KKAM http://dit.unitn.it/~bouquet/foaf.rdf

13 Untitled document 38 facts | 2010-12-23
http://semanticweb.org/wiki/Special:ExportRDF/Giova...

14 Untitled document 38 facts | 2010-12-28



Search interface type: Simple Advanced Guru Query Language Documentation

keyword(s) stefan decker

SEARCH Group By Dataset: ☑

Quick filters (All options)

Time range:
*Any date*
Today Yesterday Last week
Last month Last year

Format:
*Any format*
RDF RDFA MICRODATA
MICROFORMAT XFN HCARD
HCALENDAR HLISTING
HRESUME LICENSE GEO ADR

Predicate: ?

Class: ?
person OR publication

Ontology: ?

Domain: ?

**Sindice search:stefan decker class:(person OR publication)** found 1,286 documents (in 0.04 seconds) showing first 10 datasets out of 100+

www.bibsonomy.org (294)
+ See all results from: www.bibsonomy.org dataset

dblp.l3s.de (127)
+ See all results from: dblp.l3s.de dataset

dblp.rkbexplorer.com (108)
+ See all results from: dblp.rkbexplorer.com dataset

www.aifb.kit.edu (64)
+ See all results from: www.aifb.kit.edu dataset

oai.rkbexplorer.com (56)
+ See all results from: oai.rkbexplorer.com dataset

sw.deri.org (50)
+ See all results from: sw.deri.org dataset

www.sembase.at (48)
+ See all results from: www.sembase.at dataset

data.semanticweb.org (47)
+ See all results from: data.semanticweb.org dataset

twitter.com (44)
+ See all results from: twitter.com dataset

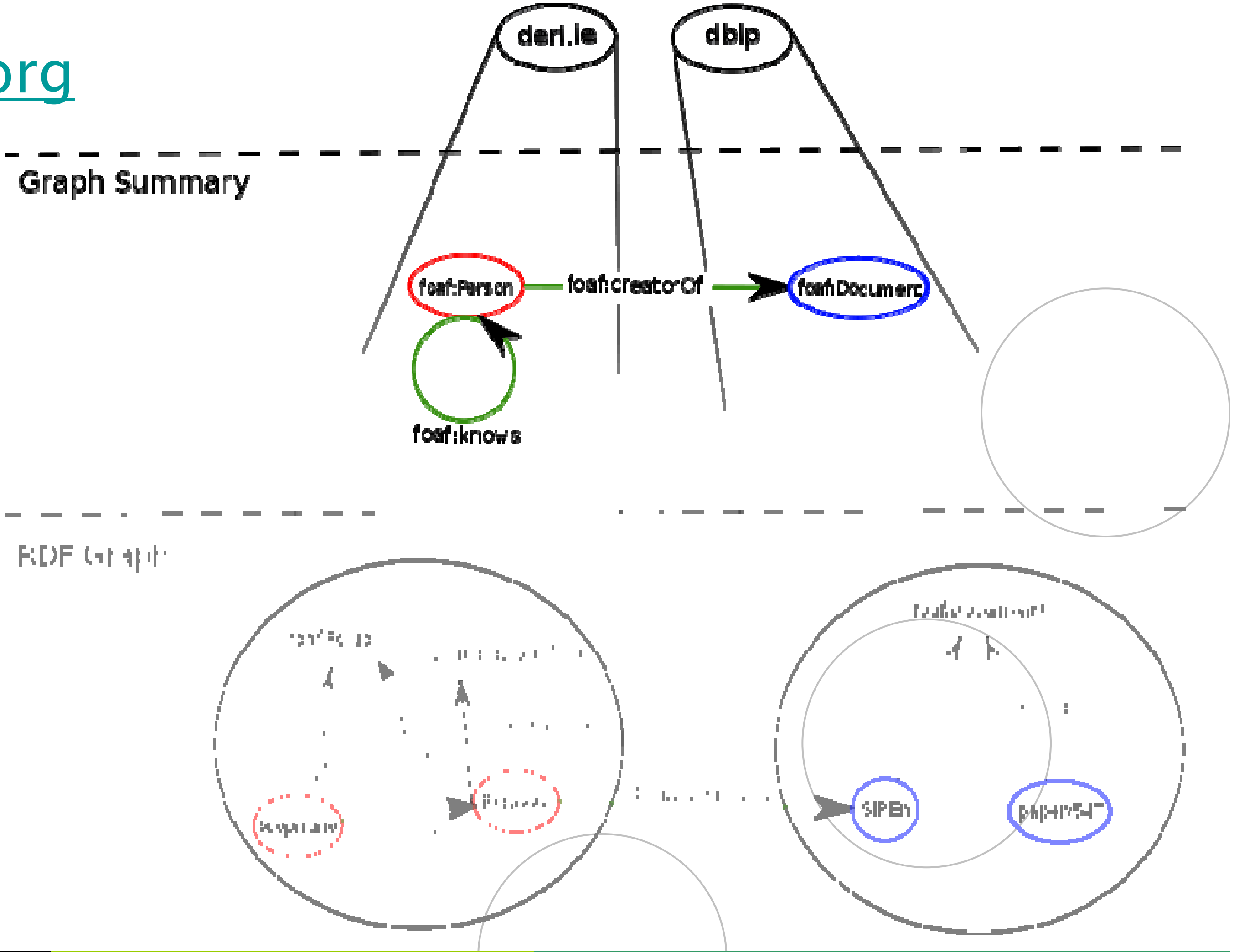
www.deri.ie (38)
+ See all results from: www.deri.ie dataset

Searching on 298.04 million documents.

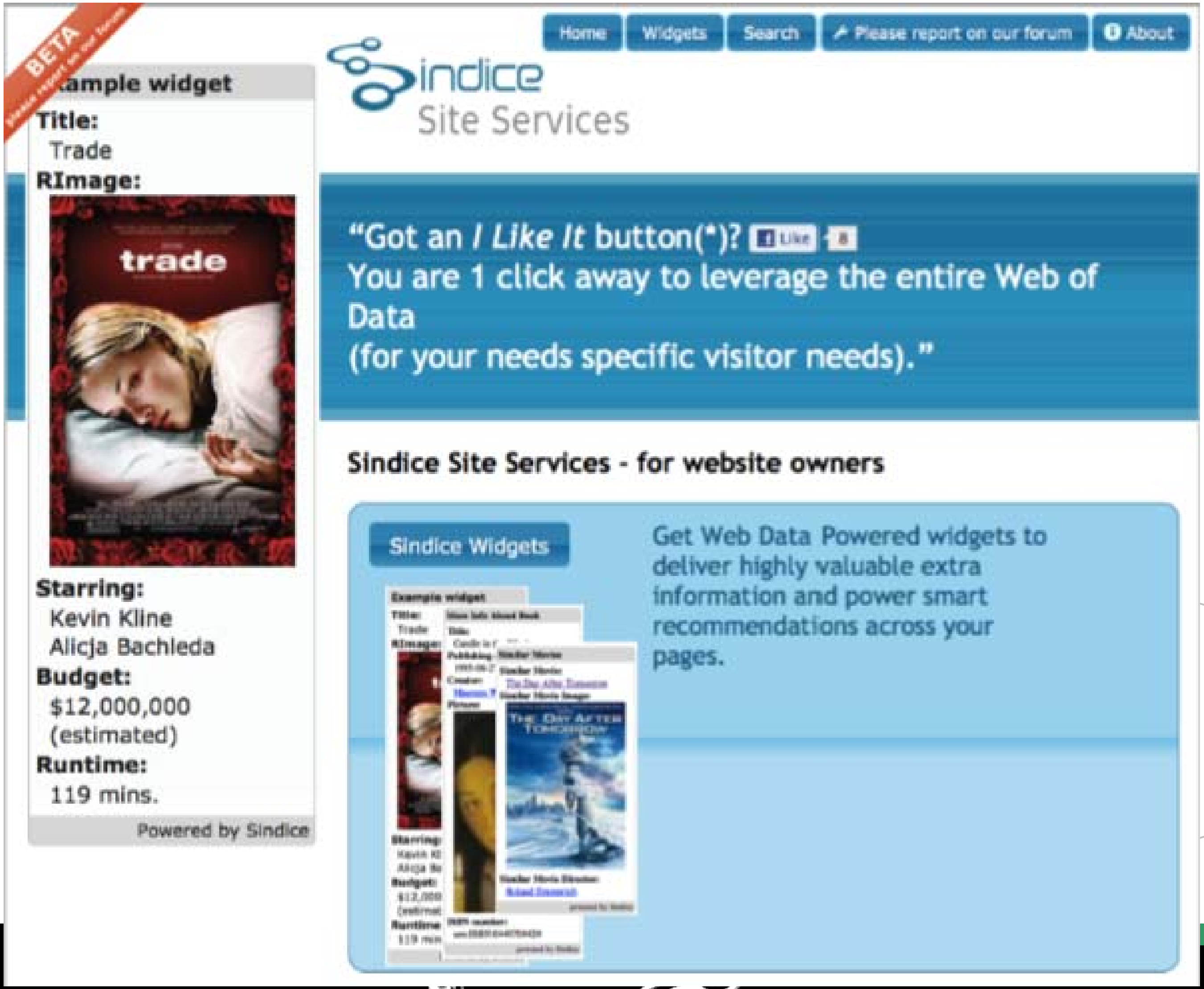
← Previous 1 2 3 4 5 6 7 8 9 10 Next →

# Analytics Example

- [Bibsonomy.org]
- [L3s.de]
- [Deri.ie]



# Building on Sindice (2): Site Services



Example widget

**Title:**
Trade

**RImage:**

**Starring:**
Kevin Kline
Alicja Bachleda

**Budget:**
$12,000,000 (estimated)

**Runtime:**
119 mins.

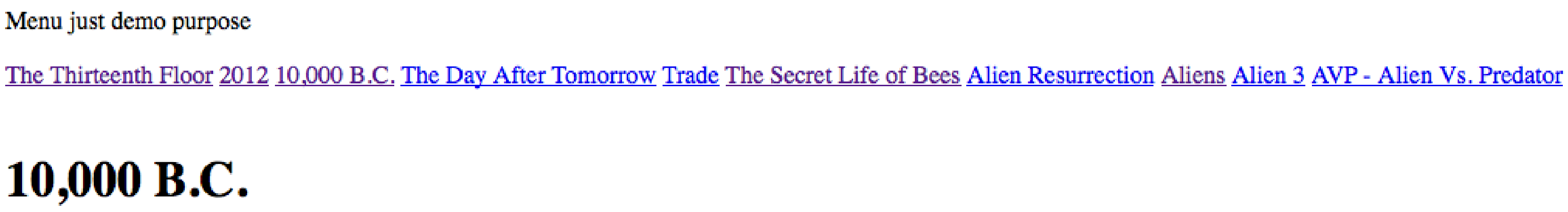
Powered by Sindice

"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**

Get Web Data Powered widgets to deliver highly valuable extra information and power smart recommendations across your pages.

# World most unfortunate DVD store

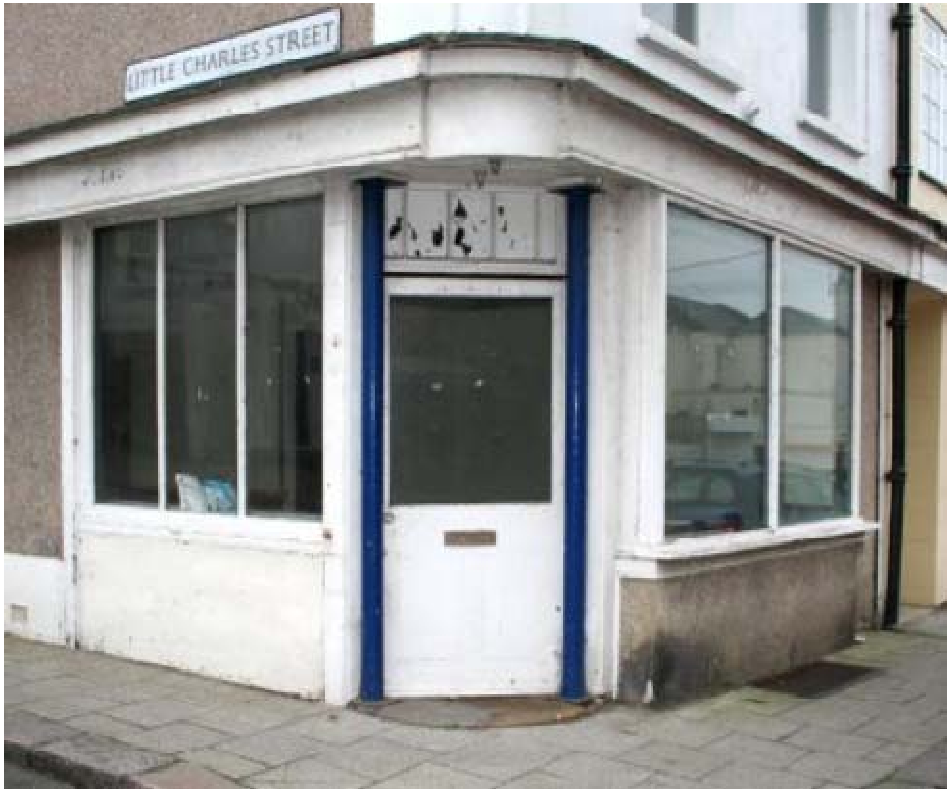


Menu just demo purpose

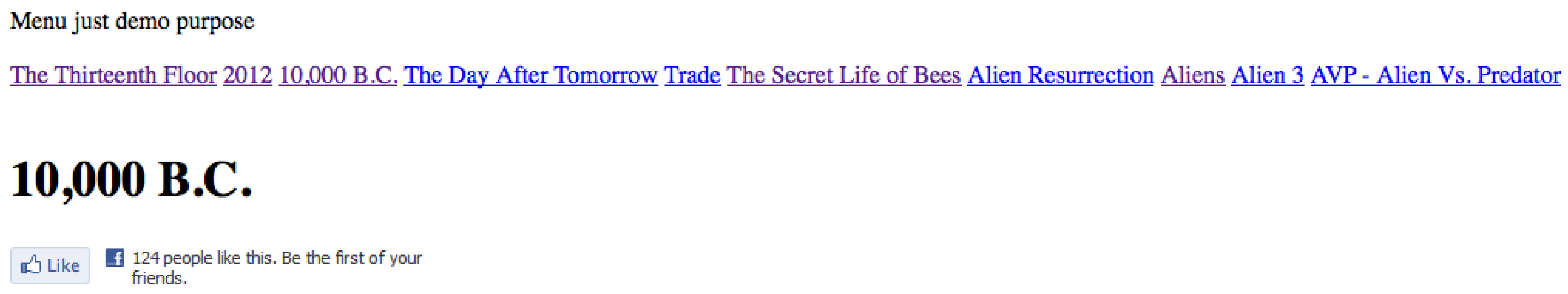
The Thirteenth Floor 2012 10,000 B.C. The Day After Tomorrow Trade The Secret Life of Bees Alien Resurrection Aliens Alien 3 AVP - Alien Vs. Predator

## 10,000 B.C.

# Ugly Shop



# Same site, but “I like it” more



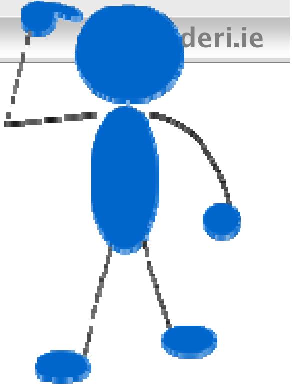
Menu just demo purpose

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

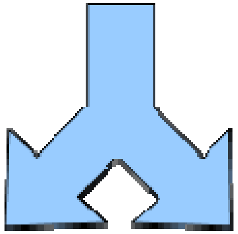
## 10,000 B.C.

124 people like this. Be the first of your friends.

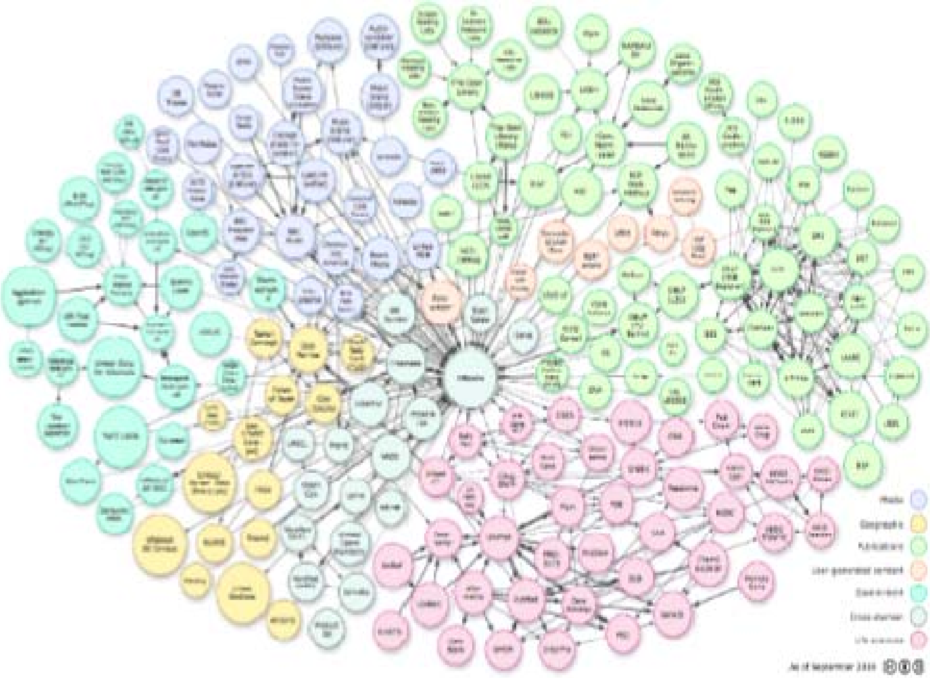
# Just a small code addition! (?)



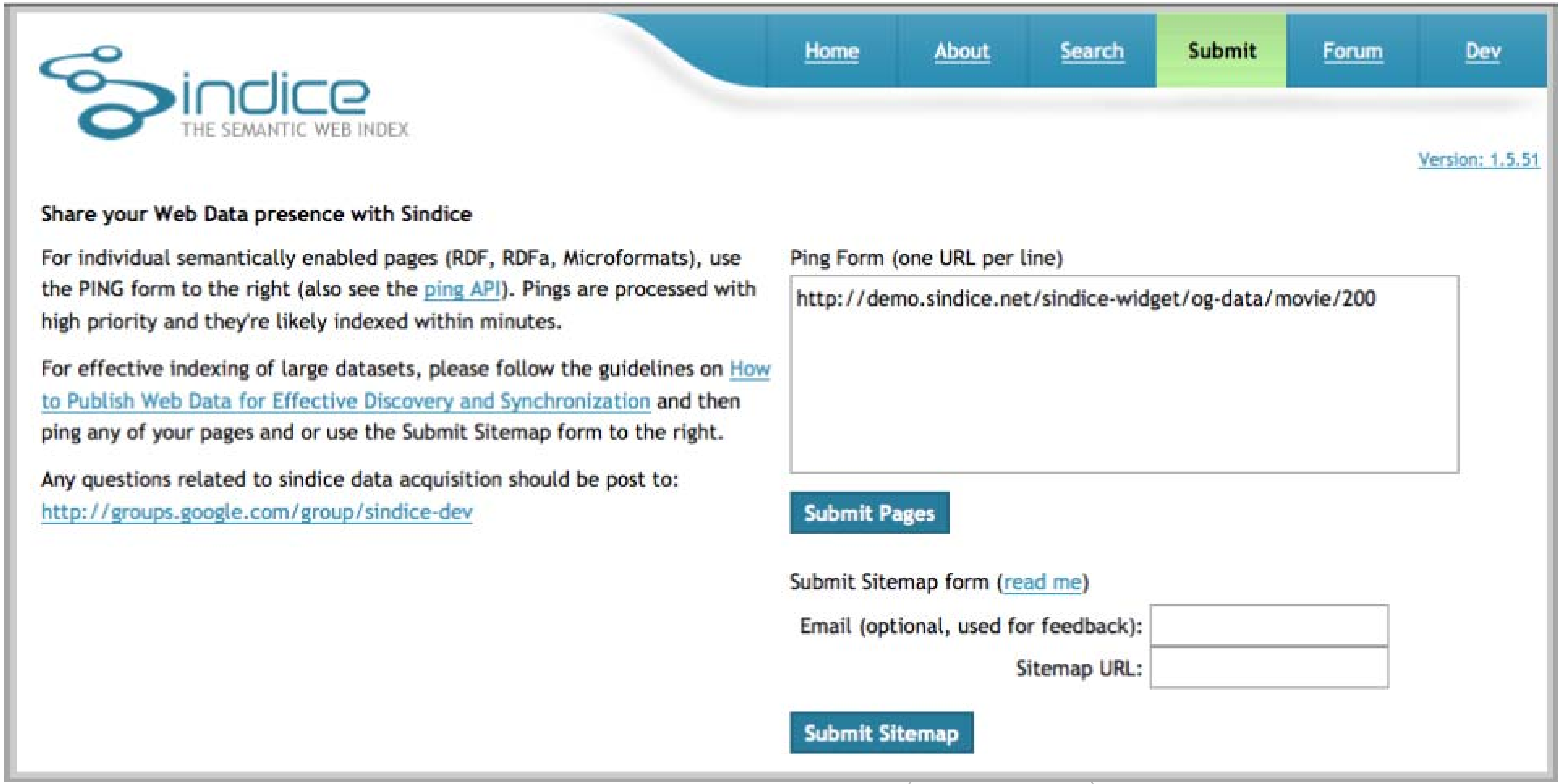
```
<meta property="og:url" content="http://demo.sindice.net/..." />
<meta property="og:title" content=“10000 BC />
<meta property="og:type" content="movie" />
<meta property="og:image" content="http://demo.sindice.net/..."/>
```



GoodRelations
The Web Ontology for E-Commerce
facebook
Drupal
SIOC
DBpedia
BEST BUY
Freebase
LinkedIn



# 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 Sitemap form (read me)

Email (optional, used for feedback):

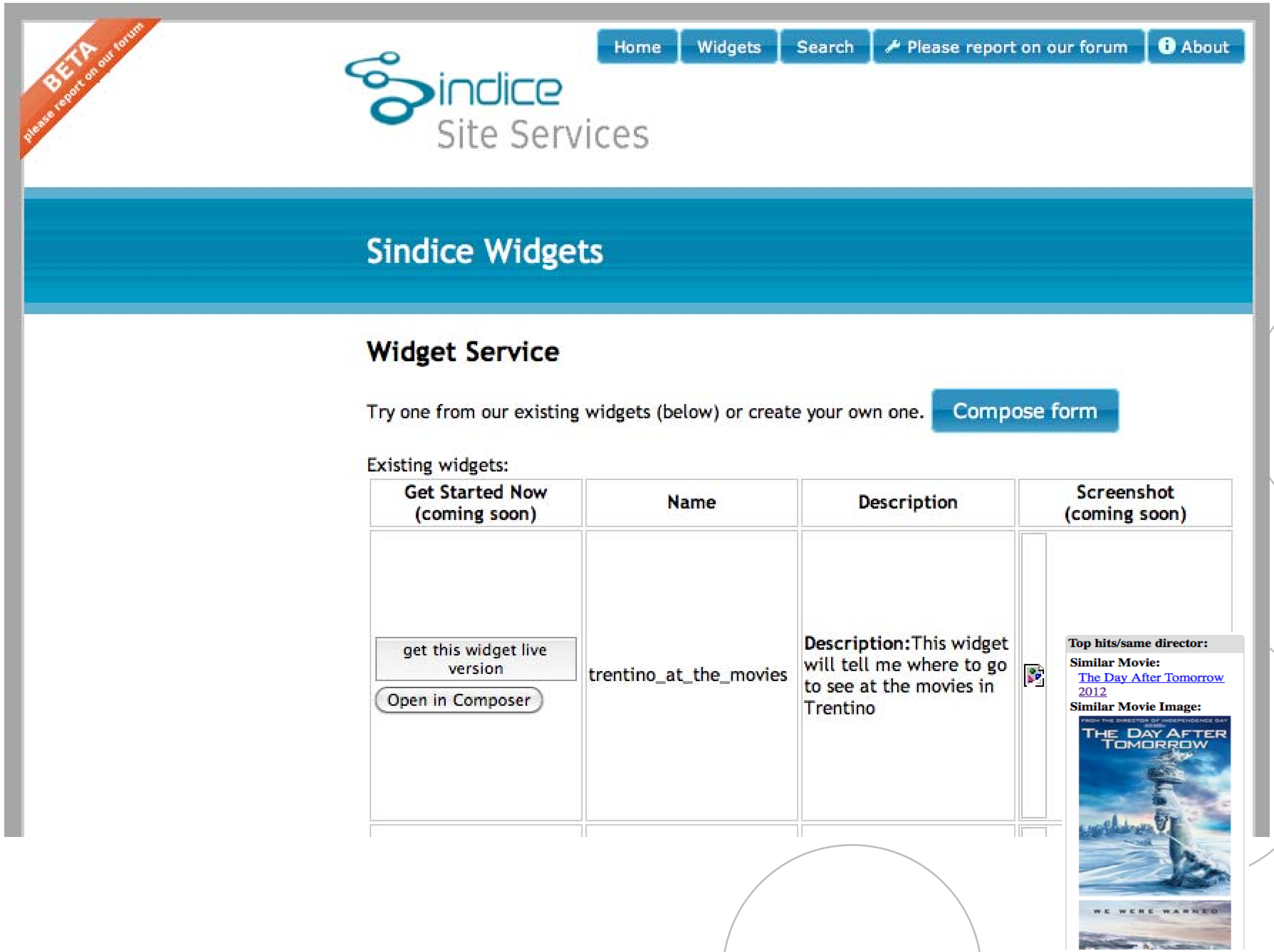
Sitemap URL:

# .. Which will happily acquire your pages

The Day After Tomorrow

## .. And keep itself in sync

# Pick a widget from our “Store”


BETA
please report on our forum
Home
Widgets
Search
Please report on our forum
About
sindice
Site Services
Compose form
get this widget live version
Open in Composer

## Sindice Widgets

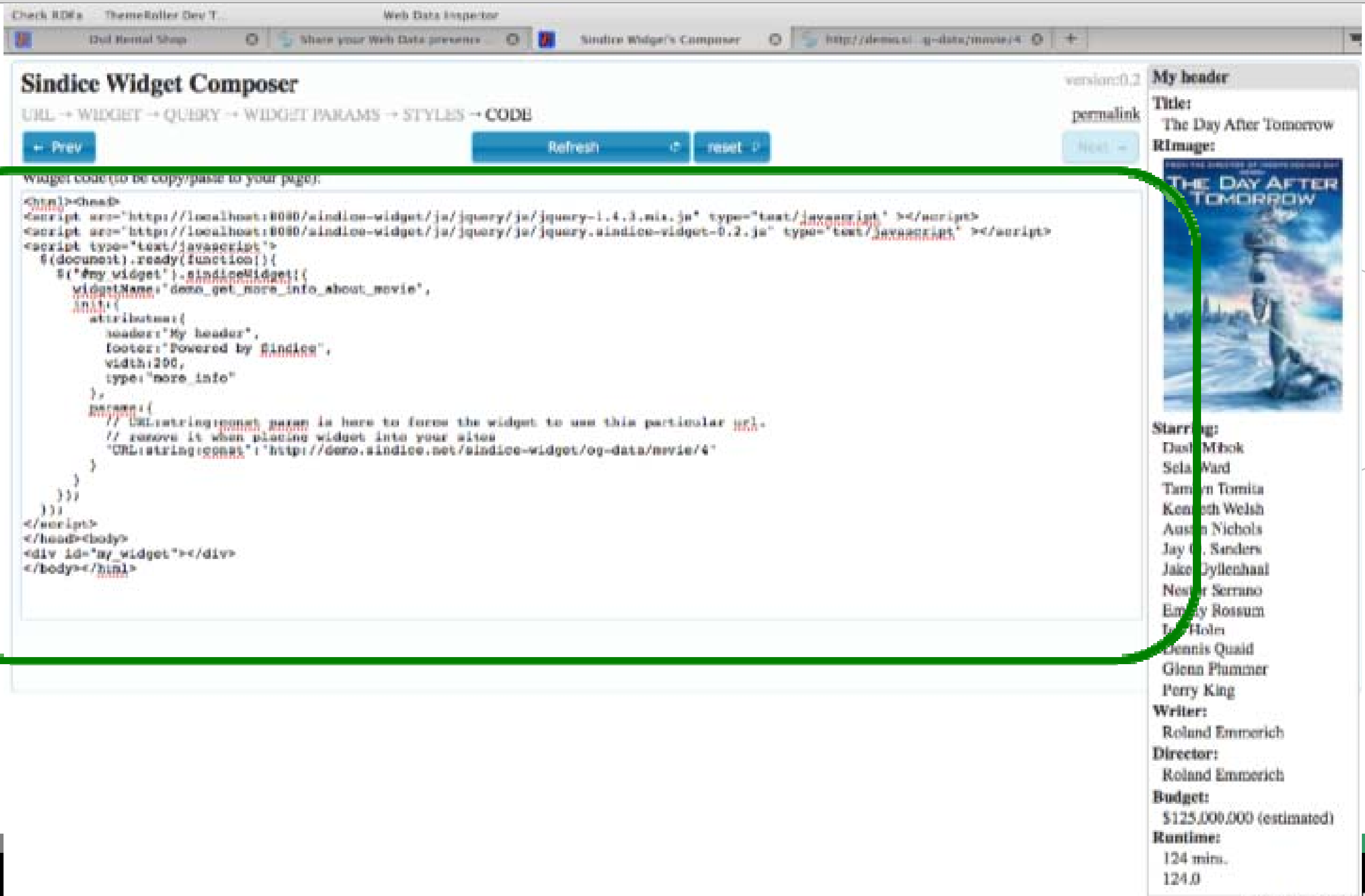
### Widget Service

Try one from our existing widgets (below) or create your own one.

Existing widgets:

| Get Started Now (coming soon) | Name | Description | Screenshot (coming soon) |
|---|---|---|---|
| get this widget live version<br>Open in Composer | trentino_at_the_movies | Description:This widget will tell me where to go to see at the movies in Trentino | Top hits/same director:<br>Similar Movie:<br>The Day After Tomorrow<br>2012<br>Similar Movie Image: |

# Copypaste Tag into your HTML



Check RDFa
ThemeRoller Dev T...
Web Data Inspector
Dvd Rental Shop
Share your Web Data presence
Sindice Widget's Composer
http://demo.si...g-data/movie/4

## Sindice Widget Composer

version:0.2

URL → WIDGET → QUERY → WIDGET PARAMS → STYLES → CODE

permalink

Widget code (to be copy/paste to your page):

```
<html><head>
<script src="http://localhost:8080/sindice-widget/js/jquery/js/jquery-1.4.3.min.js" type="text/javascript" ></script>
<script src="http://localhost:8080/sindice-widget/js/jquery/js/jquery.sindice-widget-0.2.js" type="text/javascript" ></script>
<script type="text/javascript">
  $(document).ready(function(){
    $('#my_widget').sindiceWidget({
      widgetName:'demo_get_more_info_about_movie',
      init:{
        attributes:{
          header:'My header',
          footer:'Powered by Sindice',
          width:200,
          type:'more_info'
        },
        params:{
          // URL:string:const param is here to force the widget to use this particular url.
          // remove it when placing widget into your sites
          'URL:string:const':'http://demo.sindice.net/sindice-widget/og-data/movie/4'
        }
      }
    });
  });
</script>
</head><body>
<div id="my_widget"></div>
</body></html>
```

**My header**

**Title:**
The Day After Tomorrow

**RImage:**

**Starring:**
Dash Mihok
Sela Ward
Tamlyn Tomita
Kenneth Welsh
Austin Nichols
Jay O. Sanders
Jake Gyllenhaal
Nestor Serrano
Emmy Rossum
Ian Holm
Dennis Quaid
Glenn Plummer
Perry King

**Writer:**
Roland Emmerich

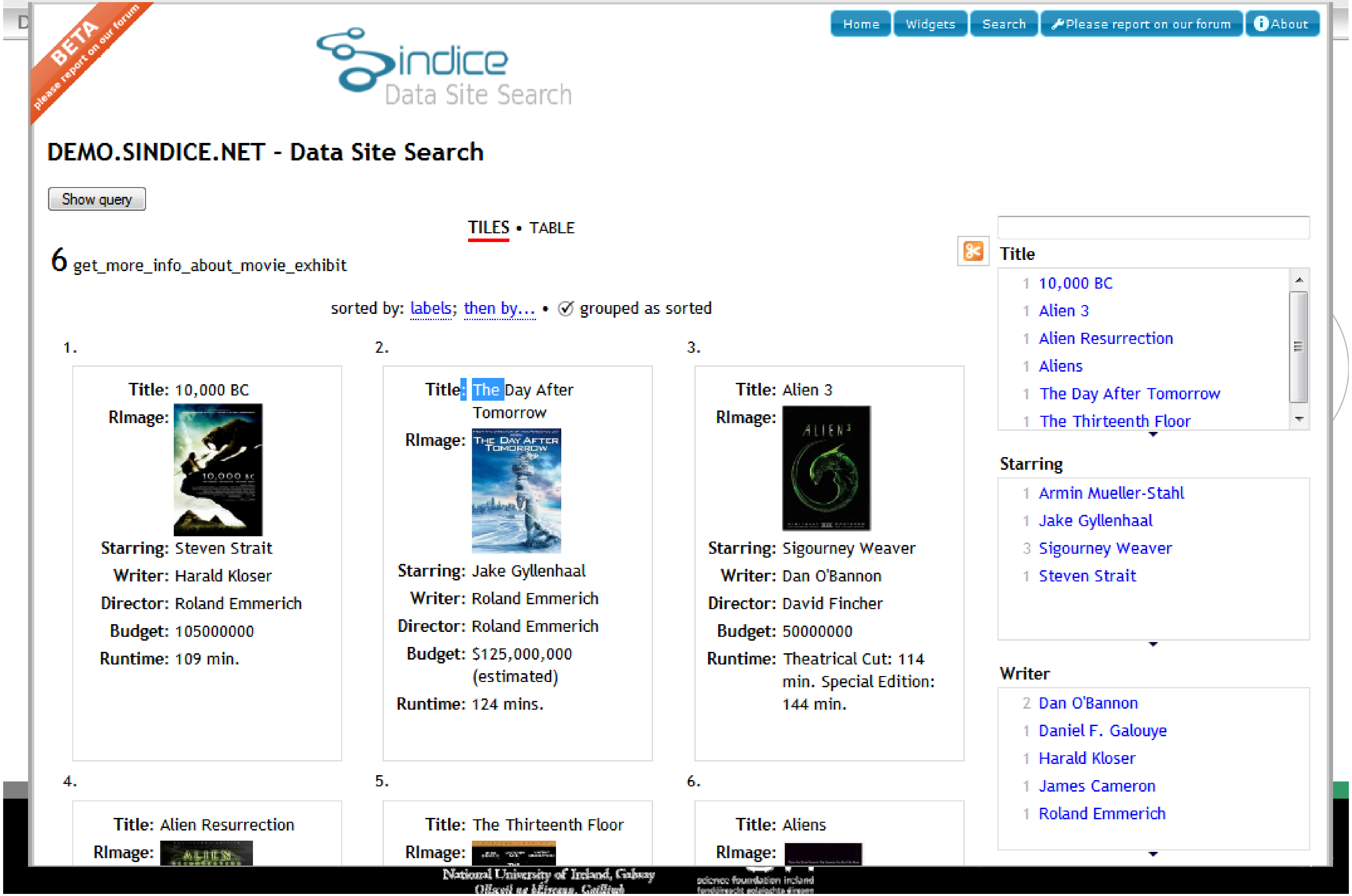
**Director:**
Roland Emmerich

**Budget:**
$125,000,000 (estimated)

**Runtime:**
124 mins.
124.0

Powered by Sindice

# Sindice.com/YOURSITE/movies/search



DEMO.SINDICE.NET - Data Site Search

6 get_more_info_about_movie_exhibit

sorted by: labels; then by... • ✓ grouped as sorted

1.

Title: 10,000 BC
RImage:
Starring: Steven Strait
Writer: Harald Kloser
Director: Roland Emmerich
Budget: 105000000
Runtime: 109 min.

2.

Title: The Day After Tomorrow
RImage:
Starring: Jake Gyllenhaal
Writer: Roland Emmerich
Director: Roland Emmerich
Budget: $125,000,000 (estimated)
Runtime: 124 mins.

3.

Title: Alien 3
RImage:
Starring: Sigourney Weaver
Writer: Dan O'Bannon
Director: David Fincher
Budget: 50000000
Runtime: Theatrical Cut: 114 min. Special Edition: 144 min.

4.

Title: Alien Resurrection
RImage:

5.

Title: The Thirteenth Floor
RImage:

6.

Title: Aliens
RImage:

Title

| | |
|---|---|
| 1 | 10,000 BC |
| 1 | Alien 3 |
| 1 | Alien Resurrection |
| 1 | Aliens |
| 1 | The Day After Tomorrow |
| 1 | The Thirteenth Floor |

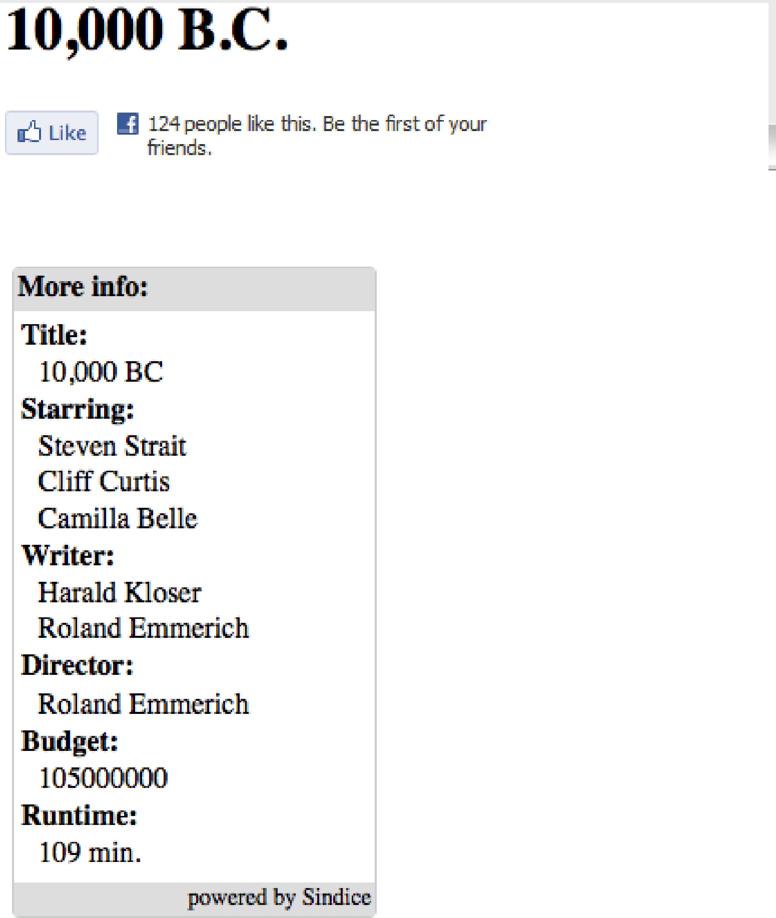
Starring

| | |
|---|---|
| 1 | Armin Mueller-Stahl |
| 1 | Jake Gyllenhaal |
| 3 | Sigourney Weaver |
| 1 | Steven Strait |

Writer

| | |
|---|---|
| 2 | Dan O'Bannon |
| 1 | Daniel F. Galouye |
| 1 | Harald Kloser |
| 1 | James Cameron |
| 1 | Roland Emmerich |

# Enjoy Services



## 10,000 B.C.

124 people like this. Be the first of your friends.

**More info:**

**Title:**
10,000 BC

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

**Writer:**
Harald Kloser
Roland Emmerich

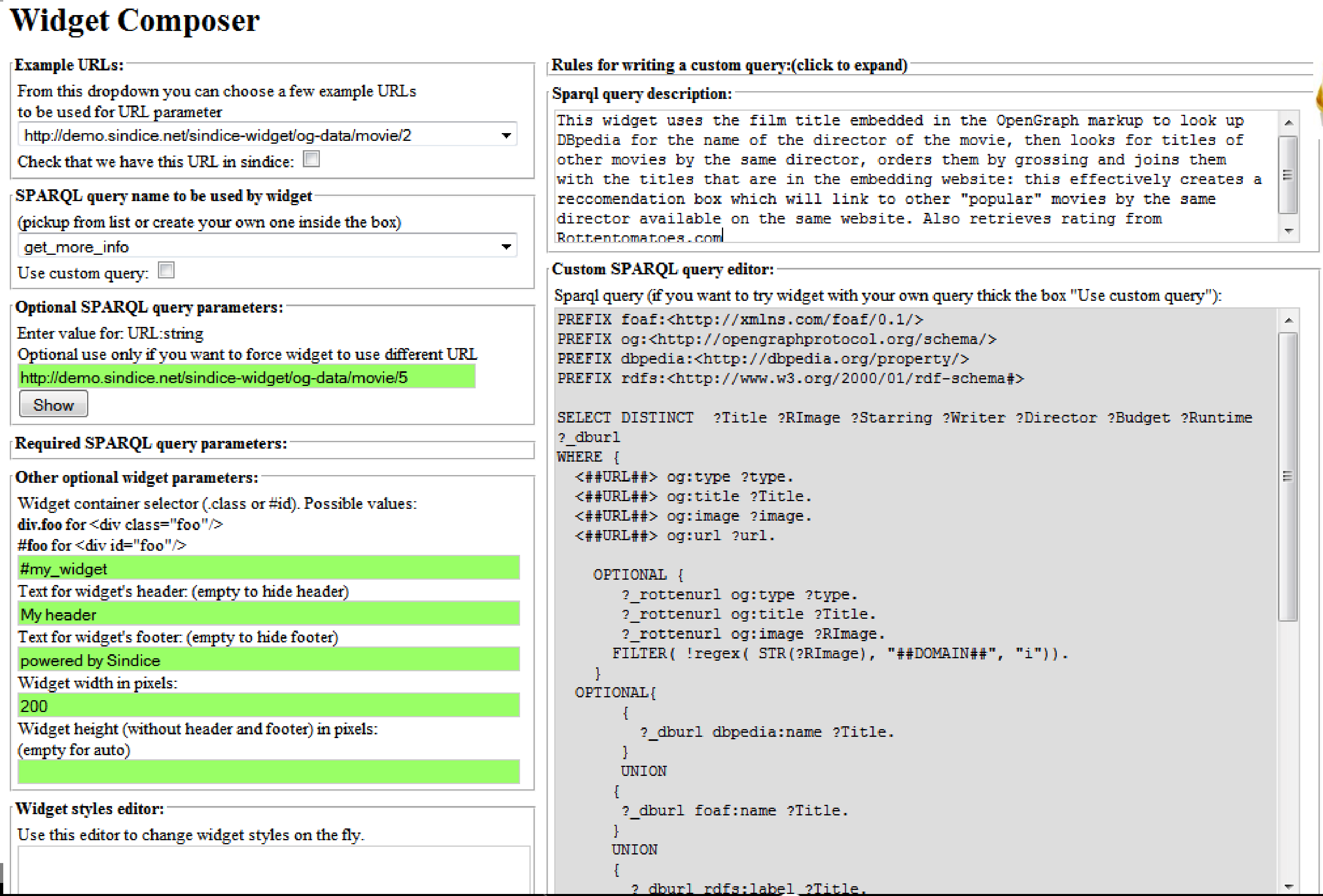
**Director:**
Roland Emmerich

**Budget:**
105000000

**Runtime:**
109 min.

powered by Sindice

# Experts create and $ from widgets



## Widget Composer

**Example URLs:**

From this dropdown you can choose a few example URLs to be used for URL parameter

http://demo.sindice.net/sindice-widget/og-data/movie/2

Check that we have this URL in sindice: □

**SPARQL query name to be used by widget**

(pickup from list or create your own one inside the box)

get_more_info

Use custom query: □

**Optional SPARQL query parameters:**

Enter value for: URL:string

Optional use only if you want to force widget to use different URL

http://demo.sindice.net/sindice-widget/og-data/movie/5

**Required SPARQL query parameters:**

**Other optional widget parameters:**

Widget container selector (.class or #id). Possible values:
div.foo for <div class="foo"/>
#foo for <div id="foo"/>

#my_widget

Text for widget's header: (empty to hide header)

My header

Text for widget's footer: (empty to hide footer)

powered by Sindice

Widget width in pixels:

200

Widget height (without header and footer) in pixels: (empty for auto)

**Widget styles editor:**

Use this editor to change widget styles on the fly.

**Rules for writing a custom query:(click to expand)**

**Sparql query description:**

This widget uses the film title embedded in the OpenGraph markup to look up DBpedia for the name of the director of the movie, then looks for titles of other movies by the same director, orders them by grossing and joins them with the titles that are in the embedding website: this effectively creates a reccomendation box which will link to other "popular" movies by the same director available on the same website. Also retrieves rating from Rottentomatoes.com

**Custom SPARQL query editor:**

Sparql query (if you want to try widget with your own query thick the box "Use custom query"):

```
PREFIX foaf:<http://xmlns.com/foaf/0.1/>
PREFIX og:<http://opengraphprotocol.org/schema/>
PREFIX dbpedia:<http://dbpedia.org/property/>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

SELECT DISTINCT  ?Title ?RImage ?Starring ?Writer ?Director ?Budget ?Runtime
?_dburl
WHERE {
  <##URL##> og:type ?type.
  <##URL##> og:title ?Title.
  <##URL##> og:image ?image.
  <##URL##> og:url ?url.

    OPTIONAL {
       ?_rottenurl og:type ?type.
       ?_rottenurl og:title ?Title.
       ?_rottenurl og:image ?RImage.
      FILTER( !regex( STR(?RImage), "##DOMAIN##", "i")).
    }
  OPTIONAL{
       {
         ?_dburl dbpedia:name ?Title.
       }
       UNION
      {
       ?_dburl foaf:name ?Title.
      }
      UNION
      {
        ?_dburl rdfs:label ?Title.
```

# Crowdsourcing rich annotations

FP7-SME-2010-1
SEMLIB 262301

**Part of**

**SemLib Project**

**Semantic tools for digital libraries**

SemLib - Annotation Demo - YouTube.flv

# Observations

- RDF is very handy not per se a magic bullet
  - Allows “late data integration efforts” but at a high cost: SPARQL esily gets prohibitively expensive
  - Used naively or too directly will allow bloated, unusable data structures
- Large scale Data transformation capabilities come to the rescue. Models:
  - Cloud Powered data pipelines allow free experiementation and data trasformations while the system keeps on running and serving users.
  - An extra dimension of flexibility and preprocessing which eases the task of SPARQL late
- Now in trials with several private entities large enterprises
  - Pharmaceuticals, Publishing

# Thanks