Semantic Web Technologies

Topic: Data Cleaning

Olaf Hartig
olaf.hartig@liu.se
Terminology and Methodologies

- **Data cleaning** *(data cleansing, data scrubbing)* “deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data.”
  
  [Rahm and Do 2000]

- There are a number of methodologies, for instance:
  1. Audit the data to identify quality issues
  2. Choose methods to automatically detect and remove the issues
  3. Apply the methods
  4. Post-processing / control step [Müller and Freytag 2003]


Data Quality
More Terminology

- **Data quality**: commonly understood as “fitness for use” for a particular application or use case
  - Hence, even a dataset with quality issues may be fully useful for use cases not affected by the issue

- **Data quality assessment**: process of measuring the quality of some data and, ultimately, identifying whether the data is fit for use

- **Data quality dimensions**: accuracy, timeliness, completeness, relevancy, objectivity, believability, understandability, consistency, conciseness, etc.
  - Different authors consider different dimensions under different names, and group them into different groups
Data Quality Dimensions
(with a Focus on Semantic Web Data)

Accessibility
- Availability
- Security*
- Performance*
- Interlinking*
- Licensing*

Representation
- Rep.-Conciseness
- Interoperability
- Versatility*

Intrinsic
- Syntactic Validity
- Consistency
- Completeness

Contextual
- Relevancy
- Trustworthiness
- Understandability
- Timeliness


Aspects of access, authenticity, and retrieval of data

Depend on the context of the task at hand

Two dimensions are related

Semantic Web Technologies – Topic: Data Cleaning
Olaf Hartig
Intrinsic Dimensions

- Aspects that are independent of the user’s context
- **Syntactic validity**: degree to which a file conforms to the specification of the serialization format
- **Semantic accuracy**: degree to which data values correctly represent the real world facts
- **Consistency**: degree to which there are no logical contradictions w.r.t. the knowledge representation
- **Conciseness**: degree to which there is no redundancy of entities at the schema level and the data level
- **Completeness**: degree to which all required information is present in the data

Intrinsic Dimensions

• Aspects that are independent of the user’s context

• **Syntactic validity**: degree to which a file conforms to the specification of the serialization format

• **Semantic accuracy**: degree to which data values correctly represent the real world facts

• **Consistency**: degree to which there are no logical contradictions w.r.t. the knowledge representation

• **Conciseness**: degree to which there is no redundancy of entities at the schema level and the data level

• **Completeness**: degree to which all required information is present in the data

---

Possible metrics for syntactic validity:

- No syntax errors in the file
- Syntactically accurate data (e.g., conformance to a given schema)
- No malformed datatype literals

---

Representational Dimensions

- Capture aspects related to the design of the data
- **Representational-conciseness**: degree to which the representation of the data is compact and well formatted
- **Interoperability**: degree to which the format and structure conforms to previously returned data and to data from other sources
- **Interpretability**: degree to which data is represented using appropriate notation and whether the machine is able to process the data
- **Versatility**: availability of the data in different representations and in an internationalized way

Tools
Goal of Data Cleaning

• Fix data quality issues in given sets of (semantic) data
• Such quality issues may …
  … be in source datasets (e.g., inaccurate or wrong data items, outdated data items)
  … result from imperfections of a data integration process (e.g., data items that have been incorrectly linked with each other)
  … reveal themselves only after the data integration (e.g., duplicates, inconsistencies)
• Hence, data cleaning may be relevant both for
  – original datasets before combining/integrating, and
  – datasets resulting from an integration
Options

- Tools that allow users to identify quality issues (e.g., by highlighting outliers or similarities)
- Tools that identify quality issues (semi-)automatically
- Tools that fix these issues in an automated process
RDFUnit

- **http://rdfunit.aksw.org/**
- Test driven data-debugging framework
- Test cases are executed as SPARQL queries using a pattern-based transformation approach
  - Template: `SELECT ?s WHERE {
    ?s %P1% ?v1 .
    ?s %P2% ?v2 .
    FILTER ( ?v1 %OP% ?v2 )
  }`
  - Test case: `SELECT ?s WHERE {
    ?s dbo:deathDate ?v2 .
    FILTER ( ?v1 > ?v2 )
  }`
RDFUnit (cont’d)

• http://rdfunit.aksw.org/
• Test driven data-debugging framework
• Test cases are executed as SPARQL queries using a pattern-based transformation approach
• Test cases that can be created manually, or generated automatically (based on a schema)
  – Supported schemas: OWL, SHACL, IBM Resource Shapes, Dublin Core Set Profiles
• Tested data loaded from a specified file or accessed via a SPARQL endpoint
• Report of a test suite can be obtained as an HTML page, but also as RDF data
### RDFUnit (cont’d)

<table>
<thead>
<tr>
<th>S</th>
<th>Test</th>
<th>Errors</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-0a77ce81bec99608d28790eb695d11fa">http://databugger.aksw.org/tests#foaf-INVFUNC-0a77ce81bec99608d28790eb695d11fa</a></td>
<td>25</td>
<td>-1</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-105e1374ad211491979c95caa27ba2f5">http://databugger.aksw.org/tests#foaf-INVFUNC-105e1374ad211491979c95caa27ba2f5</a></td>
<td>53</td>
<td>786</td>
</tr>
<tr>
<td></td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-11eb481f2e37c9e1fd18066d637bc013">http://databugger.aksw.org/tests#foaf-INVFUNC-11eb481f2e37c9e1fd18066d637bc013</a></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-18fb0cf9dc8ff9ad9d42982e0434db2c">http://databugger.aksw.org/tests#foaf-INVFUNC-18fb0cf9dc8ff9ad9d42982e0434db2c</a></td>
<td>476</td>
<td>1214</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-2e2b3b0e569d5316d760bdf30f9ecf48">http://databugger.aksw.org/tests#foaf-INVFUNC-2e2b3b0e569d5316d760bdf30f9ecf48</a></td>
<td>34</td>
<td>87</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-4fe77a880206d4b9a00b9972176043b1">http://databugger.aksw.org/tests#foaf-INVFUNC-4fe77a880206d4b9a00b9972176043b1</a></td>
<td>84</td>
<td>244</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-58e73e30a1082f24e75ecb7c394415d9">http://databugger.aksw.org/tests#foaf-INVFUNC-58e73e30a1082f24e75ecb7c394415d9</a></td>
<td>21219</td>
<td>366471</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-9e12004a97dd6757449f9a1acf86b2a0">http://databugger.aksw.org/tests#foaf-INVFUNC-9e12004a97dd6757449f9a1acf86b2a0</a></td>
<td>165</td>
<td>482</td>
</tr>
<tr>
<td></td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-a81976fee7973a3c722c1cedc2ede84f">http://databugger.aksw.org/tests#foaf-INVFUNC-a81976fee7973a3c722c1cedc2ede84f</a></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-b009723769eb05dc5b5d67594816a6da">http://databugger.aksw.org/tests#foaf-INVFUNC-b009723769eb05dc5b5d67594816a6da</a></td>
<td>69</td>
<td>168</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-b6b5b018064e92966bd79a6648b369a7">http://databugger.aksw.org/tests#foaf-INVFUNC-b6b5b018064e92966bd79a6648b369a7</a></td>
<td>2474</td>
<td>21301</td>
</tr>
<tr>
<td></td>
<td><a href="http://databugger.aksw.org/tests#foaf-INVFUNC-ece13a3f9c3919a10d56b18599412cc0">http://databugger.aksw.org/tests#foaf-INVFUNC-ece13a3f9c3919a10d56b18599412cc0</a></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-OWLCARD-0cab7cf9453873d6fdd60fac66544246">http://databugger.aksw.org/tests#foaf-OWLCARD-0cab7cf9453873d6fdd60fac66544246</a></td>
<td>1</td>
<td>7566</td>
</tr>
<tr>
<td>F</td>
<td><a href="http://databugger.aksw.org/tests#foaf-OWLCARD-28319b6c1b670d59d90438819f7e3b4">http://databugger.aksw.org/tests#foaf-OWLCARD-28319b6c1b670d59d90438819f7e3b4</a></td>
<td>1</td>
<td>484</td>
</tr>
</tbody>
</table>
Sieve

- Uses metadata to assess data quality of RDF datasets and to filter the data [http://sieve.wbsg.de/](http://sieve.wbsg.de/)

- Input:
  - a dataset, given as a set of Named Graphs
  - provenance data associated with these graphs

- Main functionality:
  - computes various, configurable quality scores for the graphs (based on the provenance data)
  - these scores are represented as RDF data

- Data fusion component
  - merges parts of the data of the Named Graphs
  - filters out some data based on the quality scores
Sieve Configuration Example

```
<QualityAssessment name="Recent and Reputable is Best">
  <AssessmentMetric id="sieve:reputation">
    <ScoringFunction class="ScoredList">
      <Param name="list" value="http://en.wikipedia.org
                     http://es.wikipedia.org
                     http://fr.wikipedia.org"/>
    </ScoringFunction>
  </AssessmentMetric>

  <AssessmentMetric id="sieve:recency">
    <ScoringFunction class="TimeCloseness">
      <Param name="timeSpan" value="50000"/>
      <Input path="?GRAPH/ldif:lastUpdate"/>
    </ScoringFunction>
  </AssessmentMetric>
</QualityAssessment>
```
Generic ”Data Wrangling” Tools

“Data wrangling is the process of taking data in its native format and making it usable for analysis.” –https://www.trifacta.com/

- OpenRefine (formerly Google Refine, open source)
  - http://openrefine.org/

- Trifacta Data Wrangler (commercial)
  - https://www.trifacta.com/products/wrangler/

- Tamr (commercial)
Options

• Tools that allow users to identify quality issues (e.g., by highlighting outliers or similarities)
• Tools that identify quality issues (semi-)automatically
• Tools that fix these issues in an automated process