Web Data Integration

Data Quality Assessment and Data Fusion
The Data Integration Process

Data Collection / Extraction

Schema Mapping
Data Translation

Identity Resolution

Data Quality Assessment
Data Fusion
Final Exam HWS2022 (IE670, 3 ECTS)

- **Date and Time**
  - Thursday, 15.12.2022, offline, time and room will be announced

- **Format**
  - 6 open questions that show that you have understood the content of the lecture (5 points per question)
  - All lecture slide sets are relevant, including
    - pro and cons of web data publication mechanisms
    - XML syntax and DTDs
    - XPath or SPARQL query (one question)
    - pro and cons of schema matching methods + data samples
    - blocking, matching rules, learning entity matching rules,
    - strength and weaknesses of different similarity measures
    - data fusion, conflict resolution methods, evaluation measures, profiling
  - We want precise answers, not all you know about the topic
  - Three example questions and answers are provided on the course webpage
Outline

1. Introduction
2. Data Profiling
3. Data Provenance
4. Data Quality Assessment
5. Data Fusion
   1. Slot Filling and Conflict Resolution
   2. Conflict Resolution Functions
   3. Evaluation of Fusion Results
   4. Case Studies
1. Introduction

Information providers on the Web have
- different levels of knowledge
- different views of the world
- different intentions

Therefore,
1. information on the Web is partly wrong, biased, outdated, incomplete, and inconsistent.
2. every piece of information on the Web needs to be considered as a claim by somebody at some point in time and not as a fact.
3. the information consumer needs to make up her mind which claims to use for a certain task.
Example: Area and Population of Monaco

Area: Different claims and different conversions

- en.wikipedia.org: 2.02 sq km, 0.78 sq miles
- www.state.gov: 1.95 sq km, 0.8 sq miles
- www.atlapedia.com: 1.94 sq km, 1 sq mile

(1.95 sq km = 0.753 sq miles)

Population: Different claims and vague meta-information

<table>
<thead>
<tr>
<th>Value</th>
<th>Meta-information</th>
<th>Webpage</th>
</tr>
</thead>
<tbody>
<tr>
<td>39,042</td>
<td>(2019 latest UN estimate)</td>
<td><a href="https://www.worldometers.info/world-population/monaco-population/">https://www.worldometers.info/world-population/monaco-population/</a></td>
</tr>
</tbody>
</table>

Source: Peter Bunemann
Definition: Data Conflict

Multiple records that describe the same real-world entity provide **different values for the same attribute**.

<table>
<thead>
<tr>
<th>DB1</th>
<th>DB2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chris Miller</td>
<td>Christian Miller</td>
</tr>
<tr>
<td>12/20/1982</td>
<td>2/20/1982</td>
</tr>
<tr>
<td>Bardon Street, Melville</td>
<td>7 Bardon St., Melville</td>
</tr>
</tbody>
</table>

Reasons for data conflicts:

1. **Data creation**: Typos, measurement errors, erroneous information extraction
2. **Data currency**: Different points in time, missing updates
3. **Data semantics**: Different definitions of concepts (like population or GDP)
4. **Data representation**: Different coding of values ("Mrs." vs. "2")
5. **Data integration**: Wrong data translation or identity resolution
6. **Actual disagreement** of data providers: Subjective attributes
**Definition: Data Fusion**

Given multiple records that describe the same real-world entity, create a single record by resolving data conflicts.

<table>
<thead>
<tr>
<th>DB1</th>
<th>Chris Miller</th>
<th>NULL</th>
<th>Bardon Street, Melville</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB2</td>
<td>Christian Miller</td>
<td>12/20/1982</td>
<td>7 Bardon St., Melwille</td>
</tr>
<tr>
<td>New</td>
<td>Christian Miller</td>
<td>12/20/1982</td>
<td>7 Bardon St., Melville</td>
</tr>
</tbody>
</table>

- **Goal:** Create a high-quality record.
- But what does high data quality mean?
Data Quality

Data quality is a multi-dimensional construct which measures the **fitness for use** of data for a **specific task**.

Fitness for use

1. **has many dimensions**
   - accuracy, timeliness, completeness, understandability, …

2. **is task-dependent**
   - higher quality requirements when you invest one million €

3. **is subjective**
   - some people are more paranoid than others
Data Quality Assessment

- **Content-based Metrics**
  - use information to be assessed itself as quality indicator
  - examples: voting, constraints and consistency rules, statistical outlier detection

- **Provenance-based Metrics**
  - employ provenance meta-information about the circumstances in which information was created as quality indicator
  - examples: “Disbelieve everything a vendor says about its competitor” or “Do not use information that is older than one week”

- **Rating-based Metrics**
  - rely on explicit or implicit ratings about information itself, information sources, or information providers
  - examples: “Only read news articles having at least 100 Facebook likes”, “Accept recommendations from a friend on restaurants, but distrust him on computers”, “Prefer content from websites having a high PageRank”
Summary: Elements of the Data Fusion Process

1. Input Data
2. Grouped Data
3. Fused Data

- Provenance Metadata
- Ratings, other knowledge

Data Quality Assessment
Apply Conflict Resolution Function
Outline

1. Introduction
2. Data Profiling
3. Data Provenance
4. Data Quality Assessment
5. Data Fusion
   1. Slot Filling and Conflict Resolution
   2. Conflict Resolution Functions
   3. Evaluation of Fusion Results
   4. Case Studies
2. Data Profiling

Data profiling refers to the activity of calculating statistics and creating summaries of a data source or data lake.

- profiling lays the foundation for recognizing data quality problems
- manual exploration (data gazing) should be supported with profiling results

## 2.1 Single Column Profiling: Metrics

<table>
<thead>
<tr>
<th>Category</th>
<th>Task</th>
<th>Task Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardinalities</td>
<td>num-rows</td>
<td>Number of rows</td>
</tr>
<tr>
<td></td>
<td>null values</td>
<td>Number or percentage of null values</td>
</tr>
<tr>
<td></td>
<td>distinct</td>
<td>Number of distinct values</td>
</tr>
<tr>
<td></td>
<td>uniqueness</td>
<td>Number of distinct values divided by number of rows</td>
</tr>
<tr>
<td>Value Distributions</td>
<td>histogram</td>
<td>Frequency histograms (equi-width, equi-depth, etc.)</td>
</tr>
<tr>
<td></td>
<td>extremes</td>
<td>Minimum and maximum values in a numeric column</td>
</tr>
<tr>
<td></td>
<td>constancy</td>
<td>Frequency of most frequent value divided by number of rows</td>
</tr>
<tr>
<td></td>
<td>quartiles</td>
<td>Three points that divide (numeric) values into four equal groups</td>
</tr>
<tr>
<td></td>
<td>first digit</td>
<td>Distribution of first digit in numeric values; to check Benford’s law</td>
</tr>
<tr>
<td>Data Types, Patterns, and Domains</td>
<td>basic type</td>
<td>Numeric, alphanumeric, date, time, etc.</td>
</tr>
<tr>
<td></td>
<td>data type</td>
<td>DBMS-specific data type (varchar, timestamp, etc.)</td>
</tr>
<tr>
<td></td>
<td>lengths</td>
<td>Minimum, maximum, median, and average lengths of values within a column</td>
</tr>
<tr>
<td></td>
<td>size</td>
<td>Maximum number of digits in numeric values</td>
</tr>
<tr>
<td></td>
<td>decimals</td>
<td>Maximum number of decimals in numeric values</td>
</tr>
<tr>
<td></td>
<td>patterns</td>
<td>Histogram of value patterns (Aa9...)</td>
</tr>
<tr>
<td></td>
<td>data class</td>
<td>Generic semantic data type, such as code, indicator, text, date/time, quantity, identifier</td>
</tr>
<tr>
<td></td>
<td>domain</td>
<td>Semantic domain, such as credit card, first name, city, phenotype</td>
</tr>
</tbody>
</table>

Central for judging the usefulness of attributes

A histogram says more than thousand averages
- outliers
- skewed distributions

Data types and lengths should always be reported

Good summarization of an attribute
Single Column Profiling: Examples

RapidMiner

Microsoft Power BI
Single Column Profiling: Examples

Goolge Cloud Dataprep by Trifacta

- Data type mismatch
- Most frequent values
- Most frequent value patterns
Profiling in Python

https://github.com/pandas-profiling/pandas-profiling

Correlations and missing values
Data type detection
2.2 Data Lake Profiling: Data and Schema Overlap

- **Approach:** Match data to central database
- **Example:** Profiling a corpus of 33.3 million HTML tables by matching them to the DBpedia knowledge base

- **Results**
  - 301,000 tables (1%) have matching rows and matching columns
  - 8,000,000 million values for fusion

- **Interpretation**
  - Topical bias of KB needs to be considered
  - Product tables missed

---

Data Lake Profiling: Topic Overlap

- Approaches:
  1. Train supervised classifier to categorize data sources / tables into predefined categories using textual metadata, schema-level labels, or textual content
  2. Cluster sources / tables based on textual metadata and/or textual content

- Example:
  - 100 LOD data sources manually assigned to 9 categories
  - 1000 records sampled per data source
  - 900 additional data sources classified with F1 of 0.81

Böhm, Kasneci, Naumann: Latent topics in graph-structured data. CIKM 2012.
3. Data Provenance

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Source: W3C PROV Specification

Provenance information = important data quality indicator

Outline of this Subsection

1. Simple Attribution versus Full Provenance Chains
2. Publishing Provenance Information on the Web
3. Representing Provenance Metadata together with Integrated Data
3.1 Simple Attribution versus Full Provenance Chains

1. Simple Attribution:
   • state who created a document/data item and when it was created
   • standard: Dublin Core vocabulary

2. Full Provenance Chains
   • Describe the full process of data creation / reuse / integration / aggregation
   • standard: W3C PROV Specification
   • alternative name: Data Lineage (explain why something is in a query result)

- Factors for the decision between both alternatives:
  • Will the users be interested in all the details?
    • Yes for science, investing, law suits. No for minor purchases in e-commerce
  • Can target applications understand/reason about all details?
In the context of the Web, you always know the **URL** from which you downloaded things. Some sites also give you **Last-Modified** information.

**HTTP-Response**

```
HTTP/1.1 200 OK
Date: Mon, 18 Jan 2019 20:54:26 GMT
Server: Apache/1.3.6 (UNIX)
Last-Modified: Mon, 06 Dec 2018 14:06:11 GMT
Content-length: 6345
Content-Type: text/html

<html>
  <head><title>CB CD-Shop</title></head>
  <body><h1>Willkommen beim CB CD-Shop</h1> ....
```

Which vocabularies/schemata should websites use to publish more detailed provenance information?
Dublin Core

- The Dublin Core vocabulary defines terms for representing simple attribution information
  - creator, contributor, publisher, date, rights, format, language, ...
- The terms are used in different technical contexts
  - HTML, Linked Data, proprietary library formats
  - Example of a Linked Data document:

http://dbpedia.org/data/Alec_Empire

```xml
# Metadata and Licensing Information
<http://dbpedia.org/data/Alec_Empire>
  rdfs:label "RDF document describing Alec Empire" ;
  rdf:type foaf:Document ;
  dc:publisher <http://dbpedia.org/resource/DBpedia> ;
  dc:date "2019-07-13"^^xsd:date ;

# The Document Content
<http://dbpedia.org/resource/Alec_Empire>
  foaf:name "Empire, Alec" ;
  rdf:type foaf:Person ;
  rdfs:comment "Alec Empire (born May 2, 1972) is a German musician..."@en ;
  ...
```
The W3C PROV vocabulary defines terms for representing complex provenance chains.

Example of a PROV XML document:

```xml
<prov:document>
  <!-- Entities -->
  <prov:entity prov:id="exn:article">
    <dct:title>Crime rises in cities</dct:title>
  </prov:entity>
  <!-- Agents -->
  <prov:agent prov:id="exc:derek">
    <prov:type>prov:Person</prov:type>
    <foaf:givenName>Derek Smith</foaf:givenName>
    <foaf:mbox>mailto:derek@example.org</foaf:mbox>
  </prov:agent>
  <!-- Activities -->
  <prov:activity prov:id="exc:compile1"/>
  <!-- Usage and Generation -->
  <prov:wasGeneratedBy>
    <prov:entity prov:ref="exn:article"/>
    <prov:activity prov:ref="exc:compile1"/>
  </prov:wasGeneratedBy>
  <!-- Agent's Responsibility -->
  <prov:wasAssociatedWith>
    <prov:activity prov:ref="exc:compile1"/>
    <prov:agent prov:ref="exc:derek"/>
  </prov:wasAssociatedWith>
</prov:document>
```
More Complex Example: W3C PROV

![Graph showing relationships between different entities with provenance data, including articles, datasets, and charts.]
3.3 Representing Provenance Metadata together with Integrated Data
Relational Data Model

- Alternative 1: Record-Level Provenance (coarse grained, fast queries)
- Alternative 2: Value-Level Provenance (fine grained, but slow queries)
- Alternative 3: Employ special database engine which implements extended relational data model with a pointer to provenance information for each attribute value (e.g. Stanford Trio Database)

Physicians with **Record-Level Provenance**

<table>
<thead>
<tr>
<th>Key</th>
<th>Name</th>
<th>Street</th>
<th>ProvID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1425</td>
<td>Dr. Mark Smith</td>
<td>14 Main Street</td>
<td>001</td>
</tr>
<tr>
<td>1425</td>
<td>Mark Smith</td>
<td>12 Main St.</td>
<td>002</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Provenance Table**

<table>
<thead>
<tr>
<th>ProvID</th>
<th>Source</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Physicians with **Value-Level Provenance**

<table>
<thead>
<tr>
<th>Key</th>
<th>Attribute</th>
<th>Value</th>
<th>ProvID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1425</td>
<td>Name</td>
<td>Dr. Mark Smith</td>
<td>001</td>
</tr>
<tr>
<td>1425</td>
<td>Name</td>
<td>Mark Smith</td>
<td>002</td>
</tr>
<tr>
<td>1425</td>
<td>Street</td>
<td>14 Main Street</td>
<td>001</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Represent provenance using multiple value elements and references to provenance elements.

```xml
<physician>
  <name>
    <value prov="prov01">Dr. Mark Smith</value>
    <value prov="prov02">Mark Smith</value>
  </name>
  <address>
    <street>
      <value prov="prov01">14 Main Street</value>
      <value prov="prov02">12 Main St.</value>
    </street>
    <city>...</city>
  </address>
</physician>
<provenance id="prov01">
  <source>http://www.marksmith.com/index.htm</source>
  <date>06 Nov 2018 14:06:11 GMT</date>
</provenance>
<provenance id="prov02">
  ...
</provenance>
```
- Group triples into **Named Graphs** (= set of triples that is identified by a URI)
- Provide provenance information by talking about a graph in another graph
- Named Graphs can be queried using the SPARQL keyword GRAPH

**Carroll, Bizer, Hayes, Stickler: Named Graphs. Journal of Web Semantics, 2005.**
4. Data Quality

Data quality is a multi-dimensional construct which measures the “fitness for use” of data for a specific task.

- Which quality dimensions matter depends on the task
- The required level of quality depends on the task and the user

Outline of this Subsection

4.1 Data Quality Dimensions

4.2 Data Quality Assessment
Data Quality in the Enterprise and Web Context

- **Enterprise Context**
  - the goal is to establish *procedures and rules* that guarantee high quality data production, quality monitoring, and regular data cleansing
  - pioneering research by MIT Total Data Quality Management (TDQM) program
  - consequences of low data quality:
    - US postal service: out of 100,000 mass-letters, 7,000 cannot be delivered because of wrong address
    - A.T. Kearny: 25%-40% of the operational costs result from low data quality as low quality data leads to wrong management decisions
    - SAS: Only 18% of all German companies trust their data

- **Web Context**
  - large number of data sources, but no possibility to influence data providers
  - thus, focus on *identifying the high-quality subset* of the available data
  - challenge: quality indicators are often sparse and unreliable
4.1 Data Quality Dimensions

As part of the MIT Total Data Quality Management (TDQM) program, [Wang/Strong1996] asked managers which data quality dimensions matter for their tasks:

- Fitness for use
  - Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Understandability, Consistency, Concise Representation

179 Dimensions
<table>
<thead>
<tr>
<th>Category</th>
<th>IQ Criteria</th>
<th>TDQM</th>
<th>MBIS</th>
<th>Weikum</th>
<th>DWQ</th>
<th>SCOUG</th>
<th>Chen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content-related Criteria</td>
<td>Accuracy</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Relevancy</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Value-Added</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Interpretability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Technical Criteria</td>
<td>Timeliness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Latency</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Performability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Response time</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Security</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accessibility</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Price</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Customer Support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intellectual Criteria</td>
<td>Believability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Reputation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Objectivity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instantiation related</td>
<td>Verifiability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criteria</td>
<td>Amount of data</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understandability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Concise represent.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consistent represent.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Felix Naumann
Relevancy of Data Quality Dimensions

Which quality dimensions matter depends on the task at hand.
4.2. Data Quality Assessment

Various domain-specific heuristics are used to measure data quality.

The applicability of specific heuristics depends on
1. Availability of quality indicators (like provenance information or ratings)
2. Quality of quality indicators (fake ratings, sparse provenance information)
Quality Indicators in the Web Context

- Background Information
- Meta-Information
- Information Consumer uses Piece of Information to be assessed
- Background Information
- Information Provider provides
- Other Actor

Ratings

Background Information

Background Information

Ratings
4.2.1 Assessing Data Accuracy

**Definition Accuracy:** The extent to which data is correct, reliable, and free of error.

- also called: Truth Discovery, Fake News Detection

- **Assessment Methods:**
  1. Constraint testing
  2. Outlier detection
  3. Expert- or user ratings

- Relevant quality indicators:
Constraint Testing

Match data against constraints and consistency rules in order to detect errors.

- Examples of constraints
  - the age of humans should be between 0 and 130
  - books must have at least one author

- Examples of consistency rules
  - if person is in middle school, then age is (likely) below 25
  - if area code is 131, then the city should be Edinburgh

- Rule and constraint acquisition
  - define rules and constraints manually
  - or learn from examples e.g. using association analysis (see lecture Data Mining)

Outlier Detection

An outlier is an individual data instance that is anomalous with respect to the rest of the data.

- Outliers can be considered as errors and be assigned a low quality score
- Techniques
  - statistical distributions, clustering, classification
- Challenges
  - the exact notion of an outlier is different for different application domains
  - an individual may be a outlier w.r.t. a single attribute or a combination of multiple attributes
  - natural outliers: population of Mexico City
  - normal behaviour keeps evolving over time

Ratings

Data is often filtered or ranked based on ratings provided by users or experts.

- Various scoring functions exist
  - practical systems often use simple, easily understandable functions

- Challenges:
  1. Motivate users to rate
     - data, data providers, data sources
  2. Quality of the ratings
     - fake ratings
     - clueless raters

- Events interpretable as positive ratings
  - clicks, page views
  - time spent on some page
The assessment of the timeliness of data usually requires provenance data.

- **Provenance metadata**
  - HTTP Last-Modified
  - dc:date

- **Fallbacks if no timestamps are available**
  - propagate timestamps to data without timestamps
    - e.g. two tables provide same profit for a company, only one table has a timestamp
  - use rules instead of timestamps
    - Number of children: Prefer higher value, as number of children of a person usually grows
4.2.3 Assessing Data Completeness

Definition Completeness: The extent to which data is not missing and is of sufficient breadth, depth, and scope for the task at hand.

- Two perspectives on completeness:
  - **Density**: Fraction of attributes filled
  - **Coverage**: Fraction of real-world objects represented

- Assessment:
  - Density
    - sample data source and calculate density from sample
  - Coverage
    - hard to calculate as overall number of real-world objects is unknown in many cases: countries fine; products or people problematic
    - fallback: prefer data sources that describe more entities
Definition Relevancy: The extent to which data is applicable and helpful for the task at hand.

- **Assessment:**
  - Example: TripAdvisor
    - Filter reviews based on background information about information provider
  - Example: Google
    - Rank webpages based on search terms and PageRank score
4.2.5 Assessing Believability / Trustworthiness

Definition Believability / Trustworthiness: The extent to which data is regarded as true, real, and credible.

- Subjective dimension which depends on the individual user
  - Assessment:
    - individual experience with the data
    - fallbacks:
      - corporate guidance about sources
      - trust networks
  - Explanations about the data quality assessment process
    - in order to trust data, the users must understand why the system regards data to be high quality
    - Tim Berners-Lee’s “Oh, yeah?”-button
Prototype: The WIQA - Browser

- Enables users to employ different quality assessment policies
- Can explain assessment results
Explanation about an Assessment Decision

The Triple:

**Siemens Share positive analyst report** Siemens agrees partnership with Novell unit SUSE. Siemens Business Services (SBS), the IT services arm of German technology conglomerate Siemens (nasdaq: NOVL - news - people) newly acquired can be freely copied and modified, unlike proprietary software such as Microsoft (nasdaq: MSFT - news - people) Windows. In the past months clients are asking more and more for open-source platforms. SBS said in a statement which said SUSE would have premier partner status. SBS is one of Europe's top 10 information technology service providers. Linux, once the exclusive province of a few dedicated enthusiasts, is now seen as the only serious rival to Windows and is supported by U.S. giant International Business Machines (nyse: IBM - news - people), among others. Its advocates, who include big businesses and government departments, argue it is cheaper, simpler and more secure than Windows.

**fulfils the policy:**

Use only information which has been asserted by German analysts.

**because:**

- it is stated in the document Information from Peter Smith, which is asserted by the German analyst Peter Smith.
Example Explanation

The triple:
- Siemens AG has positive analyst report: "As Siemens agrees partnership with Novell unit SUSE ..."

fulfills the policy:
- Accept only information that has been asserted by people who have received at least 3 positive ratings.

because:
- it was asserted by Peter Smith and
- Peter Smith has received positive ratings from
  - Mark Scott who works for Siemens.
  - David Brown who works for Intel.
Summary

- Data quality assessment is essential for web data integration as errors accumulate:
  1. Quality of the external data sources (everybody can publish on the Web)
  2. Quality of the integration process (wrong mappings, wrong identity resolution)
- Many data quality problems only become visible when we integrate data from multiple sources
- A wide range of different quality assessment heuristics can be used
  - content-based, provenance-based, rating-based metrics
- The applicability of the heuristics depends on
  - the availability of quality indicators (like provenance information or ratings)
  - quality of quality indicators (fake ratings, coarse grained provenance)
- Many systems only try to assess the accuracy and the timeliness of web data and ignore the other quality dimensions
5. Data Fusion

Given multiple records that describe the same real-world entity, create a single record while resolving conflicting data values.

- Goal: Create a single high-quality record.
- Two basic fusion situations: Slot Filling and Conflict Resolution
5.1 Slot Filling and Conflict Resolution

**Slot Filling:** Fill missing values (NULLs) in one dataset with corresponding values from other datasets.

Result: increased dataset density

**Conflict Resolution:** Resolve contradictions between records by applying a conflict resolution function (heuristic).

Result: increased data quality

---

**Complementary records**

<table>
<thead>
<tr>
<th>a, b, -, -</th>
<th>→</th>
<th>a, b, c, -</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, c, -</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conflicting records**

<table>
<thead>
<tr>
<th>a, b, c, -</th>
<th>→</th>
<th>a, f(b,e), c, -</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, e, -, -</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a, -, c, -</th>
<th>→</th>
<th>a, f(b,e), c, d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b, -, -</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a, e, -, d</th>
<th></th>
<th></th>
</tr>
</thead>
</table>
As final step of the identity resolution process, records are clustered using the discovered correspondences. Example with 3 data sources:

<table>
<thead>
<tr>
<th>Cluster Size</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4256</td>
</tr>
<tr>
<td>2</td>
<td>939</td>
</tr>
<tr>
<td>3</td>
<td>503</td>
</tr>
<tr>
<td>4</td>
<td>75</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
</tr>
<tr>
<td>61</td>
<td>1</td>
</tr>
</tbody>
</table>

No slot filling possible as single records with no overlap
Slot filling and conflict resolution allow the generation of improved records
Large cluster size indicates matching errors or duplicates in data sources

Cluster size distribution from matching web tables to DBpedia
- Out of 33.3 million web tables, 949,970 tables contain at least one matching row
- 42% of the clusters have a size of 2
- 16% of the clusters have a size of 3
- 39% of the clusters have a size of at least 4
- 13% of the clusters have a size of at least 11

5.2 Conflict Resolution Functions

- Conflict resolution functions are attribute-specific
  - you select or learn a specific function for each attribute that should be fused

- There is a wide range of different functions (heuristics) that fit different requirements

- Functions differ in regard to the data types, they can be applied for
  - numerical values (e.g. population of a place)
  - nominal values (e.g. name of a person)
  - value sets (e.g. actors performing in a movie)

- Two main categories of conflict resolution functions
  1. **Content-based functions** that rely only on the data values to be fused
  2. **Metadata-based functions** that rely on provenance data, ratings, or quality scores

\[ V_F = f (V_A, M_A, B) \]
### Content-based Conflict Resolution Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Explanation</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average, Median</td>
<td>Calculate average/median of all values</td>
<td>Rating</td>
</tr>
<tr>
<td>Longest, Shortest</td>
<td>Choose longest / shortest value</td>
<td>First name</td>
</tr>
<tr>
<td>Max, Min</td>
<td>Take maximal, minimal value</td>
<td>Number of children</td>
</tr>
<tr>
<td>Vote</td>
<td>Majority decision (one vote per site or page?)</td>
<td>Mayor of city</td>
</tr>
<tr>
<td>Clustered Vote</td>
<td>Choose centroid / medoid of largest cluster</td>
<td>Population of city</td>
</tr>
<tr>
<td>Weighted Vote</td>
<td>Weight sources according to the fraction of true values they provided</td>
<td>Address of a shop</td>
</tr>
<tr>
<td>Union</td>
<td>Union of all values (A ∪ B ∪ C)</td>
<td>Product Reviews</td>
</tr>
<tr>
<td>Intersection</td>
<td>Intersection of all values (A ∩ B ∩ C)</td>
<td>Movie Actors</td>
</tr>
<tr>
<td>IntersectionKSources</td>
<td>Values must appear in at least k sources</td>
<td>Movie Actors</td>
</tr>
<tr>
<td>MostComplete</td>
<td>Choose value from record that is most complete</td>
<td>Postal addresses</td>
</tr>
<tr>
<td>MostAbstract, MostSpecific</td>
<td>Use a taxonomy / ontology</td>
<td>Location</td>
</tr>
<tr>
<td>Random</td>
<td>Fallback: Choose random value</td>
<td></td>
</tr>
</tbody>
</table>
## Metadata-based Conflict Resolution Functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>FavorSources</td>
<td>Take first non-null value in particular order of sources</td>
</tr>
<tr>
<td></td>
<td>Example: Use Eurostat for GDP, alternatively use Wikipedia</td>
</tr>
<tr>
<td>MostRecent</td>
<td>Choose most recent (up-to-date) value</td>
</tr>
<tr>
<td></td>
<td>Example: Address, phone number</td>
</tr>
<tr>
<td>MostActive</td>
<td>Choose value that is most often accessed/edited</td>
</tr>
<tr>
<td></td>
<td>Example: Prefer Wikipedia page with more edits</td>
</tr>
<tr>
<td>FavorSources</td>
<td>Calculate quality of sources from ratings, take value from source with highest score or all values from sources with scores above specific threshold</td>
</tr>
<tr>
<td>basedOnRatings</td>
<td>MaxIQ</td>
</tr>
<tr>
<td></td>
<td>Choose the value with the highest quality score. Score might cover multiple quality dimensions, e.g. timeliness and believability of a source</td>
</tr>
<tr>
<td>TopkIQ</td>
<td>Choose the top K values with the highest quality scores</td>
</tr>
<tr>
<td>ClusterVoteAfterFiltering</td>
<td>Filter values using quality scores and apply clustered vote afterwards</td>
</tr>
<tr>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>
### Example: Complete Conflict Resolution Heuristic

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>ISBN</th>
<th>Price</th>
<th>Review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moby Dick</td>
<td>H. Melville</td>
<td>0766607194</td>
<td>$3.98</td>
<td></td>
</tr>
<tr>
<td>Mopy Dick</td>
<td>Herman Melville</td>
<td>0766607193</td>
<td>$5.99</td>
<td></td>
</tr>
</tbody>
</table>

- **amazon.com**: 8/31/2019
- **bn.com**: 7/20/2018

- Favor Sources: (amazon.com)
- Max Length
- Random
- Most Recent
- Union

**Note:** The names and prices are placeholders for demonstration purposes.
5.3 Evaluation of Fusion Results

1. Data Centric Evaluation Measures
   - Density
   - Consistency

2. Ground Truth Based Evaluation Measures
   - Accuracy
Density

Density measures the fraction of non-NULL values.

\[ \text{density}_{\text{Column}} = \frac{\text{|non-NULL values in column|}}{\text{|rows in table|}} \]

\[ \text{density}_{\text{Table}} = \frac{\text{|non-NULL values in table|}}{\text{|columns|*|rows|}} \]

- As a result of schema integration, translated data sets often contain many null values (empty columns)
- We are interested in the density increase after fusion
  1. Measure density of table A or column \( C_1 \)
  2. Fuse table A with table B
  3. Measure density of resulting table A' or column \( C_1' \)
A data set is consistent if it is free of conflicting information.

\[
\text{consistency}_{\text{Column}} = \frac{|\text{non-conflicting values in column}|}{|\text{real-world entities described}|}
\]

\[
\text{consistency}_{\text{Table}} = \frac{|\text{non-conflicting values in table}|}{|\text{columns}| \times |\text{real-world entities described}|}
\]

Measurement:

1. Group records that refer to same real-world entity
   - using correspondences generated by identity resolution
2. Calculate fraction of non-conflicting attribute values
   - same attribute value is provided by all data sources
Accuracy

**Accuracy:** Fraction of correct values selected by conflict resolution function.

\[
\text{accuracy} = \frac{\text{correct values}}{\text{all values}}
\]

\[
\text{error rate} = 1 - \text{accuracy}
\]

**Measurement:**

1. Gather Ground Truth
   - Manually determine correct values for a subset of the records
   - Alternative: Use/buy correct data from external provider
   - Can be tricky as this requires you or external provider to know the truth!
2. Compare values generated by fusion function with true values

How to Treat Similar Values?

- Treatment of similar values matters for calculating **consistency** and **accuracy**.

- Approach:
  1. Calculate similarity of values
     - using an appropriate similarity function (see slideset Identity Resolution)
  2. Treat similar values as equal (similarity above threshold)

- Example: Mayor of Berlin

  - Michael Müller
  - Michael Mueller M. Müller
  - Michael Müller
  - Klaus Wowereit K. Wowereit
  - Wowi
5.4. Example Data Fusion Tool: Fuz!on

Prototype developed at Hasso Plattner Institute
Manual Fusion of Record Groups in Fuz!on

### Fuzzy Fuz!on

**Additional Information**
- Automatic Fusion
- Rule-based Fusion
- Manual Fusion

**Groups 0 to 50 of 100000**

<table>
<thead>
<tr>
<th>fdb.group</th>
<th>Firstname</th>
<th>Lastname</th>
<th>Street</th>
<th>housenumber</th>
<th>postcode</th>
<th>city</th>
<th>ignore</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>31750025-01</td>
<td>Werner</td>
<td>Trimpert</td>
<td>Thomas-Man...</td>
<td>89</td>
<td>24943</td>
<td>Kiel</td>
<td>19470524</td>
<td>0461</td>
</tr>
<tr>
<td>31758055-01</td>
<td>Artur</td>
<td>Heiser</td>
<td>Kalkgrund</td>
<td>4</td>
<td>24939</td>
<td>Kiel</td>
<td>19360106</td>
<td>0461</td>
</tr>
<tr>
<td>31765505-01</td>
<td>Siegfried</td>
<td>Aswegen</td>
<td>Mürwiker Str.</td>
<td>6</td>
<td>4943</td>
<td>Flensburg</td>
<td>19250404</td>
<td>0461</td>
</tr>
<tr>
<td>31772625-01</td>
<td>M. Blankenburg</td>
<td>Harmssr.</td>
<td></td>
<td>48</td>
<td>24116</td>
<td>Kiel</td>
<td>19610727</td>
<td>0461</td>
</tr>
<tr>
<td>31780965-01</td>
<td>K Degen</td>
<td>Peter-Chr.-H...</td>
<td></td>
<td>5</td>
<td>24114</td>
<td>Flensburg</td>
<td>19630331</td>
<td>0461</td>
</tr>
<tr>
<td>31789325-01</td>
<td>Manh The Knaut</td>
<td>Wiedeberger ...</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
<td></td>
</tr>
<tr>
<td>31798345-01</td>
<td>horst Boltsmann</td>
<td></td>
<td></td>
<td>6</td>
<td>24937</td>
<td>Flensburg</td>
<td>19281225</td>
<td>0461</td>
</tr>
</tbody>
</table>

21. **Group**

<table>
<thead>
<tr>
<th>Firstname</th>
<th>Lastname</th>
<th>Street</th>
<th>housenumber</th>
<th>postcode</th>
<th>city</th>
<th>ignore</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manh The</td>
<td>Knaut</td>
<td>Wiedeberger Weg</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
</tr>
<tr>
<td>Manh The</td>
<td>KNAUT</td>
<td>Wiedeberger Weg</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
</tr>
<tr>
<td>Manh</td>
<td>Knaut</td>
<td>WIEDEBERGER WEG</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
</tr>
<tr>
<td>First</td>
<td>Mixed</td>
<td>Vote</td>
<td></td>
<td>First</td>
<td></td>
<td>First</td>
<td></td>
</tr>
<tr>
<td>Manh</td>
<td>The</td>
<td>Knaut</td>
<td>Wiedeberger Weg</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
</tr>
</tbody>
</table>

[Image of software interface showing fusion process]
5.5 Case Study: DBpedia Cross-Language Data Fusion

- Infoboxes in different Wikipedia editions contain conflicting values.
- Which value to prefer?
Cross-Lingual Data in DBpedia

- DBpedia extracts structured data from Wikipedia in 119 languages.
- DBpedia contains lots of data conflicts, inherited from Wikipedia.
- Identity resolution is solved by Wikipedia inter-language links.
- Schema heterogeneity problem is solved by community-created mappings from infoboxes to DBpedia ontology.
Which value to prefer

- maximum?
- average?
- most frequent?
- from the specific language edition?
- most recent?
- inserted by most trusted author?
- edited most times?
- combination of the above?

Goal: Fuse Data between different Language Editions

Population of Mannheim in 8 DBpedia language editions

Mannheim populationTotal
"314,931"@en
"291,458"@de
"311,969"@eu
"311,342"@fr
"308,676"@nl
"309,795"@pt
"313,174"@ru
"310,000"@sl
We extract provenance metadata from the Wikipedia revision dumps of the Top10 languages

- File size of revision dumps: > 6 TByte for English, >2 TByte for German

Extracted metadata

- Last edit timestamp of a fact
- Number of edits of a fact
- Author of the last edit
  - Author edit count
  - Author registration date

Provenance metadata

ru:Mannheim:populationTotal

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lastedit</td>
<td>2011-12-22T00:50:21Z</td>
</tr>
<tr>
<td>propeditcnt</td>
<td>3</td>
</tr>
<tr>
<td>autheditcnt</td>
<td>1136639</td>
</tr>
<tr>
<td>authregdate</td>
<td>2009-12-18T02:08:09Z</td>
</tr>
</tbody>
</table>

nl:Mannheim:populationTotal

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>lastedit</td>
<td>2007-12-09T16:41:06Z</td>
</tr>
<tr>
<td>propeditcnt</td>
<td>1</td>
</tr>
<tr>
<td>autheditcnt</td>
<td>73</td>
</tr>
<tr>
<td>authregdate</td>
<td>2007-04-05T08:54:19Z</td>
</tr>
</tbody>
</table>
Learning Conflict Resolution Functions

- **Ground Truth:** Geonames, public geographical database
- **Learning:** Choose function with smallest mean absolute error with respect to gold standard.
- Tested conflict resolution functions
  1. *Maximum*
  2. *Average*
  3. *English* – prefer values from English DBpedia
  4. *Vote* – choose the most frequent value
  5. *MostRecent* fact – last edit timestamp
  6. *MostActive* fact – number of edits of a property
  7. *MostActive* author – author edit count
  8. *MostSenior* author – author registration date
## DBpedia Case Study: Results

<table>
<thead>
<tr>
<th>Property</th>
<th>Dataset</th>
<th>Count</th>
<th>Learned Fusion Function</th>
<th>Error, %</th>
<th>Error, %, en.dbpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>populationTotal</td>
<td>cities1000-Germany *</td>
<td>7330</td>
<td>Vote (most frequent value)</td>
<td>0.3029</td>
<td>0.6796</td>
</tr>
<tr>
<td>populationTotal</td>
<td>cities1000-Netherlands</td>
<td>493</td>
<td>Maximum Value</td>
<td>2.1933</td>
<td>3.5714</td>
</tr>
<tr>
<td>populationTotal</td>
<td>countries</td>
<td>243</td>
<td>Maximum Value</td>
<td>2.1646</td>
<td>6.3485</td>
</tr>
<tr>
<td>country</td>
<td>cities1000-Italy</td>
<td>1078</td>
<td>Vote</td>
<td>0.0000</td>
<td>1.2060</td>
</tr>
<tr>
<td>country</td>
<td>cities1000-Brazil</td>
<td>1119</td>
<td>Max author edit count</td>
<td>9.8302</td>
<td>30.9205</td>
</tr>
<tr>
<td>country</td>
<td>cities1000-Germany</td>
<td>7638</td>
<td>Vote</td>
<td>0.0131</td>
<td>0.6415</td>
</tr>
</tbody>
</table>

* “cities1000” are cities with population >1000

- **Error**: Mean absolute percentage error between chosen value and ground truth
- **Error en.dbpedia**: Mean absolute percentage error between value in English DBpedia and gold standard

5.6 Case Study: Google Knowledge Vault

- uses 12 different extractors to extract 6.4 billion triples (1.6 billion unique triples) from 1 billion page Web crawl
- extracted data is fused to extend the Freebase knowledge base

Luna Dong, et al.: From Data Fusion to Knowledge Fusion. VLDB 2014.
Google Knowledge Vault

- uses probabilistic model to **iteratively** determine quality of triples, sources, and extractors
- result: 90 million triples with $p>0.9$ that were not in Freebase before

- **Knowledge-based Trust**
  - determine trustworthiness of a data source by comparing its content with a knowledge base (ground truth)
  - result: better than PageRank in identifying
    - tail websites with high trustworthiness
    - gossip websites

Summary: Data Fusion

- Data Fusion addresses **missing values** (slot filling) as well as **contradictions** (conflict resolution)

- Appropriate conflict resolution function depends on
  - data type of the values
  - availability of quality-related metadata
  - availability of overlapping data

- On the Web, we often encounter **long-tailed distributions**
  - lots of overlapping data for head entities (New York)
  - hardly any data to fuse for tail entities (some village)
  - example: Web tables matched to DBpedia
6. References

- Profiling

- Provenance

- Data Quality
  - Abedjan, et al.: Detecting data errors: where are we and what needs to be done? VLDB 2016.
References

- Data Fusion

- Data Fusion Evaluation Datasets
  - Dong: Data Sets for Data Fusion Experiments http://lunadong.com/fusionDataSets.htm