Linked Open Data Cloud

John P. McCrae, Thierry Declerck
Hitchhiker’s guide to the Linked Open Data Cloud
DBpedia

- Largest node in the linked open data cloud
  - “Nucleus for a web of open data”
- Most data is derived by parsing Wikipedia
  - E.g., https://en.wikipedia.org/wiki/C++
    - => http://dbpedia.org/resource/C++
- Uses transparent content negotiation
I want to know about “C++” and I understand RDF and HTML

curl -H "Accept: application/rdf+xml;text/html" -I \ http://dbpedia.org/resource/C++

HTTP/1.1 303 See Other
... Location: http://dbpedia.org/data/C++.xml
...
Transparent content negotiation

I want to know about “C++” and I only know HTML


Go to this location for the HTML version

HTTP/1.1 303 See Other
... Location: http://dbpedia.org/page/C++ ...

Use /resource/ URL to refer to concept
C++ (pronounced cee plus plus, /ˈsiː pləs pləs/) is a general-purpose programming language. It has imperative, object-oriented and generic programming features, while also providing facilities for low-level memory manipulation. Many other programming languages have been influenced by C++, including C#, D, Java, and newer versions of C (after 1998).

**Ontology properties**

- **dbo.abstract**
  - C++ (pronounced cee plus plus, /ˈsiː pləs pləs/) is a general-purpose programming language. It has imperative, object-oriented and generic programming features, while also providing facilities for low-level memory manipulation. It was designed with a bias toward system programming and embedded, resource-constrained and large systems, with performance, efficiency and flexibility of use as its design highlights. C++ has also been found useful in many other contexts, with key strengths being software infrastructure and resource-constrained applications, including desktop applications, servers (e.g. e-commerce, web search or SQL servers), and performance-critical applications (e.g. telephone switches or space probes). C++ is a compiled language, with implementations of it available on many platforms and provided by various organizations, including the Free Software Foundation (FSF's GCC), LLVM, Microsoft, Intel and IBM. C++ is standardized by the International Organization for Standardization (ISO), with the latest standard version ratified and published by ISO in December 2014 as :2014 (informally known as C++14). The C++ programming language was initially standardized in 1998 as ISO/IEC 14882:1998, which was then amended by the C++03, ISO/IEC 14882:2003, standard. The current C++14 standard supersedes these and C++11, with new features and an enlarged. Before the initial standardization in 1998, C++ was developed by Bjarne Stroustrup at Bell Labs since 1979, as an extension of the C language as he wanted an efficient and flexible language similar to C, which also provided high-level features for program organization. Many other programming languages have been influenced by C++, including C#, D, Java, and newer versions of C (after 1998). (en)

- **dbo.designer**
  - dbr:Bjarne_Stroustrup

- **dbo.influenced**
  - dbr:Lua_(programming_language)
  - dbr:Rust_(programming_language)
  - dbp:C99
### About: designer

An Entity of Type: Property, from Named Graph: http://dbpedia.org/resource/classes#, within Data Space: dbpedia.org

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdf:type</td>
<td>rdfs:Property</td>
</tr>
<tr>
<td></td>
<td>owl:ObjectProperty</td>
</tr>
<tr>
<td>rdfs:isDefinedBy</td>
<td><a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/</a></td>
</tr>
<tr>
<td>rdfs:label</td>
<td>designer (en)</td>
</tr>
<tr>
<td>rdfs:subPropertyOf</td>
<td>dbo:Person</td>
</tr>
<tr>
<td>owl:equivalentProperty</td>
<td>wikidata:P287</td>
</tr>
<tr>
<td>wdrs:describedby</td>
<td>dbo:data/definitions.ttl</td>
</tr>
<tr>
<td>is</td>
<td><a href="http://open.vocab.org/terms/defines_of">http://open.vocab.org/terms/defines_of</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://dbpedia.org/ontology/">http://dbpedia.org/ontology/</a></td>
</tr>
<tr>
<td>is</td>
<td><a href="http://open.vocab.org/terms/describes_of">http://open.vocab.org/terms/describes_of</a></td>
</tr>
</tbody>
</table>
C++ (Q2407)

general purpose high-level programming language
cpp | ISO/IEC 14882 | C Plus Plus

In more languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Label</th>
<th>Description</th>
<th>Also known as</th>
</tr>
</thead>
</table>
| English  | C++   | general purpose high-level programming language | cpp
|          |       |             | ISO/IEC 14882
|          |       |             | C Plus Plus          |
| German   | C++   | objektorientierte Programmiersprache | CPP
|          |       |             | Cplusplus
|          |       |             | C plus plus          |
|          |       |             | C-plus-plus          |
|          |       |             | C-Plusplus          |
|          |       |             | C++0x                |
|          |       |             | CXX                  |
| Irish    | C++   | No description defined | Cpp
| French   | C++   | langage de programmation | C++
|          |       |             | C plus plus          |

All entered languages

Statements

instance of
- object-oriented programming language
  - 1 reference
- multi-paradigm programming language

RDF Version: https://www.wikidata.org/entity/Q2407
BabelNet

Dictionary compiled from

- Wikipedia
- (Open Multilingual) WordNet
- Wiktionary
- OmegaWiki
- WikiData
LexVo

Resource: term/eng/cat

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdf:type</td>
<td>lvont:Term</td>
</tr>
<tr>
<td>lvont:language</td>
<td>lexvo:iso639-3/eng</td>
</tr>
<tr>
<td>skosxl:literalForm</td>
<td>cat ('en' language string)</td>
</tr>
<tr>
<td>rdfs:seeAlso</td>
<td><a href="http://en.wiktionary.org/wiki/cat">http://en.wiktionary.org/wiki/cat</a></td>
</tr>
<tr>
<td>lvont:means</td>
<td>lexvo:wordnet/30/noun/Caterpillar_1_06_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/noun/big_cat_1_05_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/noun/cat-o%27-nine-tails_1_06_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/noun/cat_1_05_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/noun/cat_1_18_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/noun/guy_1_18_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/noun/kat_1_06_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/verb/cat_2_35_00</td>
</tr>
<tr>
<td></td>
<td>lexvo:wordnet/30/verb/vomit_2_29_00</td>
</tr>
<tr>
<td>lvont:translation</td>
<td>lexvo:term/afr/kat</td>
</tr>
<tr>
<td></td>
<td>lexvo:term/afr/kots</td>
</tr>
<tr>
<td></td>
<td>lexvo:term/afr/kogoi</td>
</tr>
</tbody>
</table>

- Assigns URIs to words (strings in a language)
- Contains links to WordNet, FrameNet etc
- Definitions of ISO Language Codes
### Domain datasets

![BIO2RDF logo](https://www.ebi.ac.uk/bioportfolio/)

**Linked Data for the Life Sciences - Release 3**

- [website](#)  [datasets](#)  [documentation](#)

#### 100's of domain specific datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Date generated</th>
<th># of triples</th>
<th># of unique</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Affymetrix probesets [affymetrix]</strong></td>
<td>2014-08-01</td>
<td>11895348562</td>
<td>1107871027</td>
</tr>
<tr>
<td>This dataset contains the probesets used in the Affymetrix microarrays.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links: [search][5], [query][6], [example][7], [download][8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>2. A Database of Annotated Published Models [biomodels]</strong></td>
<td>2014-06-05</td>
<td>86942371</td>
<td>6679943</td>
</tr>
<tr>
<td>BioModels Database is a data resource that allows biologists to store,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>search and retrieve published mathematical models of biological interests.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><a href="http://www.ebi.ac.uk/biobismodels/">http://www.ebi.ac.uk/biobismodels/</a></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Links: [search][5], [query][6], [example][7], [download][8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioPortal is an open repository of biomedical ontologies that provides access</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>via Web services and Web browsers to ontologies developed in OWL, RDF,</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBO format and Protege frames. BioPortal functionality includes the ability</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to browse, search and visualize ontologies.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Reusing URIs
Why reuse URIs

- Data interoperability
  - Queries work over multiple datasets
- Semantic definitions allows alignments to be reasoned
- (Often) the creators of the URIs have good idea on how data should be structured
Challenges of interoperability

Differences in Granularity

Linguistic Differences
Language codes

Problem: 7,000+ languages and more dialects, but only $26^2 = 676$ codes
ISO Language Codes

- **fr-CA** Québecois
- **pms** Piedmontese
- **ang** Anglo-Saxon

3-Letter codes with region cover minority, historical languages, right?
Variability

How to tag this talk?

- en?
- en-Latn? (As it is not written in Cyrillic)
- en-BG or en-100 (As it is presented in Bulgaria)
- en-GB or en-826 (As is is composed in British English)
- en-Latn-GB?

Region subtags are used to indicate linguistic variations associated with or appropriate to a specific country, territory, or region. Typically, a region subtag is used to indicate variations such as regional dialects or usage, or region-specific spelling conventions. It can also be used to indicate that content is expressed in a way that is appropriate for use throughout a region, for instance, Spanish content tailored to be useful throughout Latin America. -- RFC 5646
Glottolog

- Identifies *languoids* (language varieties)
- Uses URLs
  http://glottolog.org/resource/languoid/id/queb1247
- More information can be found by following the link
Linked Open Vocabularies
http://lov.linkeddata.es/
ISOcat

- Effort to standardize linguistic vocabulary from ISO Technical Committee
- Standardized “Data Categories” in a “Registry”
- Discontinued in December 2014
Problems with ISOcat

According to Schuurman et al.

- Too easy to get a login
- “Out-of-control”
- Entries were copies of other entries
- “People sometimes copied an entry, just in order to make sure the original owner would not change the entry without them knowing it”
- Complexity – Too many obligatory and overly technical fields

As an alternative the CLARIN concept registry is (still) being introduced.

LexInfo

- Ontology for “associat[ing] linguistic information with respect to any level of linguistic description and expressivity to elements in an ontology”
- Expands OntoLex–Lemon with a set of general categories
LexInfo – Properties and Values

Properties and open-world (non-exhaustive) list of values
LexInfo – Verb Frames

Verb frames with formal definitions
LexInfo – Arguments

Hierarchies of arguments to be used in the frames
OLiA

Ontologies of Linguistic Annotation

- Modular architecture for describing annotation schemes:
  - **Reference Model:** Common terminology (similar to LexInfo)
  - **Annotation Model:** Describes a particular annotation scheme
  - **Linking Model:** Describes the linking between the reference and annotation
GOLD – General Ontology Linguistic Description

- Quite popular
- Defines many terms
- ‘Loose’ semantics
  - Sometimes has range and domains on properties
  - Not clear how this fits together
Submitting to the LOD Cloud
Fields

- **Identifier**
  - unique alphanumeric string

- **Title**
  - Full name in English

- **Description**
  - 2–10 sentence description in English

- **Full Download**
  - A link to the complete dataset, ideally as compressed N-Triples

- **SPARQL Endpoint**
  - If available

- **Other Download**
  - Other formats for download or partial downloads
Fields (2)

- **Example**
  - A single resource that resolves
- **Keywords**
- **Domain**
  - Defines the colour in the diagram
- **Website**
- **Contact Point**
- **Links**
  - Number of triples linking to another dataset in the cloud
- **Size**
  - Number of triples in this dataset
- **Namespace, DOI, Image (if desired)**
Salzburgerland IT

About this dataset
Salzburgerland dataset for Italian resources, built with WordLift from the website https://www.salzburgerland.com/it/

License: http://www.opendefinition.org/licenses/cc-by

Contact Details
Contact Point: WordLift Team
Website: https://www.salzburgerland.com/it/

Download Links
Full Downloads
- Download

Examples
- Download

Availability of resource

Data Facts

| Total size | 4,096 triples |
| Links to dbpedia | 176 triples |
| Links to freebase | 169 triples |
| Links to geonames-semantic-web | 39 triples |
| Links to yago | 76 triples |
| Links to wikidata | 1 triples |
Services using linked data
Service-oriented architectures

- It is implemented a self-contained operation unit.
- It is a black box for its consumers, which only need to know the interface, not the implementation.
- It may consist of other underlying services.

Interoperability is a significant challenge here
Service chains

Translation
DE => EN

Often tricky to do in practice!

Parser (EN)

Sentiment Analysis (EN)
Issues with service chains

- Services are often components of pipelines without clear usage to the end user.
- The technology readiness level of services is often quite low, with little documentation or graphical user interface.
- Services are hard to install often requiring compiling from source or specialized libraries not found in major software repositories.
Teanga

- RDF and Linked Data to provide service interoperability
- Docker to enable easy install and usage
- Attractive Web Front-End (Bootstrap, AngularJS, NodeJS)
- Graceful control of errors
Step 1: Upload Text

Enter the text below and click proceed

Le sugiero que vea la nueva película para el actor de Hollywood Brad Pitt
LAPPS Grid

Defines key vocabularies for service interoperability

- LAPPS Interchange Format (JSON–LD)
- Web Service Exchange Vocabulary
- Human–in–the–loop workflow construction using Galaxy
Twitter Datasource (Galaxy Version 1.0.0)

Type
- Mixed

Count
- 15

Query

Refer to the Twitter API help page <https://dev.twitter.com/rest/public/search> for a full description of the Twitter query syntax.

Some of the available operators are:

<table>
<thead>
<tr>
<th>Query</th>
<th>Finds tweet...</th>
</tr>
</thead>
<tbody>
<tr>
<td>watching</td>
<td>containing both “watching” and “now”. This is the default operator.</td>
</tr>
<tr>
<td>“happy hour”</td>
<td>containing the exact phrase “happy hour”.</td>
</tr>
<tr>
<td>love OR hate</td>
<td>containing either “love” or “hate” (or both).</td>
</tr>
<tr>
<td>beer -root</td>
<td>containing “beer” but not “root”.</td>
</tr>
<tr>
<td>#haiku</td>
<td>containing the hashtag “haiku”.</td>
</tr>
<tr>
<td>from:interior</td>
<td>sent from Twitter account “interior”.</td>
</tr>
<tr>
<td>to:NASA</td>
<td>a Tweet authored in reply to Twitter account “NASA”.</td>
</tr>
<tr>
<td>@NASA</td>
<td>mentioning Twitter account “NASA”.</td>
</tr>
<tr>
<td>flight :</td>
<td>containing “flight” and with a negative attitude.</td>
</tr>
<tr>
<td>traffic ?</td>
<td>containing “traffic” and asking a question.</td>
</tr>
</tbody>
</table>

License

Data provided by the Twitter Datasource are covered by the Twitter Developer Agreement and the Twitter Display Requirements.
Summary
Summary

- Linked Open Data Cloud
  - Big
  - Many relevant tools
  - Fragmented
  - Interoperability is less terrible than other systems
Thanks.

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289, co-funded by the European Regional Development Fund.