Semantic Web Technologies
Linked Open Data & Semantic Web Programming

Heiko Paulheim
Overview

• Linked Open Data
  – Principles
  – Examples
  – Vocabularies
• Microdata & schema.org
• Introduction to Semantic Web Programming with Jena
Linked Open Data

• What we've got to know up to now
  – RDF as a universal language for describing things
  – RDF Schema for describing vocabularies (i.e., classes and properties)

• Linked Open Data
  – uses those techniques
  – for providing open data

• The Linked Open Data Cloud
  – has nothing to do with cloud computing
  – is a big, freely available collection of knowledge
Why “Linked” Open Data?

```turtle
:p a :Physician .
:p :hasDegree "Dr." .
:p :hasName "Mark Smith" .
:p :hasAddress :a .
:a :street "Main Street" .
:a :number "14"^^xsd:int .
:a :city "Smalltown" .
p :hasOpeningHours [
  a rdf:Bag ;
  [ :day :Monday;
   :from "9"^^xsd:int;
   :to "11"^^xsd:int;
  ]
]
...

:s a :City .
s :name "Smalltown" .
s :lat "49.86"^^xsd:double .
s :long "8.65"^^xsd:double .
s :district "Birmingham" .
...

:d a :District .
d :name "Birmingham" .
d :pop "347891"^^xsd:int .
d :locatedIn "England" .
...
```
Why “Linked” Open Data?

- Information is scattered on the Web
  - that also holds for the Semantic Web

- HTML also has a concept for interlinking scattered information
  - known as hyperlink
  - More information at <a href="http://www.w3.org">W3C</a>

- Linked Open Data uses that principle, too
Why “Linked” Open Data?

```turtle
:p a :Physician .
:p :hasDegree "Dr." .
:p :hasName "Mark Smith" .
:p :hasAddress :a .
:a :street "Main Street" .
:a :number "14"^^xsd:int .
:a :city <http://.../smalltown> .
:p :hasOpeningHours [ a rdf:Bag ; [ :day :Monday; :from "9"^^xsd:int; :to "11"^^xsd:int; ] ] ...
```

```turtle
:s a :City .
:s :name "Smalltown" .
:s :lat "49.86"^^xsd:double .
:s :long "8.65"^^xsd:double .
:s :district <http://.../smalltown> .
```

```turtle
:d a :District .
:d :name "Birmingham" .
:d :pop "347891"^^xsd:int .
:d :locatedIn "England" . ...
```
Why “Linked” Open Data?

• Linked Open Data is RDF data
  – which is provided in a distributed manner

• URIs
  – have been used as simple identifiers so far
  – in LOD: links to data
    • resolvable!
    • "dereferencable URIs" (URLs)
    • can be used together with content negotiation, RDFa, etc.
Why “Linked” Open Data?

- Example:
  - `<#Heiko> :worksIn <http://dbpedia.org/resource/Mannheim> .`
Why “Linked” Open Data?

- Example:
  - `<#Heiko> :worksIn <http://dbpedia.org/resource/Mannheim>`.
HTML Links vs. Links in Linked Open Data

• Compare


to


• Observation:
  – Links in Linked Open Data are always explicitly typed
  – The semantics of the link is thus interpretable
    • given that the predicate is defined in a schema
Links in Linked Open Data

- Important special case: owl:sameAs*

* We don't know OWL yet, never mind, we'll get to that...
Important special case: owl:sameAs

Links two identical resources
  - This is required due to the non-unique naming assumption

One of the most commonly misused concepts in the Semantic Web...

Use:
  - Two datasets with information about the same person

Abuse:
  - A dataset with information about a person and the person's homepage
  - The Starbucks in O7 and the company Starbucks
  - The state and the city of Hamburg
  - The parliament as an institution and the parliament as a building

* We don't know OWL yet, never mind, we'll get to that...
Links in Linked Open Data

- Alternatives to abusing \texttt{owl:sameAs}:
  - General link to other resources
    \texttt{rdfs:seeAlso}
  - Link to (HTML) homepage:
    e.g., \texttt{foaf:homepage}

* We don't know OWL yet, never mind, we'll get to that...
Linking to a Schema

- Another important special case:
  - linking to a schema
  - luckily, everything is identified by a URI (also properties and classes)

:Heiko

"Heiko Paulheim".

<http://xmlns.com/foaf/0.1/name> "Heiko Paulheim".
Linking to a Schema

- btw: this also works for “built in” schemas

http://www.w3.org/1999/02/22-rdf-syntax-ns#type

:Heiko rdf:type :Person .
Four Principles of Linked Open Data

- The four Principles by Tim Berners-Lee (2006)
  1) Use URIs to identify things
  2) Use derefencable URIs
  3) Provide useful information upon derefencable URIs, use standards
  4) Add links to other datasets
What Data to Serve at a URI?

• Basic principle: provide a complete *RDF molecule* at the URI

• Definition of a complete RDF molecule:
  – All triples that have the URI as a subject or an object
  – Every blank node is connected by at least two predicates
RDF Molecules

- Avoid dead ends in browsing
RDF Molecules

- Recap: Blank Nodes for multi-valued predicates
  - avoid (potentially useless) partial information

Diagram:
- Recipe
- Sugar
- has ingredient
- ingredient
- amount
- value
- unit
- "100"
- gram
RDF Molecules: Theory and Practice

• Definition of a complete RDF molecule:
  – All triples that have the URI as a subject or an object
  – Every blank node is connected by at least two predicates

• Consequences:
  – Triples are duplicated (in the subject's and the object's molecule)
    • redundancy, depending on serving strategy
  – Molecules can become very big
In theory, all triples have to be served

Pragmatic approach:
- Which information is interesting for a user?
- For a person: the city of residence
  - but for a city: all persons who reside here?
RDF Molecules: Theory and Practice

- Example Graph

- Peter
- Julia
- Stefan
- Mannheim
- Karlsruhe
- Uni MA
- Uni KA

born: 01-12-1986
nick: Jule
knows
lives in
studies at
nearby
The Five Star Schema

- Five Star Scheme (Tim Berners-Lee, 2010)
  * Available on the web with an open license
  ** Available as machine-readable, structured data
  *** like ** plus using a non-proprietary format
  **** like*** plus using open standards by the W3C
  ***** like **** plus links to other datasets
Linked Open Data Best Practices

- as defined by Heath and Bizer, 2011

1) Provide dereferencable URIs
2) Set RDF links pointing at other data sources
3) Use terms from widely deployed vocabularies
4) Make proprietary vocabulary terms dereferencable
5) Map proprietary vocabulary terms to other vocabularies
6) Provide provenance metadata
7) Provide licensing metadata
8) Provide data-set-level metadata
9) Refer to additional access methods
The Linked Open Data Cloud

http://lod-cloud.net/
The Linked Open Data Cloud

• In numbers:
  – >1,000 Data sets
  – Several billion triples
  – Several million interlinks

• Topical domains:
  – Government
  – Publications
  – Life sciences
  – User-generated content
  – Cross-domain
  – Media
  – Geographic
  – Social web

http://lod-cloud.net/
The Linked Open Data Cloud

- Domains by number of datasets in Linked Open Data
  - As of 2019
  - Classified based on data provider tags
  - More than half of the datasets are government and life sciences
A Short History of Linked Open Data

- **May 2007:**

![Linking Open Data cloud diagram, as of May 2007.](image)

- **March 2008:**

![Linking Open Data cloud diagram, March 2008.](image)

- **August 2014:**

![Linking Open Data cloud diagram, August 2014.](image)

- **August 2017:**

![Linking Open Data cloud diagram, August 2017.](image)

- **March 2019:**

![Linking Open Data cloud diagram, March 2019.](image)

Examples: Government Data
Life Science Example: DrugBank

- ~7,700 drugs, descriptions, manufacturers, interactions...
**Linguistics Example: BabelNet**

**Keyboard**

http://babelnet.org/rdf/keyboard_n_EN

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemon:canonicalForm</td>
<td>bn: keyboard_n_EN/canonicalForm</td>
</tr>
<tr>
<td>Is lemon:entry of</td>
<td>bn: lexicon_EN</td>
</tr>
</tbody>
</table>
| rdfs:label                      | keyboard  
  - keyboard          
  - keyboard_(computer)  
  - keyboard_(computing) 
  - keyboard_(instrument) 
  - keyboard_(music)     
  - keyboard_(musical_instrument)  
  - keyboard_(typing)    |
| lemon:language                  | EN                                                                    |
| lexinfo:partOfSpeech            | lexinfo: noun                                                          |
| lemon:sense                     |                                                                      |

As Turtle | As RDF/XML | As N-Triple
Cross-Domain Example: DBpedia

- General knowledge on almost five million entities
- Hundreds of millions of triples
- Linked to ~100 other datasets
  - the most interlinked dataset

http://lod-cloud.net/
DBpedia: How It Is built
## Climate

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Record high °C (°F)</strong></td>
<td>16.4 (61.5)</td>
<td>16.1 (60.8)</td>
<td>24.1 (75.3)</td>
<td>29.1 (84.3)</td>
<td>32.2 (90.0)</td>
<td>36.8 (98.2)</td>
<td>39.0 (102.2)</td>
<td>39.8 (103.6)</td>
<td>32.6 (90.7)</td>
<td>28.2 (82.8)</td>
<td>15.7 (60.3)</td>
<td>16.5 (61.7)</td>
<td>39.8 (103.6)</td>
</tr>
<tr>
<td><strong>Average high °C (°F)</strong></td>
<td>4.7 (40.5)</td>
<td>6.7 (44.1)</td>
<td>11.6 (52.9)</td>
<td>16.2 (61.2)</td>
<td>20.6 (69.1)</td>
<td>23.7 (74.7)</td>
<td>26.1 (79.0)</td>
<td>25.6 (78.1)</td>
<td>21.2 (70.2)</td>
<td>15.3 (59.6)</td>
<td>8.9 (48)</td>
<td>5.3 (41.5)</td>
<td>15.5 (59.9)</td>
</tr>
<tr>
<td><strong>Daily mean °C (°F)</strong></td>
<td>1.8 (35.2)</td>
<td>2.6 (37)</td>
<td>5.7 (41.9)</td>
<td>10.7 (51.3)</td>
<td>15.2 (59.4)</td>
<td>18.2 (64.8)</td>
<td>20.3 (68.5)</td>
<td>19.9 (67.8)</td>
<td>15.8 (60.1)</td>
<td>10.7 (51.3)</td>
<td>5.7 (42.3)</td>
<td>2.8 (37)</td>
<td>10.85 (51.53)</td>
</tr>
<tr>
<td><strong>Average low °C (°F)</strong></td>
<td>-1.3 (25.7)</td>
<td>-2.9 (26.5)</td>
<td>-4.3 (26.1)</td>
<td>-5.0 (23.1)</td>
<td>-4.4 (25.5)</td>
<td>-3.2 (26.9)</td>
<td>-2.5 (27.9)</td>
<td>-1.3 (29.9)</td>
<td>0.7 (33.6)</td>
<td>3.6 (38.7)</td>
<td>6.7 (44.1)</td>
<td>2.5 (39.5)</td>
<td>-0.8 (32)</td>
</tr>
<tr>
<td><strong>Record low °C (°F)</strong></td>
<td>-18.7 (-1.7)</td>
<td>-18.7 (-1.7)</td>
<td>-13.6 (7.5)</td>
<td>-6.4 (20.5)</td>
<td>-0.1 (31.8)</td>
<td>4.0 (39.2)</td>
<td>4.7 (40.5)</td>
<td>5.3 (41.5)</td>
<td>2.5 (36.5)</td>
<td>-5.0 (23)</td>
<td>-9.7 (16.3)</td>
<td>-18.3 (-6.9)</td>
<td>-18.7 (-1.7)</td>
</tr>
<tr>
<td><strong>Average precipitation mm (inches)</strong></td>
<td>40.9 (1.61)</td>
<td>43.1 (1.69)</td>
<td>50.8 (1.99)</td>
<td>49.3 (1.94)</td>
<td>72.6 (2.86)</td>
<td>65.6 (2.58)</td>
<td>76.0 (2.98)</td>
<td>57.7 (2.27)</td>
<td>54.1 (2.13)</td>
<td>56.4 (2.22)</td>
<td>63.6 (2.5)</td>
<td>54.1 (2.13)</td>
<td>67.0 (2.66)</td>
</tr>
<tr>
<td><strong>Mean monthly sunshine hours</strong></td>
<td>262</td>
<td>85.6</td>
<td>124.9</td>
<td>186.2</td>
<td>214.1</td>
<td>219.1</td>
<td>235.1</td>
<td>222.1</td>
<td>164.1</td>
<td>108.6</td>
<td>59.0</td>
<td>44.9</td>
<td>1.712.2</td>
</tr>
</tbody>
</table>

Source: Data derived from Deutscher Wetterdienst[72]
DBpedia: Contents

• Data from different infoboxes (extracted from multiple languages)
• Redirects and disambiguations
• External web links
• Abstracts in multiple languages
• Instance type information
  – DBpedia Ontology
  – YAGO*
  – schema.org*
  – DOLCE**
  – ...and others

* later today
** in a few weeks
The DBpedia Ontology

- **Classes:**
  - 739 classes
  - partial hierarchy

- **Properties:**
  - ~1,100 relations
    - some with domain/range
  - ~1,700 data properties
    - i.e., literal-valued
  - a bit of hierarchy
YAGO

<table>
<thead>
<tr>
<th>&lt;Mannheim&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Reinhard_Bülikofer&gt;</td>
</tr>
<tr>
<td>&lt;Uml_Daval&gt;</td>
</tr>
<tr>
<td>&lt;Hans_Martin_Pippart&gt;</td>
</tr>
<tr>
<td>&lt;de/Klaus_May&gt;</td>
</tr>
<tr>
<td>&lt;Werner_Catel&gt;</td>
</tr>
<tr>
<td>&lt;de/Fritz_Rösting&gt;</td>
</tr>
<tr>
<td>&lt;Reinhold_Fanz&gt;</td>
</tr>
<tr>
<td>&lt;Peter_Dreher&gt;</td>
</tr>
<tr>
<td>&lt;Hans-Jürgen_Boysen&gt;</td>
</tr>
<tr>
<td>&lt;Albert_Speer&gt;</td>
</tr>
<tr>
<td>&lt;Manuel_Guldén&gt;</td>
</tr>
<tr>
<td>&lt;Caroline_Augusta_of_Bavaria&gt;</td>
</tr>
<tr>
<td>&lt;Reiner_Hollic&gt;</td>
</tr>
<tr>
<td>&lt;Ludwig_Landmann&gt;</td>
</tr>
<tr>
<td>&lt;Timo_Zahnleiter&gt;</td>
</tr>
<tr>
<td>&lt;Jochen_Hacht&gt;</td>
</tr>
<tr>
<td>&lt;Carolin_Leonhardt&gt;</td>
</tr>
<tr>
<td>&lt;de/Marc_Hemmen&gt;</td>
</tr>
<tr>
<td>&lt;wasBornIn&gt;</td>
</tr>
<tr>
<td>&lt;happenedIn&gt;</td>
</tr>
<tr>
<td>&lt;hasCitationTitle&gt;</td>
</tr>
</tbody>
</table>

- Metropolregion Rhein-Neckar
- World’s 15 Most Inventive Cities
- The Manhattan of Germany: the inside story of how the smart city is emerging
- Germany and the Second World War: Ausgabe der Klimadaten: Monatsr
- Press release announcing the merger of FEI European Jumping Champion: Partner und Freundesfahrt
- Orase Infražite (Twin cities of Minsk and Swansea - Wales: Mannheim.de
- Sepp_Herberger
- Jiangsu
- Prince-elector
- Karl_Benz
- Frankfurt
- Dennis_Siver
- Uwe_Rahn
- Jochen_Hacht
- Norwich_Duff
- SAP_Arena
- Oceanic_climate
- Ontario
- Lithuania
- The Left (Germany)
YAGO

- Also derived from Wikipedia
  - ~4.6M entities
  - ~26M statements
- Uses Wikipedia categories for typing
  - a class hierarchy of ~500,000 types
- Tries to capture time
  - i.e., statements that held true for a period of time
  - e.g., soccer players playing for teams
  - uses reification
Wikidata

- Collaboratively edited knowledge base
- Size
  - ~15M instances
  - ~66M statements
- Ontology
  - ~23k classes
  - ~1.6k properties
- Special
  - provenance information
  - i.e., evidence: where did that statement come from?
Trent Reznor (Q282722)

American musician
No aliases defined

Statements

sex or gender
male

4 references

imported from
Swedish Wikipedia

imported from
Virtual International Authority File

imported from
Italian Wikipedia

stated in
Integrated Authority File

retrieved
27 April 2014
Further Example Datasets

- Linked Movie Database
  - Movies, actors, directors...
- MusicBrainz
  - Artists, albums, ...
- Open Library
  - books, authors, publishers
- DBLP
  - computer science publications
Further Example Datasets

- **ProductDB**
  - products and manufacturers

- **NASA Data Incubator**
  - Data on all NASA missions

- **Linked Open Numbers**
  - Numbers and their names in different languages
  - roman and arabic notations, binary, hex etc.
Vocabularies

• Recap: LOD Best Practices, Principle 3:
  – Use terms from widely deployed vocabularies

• So, what are common widely deployed vocabularies?
Dublin Core

• We have already encountered this
• Usage: Metadata for resources and documents
• Namespace http://purl.org/dc/elements/1.1/
• Common prefix: dc
• defines properties, e.g.,
  – creator
  – subject
  – date
• Resources: DCMI Type Vocabulary:
  – Text
  – Image
  – Software
  – ...
FOAF (Friend of a Friend)

- Persons and their relations
- Created for personal home pages
  - but used widely beyond that
- Namespace http://xmlns.com/foaf/0.1/
- Common prefix: foaf:

- Important classes
  - Person
  - Group
  - Organization
  - Project
  - ...

- Important properties
  - name, firstName, lastName
  - phone, mbox, homepage
  - knows
  - currentProject, pastProject
  - ...
FOAF (Friend of a Friend)

Peter

foaf:nick: "Pete"
foaf:firstname: "Peter"
foaf:lastname: "Smith"
foaf:knows: Julia

Julia
"Improving the usability of integrated applications by using visualizations of linked data"@en

2011

"Heiko Paulheim"
WGS 84

- Encodes geographic data
- World Geodetic System 1984
- 3D reference model
- Namespace http://www.w3.org/2003/01/geo/wgs84_pos#
- Common prefix: geo:

- Classes:
  - SpatialThing
  - Point

- Properties:
  - latitude
  - longitude
  - altitude
  - location
Publishing Linked Open Data

- Possible variants
  - hand coded
  - from triple stores
  - from relational databases
Linked Data from Triple Stores

- Triple Store: RDF storage engine
  - e.g., Virtuoso
- Pubby: Front end for triple stores
- Supports content negotiation etc.

Diagram:
- Intelligent Agent
- Browser
- Pubby
- Triple Store
- RDF
- HTML
Linked Open Data from RDBMS

- D2R: Linked Open Data interface on relational databases
  - e.g., MySQL
Linked Open Data from RDBMS

<table>
<thead>
<tr>
<th>ID (int)</th>
<th>name (text)</th>
<th>location (int)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1327890123</td>
<td>&quot;Heiko&quot;</td>
<td>&quot;Mannheim&quot;</td>
</tr>
</tbody>
</table>

map:Person a d2rq:ClassMap;
d2rq:dataStorage map:Database1.
d2rq:class foaf:Person;
d2rq:uriPattern "http://foo.bar/p@@Person.ID@@";

map:personName a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:Person;
d2rq:property foaf:name;
d2rq:column "Person.name";
d2rq:datatype xsd:string;

map:location a d2rq:PropertyBridge;
d2rq:belongsToClassMap map:Person;
d2rq:property foaf:basedNear;
d2rq:column "Person.location";
d2rq:datatype xsd:string;

<http://foo.bar/p1327890123> a foaf:Person .
Microdata and schema.org

- We have already seen that in the first lecture

```html
<div itemscope
itemtype="http://schema.org/PostalAddress">
  <span itemprop="name">Data and Web Science Group</span>,
  <span itemprop="addressLocality">Mannheim</span>,
  <span itemprop="postalCode">68131</span>,
  <span itemprop="addressCountry">Germany</span>
</div>

_:1 a <http://schema.org/PostalAddress> .
_:1 <http://schema.org/name> "Data and Web Science Group" .
_:1 <http://schema.org/addressLocality> "Mannheim" .
_:1 <http://schema.org/addressCountry> "Germany" .
```
Microdata and schema.org

- schema.org defines (among others)
  - products
  - product offers
  - businesses and local businesses (stores, cafés, …)
  - books, movies, records
  - events
  - recipes
  - persons
  - …
## Movie

**Thing > CreativeWork > Movie**

A movie.

Usage: Between 10,000 and 50,000 domains

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>actor</td>
<td>Person</td>
<td>An actor, e.g. in tv, radio, movie, video games etc. Actors can be associated with individual items or with a series, episode, clip. Supersedes actors.</td>
</tr>
<tr>
<td>director</td>
<td>Person</td>
<td>A director of e.g. tv, radio, movie, video games etc. Directors can be associated with individual items or with a series, episode, clip. Supersedes directors.</td>
</tr>
<tr>
<td>duration</td>
<td>Duration</td>
<td>The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format.</td>
</tr>
<tr>
<td>musicBy</td>
<td>MusicGroup or Person</td>
<td>The composer of the soundtrack.</td>
</tr>
<tr>
<td>productionCompany</td>
<td>Organization</td>
<td>The production company or studio responsible for the item e.g. series, video game, episode etc.</td>
</tr>
<tr>
<td>subtitleLanguage</td>
<td>Text or Language</td>
<td>Languages in which subtitles/captions are available, in IETF BCP 47 standard format.</td>
</tr>
<tr>
<td>trailer</td>
<td>VideoObject</td>
<td>The trailer of a movie or tv/radio series, season, episode, etc.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Property</th>
<th>Expected Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>about</td>
<td>Thing</td>
<td>The subject matter of the content.</td>
</tr>
<tr>
<td>accessibilityAPI</td>
<td>Text</td>
<td>Indicates that the resource is compatible with the referenced accessibility API (WebSchemas wiki lists possible values).</td>
</tr>
<tr>
<td>accessibilityControl</td>
<td>Text</td>
<td>Identifies input methods that are sufficient to fully control the described resource (WebSchemas wiki lists possible values).</td>
</tr>
<tr>
<td>accessibilityFeature</td>
<td>Text</td>
<td>Content features of the resource, such as accessible media, alternatives and supported enhancements for accessibility (WebSchemas wiki lists possible values).</td>
</tr>
<tr>
<td>accessibilityHazard</td>
<td>Text</td>
<td>A characteristic of the described resource that is physiologically dangerous to some users. Related to WCAG 2.0 guideline 2.3 (WebSchemas wiki lists possible values).</td>
</tr>
<tr>
<td>accountablePerson</td>
<td>Person</td>
<td>Specifies the Person that is legally accountable for the CreativeWork.</td>
</tr>
<tr>
<td>aggregateRating</td>
<td>AggregateRating</td>
<td>The overall rating, based on a collection of reviews or ratings, of the item.</td>
</tr>
<tr>
<td>alternativeHeadline</td>
<td>Text</td>
<td>A secondary title of the CreativeWork.</td>
</tr>
</tbody>
</table>
Deployment of schema.org

- Main topics of schema.org:
  - Meta information on web page content (web page, blog...)
  - Business data (products, offers, ...)
  - Contact data (businesses, persons, ...)
  - (Product) reviews and ratings
- ...and a massive long tail
Growth of schema.org

- Note: schema.org is mainly used with Microdata
  - ...and Microdata is mainly used with schema.org

http://webdatacommons.org/structureddata/2016-10/stats/stats.html
Microdata/schema.org vs. Linked Open Data

• Commonalities
  – Both encode machine-interpretable knowledge
  – Schema.org uses a standard vocabulary
  – Both can be encoded as RDF
Microdata/schema.org vs. Linked Open Data

- Differences
  - Microdata is embedded in the DOM tree
    - i.e., the resulting RDF is always a set of trees
    - not a general directed graph
    - no cycles, no reification
  - Microdata uses only blank nodes and literals
Microdata/schema.org vs. Linked Open Data

- Linked Data Principles (Tim Berners-Lee 2006)
  - Use URIs as names for things
  - Use HTTP URIs that can be looked up
  - When someone looks up a HTTP URI, provide useful information using a standard

HTML5+MD is a standard

Blank nodes cannot be looked up

MD2RDF creates blank nodes

```html
<div itemprop="http://schema.org/PostalAddress">
  <span itemprop="name">Data and Web Science Group</span>
</div>
```
Microdata/schema.org vs. Linked Open Data

• Linked Data Principles (TimBL 2006)
  – Use URIs as names for things
  – Use HTTP URIs that can be looked up
  – When someone looks up a HTTP URI, provide useful information using a standard
  – Include links to other URIs

This is possible with schema.org/sameas

• Linkage within schema.org Microdata:
  – Only 0.02% of all data providers use schema.org/sameas
Microdata/schema.org vs. LOD

- Five Star Scheme (TimBL 2010)
  * Available on the web with an open license
  ** Available as machine-readable, structured data
  *** as (**), using a non-proprietary format
  **** plus: using open standards by the W3C
  ***** plus: links to other datasets

- What's the license of web data?
Intermediate Summary

• Until today, we have dealt with the Semantic Web as a vision
• Today, we have seen two incarnations of that vision
  – Linked Open Data
  – schema.org/Microdata
• Both have a lot in common
And Now for Something Completely Different
Programming for the Semantic Web

• Let's start with a simple application
  – a Hello World application for the Semantic Web
Using only Plain Java

```java
URL url = new URL("http://dbpedia.org/resource/Mannheim");
URLConnection conn = url.openConnection();
conn.addRequestProperty("Accept", "text/rdf+n3");
BufferedReader BR = new BufferedReader(
    new InputStreamReader(conn.getInputStream())
);

while(BR.ready()) {
    String triple = BR.readLine();
    StringTokenizer tokenizer = new StringTokenizer(triple, " ");
    String subject = tokenizer.nextToken();
    String predicate = tokenizer.nextToken();
    String object = tokenizer.nextToken();
    ...
}
```
Using only Plain Java

• Let's start with a simple application
  – a Hello World application for the Semantic Web

• Using plain Java is possible
  – but not very comfortable
  – there are more sophisticated frameworks
Programming with Jena

- Jena is a well-known Semantic Web programming framework
- started in 2000 at HP Labs
- Apache open source project since 2010

- Central concepts
  - Models (class Model) are RDF graphs
  - Resources (class Resource) are resources in RDF graphs

- Special features
  - Database connectors for persistence
  - Support for reasoning
  - Rule engines
  - Support for SPARQL (see next lecture)
Programming with Jena

• Reading a model from a derefencable URI

```java
model.read("http://dbpedia.org/resource/Mannheim");
```

• Navigating within a model

```java
Resource mannheim = model.getResource("http://dbpedia.org/resource/Mannheim");
Resource countryOfMannheim = model.getProperty("http://dbpedia.org/ontology/country").getResource();
```
• Working with literals

```java
Literal lit = mannheim.getProperty("http://www.w3.org/2000/01/rdf-schema#label").getLiteral();
lit.getString();
lit.getLanguage();
lit.getDatatype();
```
Programming with Jena

- Working with multi-valued relations
  - StmtIterator iter = mannheim.getProperty("http://www.w3.org/2000/01/rdf-schema#label");
  - while(iter.hasNext()) {
      Statement s = iter.next();
      RDFNode node = s.getObject();
      if(node.isLiteral())
          System.out.println(node.asLiteral().getString());
  }

creates an iterator over all triples with the subject node and the given predicate
Iterators in Jena

- Jena uses the iterator pattern quite frequently
- e.g.:

  ```java
  StmtIterator iter = mannheim.getProperty(
  "http://www.w3.org/2000/01/rdf-schema#label");
  ```

- But there is no such thing as

  ```java
  Collection<Statement> triples = 
  mannheim.getProperty(
  "http://www.w3.org/2000/01/rdf-schema#label");
  ```

- Why?
Iterators in Jena

• Data volumes in the Semantic Web can be big
• e.g., reading all triples from DBpedia
  – stored in List<Statement> would kill the main memory
  – iterators allow a more efficient memory use
Programming with Jena

• Manipulating models
  
  ```java
  pl.addProperty("http://xmlns.com/foaf/0.1/knows", p2);
  ```

• Watching model changes
  
  ```java
  class MyListener implements ModelChangedListener...
  MyListener listener = new MyListener();
  model.add(listener);
  ```
Reasoning with Jena

- Recap: we can derive information from a schema (T-Box) and data (A-box)
  
  :knows rdfs:domain :Person .
  :knows rdfs:range :Person .
  :Peter :knows :Tom .

  → :Peter a :Person .
  → :Tom a :Person .

- Jena also supports reasoning
Reasoning with Jena

• Given: a schema and some data

```java
Model schemaModel = ModelFactory.createDefaultModel();
InputStream IS = new FileInputStream("data/example_schema.rdf");
schemaModel.read(IS);

Model dataModel = ModelFactory.createDefaultModel();
IS = new FileInputStream("data/example_data.rdf");
dataModel.read(IS);

Model reasoningModel =
    ModelFactory.createRDFSModel(schemaModel, dataModel);
```

• Now, reasoningModel contains all derived facts
Reasoning with Jena

- Now, `reasoningModel` contains all derived facts
  
  ```java
  StmtIterator it = 
  reasoningModel.listStatements();
  while(it.hasNext()) {
    Statement s = it.next();
    System.out.println(s);
  }
  ```

- Output:
Programming with RDFLib (Python)

- RDFLib is a Python library for working with RDF
- initial release 4 June, 2002 by Daniel Krech
  - Now being developed by the community at github: https://github.com/RDFLib/rdflib/
- it contains parsers and serializers for
  - RDF/XML, N3, NTriples, N-Quads, Turtle, TriX, RDFa and Microdata
- graph interface which can be backed by store implementations
  - memory storage
  - persistent storage on top of the Berkeley DB
- reasoning possible (https://github.com/RDFLib/OWL-RL )
- SPARQL 1.1 implementation (see next lecture)
Programming with RDFLib (Python)

- primary interface is a Graph
  - represented as a set of 3-item triples

\[
\begin{align*}
\text{[} & \\
& (\text{subject, predicate, object}), \\
& (\text{subject1, predicate1, object1}), \\
& \ldots \\
& (\text{subjectN, predicateN, objectN}) \text{]}
\end{align*}
\]
Programming with RDFLib (Python)

• Reading a model from a derefencable URI

```python
import rdflib

# Create a new graph
g = rdflib.Graph()

# Load the model from a URI
g.load('http://dbpedia.org/resource/Mannheim')
```

• Print out all RDF triples

```python
for s, p, o in g:
    print(s, p, o)
```

• Navigating within a graph

```python
print(g.value(
    URIRef("http://dbpedia.org/resource/Mannheim"),
    URIRef("http://dbpedia.org/ontology/country")
))
```
Programming with RDFLib (Python)

- Most often reduced to basic triple matching
- `Graph.triples(subject, predicate, object)`
  - each of them can be None (similar to null in Java)

```python
for s, p, o in g.triples((None, RDF.type, FOAF.Person)):
    print("%s is a person" % s)
```

- Special functions for returning only specific parts
  - `Graph.subjects(predicate, object)` – returns only subjects
  - `Graph.predicate(subject, object)`
  - `Graph.objects(subject, predicate)`
  - `Graph.subject_objects(predicate)`
  - `Graph.subject_predicates(object)`
  - `Graph.predicate_objects(subject)`
  - `Graph.value(subject, predicate)`
    - For just one value and not a generator/iterator
Programming with RDFLib (Python)

• create URIs
  mannheim = URIRef('http://example.com/Mannheim')

• create literals
  mannheim_literal = Literal("Mannheim")

• Add triples to graph
  g.add((mannheim, RDFS.label, mannheim_literal))
  g.add((mannheim, RDFS.label, Literal("Mannheim", lang="de")))

• Serialize graph
  print(g.serialize(format='n3'))
Wrap-Up

• Today, we have seen
  – two incarnations of the Semantic Web
  – i.e., Linked Open Data
  – and Microdata/schema.org

• ...and we have learned how to write programs consuming Semantic Web data
  – Jena & RDFlib programming frameworks
  – loading RDF from files and from URLs
  – performing reasoning
Semantic Web – Architecture

here be dragons...

Semantic Web Technologies (This lecture)

Technical Foundations

Berners-Lee (2009): Semantic Web and Linked Data
Questions?