DID YOU BLINK? THE *STRUCTURED WEB* JUST ARRIVED

Michael K. Bergman[1], Coralville, Iowa USA

April 2, 2007

AI3:::Adaptive Information blog

DBpedia is the first and largest source of structured data on the Internet covering topics of general knowledge. You may have not yet heard of DBpedia, but you will. Its name derives from its springboard in Wikipedia. And it is free, growing rapidly and publicly available today.

With DBpedia, you can manipulate and derive facts on more than 1.6 million "things" (people, places, objects, entities). For example, you can easily retrieve a listing of prominent scientists born in the 1870s; or, with virtually no additional effort, a further filtering to all German theoretical physicists born in 1879 who have won a Nobel prize [1]. DBpedia is the first project chosen for showcasing by the Linking Open Data community of the Semantic Web Education and Outreach (SWEO) interest group within the W3C. That community has committed to make portions of other massive data sets — such as the US Census, Geonames, MusicBrainz, WordNet, the DBLP bibliography and many others — interoperable as well.

DBpedia has been unfortunately overlooked in the buzz of the past couple weeks surrounding Freebase. Luminaries such as Esther Dyson, Tim O'Reilly and others have been effusive about the prospects of the pending Freebase offering. And, while, according to O'Reilly Freebase may be "addictive" or from Dyson it may be that "Freebase is a milestone in the journey towards representing meaning in *computers*," those have been hard assertions to judge. Only a few have been invited (I'm not one) to test drive Freebase, now in alpha behind a sign-in screen, and reportedly also based heavily on Wikipedia. On the other hand, DBpedia, released in late January, is open for testing and demos and available today — to *all* [2].

Please don't misunderstand. I'm not trying to pit one service against the other. Both services herald a new era in the *structured Web*, the next leg on the road to the semantic Web. The data from both Freebase and DBpedia are being made freely available under either Creative Commons or the GNU Free Documentation License, respectively. Free and open data access is fortunately not a zero sum game — quite the opposite. Like other truisms regarding the network effects of the Internet, the more data that can be meaningfully intermeshed, the greater the value. Let's wish both services and all of their cousins and progeny much success!

Freebase may prove as important and revolutionary as some of these pundits predict — one never knows. Wikipedia, first released in Jan. 2001 with 270 articles and with only 10 editors, only had a mere 1000 mentions a month by July 2003. Yet today it has more than 1.7 million articles (English version) and is ranked about #10 in overall Web traffic (more here). So, while today Freebase has greater visibility, marketing savvy and buzz than DBpedia, so did virtually every other entity in Jan. 2001 compared to Wikipedia in its infancy. Early buzz is no guarantee of staying power.

What I *do* know is that DBpedia and the catalytic role it is playing in the open data movement is the kind of stuff from which success on the Internet naturally springs. What I also know is that in open source content a community is required to power a promise to its potential. Because of its promise, its open and collaborative approach, and the sheer quality of its information now, DBpedia deserves your and the Web's attention and

[1] Email: mike@mkbergman.com

awareness. But, only time will tell whether DBpedia is able to nurture a community or not and overcome current semantic Web teething problems not of its doing.

First, Some Basics

DBpedia represents data using the Resource Description Framework (RDF) model, as is true for other data sources now available or being contemplated for the semantic Web. Any data representation that uses a “triple” of *subject-predicate-object* can be expressed through the W3C’s standard RDF model. In such triples, subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. (You can think of *subjects* and *objects* as nouns, *predicates* as verbs.) Resources are given a URI (as may also be given to predicates or objects that are not specified with a literal) so that there is a single, unique reference for each item. These lookups can themselves be an individual assertion, an entire specification (as is the case, for example, when referencing the RDF or XML standards), or a complete or partial ontology for some domain or world-view. While the RDF data is often stored and displayed using XML syntax, that is not a requirement. Other RDF forms may include N3 or Turtle syntax, and variants of RDF also exist including RDFa and eRDF (both for embedding RDF in HTML) or more structured representations such as RDF-S (for *schema*; also known as RDFS, RDFs, RDFSchema, etc.).

The absolutely great thing about RDF is how well it lends itself to mapping and mediating concepts from different sources into an unambiguous semantic representation (my ‘glad‘ == your ‘happy‘ OR my ‘glad‘ *is* your ‘glad‘), leading to what Kingsley Idehen calls “data *meshups*“. Further, with additional structure (such as through RDF-S or the various dialects of OWL), drawing inferences and machine reasoning based on the data through more formal ontologies and descriptive logics is also within reach [3].

While these nuances and distinctions are important to the developers and practitioners in the field, they are of little to no interest to actual users [4]. But, fortunately for users, and behind the scenes, practitioners have dozens of converters to get data in whatever form it may exist in the wild such as XML or JSON or a myriad of other formats into RDF (or *vice versa*) [5]. Using such converters for structured data is becoming pretty straightforward. What is now getting more exciting are improved ways to extract structure from semi-structured data and Web pages or to use various information extraction techniques to obtain metadata or named entity structure from within unstructured data. This is what DBpedia did: it converted all of the inherent structure within Wikipedia into RDF, which then makes it manipulable similar to a conventional database.

And, like SQL for conventional databases, SPARQL is now emerging as a leading query framework for RDF-based “triplestores” (that is, the unique form of databases — most often indexed in various ways to improve performance — geared to RDF triples). Moreover, in keeping with the distributed nature of the Internet, distributed SPARQL “endpoints” are emerging, which represent specific data query points at IP nodes, the results of which can then be federated and combined. With the emerging toolset of “RDFizers“, in combination with extraction techniques, such endpoints should soon proliferate. Thus, Web-based data integration models can either follow the data federation approach or the consolidated data warehouse approach or any combination thereof.

The net effect is that the tools and standards now exist such that all data on the Internet can now be structured and combined and analyzed. This is huge. Let me repeat: this is HUGE. And all of us users will only benefit as practitioners continue their labors in the background. The era of the *structured Web* is now upon us.

A Short Intro to DBpedia [6]

The blossoming of Wikipedia has provided a serendipitous starting point to nucleate this emerging “Web of data“. Like Vonnegut’s ice-9, DBpedia‘s RDF representation — freely available for download and extension — offers the prospect to catalyze many semantic Web data

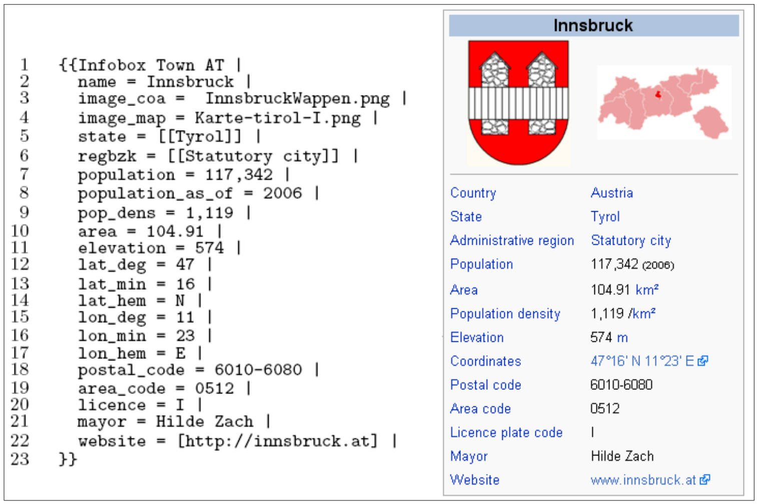
DBpedia.org
Querying Wikipedia like a Database.

sources and tools. (It is also worth study why Freebase, DBpedia, and a related effort called YAGO from the Max Planck Institute for Computer Science [7] all are using Wikipedia as a springboard.) The evidence that the ice crystals are forming comes from the literally hundreds of millions of new "facts" that were committed to be added as part of the open data initiative within days of DBpedia's release (see below).

Wikipedia articles consist mostly of free text, but also contain different types of structured information, such as infobox templates (an especially rich source of structure — see below), categorization information, images, geo-coordinates and links to external Web pages. The extraction methods are described in a paper from DBpedia developers SÃ¶ren Auer and Jens Lehmann. This effort created the initial DBpedia datasets, including extractions in English, German, French, Spanish, Italian, Portuguese, Polish, Swedish, Dutch, Japanese and Chinese.

The current extraction set is about 1.8 GB of zipped files (about 7.5 GB unzipped), consisting of an estimated 1.6 milllion entities, 8,000 relations/properties/predicates, and *91 million "facts"* (generally equated to RDF triples) over about 15 different data files. DBpedia can be downloaded as multiple datasets, with the splits and sizes in (zipped form and number of triples) shown below. Note that each download — available directly from the DBpedia site — may also include multiple language variants:

- *Articles* (217 MB zipped, 5.4 M triples) — Descriptions of all 1.6 million concepts within the English version of Wikipedia including English titles and English short abstracts (max. 500 chars long), thumbnails, links to the corresponding articles in the English Wikipedia. This is the baseline DBpedia file
- *Extended Abstracts* (447 MB, 1.6 M triples) — Additional, extended English abstracts (max. 3000 chars long)
- *External Links* (75 MB, 2.8 M triples) — Links to external web pages about a concept
- *Article Categories* (68 MB, 5.5 M triples) — Links from concepts to categories using the SKOS vocabulary
- *Additional Languages* (147 MB, 4 M triples) — Additional titles, short abstracts and Wikipedia article links in the 10 languages listed above
- *Languages Extended Abstracts* (342 MB, 1.3 M triples) — Extended abstracts in the 10 languages
- *Infoboxes* (116 MB, 9.1 M triples) — Information that has been extracted from Wikipedia infoboxes, an example of which is shown below. There are about three-quarter million English templates with more than 8,000 properties (predicates), with music, animal and plant species, films, cities and books the most prominent types. This category is a major source of the DBpedia structure (but not total content volume):



```
{{Infobox Town AT |
  name = Innsbruck |
  image_coa =  InnsbruckWappen.png |
  image_map = Karte-tirol-I.png |
  state = [[Tyrol]] |
  regbzk = [[Statutory city]] |
  population = 117,342 |
  population_as_of = 2006 |
  pop_dens = 1,119 |
  area = 104.91 |
  elevation = 574 |
  lat_deg = 47 |
  lat_min = 16 |
  lat_hem = N |
  lon_deg = 11 |
  lon_min = 23 |
  lon_hem = E |
  postal_code = 6010-6080 |
  area_code = 0512 |
  licence = I |
  mayor = Hilde Zach |
  website = [http://innsbruck.at] |
}}
```

Innsbruck

Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km²
Population density	1,119 /km²
Elevation	574 m
Coordinates	47°16′ N 11°23′ E
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at

[*Click on image for full-size pop-up*]

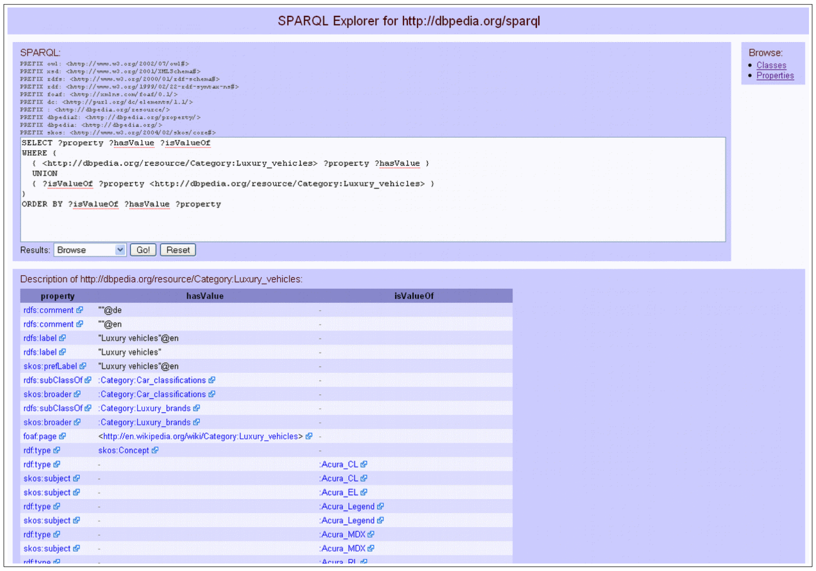
- *Categories* (10 MB, 900 K triples) — Information regarding which concepts are a category and how categories are related
- *Persons* (4 MB, 437 K triples) — Information on about 58,880 persons (date and place of birth etc.) represented using the FOAF vocabulary
- *Links to Geonames* (2 MB, 213 K triples) — Links between geographic places in DBpedia and data about them in the Geonames database
- *Location Types* (500 kb, 65 K triples) — `rdf:type` Statements for geographic locations
- *Links to DBLP* (10 kB, 200 K triples) — Links between computer scientists in DBpedia and their publications in the DBLP database
- *Links to the RDF Book Mashup* (100 kB, 9 K triples) — Links between books in DBpedia and data about them provided by the RDF Book Mashup
- *Page Links* (335 MB, around 60 M triples) — Dataset containing internal links between DBpedia instances. The dataset was created from the internal pagelinks between Wikipedia articles, potentially useful for structural analysis or data mining.

Besides direct download, DBpedia is also available via a public SPARQL endpoint at http://dbpedia.org/sparql and via various query utilities (see next).

Querying and Interacting with DBpedia

Generic RDF (semantic Web) browsers like Disco, Tabulator or the OpenLink Data Web Browser [8] can be used to navigate, filter and query data across data sources. A couple of these browsers, plus endpoints and online query services, are used below to illustrate how you can interact with the DBpedia datasets. However, that being said, it is likely that first-time users will have difficulty using portions of this infrastructure without further guidance. While data preparation and exposure via RDF standards is progressing nicely, the user interfaces, documentation, intuitive presentation, and general knowledge for standard users to use this infrastructure are lagging. Let's show this with some examples, progressing from what might be called practitioner tools and interfaces to those geared more to standard users [9].

The first option we might try is to query the data directly through DBpedia's SPARQL endpoint [10]. But, since that is generally used directly by remote agents, we will instead access the endpoint via a SPARQL viewer, as shown by this example looking for "luxury cars":



SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?property ?hasValue ?isValueOf
WHERE {
  { <http://dbpedia.org/resource/Category:Luxury_vehicles> ?property ?hasValue }
  UNION
  { ?isValueOf ?property <http://dbpedia.org/resource/Category:Luxury_vehicles> }
}
ORDER BY ?isValueOf ?hasValue ?property
```

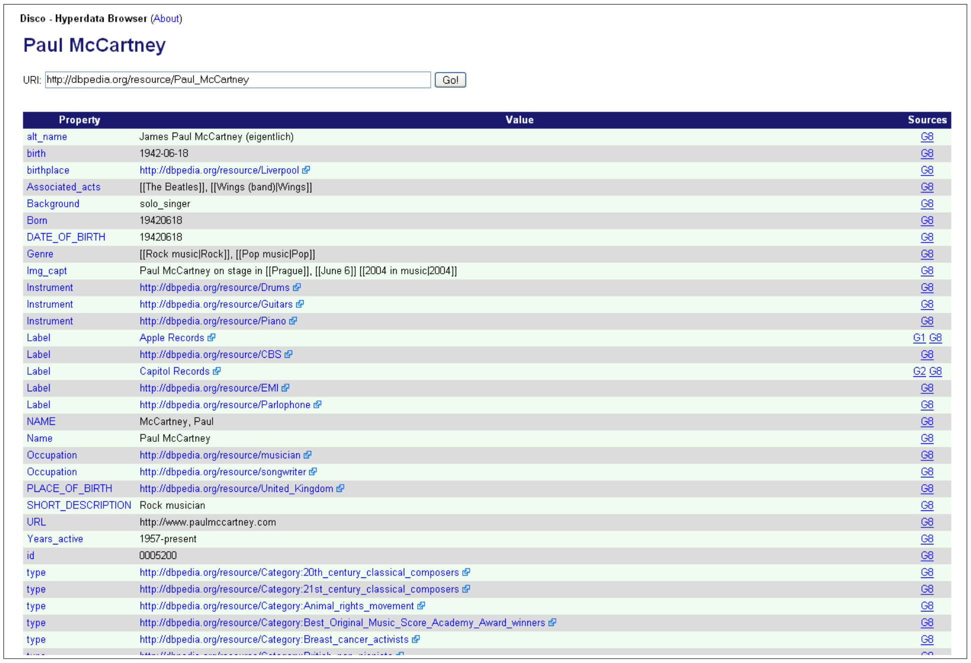
Description of http://dbpedia.org/resource/Category:Luxury_vehicles:

property	hasValue	isValueOf
rdfs:comment	""@de	-
rdfs:comment	""@en	-
rdfs:label	"Luxury vehicles"@en	-
rdfs:label	"Luxury vehicles"	-
skos:prefLabel	"Luxury vehicles"@en	-
rdfs:subClassOf	:Category:Car_classifications	-
skos:broader	:Category:Car_classifications	-
rdfs:subClassOf	:Category:Luxury_brands	-
skos:broader	:Category:Luxury_brands	-
foaf:page	<http://en.wikipedia.org/wiki/Category:Luxury_vehicles>	-
rdf:type	skos:Concept	-
rdf:type	-	:Acura_CL
skos:subject	-	:Acura_CL
skos:subject	-	:Acura_EL
rdf:type	-	:Acura_Legend
skos:subject	-	:Acura_Legend
rdf:type	-	:Acura_MDX
skos:subject	-	:Acura_MDX
rdf:type	-	:Acura_RL

[*Click on image for full-size pop-up*]

We observe a couple of things using this viewer. First, the query itself requires using the SPARQL syntax. And, the results are presented in a linked listing of RDF triples.

For practitioners, this is natural and straightforward. Indeed, other standard viewers, now using Disco as the example, also present results in the same tabular triple format (in this case for “Paul McCartney”):



Disco - Hyperdata Browser (About)

Paul McCartney

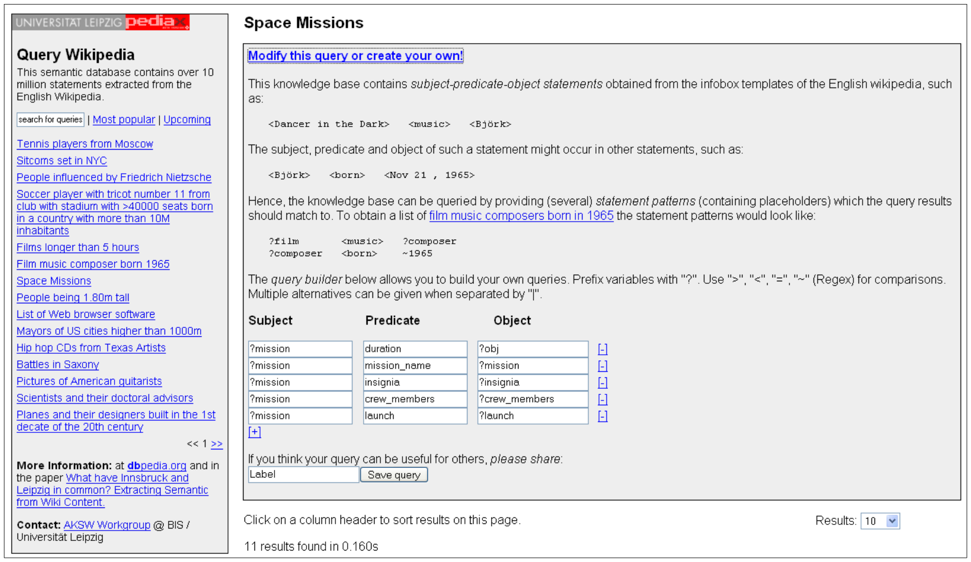
URI: http://dbpedia.org/resource/Paul_McCartney Go!

Property	Value	Sources
alt_name	James Paul McCartney (eigentlich)	G8
birth	1942-06-18	G8
birthplace	http://dbpedia.org/resource/Liverpool	G8
Associated_acts	[[The Beatles]], [[Wings (band)|Wings]]	G8
Background	solo_singer	G8
Born	19420618	G8
DATE_OF_BIRTH	19420618	G8
Genre	[[Rock music|Rock]], [[Pop music|Pop]]	G8
Img_capt	Paul McCartney on stage in [[Prague]], [[June 6]] [[2004 in music|2004]]	G8
Instrument	http://dbpedia.org/resource/Drums	G8
Instrument	http://dbpedia.org/resource/Guitars	G8
Instrument	http://dbpedia.org/resource/Piano	G8
Label	Apple Records	G1 G8
Label	http://dbpedia.org/resource/CBS	G8
Label	Capitol Records	G2 G8
Label	http://dbpedia.org/resource/EMI	G8
Label	http://dbpedia.org/resource/Parlophone	G8
NAME	McCartney, Paul	G8
Name	Paul McCartney	G8
Occupation	http://dbpedia.org/resource/musician	G8
Occupation	http://dbpedia.org/resource/songwriter	G8
PLACE_OF_BIRTH	http://dbpedia.org/resource/United_Kingdom	G8
SHORT_DESCRIPTION	Rock musician	G8
URL	http://www.paulmccartney.com	G8
Years_active	1957-present	G8
id	0005200	G8
type	http://dbpedia.org/resource/Category:20th_century_classical_composers	G8
type	http://dbpedia.org/resource/Category:21st_century_classical_composers	G8
type	http://dbpedia.org/resource/Category:Animal_rights_movement	G8
type	http://dbpedia.org/resource/Category:Best_Original_Music_Score_Academy_Award_winners	G8
type	http://dbpedia.org/resource/Category:Breast_cancer_activists	G8

[*Click on image for full-size pop-up*]

While OK for practitioners, these tools pose some challenges for standard users. First, one must know the SPARQL syntax with its SELECT statement and triples specifications. Second, *where* the resources exist (the URIs) and the specific names of the relations (predicates) must generally be known or looked up in a separate step. And, third, of course, the tabular triples result display is not terribly attractive and may not be sufficiently informative (since results are often the links to the resource, rather than the actual content of that resource).

These limits are well-known in the community; indeed, they are a reflection of RDF’s roots in machine-readable data. However, it is also becoming apparent that humans need to inspect and interact with this stuff as well, leading to consequent attempts to make interfaces more attractive and intuitive. One of the first to do so for DBpedia is the UniversitÃ¤t Leipzig’s Query Wikipedia, an example of an online query service. In this first screen, we see that the user is shielded from much of the SPARQL syntax via the three-part (*subject-predicate-object*) query entry form:



Query Wikipedia

This semantic database contains over 10 million statements extracted from the English Wikipedia.

search for queries | Most popular | Upcoming

Tennis players from Moscow
Sitcoms set in NYC
People influenced by Friedrich Nietzsche
Soccer player with tricot number 11 from club with stadium with >40000 seats born in a country with more than 10M inhabitants
Films longer than 5 hours
Film music composer born 1965
Space Missions
People being 1.80m tall
List of Web browser software
Mayors of US cities higher than 1000m
Hip hop CDs from Texas Artists
Battles in Saxony
Pictures of American guitarists
Scientists and their doctoral advisors
Planes and their designers built in the 1st decade of the 20th century

<< 1 >>

More Information: at **db**pedia.org and in the paper What have Innsbruck and Leipzig in common? Extracting Semantic from Wiki Content.

Contact: AKSW Workgroup @ BIS / Universität Leipzig

Space Missions

Modify this query or create your own!

This knowledge base contains *subject-predicate-object statements* obtained from the infobox templates of the English wikipedia, such as:

```
<Dancer in the Dark>   <music>   <Björk>
```

The subject, predicate and object of such a statement might occur in other statements, such as:

```
<Björk>   <born>   <Nov 21 , 1965>
```

Hence, the knowledge base can be queried by providing (several) *statement patterns* (containing placeholders) which the query results should match to. To obtain a list of film music composers born in 1965 the statement patterns would look like:

```
?film       <music>   ?composer
?composer   <born>    ~1965
```

The *query builder* below allows you to build your own queries. Prefix variables with "?". Use ">", "<", "=", "~" (Regex) for comparisons. Multiple alternatives can be given when separated by "|".

Subject	Predicate	Object	
?mission	duration	?obj	[-]
?mission	mission_name	?mission	[-]
?mission	insignia	?insignia	[-]
?mission	crew_members	?crew_members	[-]
?mission	launch	?launch	[-]

[+]

If you think your query can be useful for others, *please share*:

Label | Save query

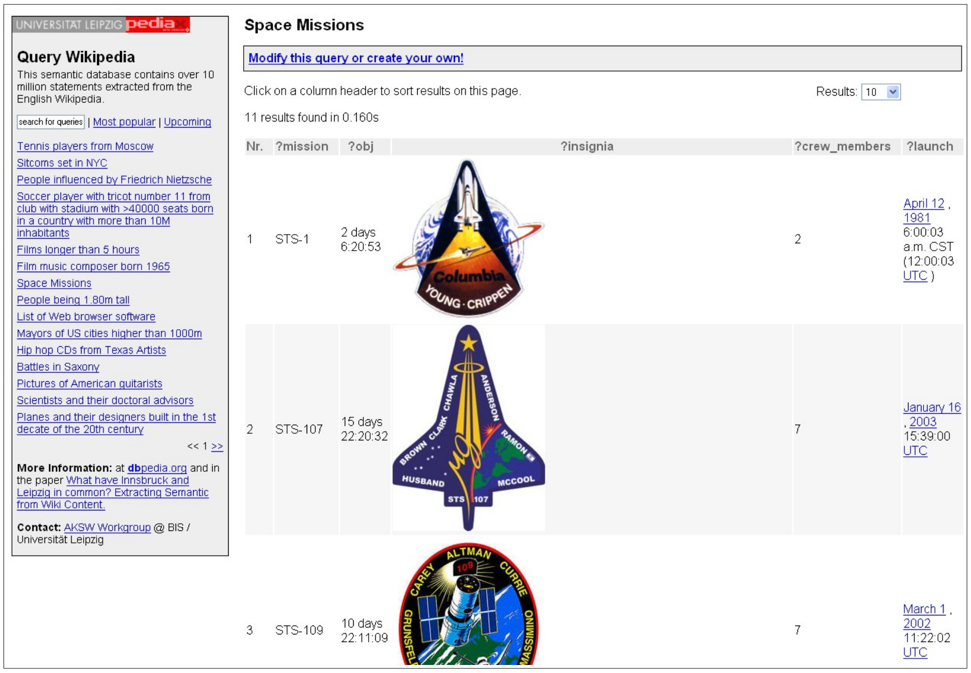
Click on a column header to sort results on this page.

Results: 10

11 results found in 0.160s

[*Click on image for full-size pop-up*]

We also see that the results can be presented in a more attractive form including image thumbnails:



Query Wikipedia

This semantic database contains over 10 million statements extracted from the English Wikipedia.

search for queries | Most popular | Upcoming

Tennis players from Moscow
Sitcoms set in NYC
People influenced by Friedrich Nietzsche
Soccer player with tricot number 11 from club with stadium with >40000 seats born in a country with more than 10M inhabitants
Films longer than 5 hours
Film music composer born 1965
Space Missions
People being 1.80m tall
List of Web browser software
Mayors of US cities higher than 1000m
Hip hop CDs from Texas Artists
Battles in Saxony
Pictures of American guitarists
Scientists and their doctoral advisors
Planes and their designers built in the 1st decade of the 20th century

<< 1 >>

More Information: at **db**pedia.org and in the paper What have Innsbruck and Leipzig in common? Extracting Semantic from Wiki Content.

Contact: AKSW Workgroup @ BIS / Universität Leipzig

Space Missions

Modify this query or create your own!

Click on a column header to sort results on this page.

Results: 10

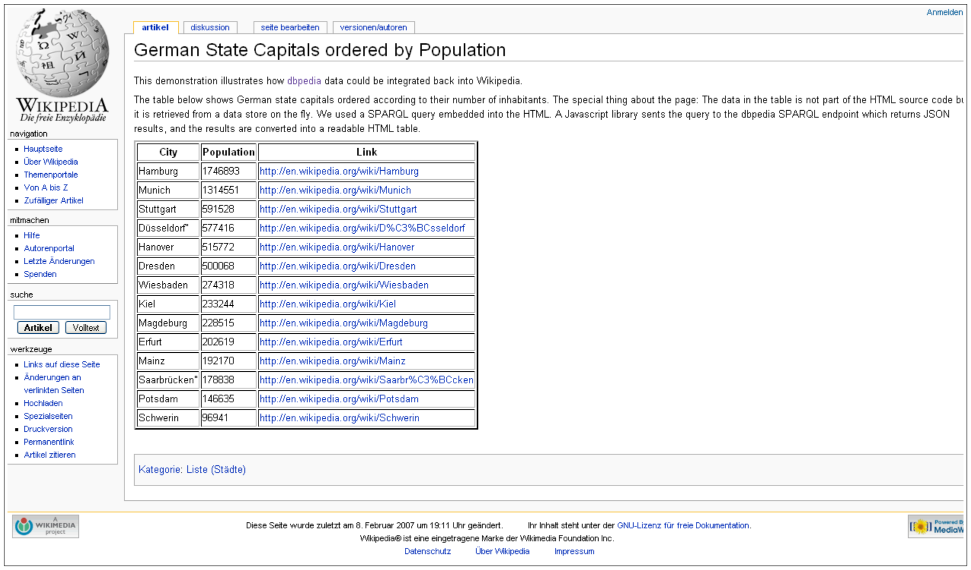
11 results found in 0.160s

Nr.	?mission	?obj	?insignia	?crew_members	?launch
1	STS-1	2 days 6:20:53		2	April 12 , 1981 6:00:03 a.m. CST (12:00:03 UTC)
2	STS-107	15 days 22:20:32		7	January 16 , 2003 15:39:00 UTC
3	STS-109	10 days 22:11:09		7	March 1 , 2002 11:22:02 UTC

[*Click on image for full-size pop-up*]

Use of dropdown lists could help this design still further. Better tools and interface design is an active area of research and innovation (see below).

However, remember the results of such data queries are themselves machine readable, which of course means that the results can be embedded into existing pages and frameworks or combined ("*meshed up*") with still other data in still other presentation and visualization frameworks [11]. For example, here is one DBpedia results set on German state capitals presented in the familiar Wikipedia (actually, MediaWiki) format:



German State Capitals ordered by Population

This demonstration illustrates how dbpedia data could be integrated back into Wikipedia.

The table below shows German state capitals ordered according to their number of inhabitants. The special thing about the page: The data in the table is not part of the HTML source code bu it is retrieved from a data store on the fly. We used a SPARQL query embedded into the HTML. A Javascript library sents the query to the dbpedia SPARQL endpoint which returns JSON results, and the results are converted into a readable HTML table.

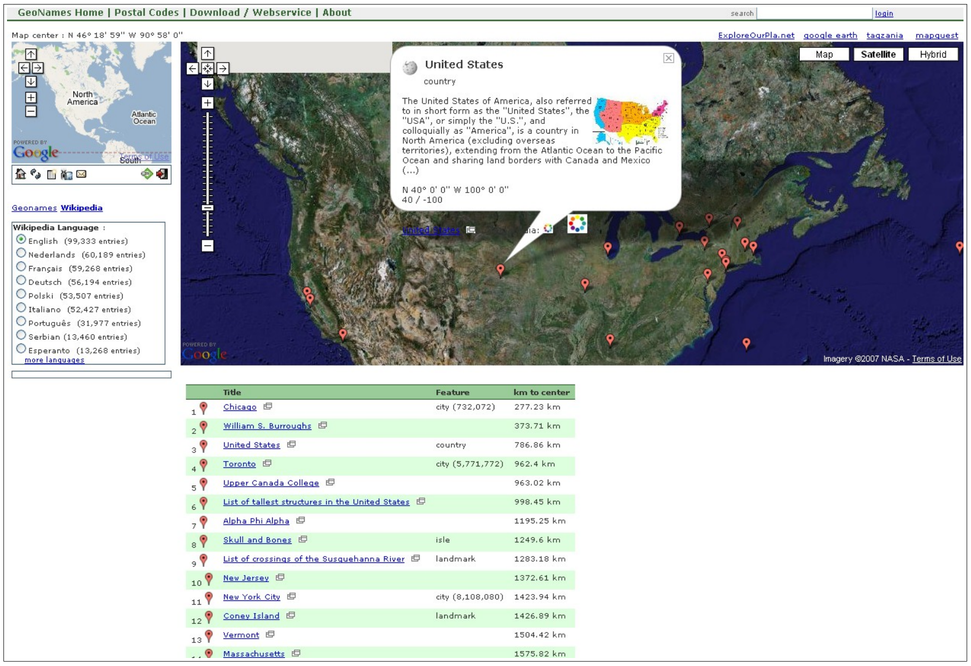
City	Population	Link
Hamburg	1746893	http://en.wikipedia.org/wiki/Hamburg
Munich	1314551	http://en.wikipedia.org/wiki/Munich
Stuttgart	591528	http://en.wikipedia.org/wiki/Stuttgart
Düsseldorf"	577416	http://en.wikipedia.org/wiki/D%C3%BCsseldorf
Hanover	515772	http://en.wikipedia.org/wiki/Hanover
Dresden	500068	http://en.wikipedia.org/wiki/Dresden
Wiesbaden	274318	http://en.wikipedia.org/wiki/Wiesbaden
Kiel	233244	http://en.wikipedia.org/wiki/Kiel
Magdeburg	228515	http://en.wikipedia.org/wiki/Magdeburg
Erfurt	202619	http://en.wikipedia.org/wiki/Erfurt
Mainz	192170	http://en.wikipedia.org/wiki/Mainz
Saarbrücken"	178838	http://en.wikipedia.org/wiki/Saarbr%C3%BCcken
Potsdam	146635	http://en.wikipedia.org/wiki/Potsdam
Schwerin	96941	http://en.wikipedia.org/wiki/Schwerin

Kategorie: Liste (Städte)

Diese Seite wurde zuletzt am 8. Februar 2007 um 19:11 Uhr geändert. Ihr Inhalt steht unter der GNU-Lizenz für freie Dokumentation.
Wikipedia® ist eine eingetragene Marke der Wikimedia Foundation Inc.
Datenschutz Über Wikipedia Impressum

[*Click on image for full-size pop-up*]

Or, here is a representative example of what DBpedia data might look like when the addition of the planned Geonames dataset is completed:



Map center : N 46° 18' 59" W 90° 58' 0"

United States

country

The United States of America, also referred to in short form as the "United States", the "USA", or simply the "U.S.", and colloquially as "America", is a country in North America (excluding overseas territories), extending from the Atlantic Ocean to the Pacific Ocean and sharing land borders with Canada and Mexico (...)

N 40° 0' 0" W 100° 0' 0"
40 / -100

Wikipedia Language :
- ● English (99,333 entries)
- ○ Nederlands (60,189 entries)
- ○ Français (59,268 entries)
- ○ Deutsch (56,194 entries)
- ○ Polski (53,507 entries)
- ○ Italiano (52,427 entries)
- ○ Português (31,977 entries)
- ○ Serbian (13,460 entries)
- ○ Esperanto (13,268 entries)

more languages

	Title	Feature	km to center
1	Chicago	city (732,072)	277.23 km
2	William S. Burroughs		373.71 km
3	United States	country	786.86 km
4	Toronto	city (5,771,772)	962.4 km
5	Upper Canada College		963.02 km
6	List of tallest structures in the United States		998.45 km
7	Alpha Phi Alpha		1195.25 km
8	Skull and Bones	isle	1249.6 km
9	List of crossings of the Susquehanna River	landmark	1283.18 km
10	New Jersey		1372.61 km
11	New York City	city (8,108,080)	1423.94 km
12	Coney Island	landmark	1426.89 km
13	Vermont		1504.42 km
	Massachusetts		1575.82 km

[*Click on image for full-size pop-up*]

Some of these examples, indeed because of the value of the large-scale RDF data that DBpedia now brings, begin to glaringly expose prior weaknesses in tools and user interfaces that were hidden when only applied to toy datasets. With the massive expansion to still additional datasets, plus the interface innovations of Freebase and Yahoo! Pipes and elsewhere, we should see rapid improvements in presentation and usability.

Extensibility and Relation to Open Data

In addition to the Wikipedia, DBLP bibliography and RDF book mashup data already in DBpedia, significant commitments to new datasets and tools have been quick to come, including the addition of full-text search [12], with many more candidate datasets also identified. A key trigger for this interest was the acceptance of the Linking Open Data project, itself an outgrowth of DBpedia proposed by Chris Bizer and Richard Cyganiak, as one of the kick-off community projects of the W3C's Semantic Web Education and Outreach (SWEO) interest group. Some of the new datasets presently being integrated into the system — with many notable advocates standing behind them — include:

- Geonames data, including its ontology and 6 million geographical places and features, including implementation of RDF data services
- 700 million triples of U.S. Census data from Josh Tauberer
- Revyu.com, the RDF-based reviewing and rating site, including its links to FOAF, the Review Vocab and Richard Newman's Tag Ontology
- The "RDFizers" from MIT Simile Project (not to mention other tools), plus 50 million triples from the MIT Libraries catalog covering about a million books
- GEMET, the GEneral Multilingual Environmental Thesaurus of the European Environment Agency
- And, WordNet through the YAGO project, and its potential for an improved hierarchical ontology to the Wikipedia data.

Additional candidate datasets of interest have also been identified by the SWEO interest group and can be found on this page:

- Gene Ontology database and its 6 million annotations
- Gene fruitfly embryogenesis images from the Berkeley Drosophila Genome Project
- Roller Blog entries using the Atom/OWL vocabulary
- Various semantic Web interest group and conference materials
- Various FOAF-enabled profiles
- The UniProt protein database with its 300 million triples
- OpenGuides, a network of wiki-based city guides
- dbtune, an RDF-enabled version of the Magnatune music database using the Music Ontology
- The SKOS Data Zone
- Other MIT Simile data collections, including the CIA's World Factbook, Library of Congress' Thesaurus of Graphic Materials, National Cancer Institute's cancer thesaurus, W3C's technical reports
- The RDF version of the DMOZ Open Directory Project
- GovTrack.us of U.S. Congress legislator and voting records
- Chef Moz restaurant and review guides from DMOZ
- DOAP Store and its DOAP project descriptions, and
- MusicBrainz.

Note the richness within music and genetics, two domains that have been early (among others not yet listed) to embrace semantic Web approaches. Clearly, this listing, itself only a few weeks old, will grow rapidly. And, as has been noted, the availability of a wide variety of RDFizers and other tools should cause such open data to continue to grow explosively.

There are also tremendous possibilities for different presentation formats of the results data, as the MediaWiki and Geonames examples above showed. Presentation options include calendars, timelines, lightweight publishing systems (such as Exhibit, which already has a DBpedia example), maps, data visualizations, PDAs and mobile devices, notebooks and annotation systems.

Current Back-end Infrastructure

The online version of DBpedia and its SPARQL endpoint is managed by the open source Virtuoso middleware on a server provided by OpenLink Software. This software also hosts the various third-party browsers and query utilities noted above. These OpenLink programs themselves are some of the more remarkable tools available to the semantic Web community, are open source, and are also very much "under the radar."

Just as Wikipedia helps provide the content grist for springboarding the *structured Web*, such middleware and conversion tools are another part of the new magic essential to what is now being achieved via DBpedia and Freebase. I will be discussing further the impressive OpenLink Software tools and back-end infrastructure in detail in an upcoming posting.

Structured Data: Foundation to the Road of the Semantic Web

We thus have an exciting story of a new era — the *structured Web* — arising from the RDF exposure of large-scale structured data and its accompanying infrastructure of converters, middleware and datastores. This *structured Web* provides the very foundational roadbed underlying the semantic Web.

Yet, at the same time as we begin traveling portions of this road, we can also see more clearly some of the rough patches in the tools and interfaces needed by non-practitioners. I suspect much of the excitement deriving from both Yahoo! Pipes and Freebase comes from the fact that their interfaces are "fun" and "addictive." (Excitement also comes from allowing users with community rights to mold ontologies or to add structured content of their own.) Similar innovations will be needed to smooth over other rough spots in query services and use of the SPARQL language, as well as in lightweight forms of results and dataset presentation. And still even greater challenges in mediating semantic heterogeneities lie much further down this road, which is a topic for another day.

I suspect this new era of the *structured Web* also signals other transitions and changes. Practitioners and the researchers who have long labored in the laboratories of the semantic Web need to get used to business types, marketers and promoters, and (even) the general public crowding into the room and jostling the elbows. Explication, documentation and popularization will become more important; artists and designers need to join the party. While we've only seen it in limited measure to date, venture interest and money will also begin flooding into the room, changing the dynamics and the future in unforeseeable ways. I suspect many who have worked the hardest to bring about this exciting era may look back ruefully and wonder why they ever despaired that the broader world didn't "get it" and why it was taking so long for the self-evident truth of the semantic Web to become real. It now is.

But, like all Thermidorean reactions, this, too, shall pass. Whether we have to call it "Web 3.0" or "Web 4.0" or even "Web 98.6", or any other such silliness for a period of time in order to broaden its appeal, we should accept this as the price of progress and do so. We really have the standards and tools at hand; we now need to get out front as quickly as possible to get the RDFized data available. We are at the tipping point for accelerating the network effects deriving from structured data. With just a few more blinks of the eye, the semantic Web will have occurred before we know it.

DBpedia Is A Winner

The DBpedia team of Chris Bizer, Richard Cyganiak and Georgi Kobilarov (Freie UniversitÃ¤t Berlin), of Sören Auer, Jens Lehmann and Jörg Schüppel (Universität Leipzig), and of Orri Erling and Kingsley Idehen (OpenLink Software) all deserve our thanks for showing how "RDFizing" available data and making it accessible as a SPARQL endpoint can move from toy semantic Web demos to actually useful data. Chris and Richard also deserve kudos for proposing the Linking Open Data initiative with DBpedia as its starting nucleus.

Acknowledgements

This article was originally posted on the *AI3:::Adaptive Information* Web site at http://www.mkbergman.com/354/did-you-blink-the-structured-web-just-arrived/. This version has been edited and reformatted slightly for PDF distribution. We thank Cognonto Corporation for making this content freely available.

[1] BTW, the answer is Albert Einstein and Max von Laue; another half dozen German physicists received Nobels within the surrounding decade.

[2] Moreover, just for historical accuracy, DBpedia was also the first released, being announced on January 23, 2007.

[3] It is this potential ultimate vision of autonomous "intelligent" agents or bots working on our behalf getting and collating relevant information that many confusedly equate to what is meant to constitute the "Semantic Web" (both capitalized). While the vision is useful and reachable, most semantic Web researchers understand the semantic Web to be a basic infrastructure, followed by an ongoing process to publish and harness more and more data in machine-processable ways. In this sense, the "semantic Web" can be seen more as a *journey* than a *destination*.

[4] Like plumbing, the semantic Web is best hidden and only important as a means to an end. The originators of Freebase and other entities beginning to tap into this infrastructure intuitively understand this. Some of the challenges of education and outreach around the semantic Web are to emphasize benefits and delivered results rather than describing the washers, fittings and sweated joints of this plumbing.

[5] A great listing of such "RDFizers" can be found on MIT's Simile Web site.

[6] This introduction to DBpedia borrows liberally from its own Web site; please consult it for more up-to-date, complete, and accurate details.

[7] YAGO ("yet another great ontology") was developed by Fabian M. Suchanek, Gjergji Kasneci and Gerhard Weikum at the Max-Plack-Institute for Computer Science at Saarbrücken University. A WWW 2007 paper describing the project is available, plus there is an online query point for the YAGO service and the data can be downloaded. YAGO combines Wikipedia and WordNet to derive a rich hierarchical ontology with high precision and recall. The system contains only 14 relations (predicates), but they are more specific and not directly comparable to the DBpedia properties. YAGO also uses a slightly different data model, convertible to RDF, that has some other interesting aspects.

[8] See also this **AI3** blog's **Sweet Tools** and sort by browser to see additional candidates.

[9] A nice variety of other query examples in relation to various tools is available from the DBpedia site.

[10] According to the Ontoworld wiki, "A SPARQL endpoint is a conformant SPARQL protocol service as defined in the SPROT specification. A SPARQL endpoint enables users (human or other) to query a knowledge base via the SPARQL language. Results are typically returned in one or more machine-processable formats. Therefore, a SPARQL endpoint is mostly conceived as a machine-friendly interface towards a knowledge base."

[11] Please refer to the Developers Guide to Semantic Web Toolkits or **Sweet Tools** to find a development toolkit in your preferred programming language to process DBpedia data.

[12] Actually, the pace of this data increase is amazing. For example, in the short couple of days I was working on this write-up, I received an email from Kingsley Idehen noting that OpenLink Software had created another access portal to the data that now included DBpedia, 2005 US Census Data, all of the data crawled by PingTheSemanticWeb, and Musicbrainz (see above), plus full-text indexing of same! BTW, this broader release can be found at http://dbpedia2.openlinksw.com:8890/isparql. Whew! It's hard to keep up.