Knowledge Graphs
Data Quality and Linking
The Journey Ends Here

here be dragons...

Knowledge Graph Technologies
(This lecture)

Technical Foundations

Berners-Lee (2009): Semantic Web and Linked Data
Before You Go...

Berners-Lee (2009): Semantic Web and Linked Data
Before You Go...

• We’ve learned about
  – standards
  – methods
  – datasets

• You’ve played with
  – datasets
  – tools

• Now, let’s be serious…
  – how good is that data, actually?
Previously on Knowledge Graphs

• 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

how well are they followed in practice?
Studies of Best Practice Conformance

• Hogan et al.: *An empirical survey of Linked Data conformance*, 2012
  – top-level view

• Schmachtenberg et al.: *Adoption of the Linked Data Best Practices in Different Topical Domains*, 2014
  – domain-specific view
1) Provide Dereferenceable URIs

- Metric: how many URIs used are actually dereferenceable?
  - i.e., do not link to HTTP 404 (possible bias: study time)
  - provide RDF

- Hogan et al.: ~70% of URIs are dereferenceable in above sense
2) Set RDF links pointing at other data sources

- Schmachtenberg et al.:
  - ~55% of all datasets link to at least one other dataset
  - There are some hubs as link targets
    - DBpedia (~200 datasets)
    - geonames.org (~140 datasets)

- Hogan et al.:
  - on average, a dataset links to 20.4 (±38.2) other datasets
2) Set RDF links pointing at other data sources

- Are all links owl:sameAs?
  - Schmachtenberg et al.: domain-specific differences

Table 3: Top three linking predicates per category. The percentage is relative to number of datasets within the category which set outgoing links.

<table>
<thead>
<tr>
<th>category</th>
<th>predicate</th>
<th>usage</th>
<th>category</th>
<th>predicate</th>
<th>usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>social networking</td>
<td>foaf:knows</td>
<td>59.87%</td>
<td>life sciences</td>
<td>owl:sameAs</td>
<td>57.69%</td>
</tr>
<tr>
<td>social networking</td>
<td>foaf:based_near</td>
<td>35.79%</td>
<td>life sciences</td>
<td>rdfs:seeAlso</td>
<td>38.46%</td>
</tr>
<tr>
<td>social networking</td>
<td>sioc:follows</td>
<td>34.11%</td>
<td>life sciences</td>
<td>dct:creator</td>
<td>19.23%</td>
</tr>
<tr>
<td>publications</td>
<td>owl:sameAs</td>
<td>32.20%</td>
<td>government</td>
<td>dct:publisher</td>
<td>47.12%</td>
</tr>
<tr>
<td>publications</td>
<td>dct:language</td>
<td>25.42%</td>
<td>government</td>
<td>dct:spatial</td>
<td>29.81%</td>
</tr>
<tr>
<td>publications</td>
<td>rdfs:seeAlso</td>
<td>23.73%</td>
<td>government</td>
<td>owl:sameAs</td>
<td>25.00%</td>
</tr>
<tr>
<td>user-generated content</td>
<td>owl:sameAs</td>
<td>52.94%</td>
<td>geographic</td>
<td>owl:sameAs</td>
<td>59.09%</td>
</tr>
<tr>
<td>user-generated content</td>
<td>rdfs:seeAlso</td>
<td>23.53%</td>
<td>geographic</td>
<td>skos:exactMatch</td>
<td>36.36%</td>
</tr>
<tr>
<td>user-generated content</td>
<td>dct:source</td>
<td>17.65%</td>
<td>geographic</td>
<td>skos:closeMatch</td>
<td>22.73%</td>
</tr>
<tr>
<td>media</td>
<td>owl:sameAs</td>
<td>76.47%</td>
<td>crossdomain</td>
<td>owl:sameAs</td>
<td>76.92%</td>
</tr>
<tr>
<td>media</td>
<td>rdfs:seeAlso</td>
<td>23.53%</td>
<td>crossdomain</td>
<td>rdfs:seeAlso</td>
<td>53.85%</td>
</tr>
<tr>
<td>media</td>
<td>foaf:based_near</td>
<td>17.65%</td>
<td>crossdomain</td>
<td>dct:creator</td>
<td>23.08%</td>
</tr>
</tbody>
</table>
3) Use terms from widely deployed vocabularies

- Schmachtenberg et al.: most used vocabularies

  ![Table 5: Vocabularies used by more than 5% of all datasets.](image)

- Hogan et al.: on average, 6.6k classes and properties are shared between at least two datasets
3) Use terms from widely deployed vocabularies

- Linked Open Vocabularies (LOV) project
  - analyze usage of vocabularies

https://lov.linkeddata.es/dataset/lov/
4) Make proprietary vocabulary terms dereferencable

- Schmachtenberg et al.:
  - ~23% of all datasets use proprietary vocabularies
  - ~58% of all vocabularies are proprietary

Table 6: Proprietary vocabularies with dereferencability per category and quota of vocabularies linking to others

<table>
<thead>
<tr>
<th>category</th>
<th>different prop. vocabs. used (% of all prop. vocabs.)</th>
<th># of datasets using prop. vocab. (% of all datasets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>social networking</td>
<td>128 (33.86%)</td>
<td>83 (15.99%)</td>
</tr>
<tr>
<td>publications</td>
<td>58 (15.34%)</td>
<td>35 (33.65%)</td>
</tr>
<tr>
<td>government</td>
<td>48 (12.70%)</td>
<td>35 (18.82%)</td>
</tr>
<tr>
<td>cross-domain</td>
<td>55 (14.55%)</td>
<td>16 (36.36%)</td>
</tr>
<tr>
<td>geographic</td>
<td>24 (6.34%)</td>
<td>16 (39.02%)</td>
</tr>
<tr>
<td>life sciences</td>
<td>35 (9.25%)</td>
<td>26 (29.21%)</td>
</tr>
<tr>
<td>media</td>
<td>22 (5.82%)</td>
<td>21 (56.76%)</td>
</tr>
<tr>
<td>user-gen. cnt.</td>
<td>30 (7.93%)</td>
<td>26 (47.27%)</td>
</tr>
<tr>
<td>Total</td>
<td>378 (58.24%)</td>
<td>241 (23.17%)</td>
</tr>
</tbody>
</table>
4) Make proprietary vocabulary terms dereferencable

- Schmachtenberg et al.:
  - less than 20% of all vocabularies are fully derefencable

- Common reasons:
  - use of deprecated terms
  - namespace hijacking

Table 6: Proprietary vocabularies with dereferencability per category and quota of vocabularies linking to others

<table>
<thead>
<tr>
<th>category</th>
<th>different prop. vocabs. used (% of all prop. vocabs.)</th>
<th># of datasets using prop. vocabs. (% of all datasets)</th>
<th>Dereferencability</th>
<th># of vocab linking (quota)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>full</td>
<td>partial</td>
</tr>
<tr>
<td>social networking</td>
<td>128 (33.86%)</td>
<td>83 (15.99%)</td>
<td>16.41%</td>
<td>6.25%</td>
</tr>
<tr>
<td>publications</td>
<td>58 (15.34%)</td>
<td>35 (33.65%)</td>
<td>20.69%</td>
<td>6.90%</td>
</tr>
<tr>
<td>government</td>
<td>48 (12.70%)</td>
<td>35 (18.82%)</td>
<td>20.83%</td>
<td>12.50%</td>
</tr>
<tr>
<td>cross-domain</td>
<td>55 (14.55%)</td>
<td>16 (36.36%)</td>
<td>27.27%</td>
<td>10.91%</td>
</tr>
<tr>
<td>geographic</td>
<td>24 (6.34%)</td>
<td>16 (39.02%)</td>
<td>20.83%</td>
<td>4.17%</td>
</tr>
<tr>
<td>life sciences</td>
<td>35 (9.25%)</td>
<td>26 (29.21%)</td>
<td>28.57%</td>
<td>5.71%</td>
</tr>
<tr>
<td>media</td>
<td>22 (5.82%)</td>
<td>21 (56.76%)</td>
<td>0.00%</td>
<td>9.09%</td>
</tr>
<tr>
<td>user-gen. ent.</td>
<td>30 (7.93%)</td>
<td>26 (47.27%)</td>
<td>13.33%</td>
<td>10.00%</td>
</tr>
<tr>
<td>Total</td>
<td>378 (58.24%)</td>
<td>241 (23.17%)</td>
<td>19.25%</td>
<td>8.00%</td>
</tr>
</tbody>
</table>
5) Map proprietary vocabulary terms to other vocabularies

• Schmachtenberg et al.:
  – only a small fraction of proprietary vocabularies are linked :-(

Table 7: Predicates used to link terms between different vocabularies.

<table>
<thead>
<tr>
<th>term</th>
<th>% of vocabularies</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:range</td>
<td>9.52%</td>
</tr>
<tr>
<td>rdfs:subClassOf</td>
<td>8.47%</td>
</tr>
<tr>
<td>rdfs:subPropertyOf</td>
<td>6.88%</td>
</tr>
<tr>
<td>rdfs:domain</td>
<td>5.29%</td>
</tr>
<tr>
<td>rdfs:isContainedBy</td>
<td>3.70%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>term</th>
<th>% of vocabularies</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:seeAlso</td>
<td>1.59%</td>
</tr>
<tr>
<td>owl:inverseOf</td>
<td>1.32%</td>
</tr>
<tr>
<td>owl:equivalentClass</td>
<td>1.32%</td>
</tr>
<tr>
<td>swivt:type</td>
<td>1.06%</td>
</tr>
<tr>
<td>owl:equivalentProperty</td>
<td>0.79%</td>
</tr>
</tbody>
</table>
6) Provide provenance metadata

- Hogan et al.:
  - ~41% of all datasets provide (provenance) metadata
- Schmachtenberg et al.:
  - ~35% of all datasets provide provenance metadata
  - most used vocabulary is Dublin Core

Table 8: Provenance vocabulary usage and license vocabulary usage by category

<table>
<thead>
<tr>
<th>Category</th>
<th>Any prov-vocab</th>
<th>Dublin Core</th>
<th>Admin</th>
<th>prv/prov</th>
</tr>
</thead>
<tbody>
<tr>
<td>social networking</td>
<td>169 (32.56%)</td>
<td>56.21%</td>
<td>58.58%</td>
<td>1.18%</td>
</tr>
<tr>
<td>publications</td>
<td>39  (37.50%)</td>
<td>94.87%</td>
<td>5.13%</td>
<td>2.56%</td>
</tr>
<tr>
<td>government</td>
<td>77  (41.40%)</td>
<td>100.00%</td>
<td>0.00%</td>
<td>1.30%</td>
</tr>
<tr>
<td>life sciences</td>
<td>21  (23.60%)</td>
<td>100.00%</td>
<td>0.00%</td>
<td>2.56%</td>
</tr>
<tr>
<td>cross-domain</td>
<td>8   (18.18%)</td>
<td>100.00%</td>
<td>12.50%</td>
<td>0.00%</td>
</tr>
<tr>
<td>geographic</td>
<td>4   (9.76%)</td>
<td>100.00%</td>
<td>0.00%</td>
<td>25.00%</td>
</tr>
<tr>
<td>user-gen. content</td>
<td>11  (20.00%)</td>
<td>90.91%</td>
<td>54.55%</td>
<td>0.00%</td>
</tr>
<tr>
<td>media</td>
<td>5   (13.51%)</td>
<td>100%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Total</td>
<td>372 (35.77%)</td>
<td>28.37%</td>
<td>10.77%</td>
<td>0.77%</td>
</tr>
</tbody>
</table>
7) Provide licensing metadata

- Hogan et al.: ~14% of all datasets provide licensing metadata
- Schmachtenberg et al.: ~8% of all datasets provide licensing metadata
8) Provide data-set-level metadata

- Schmachtenberg et al.:
  - Issue: referral and discovery
  - methods: inline, link, /.well-known/void
  - in total, ~14% provide data-set-level metadata

Table 9: Percentage of datasets using the VoID vocabulary and percentage of datasets offering alternative access methods

<table>
<thead>
<tr>
<th>Category</th>
<th>VOID (%)</th>
<th>link (%)</th>
<th>well-known (%)</th>
<th>inline (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>social networking</td>
<td>6 (1.16%)</td>
<td>0.58%</td>
<td>0.19%</td>
<td>0.58%</td>
</tr>
<tr>
<td>publications</td>
<td>14 (13.46%)</td>
<td>6.73%</td>
<td>2.88%</td>
<td>5.77%</td>
</tr>
<tr>
<td>life sciences</td>
<td>29 (32.58%)</td>
<td>19.10%</td>
<td>4.49%</td>
<td>12.36%</td>
</tr>
<tr>
<td>government</td>
<td>75 (40.32%)</td>
<td>6.99%</td>
<td>3.23%</td>
<td>31.18%</td>
</tr>
<tr>
<td>user-gen. content</td>
<td>6 (10.91%)</td>
<td>5.45%</td>
<td>0.00%</td>
<td>5.45%</td>
</tr>
<tr>
<td>geographic</td>
<td>15 (36.59%)</td>
<td>14.63%</td>
<td>12.20%</td>
<td>12.20%</td>
</tr>
<tr>
<td>cross-domain</td>
<td>5 (11.36%)</td>
<td>9.09%</td>
<td>2.27%</td>
<td>2.27%</td>
</tr>
<tr>
<td>media</td>
<td>2 (5.41%)</td>
<td>2.70%</td>
<td>0.00%</td>
<td>2.70%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>140 (13.46%)</td>
<td>4.62%</td>
<td>1.44%</td>
<td>8.27%</td>
</tr>
</tbody>
</table>
9) Refer to additional access methods

- Schmachtenberg et al.:
  - SPARQL and dump download are rarely *referred to*
  - This does not mean that they don’t exist...

### Table 9: Percentage of datasets using the VoID vocabulary and percentage of datasets offering alternative access methods

<table>
<thead>
<tr>
<th>Category</th>
<th>VOID (1.16%)</th>
<th>link</th>
<th>well-known</th>
<th>inline</th>
<th>alt. access</th>
<th>SPARQL</th>
<th>Dump</th>
</tr>
</thead>
<tbody>
<tr>
<td>social networking</td>
<td>6 (1.16%)</td>
<td>0.58%</td>
<td>0.19%</td>
<td>0.58%</td>
<td>6 (1.16%)</td>
<td>1.16%</td>
<td>0.39%</td>
</tr>
<tr>
<td>publications</td>
<td>14 (13.46%)</td>
<td>6.73%</td>
<td>2.88%</td>
<td>5.77%</td>
<td>10 (10.58%)</td>
<td>9.62%</td>
<td>3.85%</td>
</tr>
<tr>
<td>life sciences</td>
<td>29 (32.58%)</td>
<td>19.10%</td>
<td>4.49%</td>
<td>12.36%</td>
<td>19 (21.35%)</td>
<td>20.22%</td>
<td>16.85%</td>
</tr>
<tr>
<td>government</td>
<td>75 (40.32%)</td>
<td>6.99%</td>
<td>3.23%</td>
<td>31.18%</td>
<td>61 (32.80%)</td>
<td>30.11%</td>
<td>30.65%</td>
</tr>
<tr>
<td>user-gen. content</td>
<td>6 (10.91%)</td>
<td>5.45%</td>
<td>0.00%</td>
<td>5.45%</td>
<td>3 (5.45%)</td>
<td>5.45%</td>
<td>1.82%</td>
</tr>
<tr>
<td>geographic</td>
<td>15 (36.59%)</td>
<td>14.63%</td>
<td>12.20%</td>
<td>12.20%</td>
<td>8 (19.51%)</td>
<td>12.20%</td>
<td>12.20%</td>
</tr>
<tr>
<td>cross-domain media</td>
<td>5 (11.36%)</td>
<td>9.09%</td>
<td>2.27%</td>
<td>2.27%</td>
<td>4 (9.09%)</td>
<td>4.55%</td>
<td>6.82%</td>
</tr>
<tr>
<td>Total</td>
<td>140 (13.46%)</td>
<td>4.62%</td>
<td>1.44%</td>
<td>8.27%</td>
<td>48 (5.89%)</td>
<td>4.54%</td>
<td>3.80%</td>
</tr>
</tbody>
</table>
9) Refer to additional access methods

- Study by Hertling & Paulheim (2013)
  - sample random URIs from large Linked Data corpus
  - try to discover a SPARQL endpoint, e.g., by
    - using /.well-known/void
    - using inline links
    - using external catalogs (!)

---

**Table 1.** Results on different strategies for finding SPARQL endpoints on 10,000 random URIs, reporting both the number of URIs for which any SPARQL endpoint was found, as well as the number of URIs for which a valid SPARQL endpoint was found. The numbers in parentheses denote the total number of endpoints found.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Datahub Catalog (/all)</th>
<th>/.well-known/void (standard)</th>
<th>Link to VoID</th>
</tr>
</thead>
<tbody>
<tr>
<td># found</td>
<td>7,389 (26,124)</td>
<td>110 (392)</td>
<td>9 (9)</td>
</tr>
<tr>
<td># valid</td>
<td>1,375 (2,978)</td>
<td>53 (106)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>
More Indicators

  - also includes performance
  - latency, throughput, ...

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Abr</th>
<th>Metric</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>A1</td>
<td>accessibility of the SPARQL endpoint and the server</td>
<td>checking whether the server responds to a SPARQL query [18]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>accessibility of the RDF dumps</td>
<td>checking whether an RDF dump is provided and can be downloaded [18]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>dereferenceability of the URI</td>
<td>checking (i) for dead or broken links i.e. when an HTTP-GET request is sent, the status code 404 Not Found is not returned (ii) that useful data (particularly RDF) is returned upon lookup of a URL (iii) for changes in the URI i.e. the compliance with the recommended way of implementing redirections using the status code 303 See Other [18,30]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>A4</td>
<td>no misreported content types</td>
<td>detect whether the HTTP response contains the header field stating the appropriate content type of the returned file e.g. application/rdf+xml [50]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>A5</td>
<td>dereferenced forward-links</td>
<td>dereferencability of all forward links: all available triples where the local URI is mentioned in the subject (i.e. the description of the resource) [31]</td>
<td>QN</td>
</tr>
<tr>
<td>Licensing</td>
<td>L1</td>
<td>machine-readable indication of a license</td>
<td>detection of the indication of a license in the VoID description or in the dataset itself [18,31]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>L2</td>
<td>human-readable indication of a license</td>
<td>detection of a license in the documentation of the dataset [18,31]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>L3</td>
<td>specifying the correct license</td>
<td>detection of whether the dataset is attributed under the same license as the original [18]</td>
<td>QN</td>
</tr>
<tr>
<td>Interlinking</td>
<td>I1</td>
<td>detection of good quality interlinks</td>
<td>(i) detection of (a) interlinking degree, (b) clustering coefficient, (c) centrality, (d) open sameAs chains and (e) description richness through sameAs by using network measures [25], (ii) via crowdsourcing [1,65]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>I2</td>
<td>existence of links to external data providers</td>
<td>detection of the existence and usage of external URIs (e.g. using owl:sameAs links) [31]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>I3</td>
<td>dereferenced back-links</td>
<td>detection of all local in-links or back-links: all triples from a dataset that have the resource’s URI as the object [31]</td>
<td>QN</td>
</tr>
<tr>
<td>Security</td>
<td>S1</td>
<td>usage of digital signatures</td>
<td>by signing a document containing an RDF serialization, a SPARQL result set or signing an RDF graph [13,18]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>S2</td>
<td>authenticity of the dataset</td>
<td>verifying authenticity of the dataset based on a provenance vocabulary such as author and his contributors, the publisher of the data and its sources (if present in the dataset) [18]</td>
<td>QL</td>
</tr>
<tr>
<td>Performance</td>
<td>P1</td>
<td>usage of slash-URIs</td>
<td>checking for usage of slash-URIs where large amounts of data is provided [18]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>P2</td>
<td>low latency</td>
<td>(minimum) delay between submission of a request by the user and reception of the response from the system [18]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>P3</td>
<td>high throughput</td>
<td>(maximum) no. of answered HTTP-requests per second [18]</td>
<td>QN</td>
</tr>
<tr>
<td></td>
<td>P4</td>
<td>scalability of a data source</td>
<td>detection of whether the time to answer an amount of ten requests divided by ten is not longer than the time it takes to answer one request [18]</td>
<td>QN</td>
</tr>
</tbody>
</table>
Linked Data Conformance vs. Quality

- So far, we’ve looked at conformance
  - i.e., following standards and best practices
  - technical dimension
  - can be evaluated automatically

- Quality
  - i.e., how complete/correct/… is the data
  - content dimension
  - hard to evaluate automatically
Quality of Knowledge Graphs

- Färber et al.: *Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO*. SWJ 9(1), 2018
  - internal validation
    - e.g., schema violations
  - proxy metrics
    - e.g., timeliness measured by frequency of dataset updates
      → does not necessarily imply more recent data
  - manual evaluation
    - e.g., semantic validity
Issues with Automatic Evaluation

• Where to find a gold standard?
  – e.g., sample 1k population figures from DBpedia
  – check whether they are correct

• Open World Assumption
  – ~60% of all persons in DBpedia do not have a deathDate
  – so?

• …
Issues with Automatic Evaluation

• So, we need human experts!
  – however, human evaluation is often expensive
  – more complex problems
    are hard to specify as microtasks

Regarding Amazon Mechanical Turk Project (HIT Type) 3SKITTYVZ4PF45V9L1Q331NN98X367

Message from Amazon Mechanical Turk user:

Worker ID: A20EIM01BOUDEB
HIT Set ID: 3SKITTYVZ4PF45V9L1Q331NN98X367
HIT Title: Decide if two wiki pages describe the same thing
HIT Description: The wiki topics are Runescape(Gaming), Marvel (Comics) and Star Trek(TV)

hello, I believe you must be off a decimal point. you mean 1.50 not .15 right?

Greetings from Amazon Mechanical Turk,

The message above was sent by an Amazon Mechanical Turk user.
Please review the message and respond to it as you see fit.

Sincerely,
Amazon Mechanical Turk
https://requester.mturk.com
Example: Crowd Evaluation of DBpedia

Example: Crowd Evaluation of DBpedia


• From the paper: “Considering the HIT granularity, we paid 0.04 US dollar per 5 triples.”

• DBpedia (en): 176M statements
• Total cost of validation with this approach: 1.4M USD!
Intermediate Summary

- The Quality of Linked Open Data is far from perfect
  - conformance
  - content

- Improving the quality is an active field of research
  - Survey 2017: >40 approaches
  - since then: a lot of work in KG embeddings
And now for something completely different

• Let’s jump back to the best practices one last time
Previously on Knowledge Graphs

• 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
Previously on Knowledge Graphs

- Integrate data from different sources
- Make connections between entities in those sources
- Facilitate cross data source queries
- Overcome data silos

Why do we need Links?

• Task:
  – Find contact data for Dr. Mark Smith
  – Input: various knowledge graphs

• Problems:
  – Every knowledge graph uses its own identifiers (by design)
  – Every knowledge graph may use its own vocabulary
  – Some reuse vocabularies, some don’t

Knowledge Graph 1
:p a foaf:Person .
:p foaf:name "Mark Smith" .
:p bar:profession bar:Physician .
...

Knowledge Graph 2
:q a foo:Human .
:q foo:called "Mark Smith" .
:q foo:worksAs foo:MedDoctor .
...
How do we Create the Links?

- Technically, links can be added with OWL statements
- We know:
  - `owl:sameAs`, `owl:equivalentClass`, `owl:equivalentProperty`

```
:p a foaf:Person .
:p owl:sameAs foo:q .
:p foaf:name "Mark Smith" .
:p bar:profession bar:Physician .
:q a foo:Human .
:q owl:sameAs bar:p .
:q foo:called "Mark Smith" .
foo:called
owl:EquivalentProperty
foo:worksAs .

Knowledge Graph 1

Knowledge Graph 2
```
How do we Create the Links?

• Remember
  – The LOD cloud
  – >1,200 datasets

• Pairwise interlinking?
How do we Create the Links?

- Datasets with millions of entities...

**Data Facts**

<table>
<thead>
<tr>
<th>Link Description</th>
<th>Number of Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total size</td>
<td>9,500,000,000 triples</td>
</tr>
<tr>
<td>Namespace</td>
<td><a href="https://dbpedia.org/resource/">https://dbpedia.org/resource/</a></td>
</tr>
<tr>
<td>Links to 2000-us-census-rdf</td>
<td>12,529 triples</td>
</tr>
<tr>
<td>Links to dbtune-musikbrainz</td>
<td>22,981 triples</td>
</tr>
<tr>
<td>Links to education-data-gov-uk</td>
<td>1,697 triples</td>
</tr>
<tr>
<td>Links to aunts</td>
<td>3,600 triples</td>
</tr>
<tr>
<td>Links to flickr-wrapepr</td>
<td>8,800,000 triples</td>
</tr>
<tr>
<td>Links to freebase</td>
<td>3,400,000 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-dallymed</td>
<td>40 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-dibip</td>
<td>196 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-diseasome</td>
<td>1,943 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-drugbank</td>
<td>729 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-eurostat</td>
<td>157 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-project-gutenberg</td>
<td>2,510 triples</td>
</tr>
<tr>
<td>Links to fu-berlin-slider</td>
<td>751 triples</td>
</tr>
<tr>
<td>Links to geonames-semantic-web</td>
<td>66,547 triples</td>
</tr>
<tr>
<td>Links to geoespecies</td>
<td>15,972 triples</td>
</tr>
<tr>
<td>Links to italiann-public-schools-linkeddata-it</td>
<td>5,822 triples</td>
</tr>
<tr>
<td>Links to linkedgeo-data</td>
<td>99,075 triples</td>
</tr>
<tr>
<td>Links to linkedenclb</td>
<td>13,800 triples</td>
</tr>
<tr>
<td>Links to nytimes-linked-open-data</td>
<td>10,359 triples</td>
</tr>
<tr>
<td>Links to opencyc</td>
<td>20,362 triples</td>
</tr>
<tr>
<td>Links to rdf-book-mashup</td>
<td>9,078 triples</td>
</tr>
<tr>
<td>Links to reference-data-gov-uk</td>
<td>22 triples</td>
</tr>
<tr>
<td>Links to revuu</td>
<td>6 triples</td>
</tr>
</tbody>
</table>
Tool Support

• A plethora of names
• Mostly used for schema level:
  – Ontology matching/alignment/mapping
  – Schema matching/mapping
• Mostly used for the instance level:
  – Instance matching/alignment
  – Interlinking
  – Link discovery
Automating Interlinking

- Given two input ontologies/knowledge graphs
  - And optional: a set of existing interlinks/mappings
- Provide a target set of interlinks/mappings
Automatic Interlinking

• Automatic interlinking is usually *heuristic*
  – i.e., not exact

• Most approaches provide confidence scores

• General format: \(<e_1, e_2, \text{relation}, \text{score}>\)
  
  \(<\text{dbpedia:University_of_Mannheim, wd:Q317070, owl:sameAs, 0.96}>\)

• Relations may include
  – equality (owl:sameAs, owl:equivalentClass, owl:equivalentProperty)
  – specialization (rdfs:subClassOf, rdfs:subPropertyOf)

• Actively researched, but not yet finally solved
  – complex relations
Summary and Takeaways

- Over the years, a large variety of approaches has been developed

![Ontology Matching Techniques Diagram](Euzenat & Shvaiko: Ontology Matching (2007))
Basic Interlinking Techniques

• Element vs. structural
  – Element level: only consider single elements in isolation
  – Structure based: exploit structure
    • e.g., class/property inheritance

• Syntactic vs. external vs. semantic
  – Syntactic: only use knowledge graphs themselves
  – External: use external sources of knowledge (e.g., dictionaries)
  – Semantic: exploit ontology semantics, e.g., by reasoning
Sources for Interlinking Signals

• Some knowledge graphs have “speaking” URIs, some don’t
  – http://dbpedia.org/resource/Germany, but
  – https://www.wikidata.org/wiki/Q183

• Most knowledge graphs have labels and textual descriptions
  – rdfs:label
  – skos:preferredLabel, skos:altLabel, …
  – rdfs:comment

• Proprietary string labels
  – dbo:abstract
  – https://www.wikidata.org/wiki/Property:P2561 (“name”)
  – …
Simple String Based Metrics

• String equality
  – e.g. foo:University_of_Mannheim, bar:University_of_Mannheim

• Common prefixes
  – e.g. foo:United_States, bar:United_States_of_America

• Common postfixes
  – e.g. foo:Barack_Obama, bar:Obama

• Typical usage of prefixes/postfixes: |common|/max(length)
  – foo:United_States, bar:United_States_of_America → 12/22
  – foo:Barack_Obama, bar:Obama → 5/12
Edit Distance

• Notion: minimal number of basic edit operations needed to get from one string to the other
  – insert character
  – delete character
  – change character
• Can handle:
  – alternate spellings, small typos and variations
  – matches in different, but similar languages

• Example:
  – Universität Mannheim, University of Mannheim
  – Universität of Mannheim
  \[\rightarrow \text{edit distance 5/20} \rightarrow \text{similarity score} = 3/4\]
N-gram based Similarity

• Problem: word order
  – e.g., University_of_Mannheim vs. Mannheim_University
  – prefix/postfix similarity: 0, edit distance similarity 5/11

• n-gram similarity
  – how many substrings of length n are common?
  – divided by no. of n-grams in longer string

• Example above with n=3
  – common: Uni, niv, ive, ver, ers, rsi, sit, ity, Man, ann, nnh, nhe, hei, eim
  – not common: ty_, y_o, _of, of_, f_M, _Ma, im_, m_U, _Un

• Similarity: 14/(14+9) = 14/25
Typical Preprocessing Techniques

• Unifying whitespace
  – University_of_Mannheim → University of Mannheim
  – UniversityOfMannheim → University Of Mannheim

• Unifying capitalization
  – University of Mannheim → university of mannheim

• Tokenization
  – university of mannheim → \{university, of, mannheim\}
  – similarity then becomes (average, maximum, …)
    similarity among token sets
  – also allows for other metrics, such as Jaccard overlap
Language-specific Preprocessing

- **Stopword Removal**
  - University of Mannheim → University Mannheim

- **Stemming**
  - German Universities → German Universit
  - Universities in Germany → Universit in German

- Usually, whole preprocessing pipelines are applied
  - e.g., stemming, stopword removal, tokenization, averaging the maximum edit distance similarity

- As above:
  - avg (max(similarity))\([\{\text{German, Universit}\}, \{\text{Universit, German}\}]\) = 1.0
Using External Knowledge

- e.g., linguistic resources (Wiktionary, BabelNet, ...)

Proper noun [edit]

New York

1. The largest city in New York State, a metropolis. 
   *New York is a former capital of the United States.*

2. A state of the United States of America and New York City.
   *The capital of New York is Albany, not New York City.*

3. A county of New York State, coterminous with the city.

Synonyms [edit]

- (state): the Empire State, New York State, NY
- (city): Big Apple (informal), New Amsterdam (Dutch East India Company)

Using External Knowledge

**New York City** - *New York* - Greater New York - Big Apple - the five boroughs

The largest city in New York State and in the United States; located in southeastern New York at the mouth of the Hudson river; a major financial and cultural center.
From Matching Literals to Matching Entities

• Exploiting properties
  – e.g., person: birth date
  – e.g., place: coordinates
  – e.g., movie: director
  – ...

• Usually, a mix of measures
  – e.g., person: name similarity + equal birthdate
  – e.g., place: name similarity + coordinates w/in range
  – e.g., movie: name similarity + director name similarity
  – ...

Preprocessing and Matching Pipelines

- Example tool: Silk Workbench
Schema Matching

- similar to interlinking
- typical approach: start with anchors based on string matching
- other signals
  - e.g., exploiting class/subclass similarity
  - e.g., exploiting property domain/range
  - using reasoning to determine validity
Schema Matching

- similar to interlinking
- typical heuristics include
  - classes appearing in the domain/range of matched properties are similar
Schema Matching

• similar to interlinking
• typical heuristics include
  – properties having matched domains/ranges are similar
Schema Matching

• similar to interlinking
• typical heuristics include
  – superclasses of mapped classes are similar
Schema Matching

• similar to interlinking
• typical heuristics include
  – pairs of classes along paths are similar (bounded path matching)
Instance based Matching

- Assumption: instances are already matched
  - either explicitly or heuristically
- Using, e.g., Jaccard
  - \(|\text{ex1:Human} \cap \text{ex2:Person}| / |\text{x1:Human} \cup \text{ex2:Person}|\)
  - example below: 18/23 → confidence ~0.78
- Finds non-trivial matches
  - e.g., dbpedia:Park ↔ yago:ProtectedArea
Enforcing 1:1 Mappings

• Assumption
  – each element can only be mapped to one other element
  – very often used in matching and linking

• Example:
  – stable marriage problem
  – try to find best matching partner for each element

```
<table>
<thead>
<tr>
<th>Product</th>
<th>Article</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile</td>
<td>Electronics</td>
</tr>
<tr>
<td>Accessory</td>
<td>Grocery</td>
</tr>
</tbody>
</table>
```

Weights:
- Product → Article: 0.9
- Product → Mobile: 0.5
- Product → Accessory: 0.3
- Mobile → Electronics: 0.6
- Mobile → Grocery: 0.4
- Accessory → Grocery: 0.2
Schema Matching

• Refining a matching with reasoning
  – i.e., is the matching consistent with the ontology
Schema Matching

• Refining a matching with reasoning
  – i.e., is the matching consistent with the ontology

```
:Mobile
  rdfs:subClassOf  :Product .
:Accessory
  rdfs:subClassOf  :Product .
:Mobile  owl:disjointWith
  :Accessory.

:ElectronicProduct
  rdfs:subClassOf  :Article .
:Grocery
  rdfs:subClassOf  :Article .

ex1:Product  owl:equivalentClass
  ex2:ElectronicProduct.
ex1:Accessory
  owl:equivalentClass  ex2:Article .
```
Reasoning on Mappings

• Reasoning:

```rdfs
ex1:Mobile rdfs:subClassOf ex1:Product .
→ ex1:Mobile rdfs:subClassOf ex2:ElectronicProduct .
+ ex2:ElektronicProduct rdfs:subClassOf ex2:Article .
→ ex1:Mobile rdfs:subClassOf ex2:Article .
+ ex2:Article owl:equivalentClass ex1:Accessory .
→ ex1:Mobile rdfs:subClassOf ex1:Accessory .
```

• And

```rdfs
ex1:Mobile owl:disjointWith ex1:Accessory .
```

• The mapping is contradictory!
  – Solution: remove a mapping element
  – e.g. by lowest confidence
Matcher Combination

- Chaining
Matcher Combination

- Parallel execution

![Diagram showingMatcher Combination and Parallel execution](Diagram.png)
Matcher Combination

• Iterative execution
Evaluating Matchers

- Typical measures: recall, precision, F1
  - Scenario: reference alignment (gold standard) R, matcher found M
- Recall \( r = \frac{|R \cap M|}{|R|} \)
- Precision \( p = \frac{|R \cap M|}{|R|} \)
- \( F1 = \) harmonic mean of \( r \) and \( p \)
  - i.e., \( 2r*p / (r+p) \)
OAEI: an Annual Competition for Matching

- Different Tracks
  - started 2014
  - tracks usually repeated over the years
    - track progress in the field
- Different focus
  - domains
  - scalability
  - schema-instance
  - interactive
**Track Example: Knowledge Graphs**

- Uses data from DBkWik
  - different graphs extracted from Wikis
  - (partial) gold standard: explicit links

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ALOD2Vec</td>
<td>0.13</td>
<td>20.01</td>
<td>0.00</td>
<td>0.80</td>
<td>0.67</td>
<td>76.6</td>
<td>0.94</td>
<td>0.95</td>
<td>97.97</td>
<td>4039.3</td>
<td>0.91</td>
<td>(0.91) (0.87) (0.87) (0.83)</td>
</tr>
<tr>
<td>AML</td>
<td>0.50</td>
<td>23.60</td>
<td>0.90</td>
<td>0.85</td>
<td>0.89</td>
<td>51.6</td>
<td>0.92</td>
<td>0.70</td>
<td>57.57</td>
<td>8802.8</td>
<td>0.90</td>
<td>(0.90) (0.85) (0.85) (0.80)</td>
</tr>
<tr>
<td>ATBox</td>
<td>0.16</td>
<td>25.60</td>
<td>0.97</td>
<td>0.87</td>
<td>0.79</td>
<td>78.8</td>
<td>0.97</td>
<td>0.96</td>
<td>95.95</td>
<td>4858.8</td>
<td>0.89</td>
<td>(0.89) (0.84) (0.84) (0.80)</td>
</tr>
<tr>
<td>baselineAllLabel</td>
<td>0.10</td>
<td>16.41</td>
<td>0.00</td>
<td>0.74</td>
<td>0.59</td>
<td>47.8</td>
<td>0.99</td>
<td>0.79</td>
<td>68.68</td>
<td>4674.8</td>
<td>0.89</td>
<td>(0.89) (0.84) (0.84) (0.80)</td>
</tr>
<tr>
<td>baselineLabel</td>
<td>0.10</td>
<td>16.41</td>
<td>0.00</td>
<td>0.74</td>
<td>0.59</td>
<td>47.8</td>
<td>0.99</td>
<td>0.79</td>
<td>68.68</td>
<td>3614.8</td>
<td>0.95</td>
<td>(0.95) (0.81) (0.81) (0.71)</td>
</tr>
<tr>
<td>DESKMatcher</td>
<td>0.13</td>
<td>91.07</td>
<td>0.76</td>
<td>0.71</td>
<td>0.66</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>3820.8</td>
<td>0.94</td>
<td>(0.94) (0.82) (0.82) (0.74)</td>
</tr>
<tr>
<td>LogMap</td>
<td>2.55</td>
<td>24.00</td>
<td>0.95</td>
<td>0.84</td>
<td>0.76</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>29190.40</td>
<td>0.40</td>
<td>0.54 (0.54) (0.54) (0.54)</td>
</tr>
<tr>
<td>LogMapBio</td>
<td>4.35</td>
<td>24.00</td>
<td>0.95</td>
<td>0.84</td>
<td>0.76</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>LogMapKM</td>
<td>2.49</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>29190.40</td>
<td>0.40</td>
<td>0.54 (0.54) (0.54) (0.54)</td>
</tr>
<tr>
<td>LogMapKG</td>
<td>2.47</td>
<td>24.00</td>
<td>0.95</td>
<td>0.84</td>
<td>0.76</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>29190.40</td>
<td>0.40</td>
<td>0.54 (0.54) (0.54) (0.54)</td>
</tr>
<tr>
<td>LogMapLt</td>
<td>0.07</td>
<td>23.00</td>
<td>0.56</td>
<td>0.70</td>
<td>0.43</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>8653.8</td>
<td>0.73</td>
<td>(0.91) (0.87) (0.84) (0.82)</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>0.30</td>
<td>22.41</td>
<td>0.80</td>
<td>0.80</td>
<td>0.67</td>
<td>80.0</td>
<td>0.94</td>
<td>0.95</td>
<td>97.97</td>
<td>4968.2</td>
<td>0.91</td>
<td>(0.91) (0.87) (0.87) (0.83)</td>
</tr>
</tbody>
</table>

*Aggregated results per matcher, divided into class, property, instance, and overall alignments. Time is displayed as HH:MM:SS. Column #testcases indicates the number of testcases where the tool is able to generate (non empty) alignments. Column size indicates the averaged number of system correspondences. Two kinds of results are reported: (1) those not distinguishing empty and erroneous (or not generated) alignments, and (2) those considering only non empty alignments (value between parenthesis).*
Track Example: Knowledge Graphs

• Generally, performance is high (F1>0.9) on many OAEI tracks
• So, what keeps us from interlinking the entire LOD cloud?
  – Performance is one issue, but...
The Golden Hammer Bias

• Challenge:
  – OAEI setup expect two related KGs
  – In the general case, this cannot be taken for granted
  • Manual pre-inspection for every pair is infeasible
  – Experiments with unrelated KGs:

<table>
<thead>
<tr>
<th>Matcher</th>
<th>mcu lyrics</th>
<th>memoryalpha lyrics</th>
<th>starwars lyrics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>matches</td>
<td>precision</td>
<td>matches</td>
</tr>
<tr>
<td>AML</td>
<td>2,642</td>
<td>0.12</td>
<td>7,691</td>
</tr>
<tr>
<td>baselineAltLabel</td>
<td>588</td>
<td>0.44</td>
<td>1,332</td>
</tr>
<tr>
<td>baselineLabel</td>
<td>513</td>
<td>0.54</td>
<td>1,006</td>
</tr>
<tr>
<td>FCAMap-KG</td>
<td>755</td>
<td>0.40</td>
<td>2,039</td>
</tr>
<tr>
<td>LogMapKG</td>
<td>29,238</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td>LogMapLt</td>
<td>2,407</td>
<td>0.08</td>
<td>7,199</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>971</td>
<td>0.12</td>
<td>3,457</td>
</tr>
</tbody>
</table>

See: ESWC 2020 Paper on OAEI Knowledge Graph Track
Challenges in Matching

• Usage of external resources
  – which are useful for which task? automatic selection?
  – embeddings?

• Automatic matcher combination & parameterization
  – analogy: AutoML

• Scalability
  – more or less solved for large pairs
  – open for large number of datasets

• Robustness
  – almost all of the OAEI tasks have a positive outcome bias (aka as “Golden Hammer Bias”)
Summary and Takeaways

• Data Quality of public Knowledge Graphs / Linked Open Data
  – Conformance and Content
  – Both have weak spots
  – An active research area

• Matching
  – Schema and instance matching
  – Typical measures, heuristics, preprocessing
  – Still: no one size fits all matcher
    • we are far from full automation
    • deep learning and embeddings have also not brought the ultimate weapon
Recommendations for Upcoming Semesters

- Information Retrieval and Web Search (FSS), Prof. Ponzetto
- Web Mining (FSS), Prof. Bizer
- Web Data Integration (HWS), Prof. Bizer
- Relational Learning (HWS), Prof. Stuckenschmidt
- Text Analytics (HWS), Prof. Strohmaier
Special Recommendation

- Seminar “Knowledge Graphs and Large Language Models”

Pan et al. (2023): Unifying Large Language Models and Knowledge Graphs: A Roadmap
Coming up Next

• ...your presentations!
• Mind the submission of your reports (8.12.)
• Prepare for (max) 6 minutes presentation + 3 minutes questions
  – present a consistent story
  – focus on key issues and lessons learned
  – demonstrations are appreciated
    • but make sure you stick to the time limit!
Questions?