

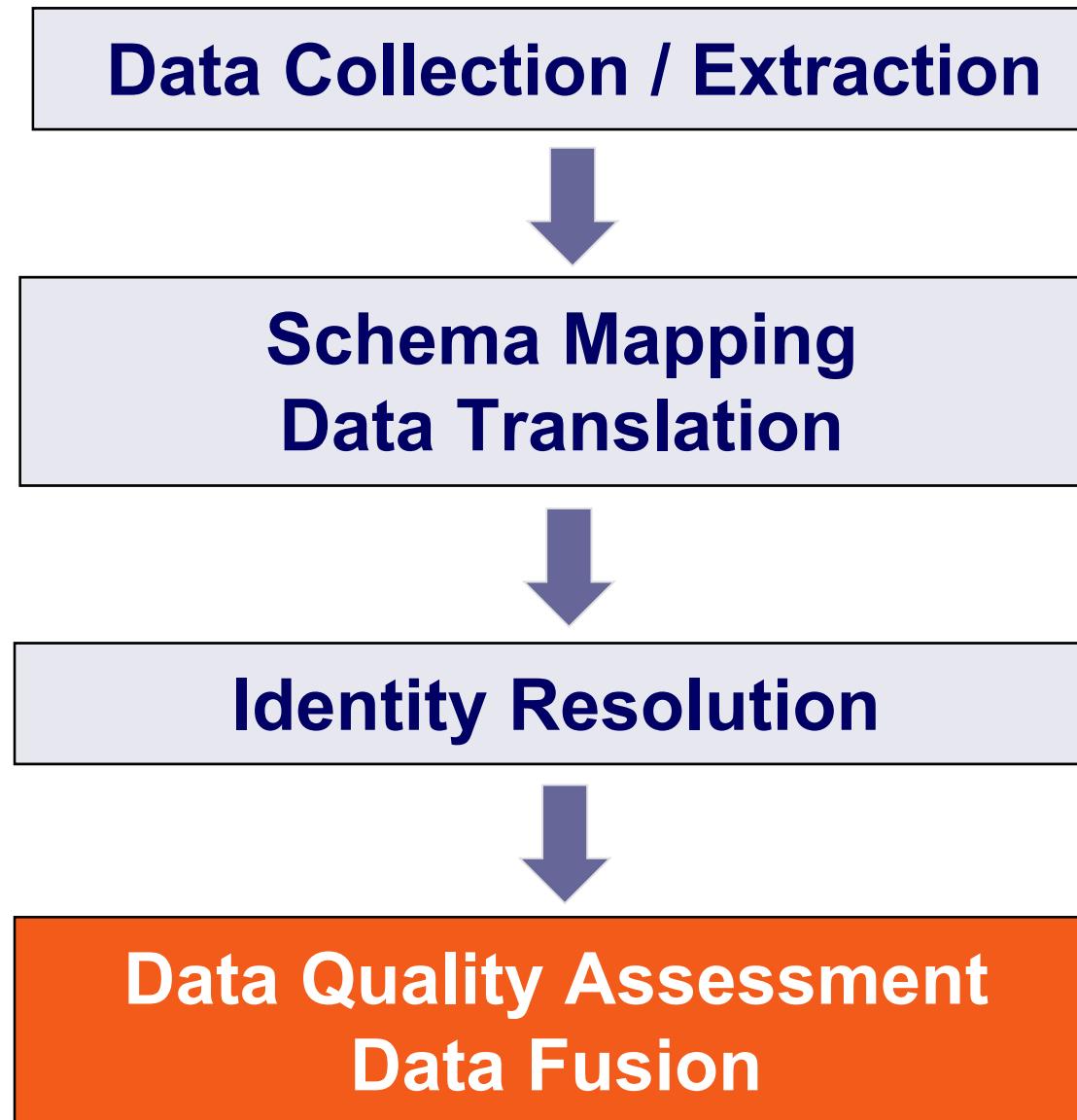
Web Data Integration

Data Quality Assessment

and Data Fusion



The Data Integration Process



Schedule for Today

1. Information about Final Exam (IE670)
2. Lecture Evaluation (IE670)
3. Q&A: Data Quality Assessment and Data Fusion

Final Exam HWS2021 (IE670, 3 ECTS)

- Date and Time
 - Tuesday, 14.12.2021, 15:00 -16:00
 - Location: Offline, 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)
 - schema matching methods
 - 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

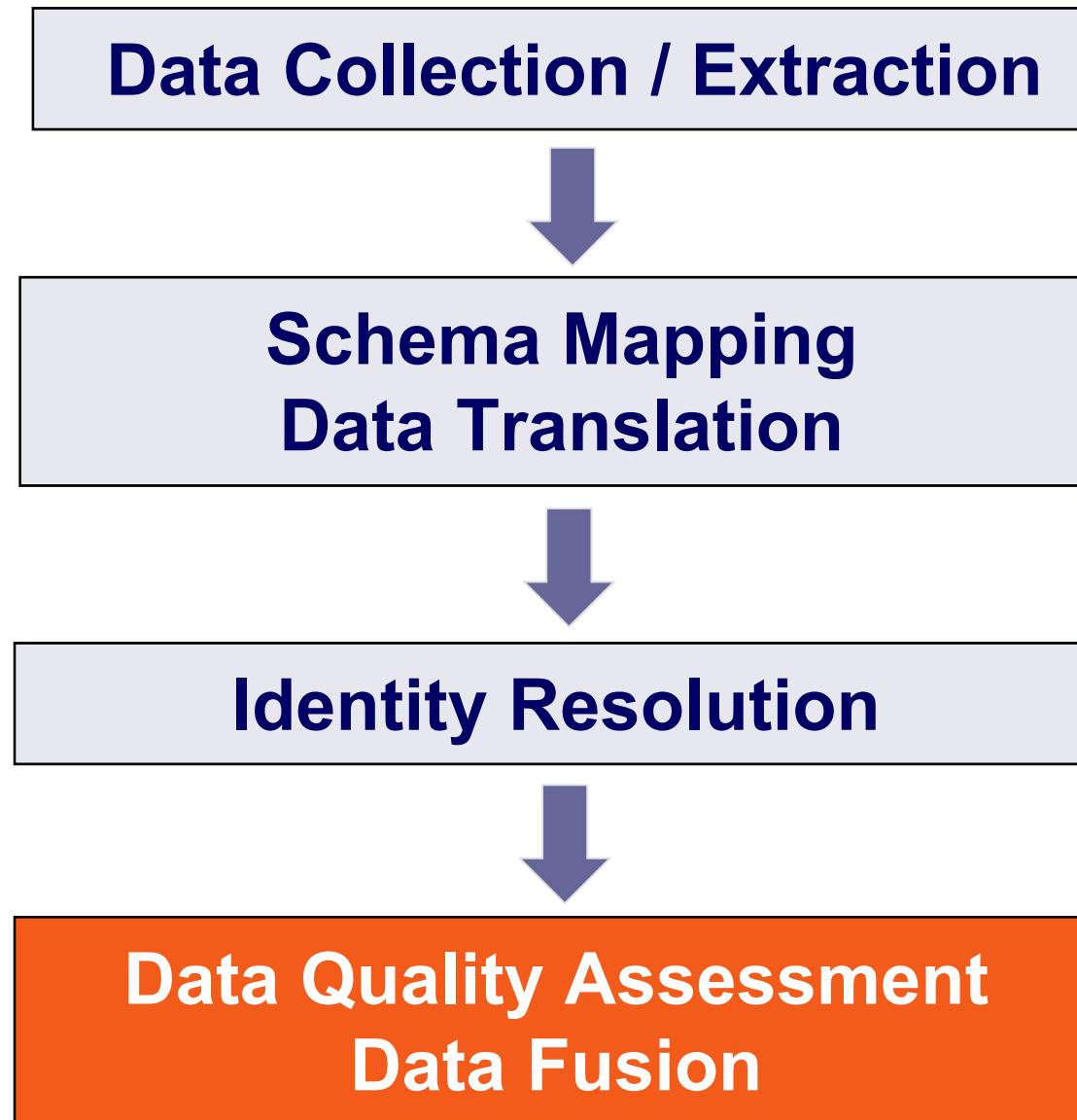
Lecture Evaluation (IE670)

- Please take 10 minutes to answer the evaluation questions concerning the lecture and exercise (IE670)
- Separate evaluation for student projects (IE683) on December 8th



<https://evasys.uni-mannheim.de/evasys/online.php?p=Z4JF2>

The Data Integration Process



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

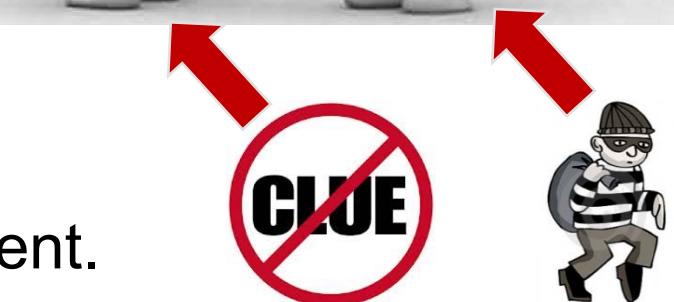
Information providers on the Web have

- 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 information on the Web is partly wrong, biased, outdated, incomplete, and inconsistent.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

$$(1.95 \text{ sq km} = 0.753 \text{ sq miles})$$



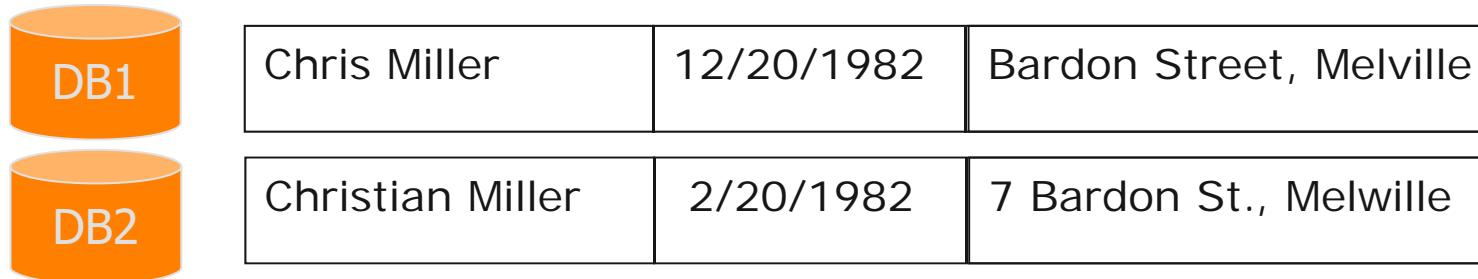
Population: Different claims and vague meta-information

Value	Meta-information	Webpage
30,727	(July 2018 est.)	http://www.cia.gov/cia/publications/factbook/geos/mn.html
38,897	(2016 census)	https://en.wikipedia.org/wiki/Monaco , reference pointing at statistics from 2009
39,042	(2019 latest UN estimate)	https://www.worldometers.info/world-population/monaco-population/

Source: Peter Bunemann

Definition: Data Conflict

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



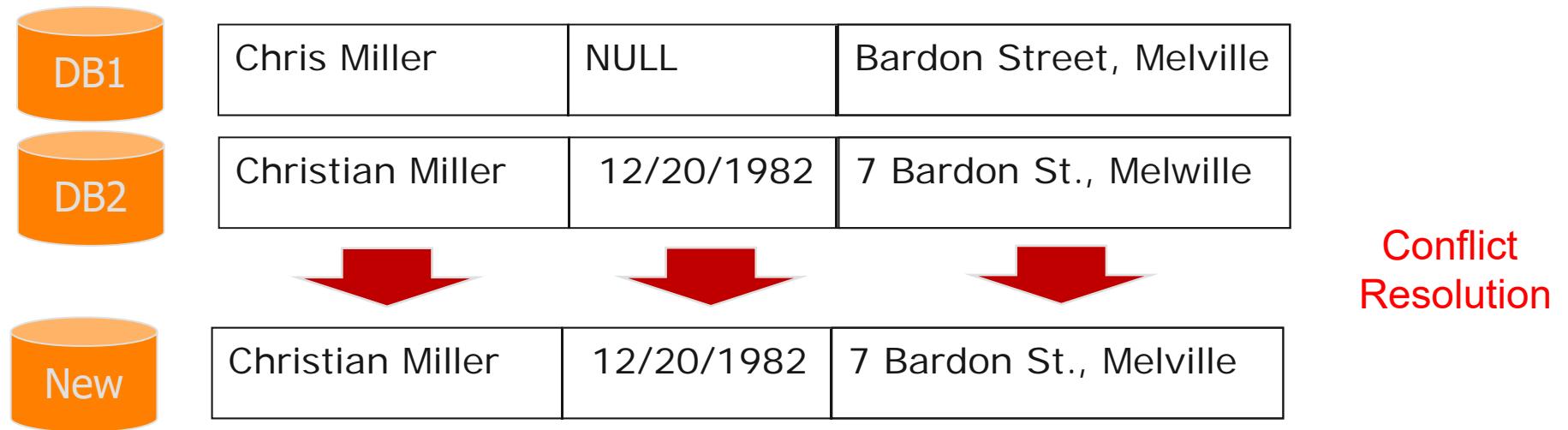
DB1	Chris Miller	12/20/1982	Bardon Street, Melville
DB2	Christian Miller	2/20/1982	7 Bardon St., Melville

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 (like cuteness)

Definition: Data Fusion

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



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

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

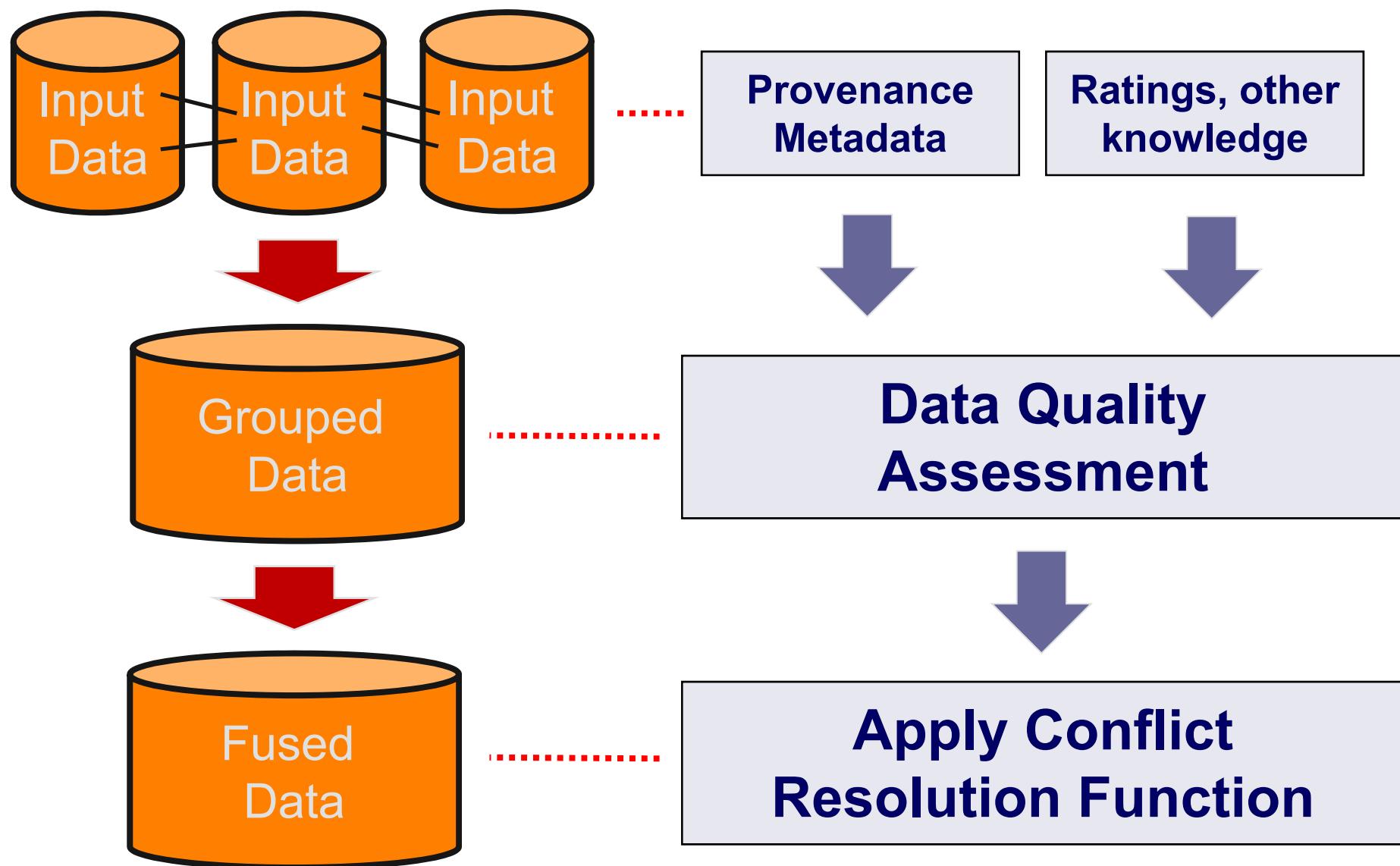
Fitness for use

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”

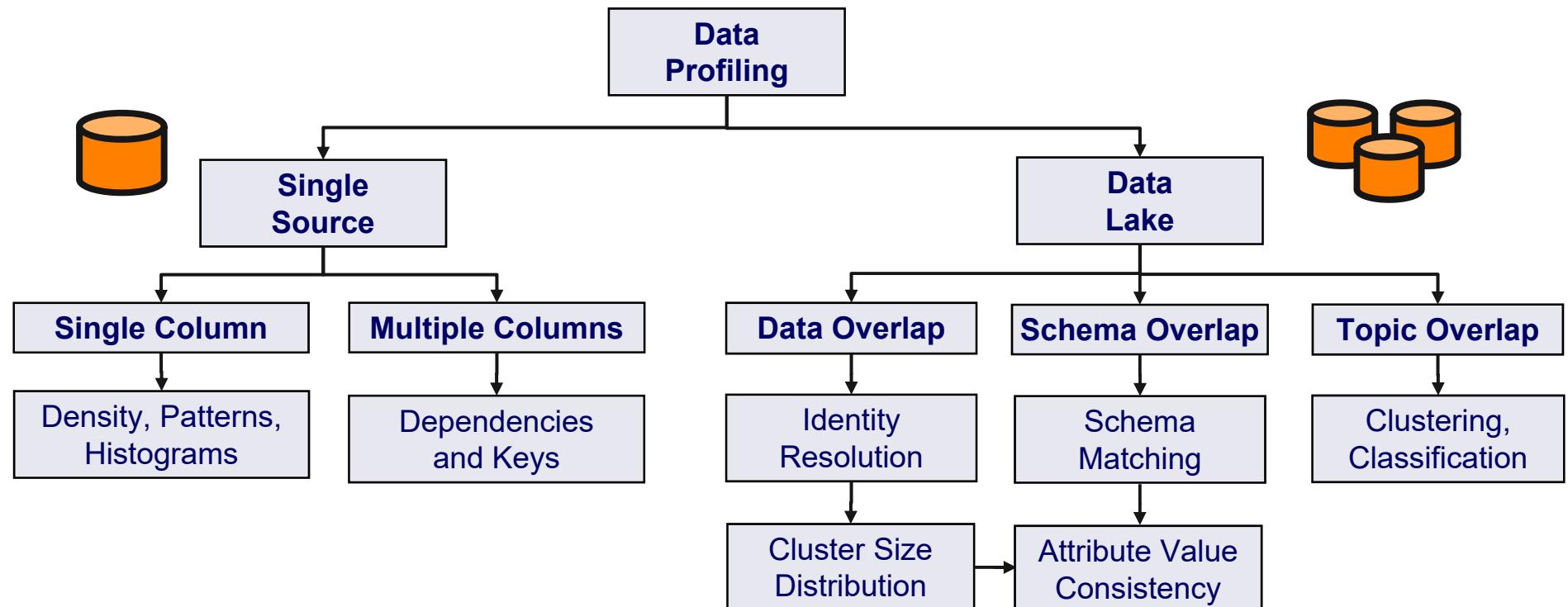
Summary: Elements of the Data Fusion Process



2. Data Profiling

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

- manual exploration (data gazing) should be supported with profiling results
- crucial when new data sets arrive or new people work with existing data lakes



Abedjan, et al.: Data Profiling. Morgan & Cleypool Synthesis Lecture in Computer Science, 2018.

2.1 Single Column Profiling: Metrics

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

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

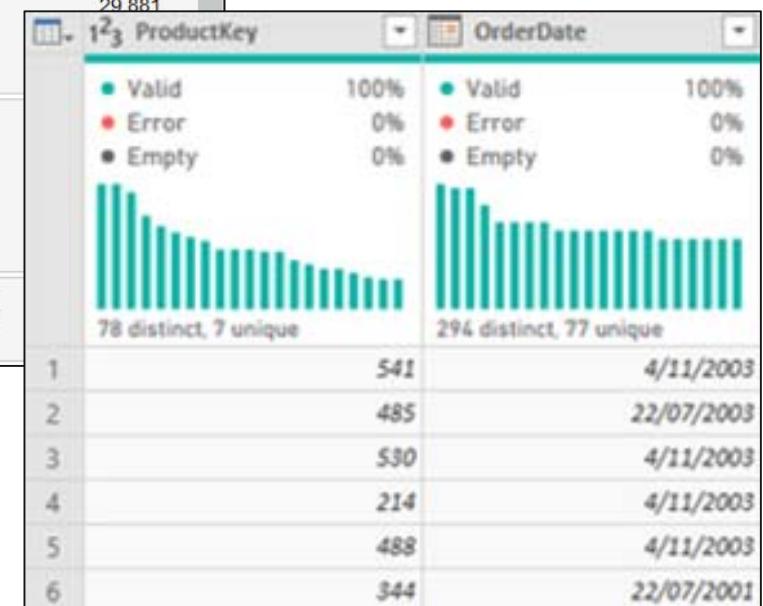
Advanced column profiling

Single Column Profiling: Examples

RapidMiner

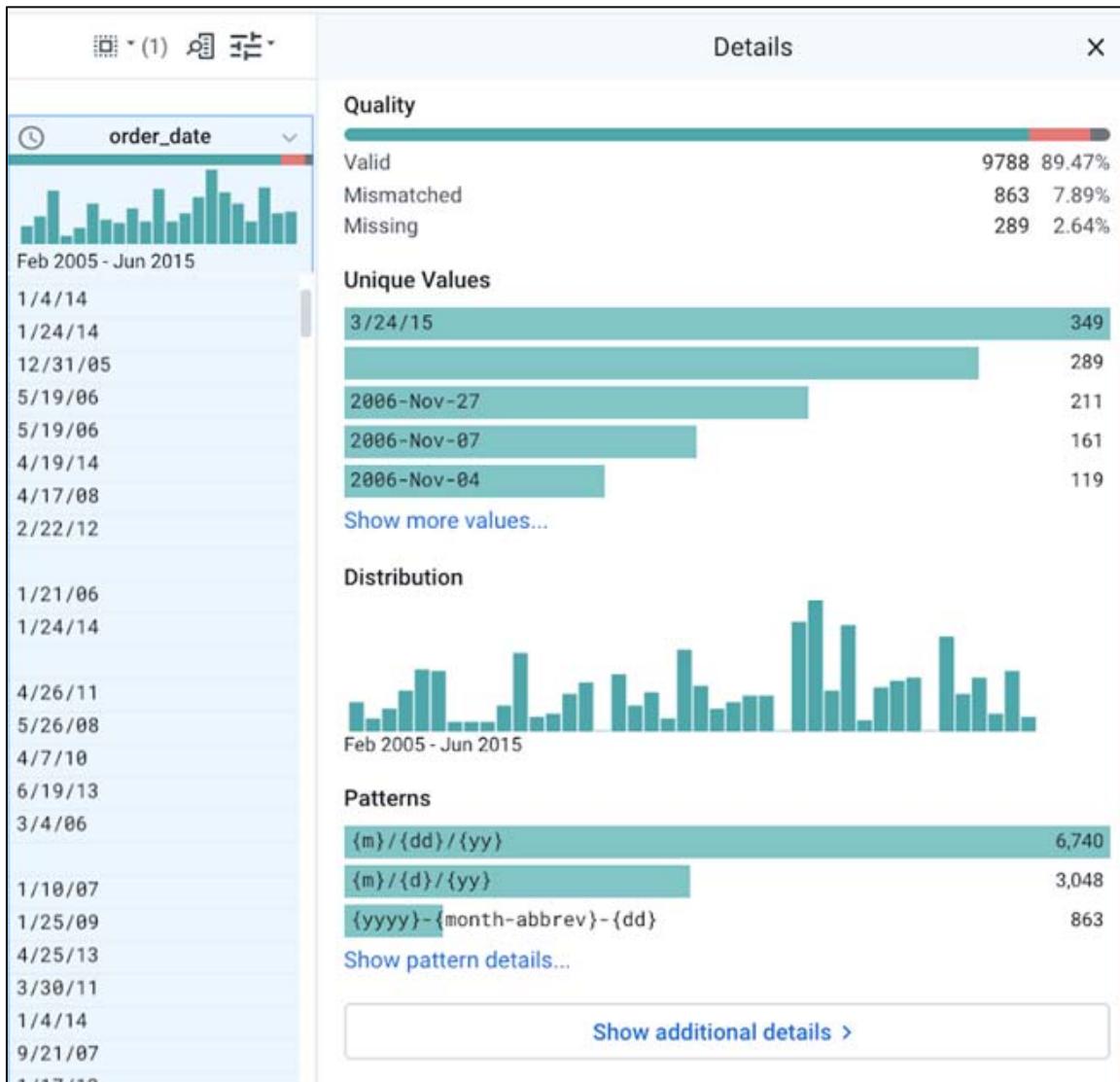


Microsoft Power BI



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>

The screenshot shows the pandas-profiling interface for a CSV file named 'anime_planet.csv'. The interface includes a code editor at the top with the following Python code:

```
1 import pandas as pd
2 import pandas_profiling
3
4 csv_file = '../datasets/anime/anime_planet.csv'
5
6 data_frame = pd.read_csv(csv_file)
7 pandas_profiling.ProfileReport(data_frame)
```

Below the code are five tabs: Overview (selected), Variables, Correlations, Missing values, and Sample. A red arrow points from the 'Correlations and missing values' annotation to the 'Correlations' tab.

The 'Dataset info' section contains the following statistics:

Number of variables	8
Number of observations	4000
Missing cells	278 (0.9%)
Duplicate rows	0 (0.0%)
Total size in memory	1.8 MiB
Average record size in memory	467.5 B

The 'Variables types' section shows the count of categorical (CAT) and numerical (NUM) variables:

CAT	6
NUM	2

Red arrows point from the 'Data type detection' annotation to the 'CAT' and 'NUM' counts.

The 'Warnings' section lists several issues:

- Episodes has a high cardinality: 128 distinct values (Warning)
- Genres has a high cardinality: 1828 distinct values (Warning)
- Producers has 272 (6.8%) missing values (Missing)
- Producers has a high cardinality: 176 distinct values (Warning)
- subject_id has a high cardinality: 4000 distinct values (Warning)
- Title has a high cardinality: 3983 distinct values (Warning)

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

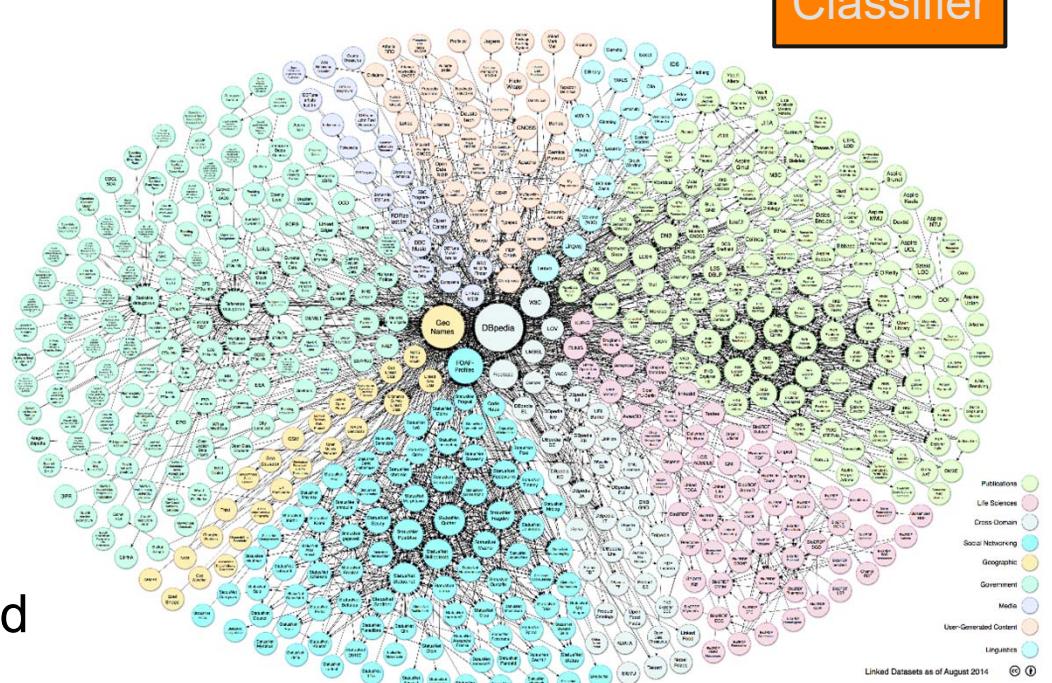
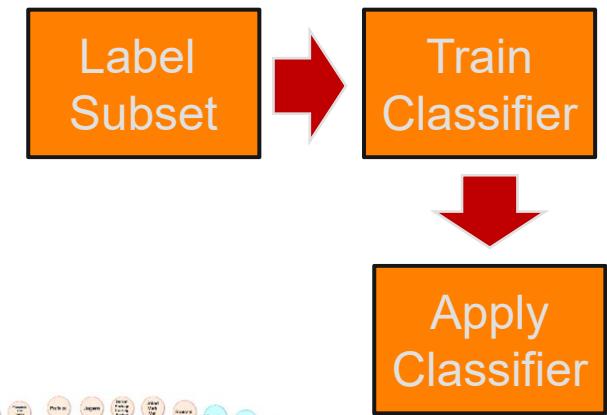


DBpedia Class	Number of Tables/Values			V _c	Data Type	Type	Reference
	T ₀	T _c	V _c				
+ Person	265 685	103 801	4 176 370	2 117 793	1 588 475	266 628	203 474
- Athlete	243 322	95 916	3 861 641	2 084 017	1 435 775	163 771	178 078
- Artist	9 981	2 356	18 886	3	11 527	3 499	3 857
- Politician	3 701	1 388	18 505	10	7 725	3 393	7 377
- Office Holder	2 178	1 435	131 633	30	66 762	59 332	5 509
+ Organisation	194 317	36 402	573 633	99 714	187 370	100 710	185 839
- Company	97 891	6 943	203 899	58 621	83 001	34 665	27 612
- SportsTeam	50 043	2 722	31 866	2 206	22 368	43	7 249
- Educational	25 737	14 415	238 365	38 056	64 578	13 334	122 397
- Institution							
- Broadcaster	14 515	11 315	93 042	564	13 095	52 186	27 197
Work	269 570	127 677	2 284 916	109 265	1 354 923	33 091	787 637
+ MusicalWork	138 676	80 880	1 131 167	64 545	396 940	7 610	662 072
+ Film	43 163	9 725	256 425	10 844	198 913	14 382	32 286
+ Software	39 382	23 829	486 868	418	414 092	9 194	63 164
Place	133 141	24 341	859 995	413 375	273 510	84 111	88 999
+ PopulatedPlace	119 361	21 486	787 854	405 406	257 780	57 064	67 604
- Country	36 009	6 556	208 886	93 107	66 492	31 793	17 494
- Settlement	17 388	2 672	17 585	4 492	6 662	2 444	3 987
Species	14 247	4 893	83 359	-	7 902	38 682	36 775
Σ	949 970	301 450	8 037 562	2 751 105	3 437 420	536 526	1 312 511

Hassanzadeh, et al.: Understanding a Large Corpus of Web Tables through Matching with Knowledge Bases. OM. 2015.
Ritze, et al. Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases. WWW 2016.

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.

Meusel, Spahiu, Bizer, Paulheim: Towards automatic topical classification of LOD datasets. LDOW 2015.

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

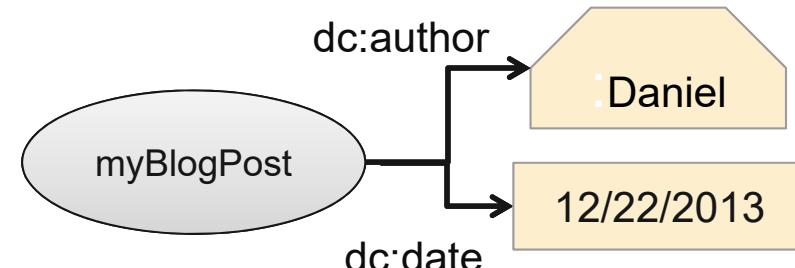
Outline of this Subsection

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

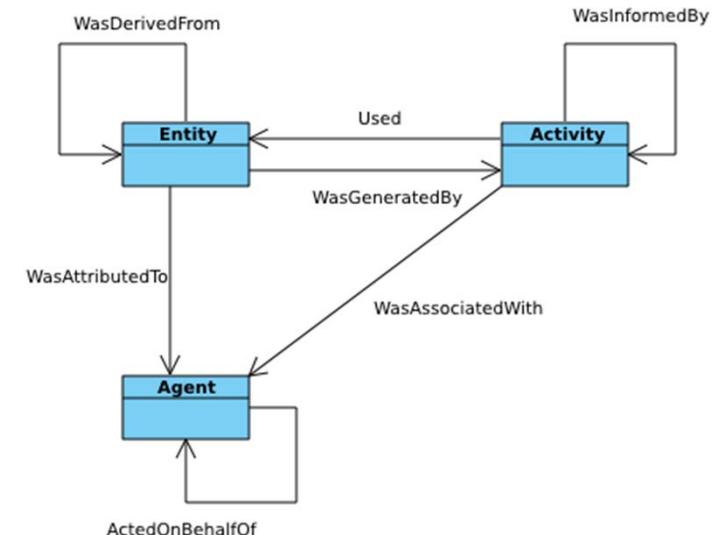
1. Simple Attribution:

- 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?

3.2 Publishing Provenance Information on the Web

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

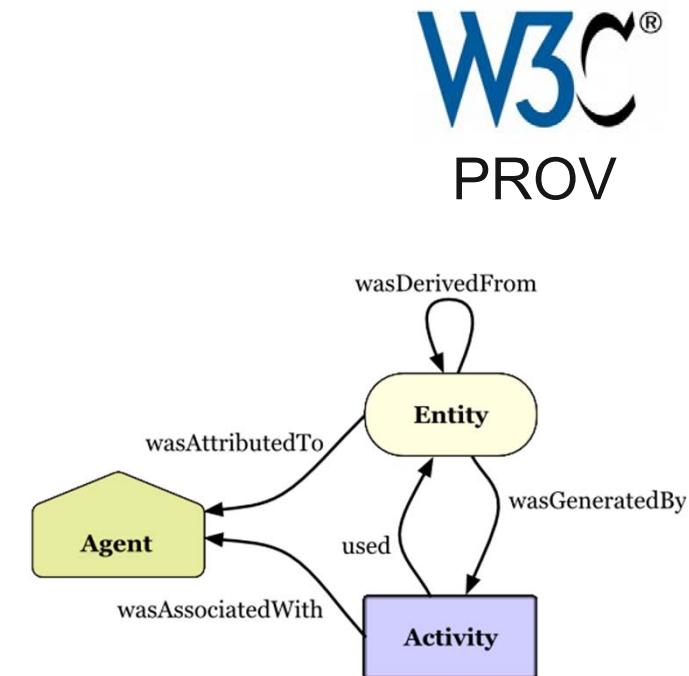
```
# Metadata and Licensing Information
<http://dbpedia.org/data/Alec_Empire>
    rdfs:label "RDF document describing Alec Empire" ;
    rdf:type foaf:Document ;
    dc:publisher <http://dbpedia.org/resource/DBpedia> ;
    dc:date "2019-07-13"^^xsd:date ;
    dc:rights <http://en.wikipedia.org/wiki/WP:GFDL>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

# 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 ;
...
...
```

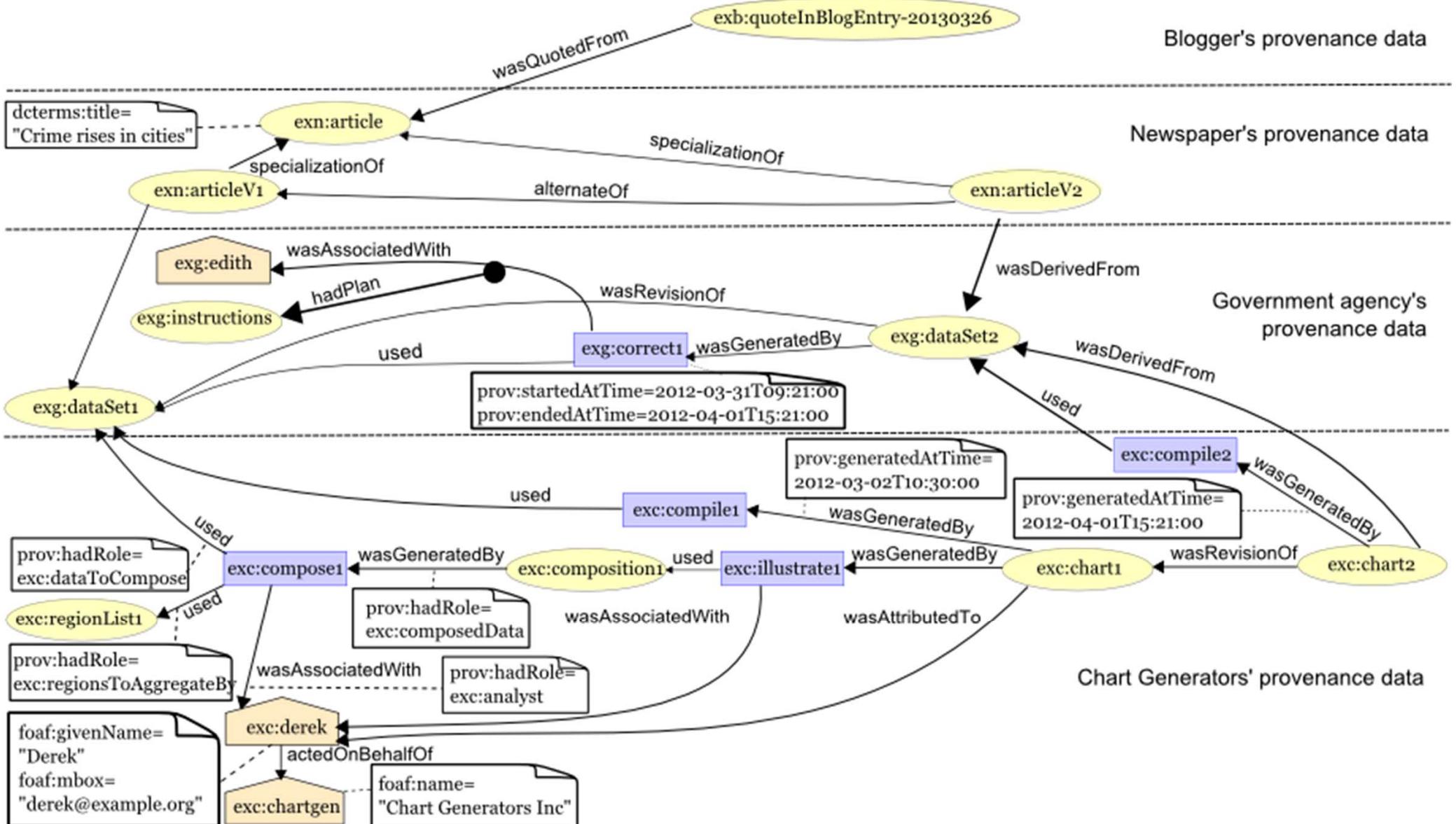
W3C PROV

- The W3C PROV vocabulary defines terms for representing **complex provenance chains**
- Example of a PROV XML document:

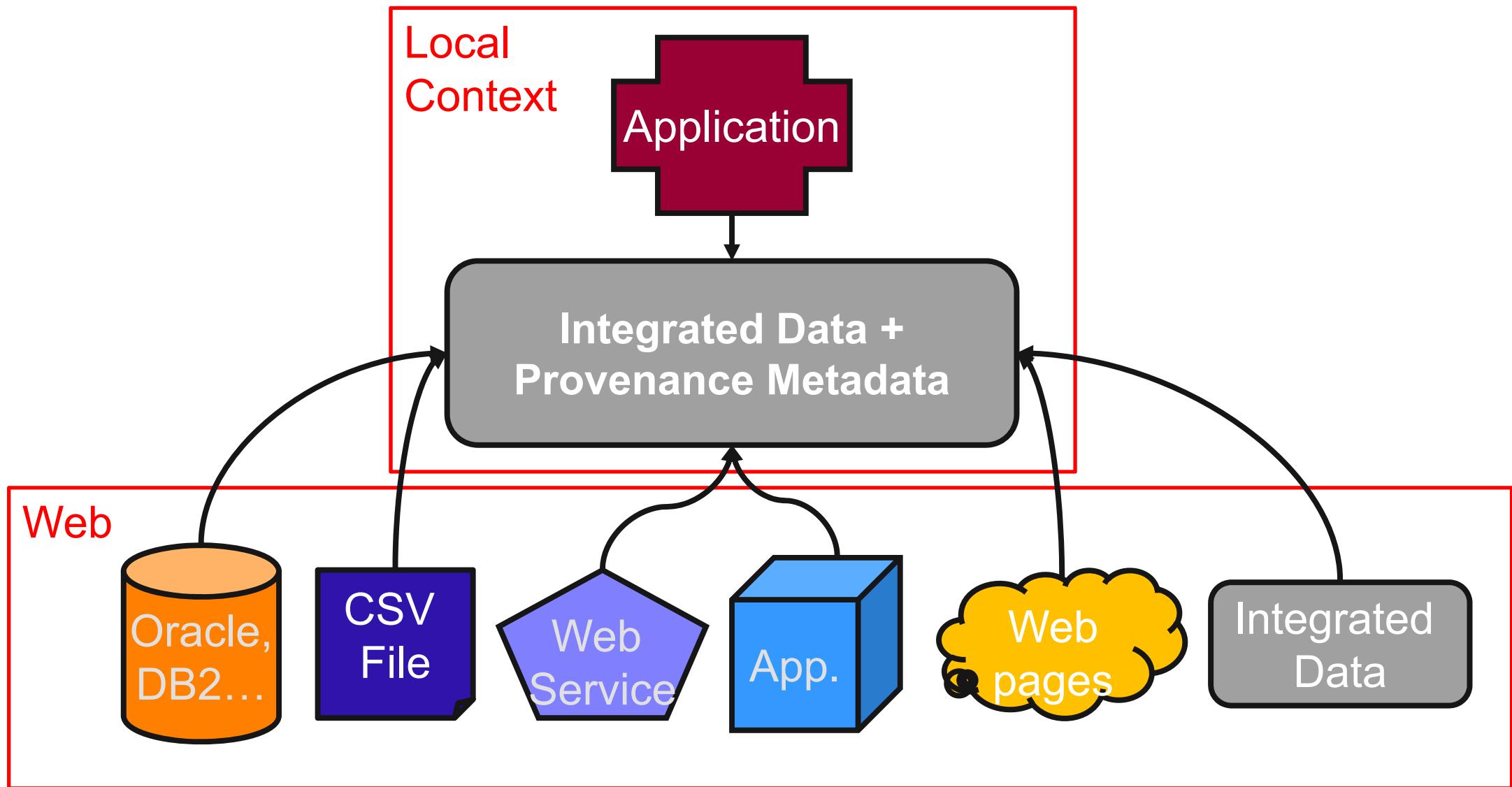
```
<prov:document>
  <!-- Entities -->
  <prov:entity prov:id="exn:article">
    <dct:title>Crime rises in cities</dct:title>
  </prov:entity>
  <!-- Agents -->
  <prov:agent prov:id="exc:derek">
    <prov:type>prov:Person</prov:type>
    <foaf:givenName>Derek Smith</foaf:givenName>
    <foaf:mbox>mailto:derek@example.org</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>
...
...
```



More Complex Example: W3C PROV



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**

<u>Key</u>	Name	Street	<u>ProvID</u>
1425	Dr. Mark Smith	14 Main Street	001
1425	Mark Smith	12 Main St.	002
...

Physicians with **Value-Level Provenance**

<u>Key</u>	<u>Attribute</u>	Value	<u>ProvID</u>
1425	Name	Dr. Mark Smith	001
1425	Name	Mark Smith	002
1425	Street	14 Main Street	001
...

Provenance Table

<u>ProvID</u>	Source	Date
001	www.mark-smith.com	12/6/2018 18:42:12
002	www.doc-find.com	12/1/2018 12:21:54
...

XML Data Model

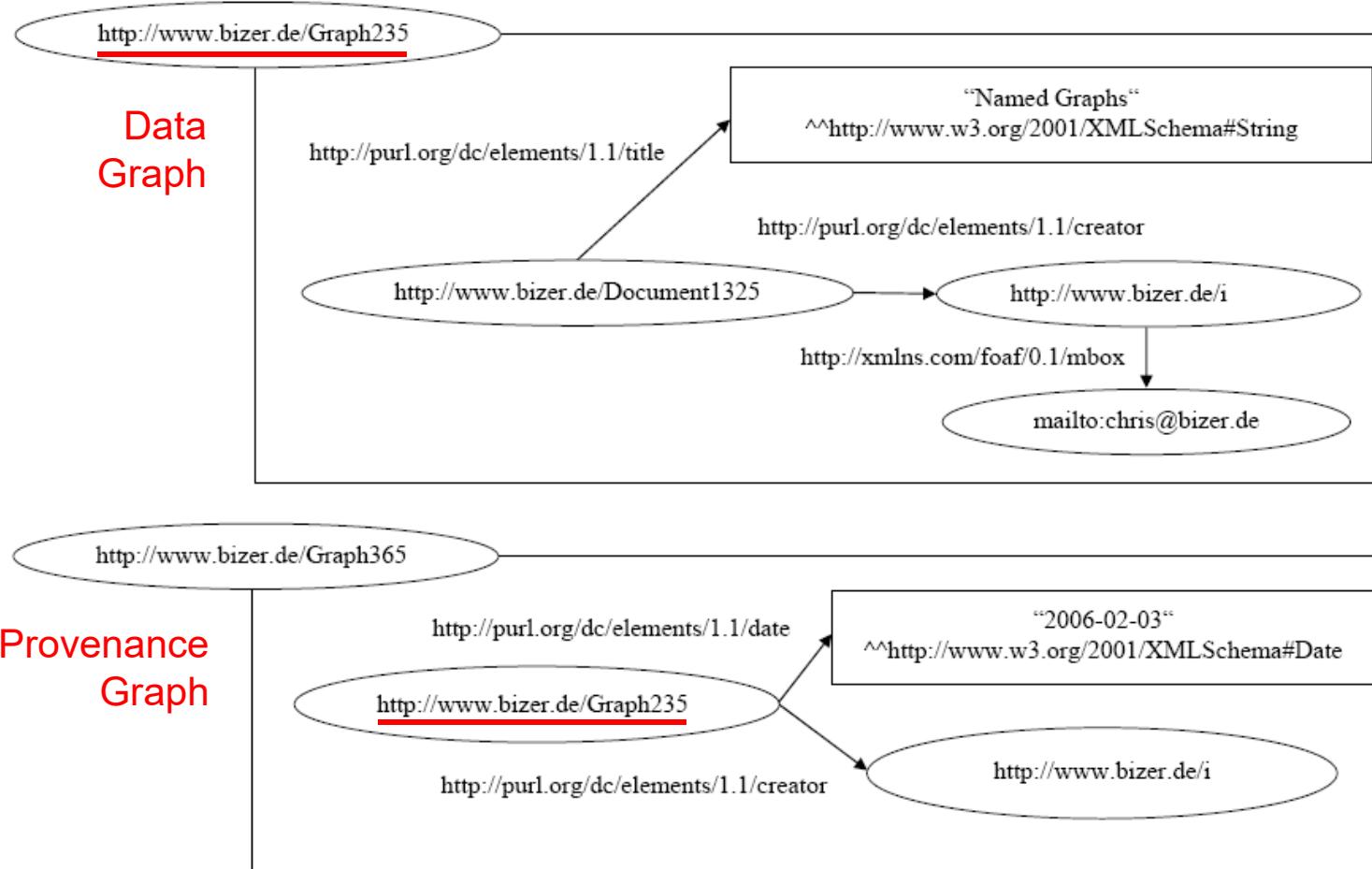
Represent provenance using multiple **value elements** and references to **provenance elements**.

```
<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">
  ...

```

RDF Data Model

- 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

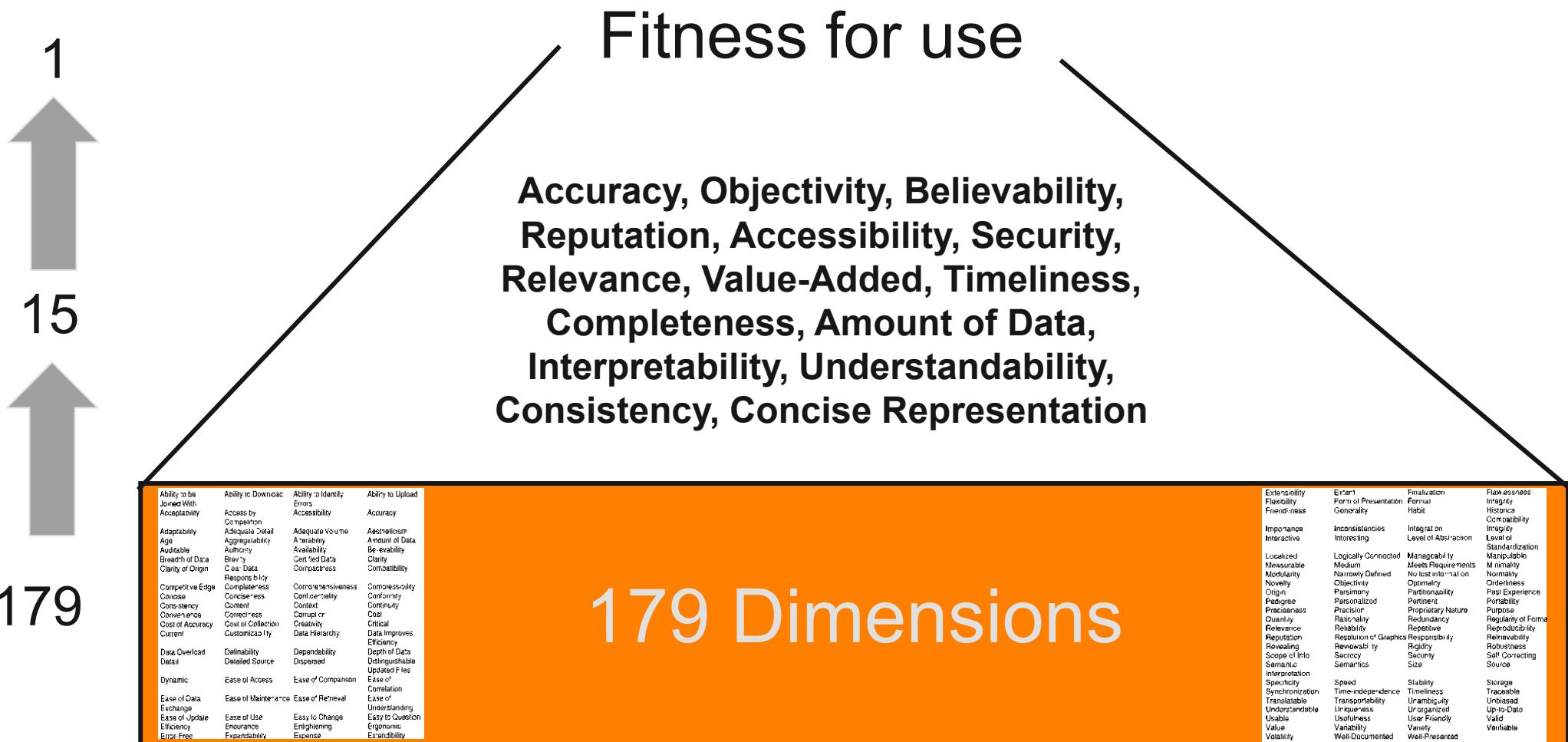
- the goal is to establish **procedures and rules** that guarantee high quality data production, quality monitoring, and regular data cleansing
- pioneering research by MIT Total Data Quality Management (TDQM) program
- consequences of low data quality:
 - US postal service: out of 100.000 mass-letters, 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 spare 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:

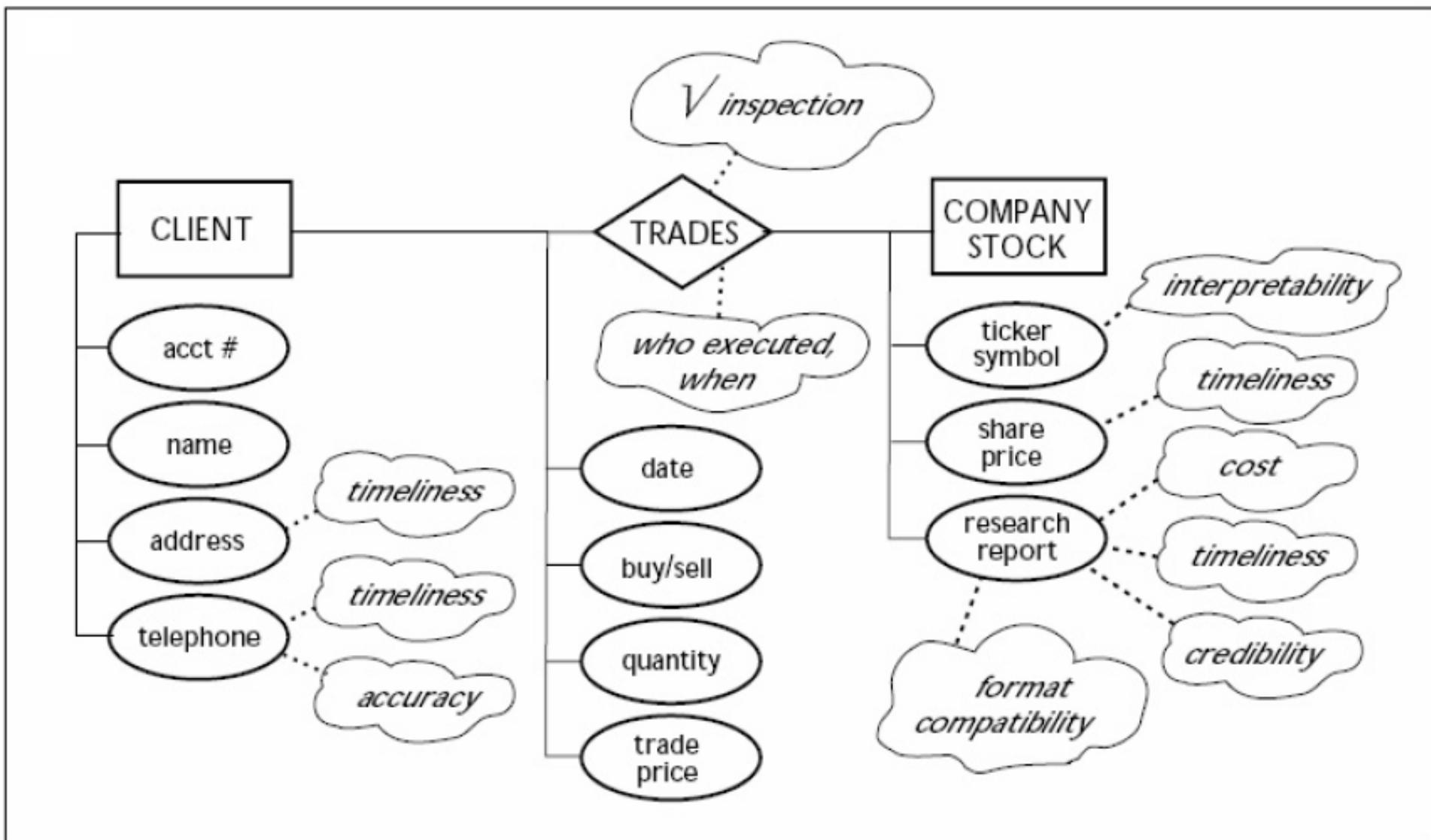


Category	IQ Criteria	TDQM	MBIS	Weikum	DWQ	SCOUG	Chen
Content-related Criteria	Accuracy	Yes	Yes	Yes	Yes	Yes	Yes
	Documentation					Yes	
	Relevancy	Yes	Yes		Yes		Yes
	Value-Added	Yes				Yes	
	Completeness	Yes	Yes	Yes	Yes	Yes	
	Interpretability	Yes			Yes		Yes
Technical Criteria	Timeliness	Yes	Yes	Yes	Yes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

Vrxufh#Thd{#Q dxp dqg

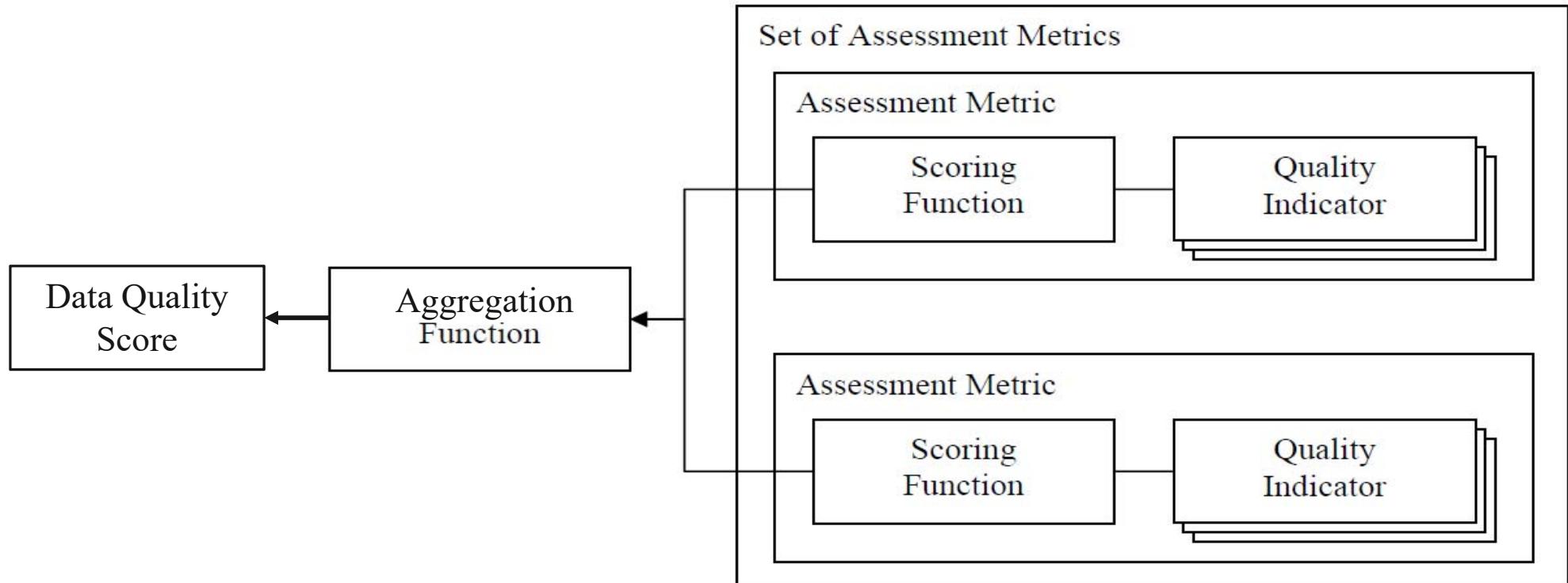
Relevancy of Data Quality Dimensions

Which quality dimensions matter depends on the task at hand.



4.2. Data Quality Assessment

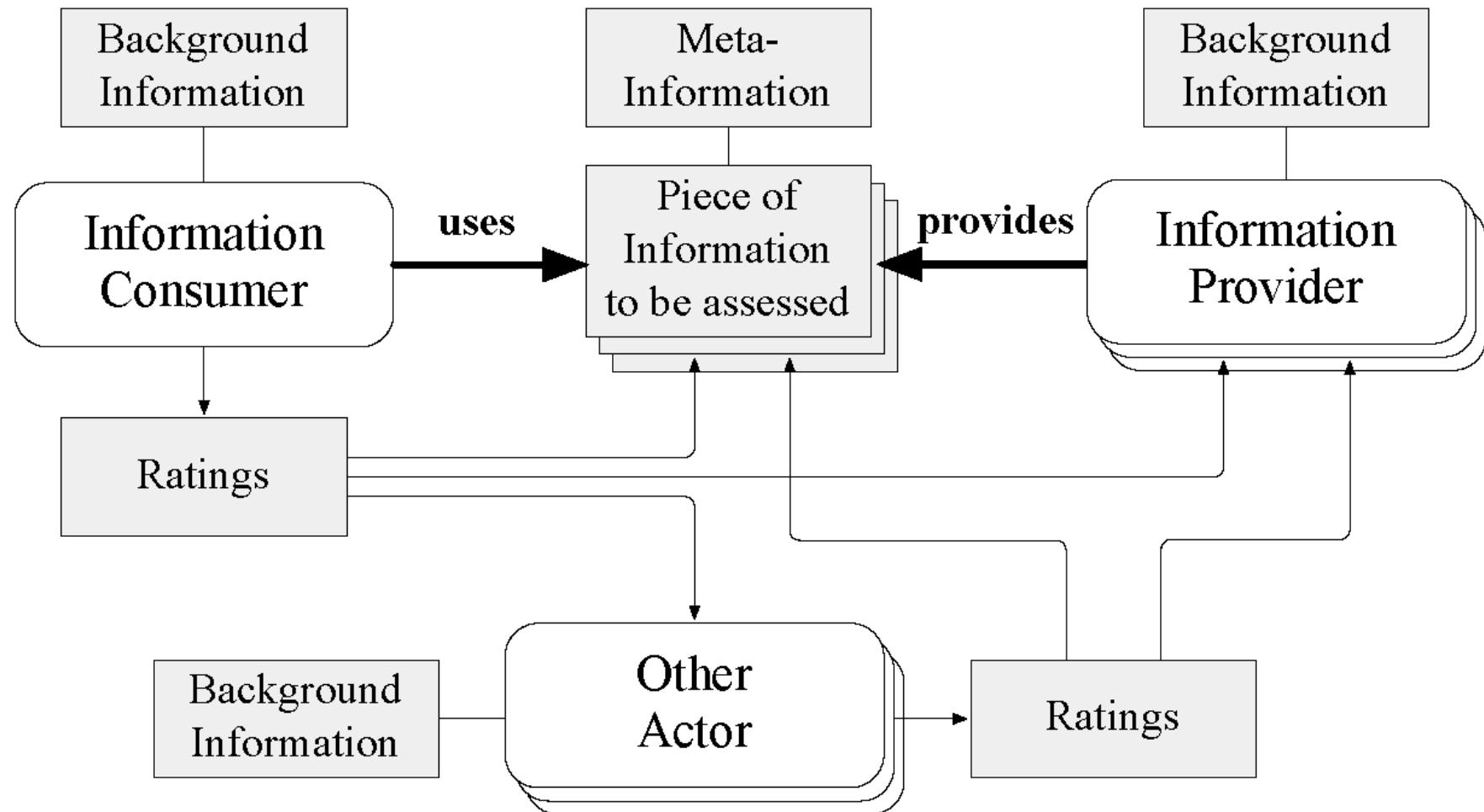
Various domain-specific heuristics are used to measure data quality.



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)

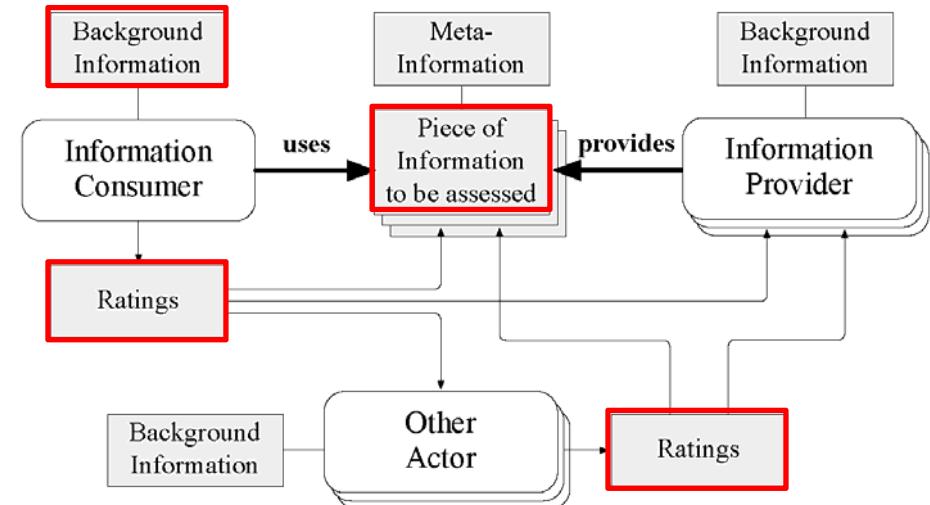
Quality Indicators in the Web Context



4.2.1 Assessing Data Accuracy

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

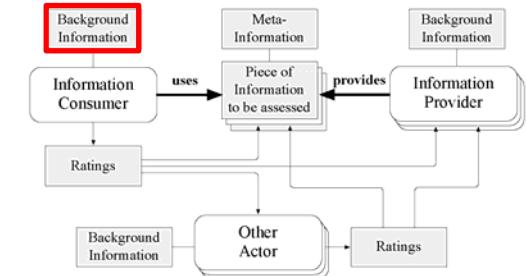
- also called: **Truth Discovery, Fake News Detection**
- **Assessment Methods:**
 1. Constraint testing
 2. Outlier detection
 3. Expert- or user ratings
- **Relevant quality indicators:**



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)

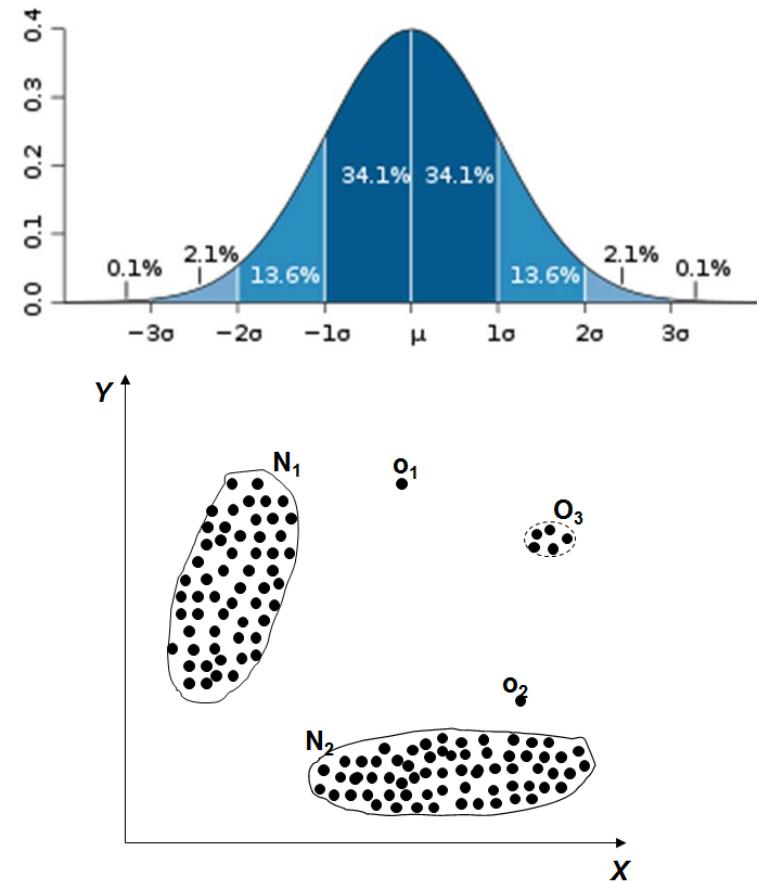


Fan, Geerts: Foundations of Data Quality Management. Morgan & Claypool, 2012.

Outlier Detection

An outlier is a 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

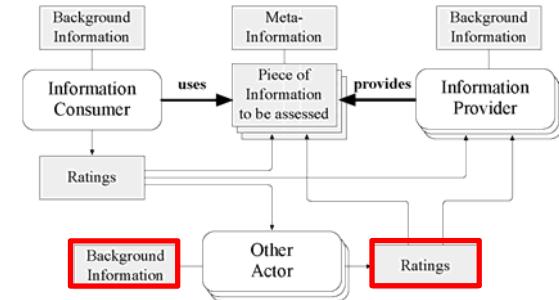


Chandola, et al.: Anomaly Detection: A Survey. ACM Computing Surveys, 2009.

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 Place Luxury Boutique Villas
15/4 Moo 2 | Sairee, Koh Tao 84360, Thailand
Hotel website E-mail hotel 00 66 87 887 5066 Hotel amenities
298 reviews sorted by Date Rating ▲ English first

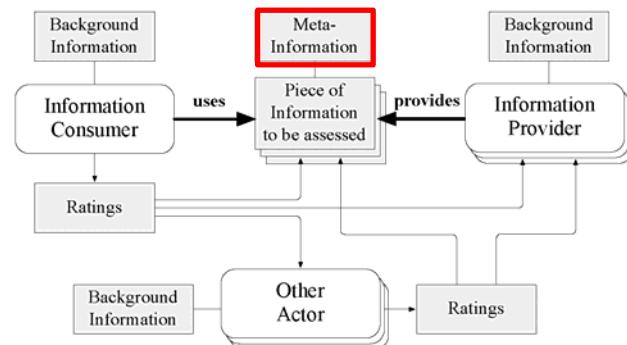
"It's so nice you don't need to leave"
Reviewed October 11, 2011
The reviews below complement this place enough - so I don't need to repeat everything that's been said! Suffice to say, I agree with the general consensus. Particular highlights are the AMAZING view, the private pool and the entertainment system (combined with a selection of tried and tested take out menus) which mean you never have to leave, which we...
More ▾
Was this review helpful? Yes 2 Problem with this review?

"Could be perfect!"
Reviewed January 13, 2012
After an 8000 mile journey by air and a treacherous sea crossing to Koh Tao, we were delighted to meet the owners of The Place, Guy and Bex, waiting for us at the dockside on our arrival. Together with another couple, we and our luggage were bundled into a pickup truck, owned by Mr Bear and we made the 10...
More ▾
Was this review helpful? Yes 8 Problem with this review?

4.2.2 Assessing Data Timeliness

Definition Timeliness: The extent to which the age of the data is appropriate for the task at hand.

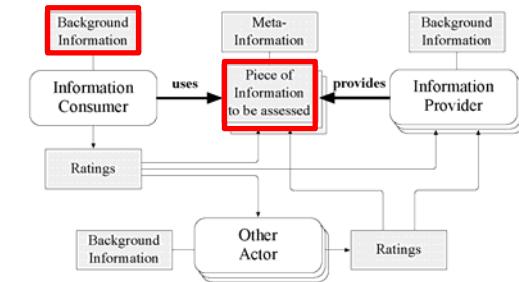
- The assessment of the timeliness of data usually requires provenance data.
- Provenance metadata
 - HTTP Last-Modified
 - dc:date
- Fallbacks if no timestamps are available
 - propagate timestamps to data without timestamps
 - e.g. two tables provide same profit for a company, only one table has a timestamp
 - Zhang, Chakrabarti: InfoGather+, SIGMOD 2013.
 - 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

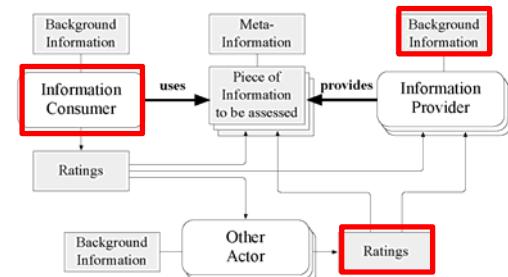
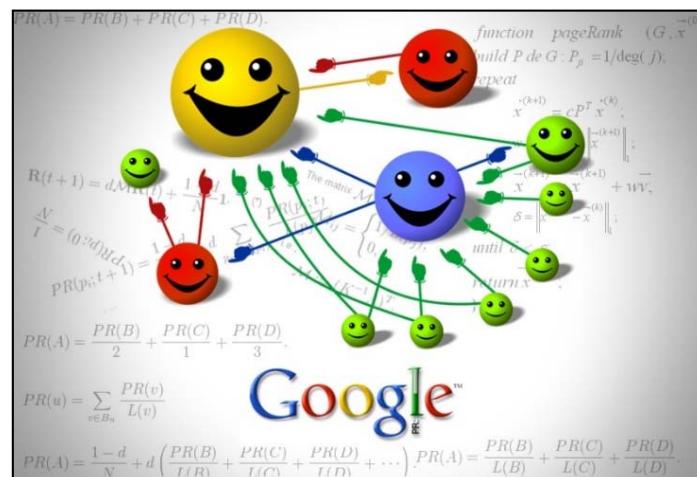


4.2.4 Assessing Data Relevancy

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



The screenshot shows a TripAdvisor review page for a community. At the top, it says "31 reviews from our community" and "Write a Review". Below that, there's a "Traveler rating" section with five categories: Excellent, Very good, Average, Poor, and Terrible, each with a corresponding bar chart and a count of 0, 2, 0, 3, and 3 respectively. To the right, there's a "Trip type" section showing "Family reviews (8)", "Couples reviews (12)", "Solo travel reviews (1)", "Business reviews (0)", and "Friends reviews (6)". A red arrow points from the "Terrible" rating section towards the reviews below. The reviews are sorted by Date and Rating. One review is highlighted with a red arrow pointing to it:

"nightmare in koh tao"
Reviewed January 10, 2010
We had booked three nights at the black tip resort and it was terrible. The staff is always in a bad mood, never polite or helpful. The beach was full of garbage and gazoll . we could not swim and it was worst everyday. The restaurant staff would never smile or say hi . Well, having travel many times in...

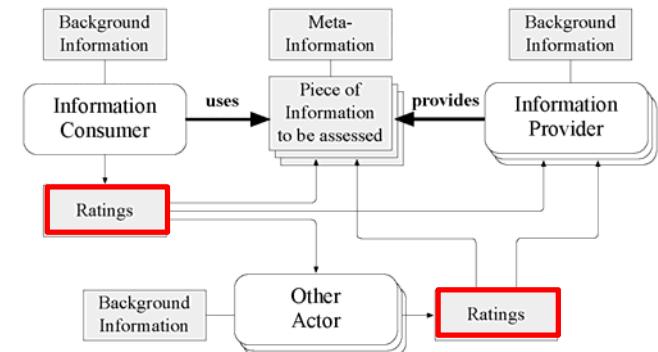
salman_10 geneva switzerland 1 review 2 helpful votes

Was this review helpful? Yes [2] Problem with this review?

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

The screenshot shows the WIQA Browser interface running in Mozilla Firefox. The title bar reads "WIQA Browser - Mozilla Firefox". The menu bar includes Datei, Bearbeiten, Ansicht, Gehe, Lesezeichen, Extras, and Hilfe. The address bar shows the URL: <http://127.0.0.1:1978/piggy-bank/e1eb9ba7fe10653332021055d7562c83/default?command=browse&policyURI=Information+from+German+analysts&=.%40lwq.project.Proj>. The main content area is titled "WIQA Browser" and shows the date "19.07.2006 14:35:50". A "filter criterion" section contains a single item: "is a: Share (remove) [add more]". Below this, there are tabs for "Order" and "Commands". A search bar says "Type here to search". On the right, a "Policy Selection Panel" lists policies such as "is a", "name", "discussion forum posting", etc. A large "Policy: Information from German analysts" section is visible. The bottom part of the screen shows search results for ISIN codes: "urn:ISIN:DE0007236101" and "urn:ISIN:US4581401001", each with sections for "emitted by", "is a", and "positive/negative analyst report". A prominent "Oh, yeah? Button" is overlaid on the results for the first ISIN code. Logos for "veritas institut für beratung" and "Simile" are at the bottom right.

Explanation about an Assessment Decision

WIQA Browser - Mozilla Firefox
Datei Bearbeiten Ansicht Gehe Lesezeichen Extras Hilfe
http://127.0.0.1:1978/piggy-bank/e1eb9ba7fe10653332021055d7562c83/default?command=browse&policyURI=Information+from+German+analysts&=-%40lwq.project.Proj Go [G]
[WIQA Browser]

WIQA Browser
19.07.2006 14:35:50

1 filter criterion
• is a:Share (remove) [add more]

Order Commands

2 items
sorted by name [A to Z]

urn:ISIN:DE0007236101

emitted by [urn:DUNS:316067164](#)
is a [Share](#)
positive analyst report [Siemens agrees partnership with Novell unit SUSE. Siemens Business Services \(SBS\), the IT services arm of German technology conglomerate Siemens <SIEGn.DE>, said on Tuesday it had agreed a partnership deal with Novell's \(nasdaq: NOVL - news - people\) newly acquired unit SUSE Linux. Linux software is open-source, meaning it can be freely copied and modified, unlike proprietary software such as Microsoft \(nasdaq: MSFT - news - people\) Windows. In the past months clients have been 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.](#)

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 unit SUSE Linux. Linux software is open-source, meaning it can be freely copied and modified, unlike proprietary software such as Microsoft \(nasdaq: MSFT - news - people\) Windows. In the past months clients have been 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.](#)

Show Referrers Delete

urn:ISIN:US458401001

is a [Share](#)
negative analyst report [Intel investiert Milliarden in Werks-Modernisierung. Der weltgrößte Chiphersteller Intel will nach Firmenangaben mit milliardenschweren Investitionen seine älteren Werke modernisieren, um ihnen die Fertigung kleinerer Mikroprozessoren zu ermöglichen. Ziel ist die Umstellung älterer Anlagen auf die Produktion von 65-Nanometer- von 90-Nanometer-Chips. Der Konzern befindet sich mitten in einem Modernisierungsprogramm über fünf Mrd. Dollar, sagte Intel-Chef Craig Barret am Sonntag zum 30. Jahrestag der Tätigkeit von Intel in Israel. Die älteren Anlagen sollen auf die Produktion von 65-Nanometer- von 90-Nanometer-Chips \(ein Nanometer ist ein Millionstel Millimeter\) umgestellt werden. Wir haben eine Menge 65-Nanometer-Investitionen. Dafür geht der größte Teil der Aufwendungen von 5 Mrd. \\$ drauf, sagte Barret. Er verwies dazu insbesondere auf die US-Werke in Phoenix, Portland und Oregon sowie die Anlage in Irland. In zwei Jahren seien noch kleinere Halbleiter möglich, sagte er. Im zweiten Halbjahr 2007 sollte die 45-Nanometer- Technologie geben, erklärte Barret. Er lehnte es jedoch ab, sich zu den Finanzergebnissen des Konzerns zu ändern. Er sagte lediglich, das Geschäft wachse weltweit. Kraftiges Wachstum sei in den Schwellenländern zu verzeichnen.](#)

Show Referrers Fertig

Explanation - Mozilla Firefox
EXPLANATION
WIQA Browser

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 <SIEGn.DE>, said on Tuesday it had agreed a partnership deal with Novell's (nasdaq: NOVL - news - people) newly acquired unit SUSE Linux. Linux software is open-source, meaning it can be freely copied and modified, unlike proprietary software such as Microsoft (nasdaq: MSFT - news - people) Windows. In the past months clients have been 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.

fulfills 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](#).

[Close](#)



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:

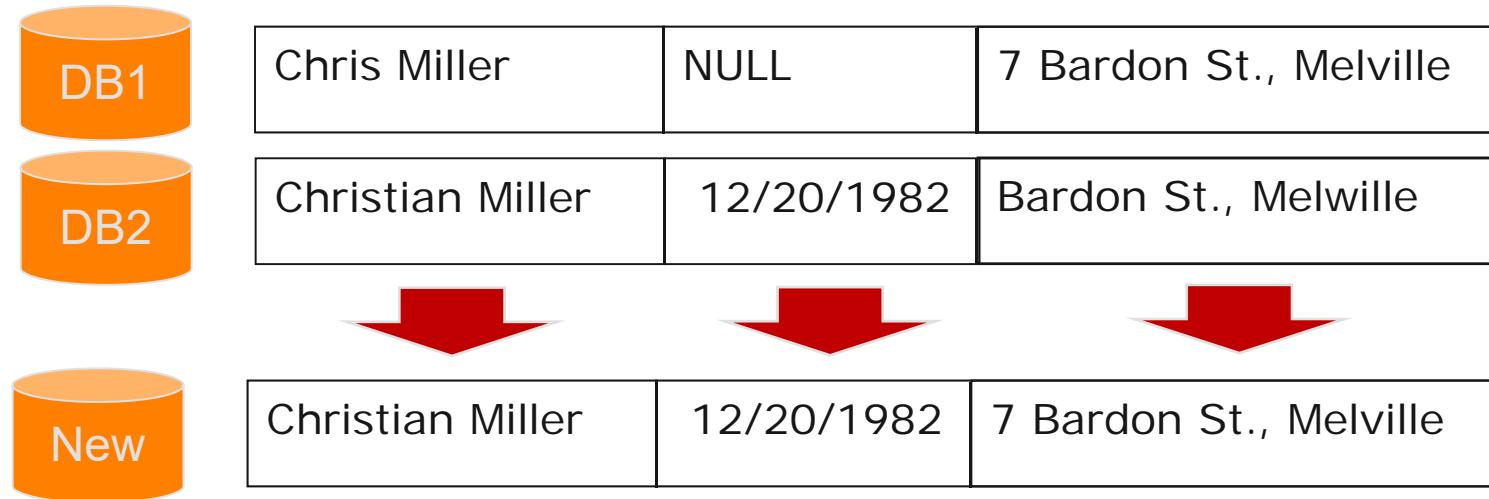
- it was asserted by Peter Smith and
- Peter Smith has received positive ratings from
 - Mark Scott who works for Siemens.
 - David Brown who works for Intel.
 - John Maynard who works for Financial Times.

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

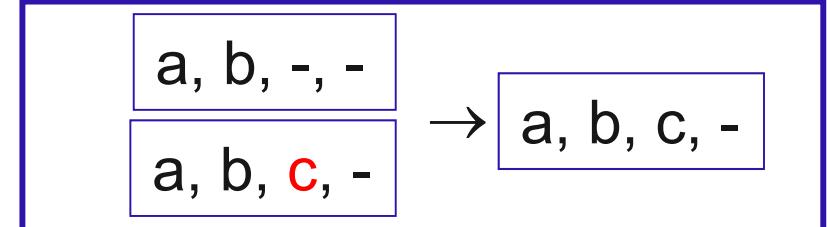
Slot Filling: Fill missing values (NULLs) in one dataset with corresponding values from other datasets.

Result: increased dataset density

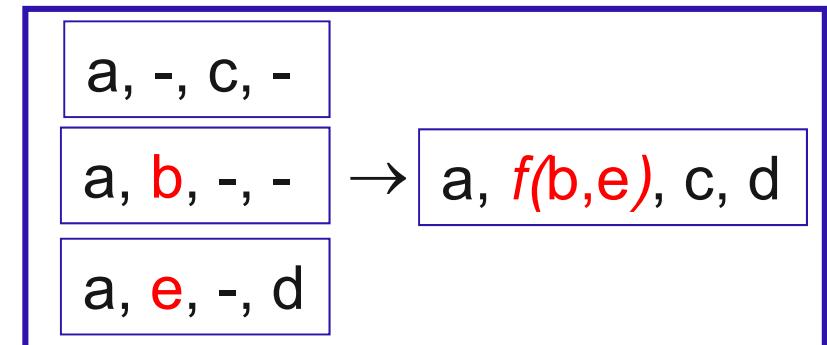
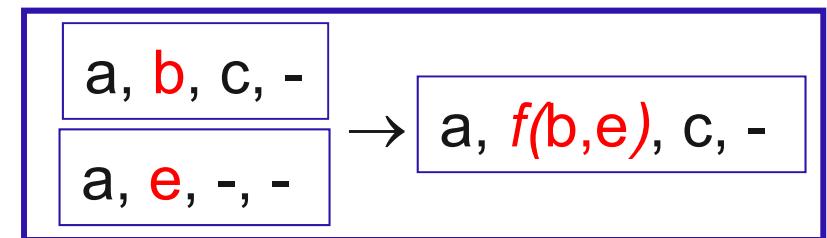
Conflict Resolution: Resolve contradictions between records by applying a conflict resolution function (heuristic).

Result: increased data quality

Complementary records



Conflicting records



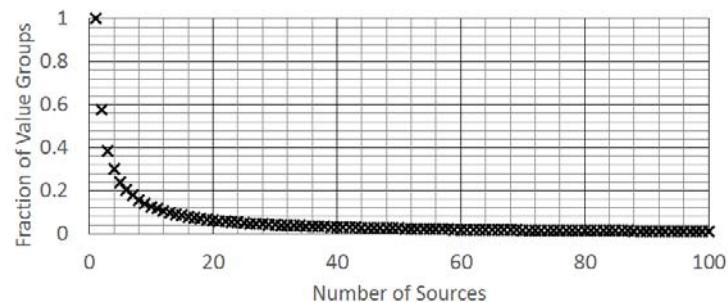
Cluster Size Distribution, Matching Errors, and Data Fusion

- As final step of the identity resolution process, records are clustered using the discovered correspondences. Example with 3 data sources:

Cluster Size	Frequency	
1	4256	No slot filling possible as single records with no overlap
2	939	
3	503	Slot filling and conflict resolution allow the generation of improved records
4	75	
5	14	
35	3	Large cluster size indicates matching errors or duplicates in data sources
61	1	

- Cluster size distribution from matching web tables to DBpedia

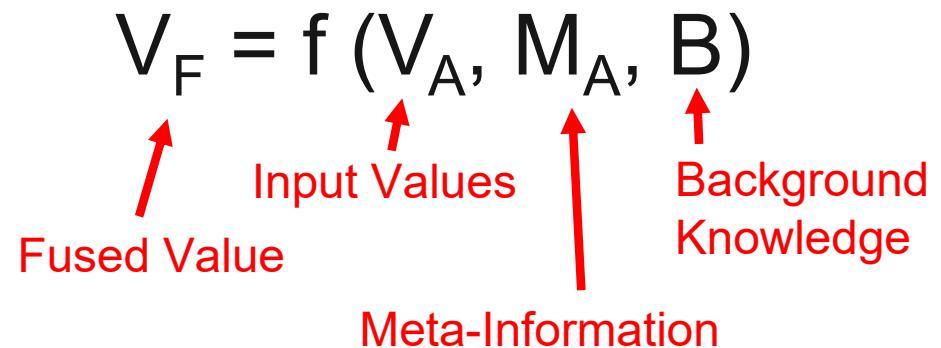
- Out of 33.3 million web tables, 949,970 tables contain at least one matching row
- 42% of the clusters have a size of 1 (slot filling)
- 16% of the clusters have a size of 2 (conf. res.)
- 39% of the clusters have a size of at least 3
- 13% of the clusters have a size of at least 10



Ritze, et al.: Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases. WWW 2016.

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. **Instance-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

$$V_F = f(V_A, M_A, B)$$


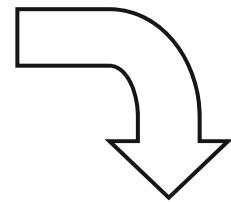
Instance-based Conflict Resolution Functions

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

Metadata-based Conflict Resolution Functions

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

Example: Complete Conflict Resolution Heuristic



0766607194	H. Melville	Moby Dick	\$3.98	Uhyhz
------------	-------------	-----------	--------	-------

Favor Sources
(amazon.com)

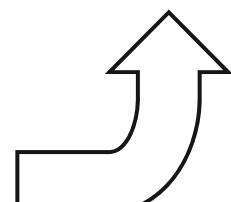
Max Length

Random

Most Recent

Union

0766607193	Herman Melville	Mopy Dick	\$5.99	
------------	-----------------	-----------	--------	--



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.

$$density_{column} = \frac{|non-NULL\ values\ in\ column|}{|rows\ in\ table|}$$

$$density_{Table} = \frac{|non-NULL\ values\ in\ table|}{|columns| * |rows|}$$

- 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 C₁
 2. Fuse table A with table B
 3. Measure density of resulting table A' or column C₁'

Consistency

A data set is consistent if it is free of conflicting information.

$$consistency_{Column} = \frac{|non-conflicting\ values\ in\ column|}{|real-world\ entities\ described|}$$

$$consistency_{Table} = \frac{|non-conflicting\ values\ in\ table|}{|columns| * |real-world\ entities\ described|}$$

Measurement:

1. Combine multiple tables using record correspondences
 - group records that refer to same real-world entity
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.

$$\text{accuracy} = \frac{|\text{correct values}|}{|\text{all values}|}$$

$$\text{error rate} = 1 - \text{accuracy}$$

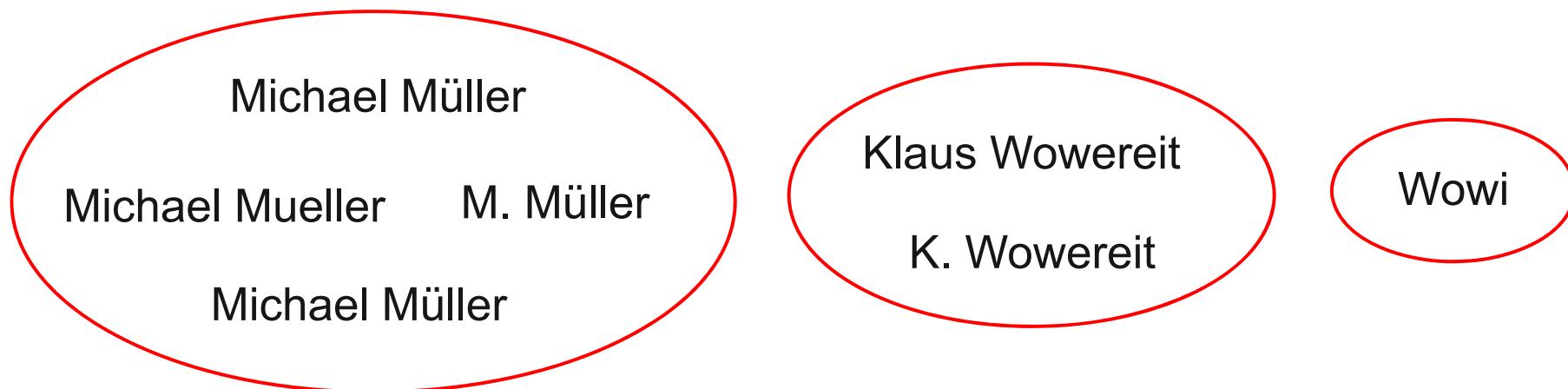
Measurement:

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

Gao, et al.: Efficient Knowledge Graph Accuracy Evaluation. VLDB Endowment, 2019.

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 Chapter Identity Resolution)
 2. Treat all values above threshold as equal
- Example: Mayor of Berlin



5.4. Example Data Fusion Tool: Fuz!on

Prototype
developed at
Hasso Plattner
Institute

Fuzzy Fuz!on

Additional Information Test/Debug

Automatic Fusion Rule-based Fusion Manual Fusion

Rule Matrix

	Firstname	Lastname	Street	housenumber	postcode	city	ignore	phone
None	66105	68111	58872	66404	63121	71285	100000	73936
Null values	5671	6402	6116	16746	12208	5643	0	26064
Case Variance	10835	12745	14563	0	0	11330	0	0
Abbreviation	7095	1170	8256	16850	12364	942	0	0
Tokenization	0	0	0	0	0	0	0	0
Substrings	2122	2091	1088	0	12307	1701	0	0
Dominance	2170	2424	2883	0	0	2434	0	0
Low edit distance	5913	7057	7101	0	0	6664	0	0
Global dominance	88	0	762	0	0	1	0	0
Undefined	1	0	359	0	0	0	0	0

Actions

Fusionsregel(n) anzeigen/erzeugen Nur aktuelle Markierung anzeigen [WEITER -->](#)

Selected Rules

Regeldefinition (Status: neu)

Spalten

Firstname
Lastname
Konflikttypen
Low edit distance

Primäre Konfliktauflösung

Vote
Minimum fraction of solution (in %) : 50
 Ignore case
 Ignore null-values

Aktionen

Übernehmen
Ausblenden
Spalte hinzufügen
Konflikttyp hinzufügen

Sekundäre Konfliktauflösung

First

Manual Fusion of Record Groups in Fuz!on

Fuzzy Fuz!on

Additional Information Test/Debug

Automatic Fusion Rule-based Fusion Manual Fusion

Groups 0 to 50 of 100000 All Groups

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

Back Next

21. Group :

Firstname	Lastname	Street	housenumber	postcode	city	ignore	phone
Manh The	Knaut	Wiedeberger Weg	37	24943	Flensburg	19280312	0461
Manh The	KNAUT	Wiedeberger Weg		24943	Flensburg	19280312	0461
Manh	Knaut	WIEDEBERGER WEG	37	24943	Flensburg	19280312	0461
First	Mixed ...	Vote	First non-null value	First	First	First	First
Manh The	Knaut	Wiedeberger Weg	37	24943	Flensburg	19280312	0461

Merge Save Configurations

5.5 Case Study: DBpedia Cross Language Data Fusion

- Infoboxes in different Wikipedia editions contain conflicting values.
- Which value to prefer?



The screenshot shows the English Wikipedia article for Mannheim. The page title is "Mannheim". Below it is a summary: "From Wikipedia, the free encyclopedia". A note states: "This article is about the city in Germany." The main content describes Mannheim as a city in southwestern Germany between Stuttgart and Karlsruhe. To the right is an infobox containing the following data:

Area	
• Total	144.96 km ² (55.97 sq mi)
Elevation	97 m (318 ft)
Population (2011-12-31) [1]	
• Total	314,931
• Density	2,200/km ² (5,600/sq mi)



The screenshot shows the German Wikipedia article for Mannheim. The page title is "Mannheim". Below it is a note: "Der Titel dieses Artikels ist mehrdeutig. Weiterlesen". The main content describes Mannheim as a "Quadratstadt und Universitätsstadt" (square city and university town) from 1720-1778. To the right is an infobox containing the following data:

Höhe:	97 m ü. NHN
Fläche:	144,96 km ²
Einwohner:	291.458 (31. Dez. 2011) [1]
Bevölkerungsdichte:	2011 Einwohner je km ²

A red arrow points from the English Wikipedia's population entry to the German Wikipedia's population entry, highlighting the conflict.

Cross-Lingual Data in DBpedia

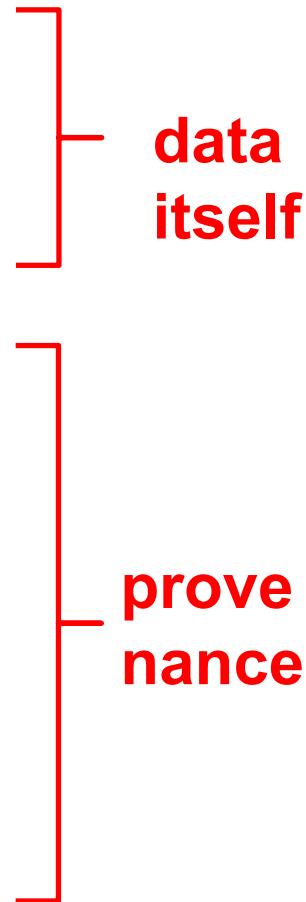
- DBpedia extracts structured data from Wikipedia in **119 languages**.
- DBpedia contains **lots of data conflicts**, inherited from Wikipedia.
- **Identity resolution is solved** by Wikipedia inter-language links.
- **Schema heterogeneity problem is solved** by community-created mappings from infoboxes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



Goal: Fuse Data between different Language Editions

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?



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
```

Provenance Metadata from the Wikipedia Revision Dumps

- 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	lastedit	2011-12-22T00:50:21Z
	propeditcnt	3
	autheditcnt	1136639
	authregdate	2009-12-18T02:08:09Z
nl:Mannheim:populationTotal	lastedit	2007-12-09T16:41:06Z
	propeditcnt	1
	autheditcnt	73
	authregdate	2007-04-05T08:54:19Z

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

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

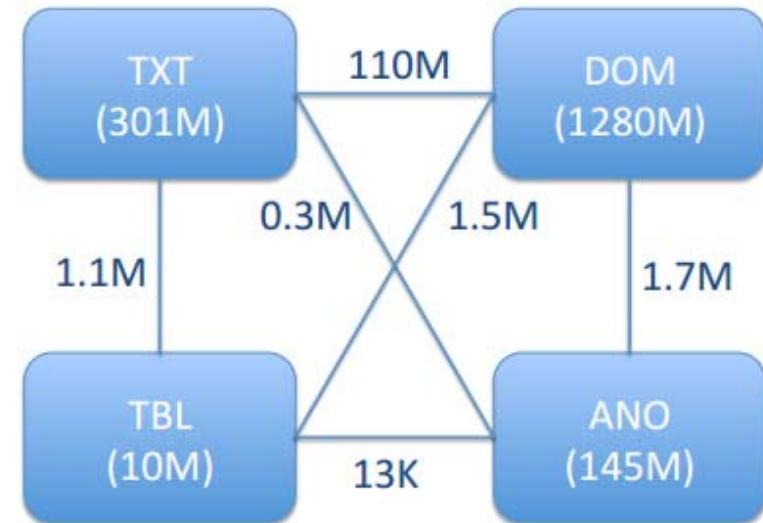
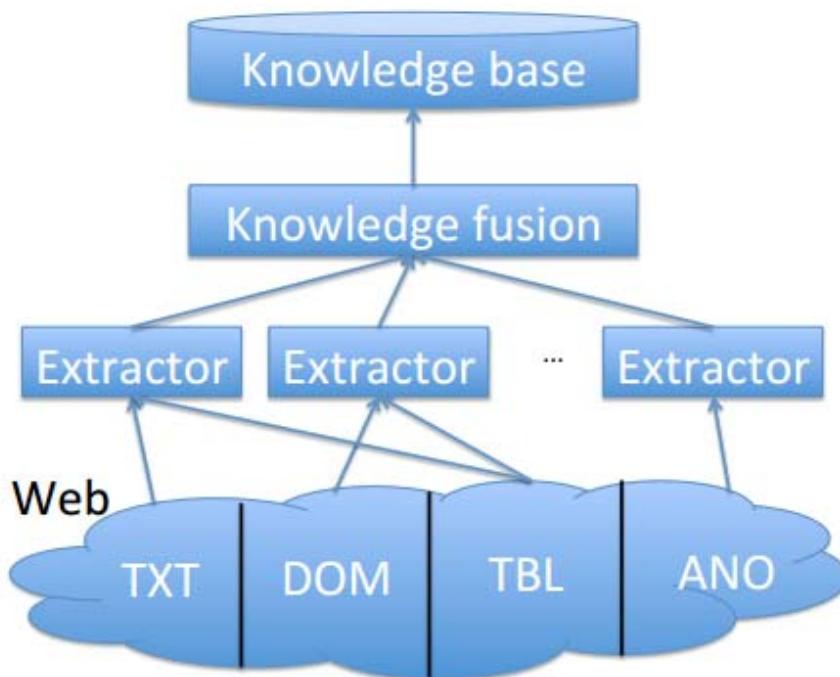
* "cities1000" are cities with population >1000

- **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

Volha Bryl, Christian Bizer: Learning Conflict Resolution Strategies for Cross-Language Wikipedia Data Fusion. 4th Workshop on Web Quality @ WWW 2014.

5.6 Case Study: Google Knowledge Vault

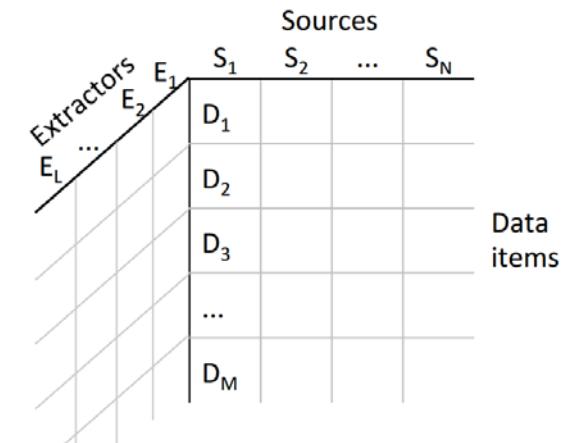
- uses 12 different extractors to extract 6.4 billion triples (1.6 billion unique triples) from 1 billion page Web crawl
- extracted data is fused to extend the Freebase knowledge base



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

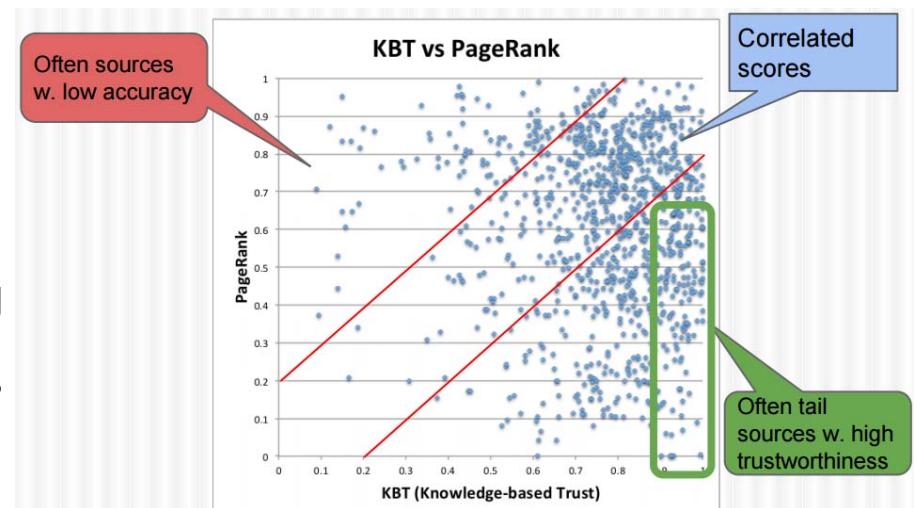
Google Knowledge Vault

- 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

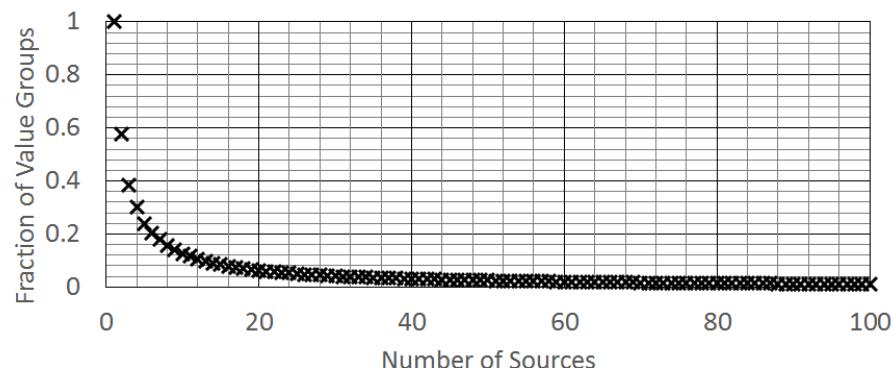
- determine trustworthiness of a data source by comparing its content with a knowledge base (ground truth)
- result: better than PageRank in identifying
 - tail websites with high trustworthiness
 - gossip websites



Luna Dong, et al.: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. SIGKDD 2014.
Luna Dong, et al.: Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. VLDB 2015.

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
 - Abedjan, Golab, Naumann, Papenbrock: Data Profiling. Morgan & Cleypool Synthesis Lecture in Computer Science, 2018.
- Provenance
 - Dublin Core Metadata Element Set. <http://dublincore.org/documents/dces/>, 2012.
 - Gil, Miles: PROV Model Primer, <http://www.w3.org/TR/prov-primer/>, 2013.
- Data Quality
 - Wang, Strong: Beyond accuracy: What data quality means to data consumers. JMIS, 1996.
 - Naumann, Rolker: Assessment Methods for Information Quality Criteria. Conference on Information Quality, 2000.
 - Abedjan, et al.: Detecting data errors: where are we and what needs to be done? VLDB 2016.
 - Fan, Geerts: Foundations of Data Quality Management. Morgan & Claypool, 2012.
 - Chandola, et al.: Anomaly Detection: A Survey. ACM Computing Surveys, 2009.

References

- Data Fusion
 - Bleiholder, Naumann: Data Fusion. ACM Computing Surveys, 2008.
 - Li, Gao, Meng, et al.: Survey on Truth Discovery. SIGKDD Explorations, 2016.
 - Dong, Srivastava: Big Data Integration. Chapter 4. Morgan & Claypool, 2015.
 - Ilyas, Chu: Data Cleansing. ACM Books, 2019.
 - Dong & Naumann: Data Fusion. Tutorial at VLDB 2009.
Slides: http://dc-pubs.dbs.uni-leipzig.de/files/dataFusion_vldb.pdf
 - Rekatsinas: Tutorial Data Integration and Machine Learning. SIGMOD 2018
Chapter ML for DF. https://thodrek.github.io/di-mi/sigmod2018/slides/05_MLforDF.pdf
 - Aggarwal: Managing and Mining Uncertain Data. Springer, 2010.
- Data Fusion Evaluation Datasets
 - Dong: Data Sets for Data Fusion Experiments
<http://lunadong.com/fusionDataSets.htm>