WWW 2017 Tutorial:
Semantic Data Management in Practice

Part 7: Cleaning

Olaf Hartig

Linköping University

olaf.hartig@liu.se

@olafhartig

Olivier Curé

University of Paris-Est Marne la Vallée

olivier.cure@u-pem.fr

@oliviercure

LINKÖPING UNIVERSITY

UNIVERSITÉ PARIS-EST MARNE-LA-VALLÉE

Terminology

- 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]

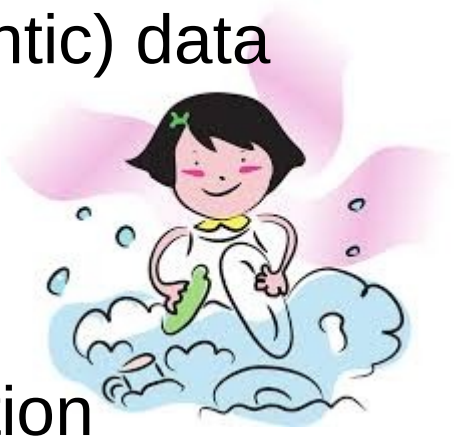
- There are a number of methodologies, for instance:
  1. Audit the data to identify quality issues
  2. Choose methods to automatically detect and remove the issues
  3. Apply the methods
  4. Post-processing / control step [Müller and Freytag 2003]

Rahm and Do: *Data Cleaning: Problems and Current Approaches*. IEEE Data Eng. Bull. 23(4): 3-13, 2000

Müller and Freytag: *Problems, Methods, and Challenges in Comprehensive Data Cleansing*. Technical Report, Humboldt-Universität zu Berlin, HUB-IB-164, 2003.

Goal

- 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  - … be in source datasets (e.g., inaccurate or wrong data items, outdated data items)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 for 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 theses issues automatically

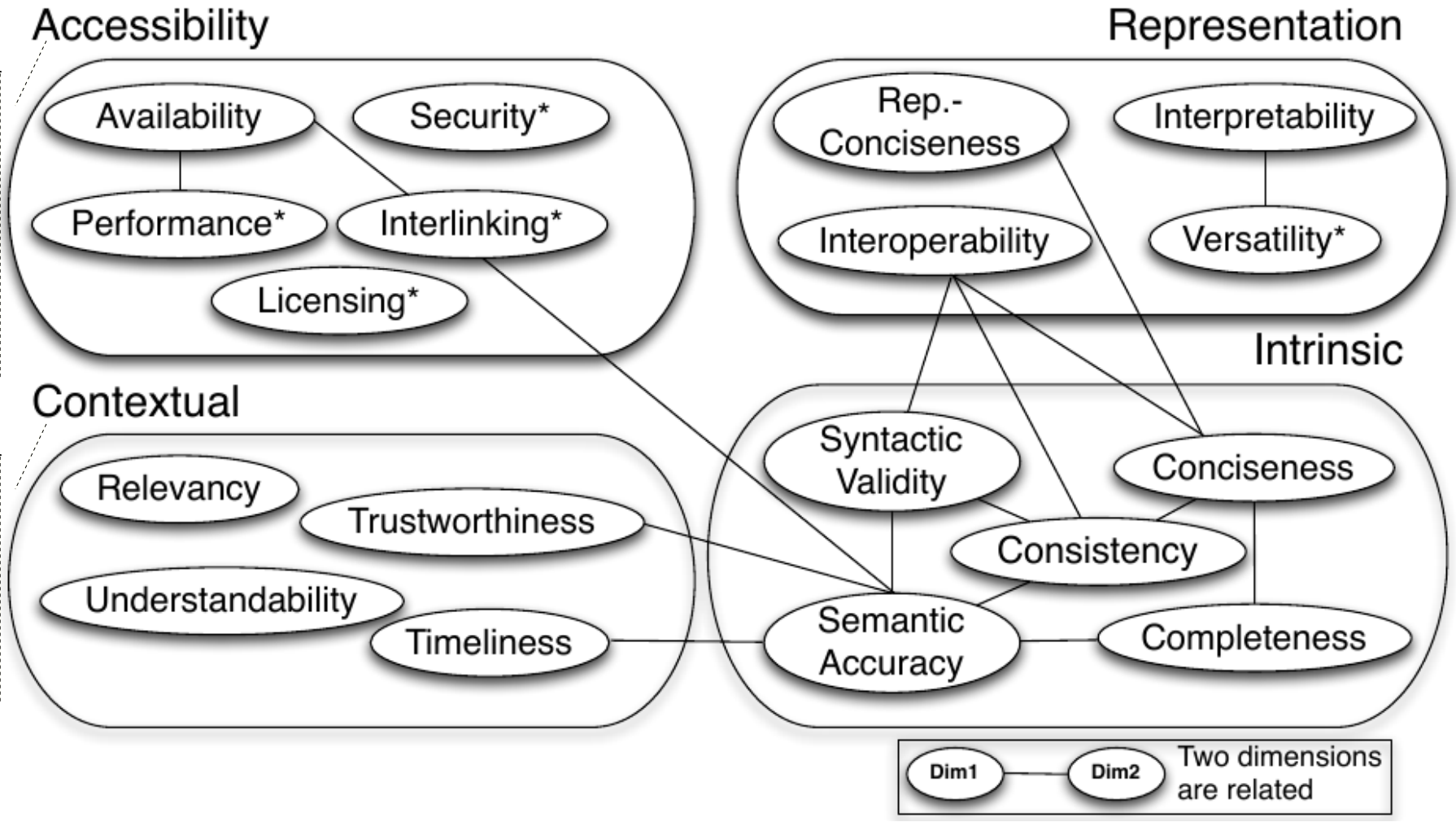
More Terminology

- 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)

Aspects of access, authenticity, and retrieval of data

Depend on the context of the task at hand



Zaveri et al.: *Quality Assessment for Linked Data: A Survey*. Semantic Web 7(1), 2016

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

Zaveri et al.: *Quality Assessment for Linked Data: A Survey*. Semantic Web 7(1), 2016

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
- Ser
cor
- Co
cor
- Co
of e
- 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

Zaveri et al.: *Quality Assessment for Linked Data: A Survey*. Semantic Web 7(1), 2016

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.

Zaveri et al.: *Quality Assessment for Linked Data: A Survey*. Semantic Web 7(1), 2016

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:birthDate ?v1 .
  ?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 generated automatically (based on a schema) and manually
  - Supported schemas: OWL, SHACL, IBM Resource Shapes, Dublin Core Set Profiles
- Tested data loaded from a specified file or is 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)

Testing Run tests Completed! (S: 0 / F:13 / T: 28 / E : 24603) Cancel

Test Results

S	Test	Errors	Prevalence
F	http://databugger.aksw.org/tests#foaf-INVFUNC-0a77ce81bec99608d28790eb695d11fa	25	-1
F	http://databugger.aksw.org/tests#foaf-INVFUNC-105e1374ad211491979c95caa27ba2f5	53	786
-	http://databugger.aksw.org/tests#foaf-INVFUNC-11eb481f2e37c9e1fd18066d637bc013	-	-
F	http://databugger.aksw.org/tests#foaf-INVFUNC-18fb0cf9dc8ff9ad9d42982e0434db2c	476	1214
F	http://databugger.aksw.org/tests#foaf-INVFUNC-2e2b3b0e569d5316d760bdf30f9ecf48	34	87
F	http://databugger.aksw.org/tests#foaf-INVFUNC-4fe77a880206d4b9a00b9972176043b1	84	244
F	http://databugger.aksw.org/tests#foaf-INVFUNC-58e73e30a1082f24e75ecb7c394415d9	21219	366471
F	http://databugger.aksw.org/tests#foaf-INVFUNC-9e12004a97dd6757449f9a1acf86b2a0	165	482
-	http://databugger.aksw.org/tests#foaf-INVFUNC-a81976fee7973a3c722c1cedc2ede84f	-	-
F	http://databugger.aksw.org/tests#foaf-INVFUNC-b009723769eb05dcb5d67594816a6dba	69	168
F	http://databugger.aksw.org/tests#foaf-INVFUNC-b6b5b018064e92966bd79a6648b369a7	2474	21301
-	http://databugger.aksw.org/tests#foaf-INVFUNC-ece13a3f9c3919a10d56b18599412cc0	-	-
F	http://databugger.aksw.org/tests#foaf-OWLCARD-0cab7cf9453873d6fdd60fac66544246	1	7566
F	http://databugger.aksw.org/tests#foaf-OWLCARD-28319b6c1b670d59d90438819fe7e3b4	1	484

Sieve

- Uses metadata to assess data quality of RDF datasets and to filter the data 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)
  - http://www.tamr.com/product/

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 theses issues automatically