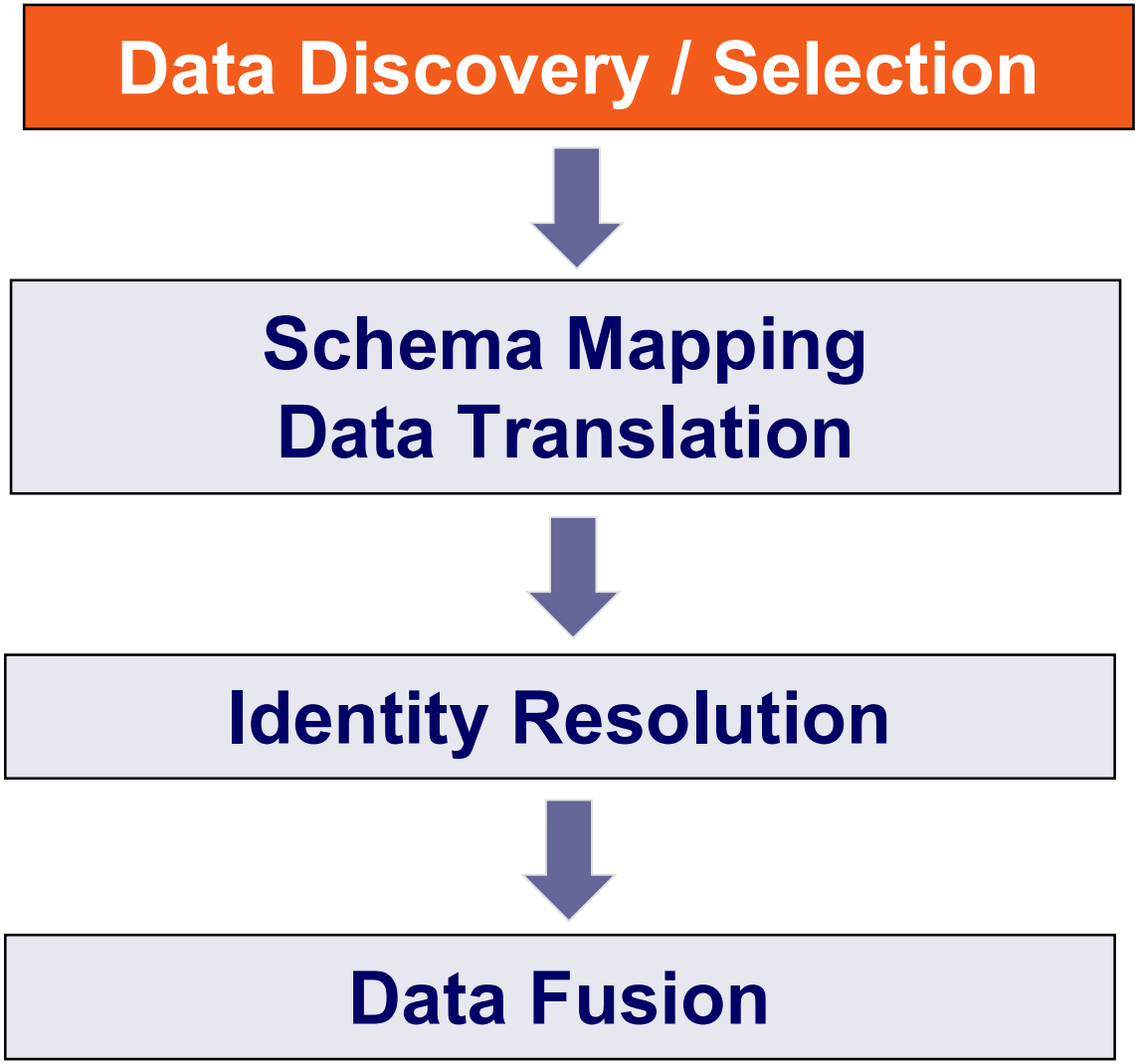
UNIVERSITY OF MANNHEIM

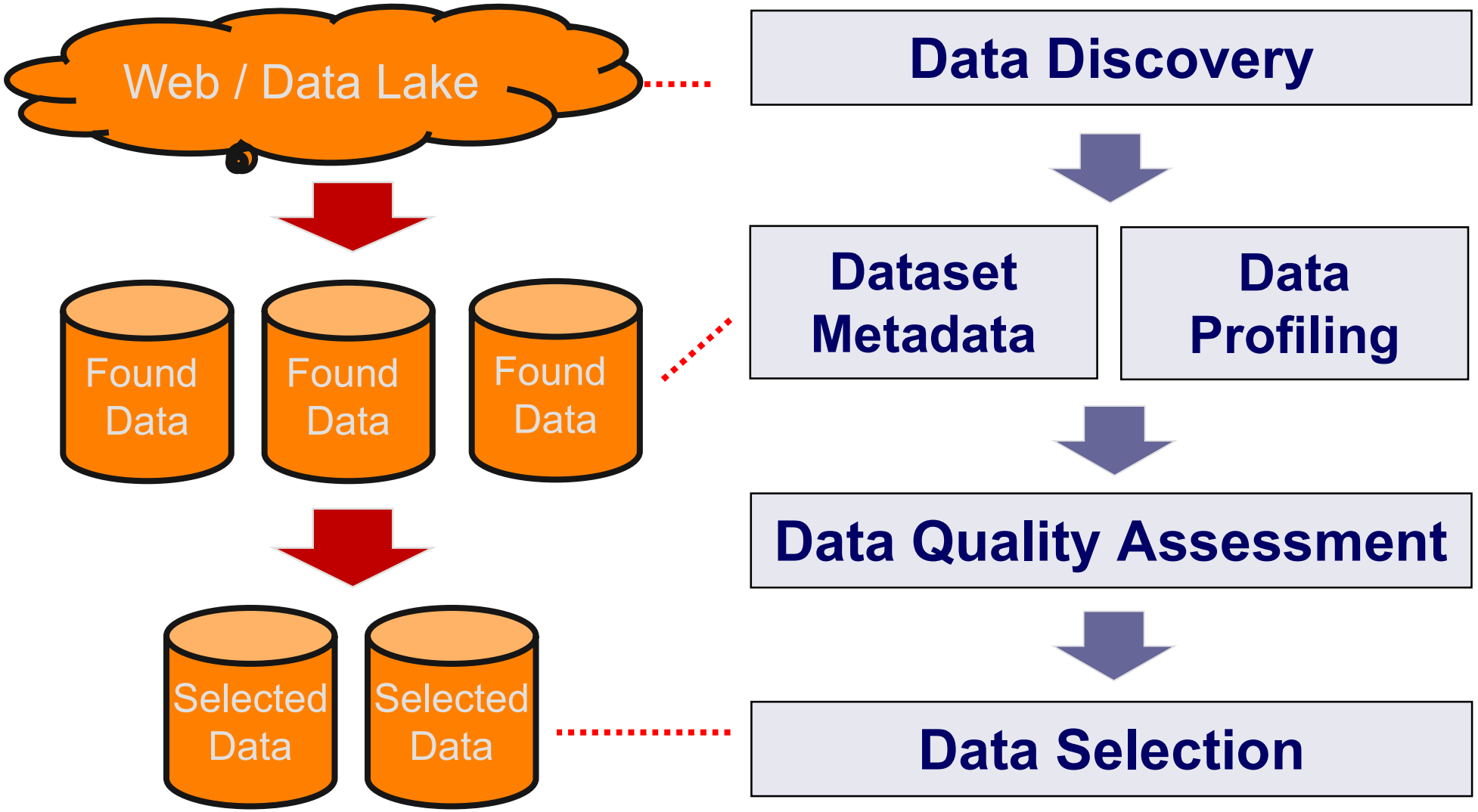
Web Data Integration

Data Discovery and Data Quality Assessment

The Data Integration Process



5.1 Data Discovery and Selection



Outline

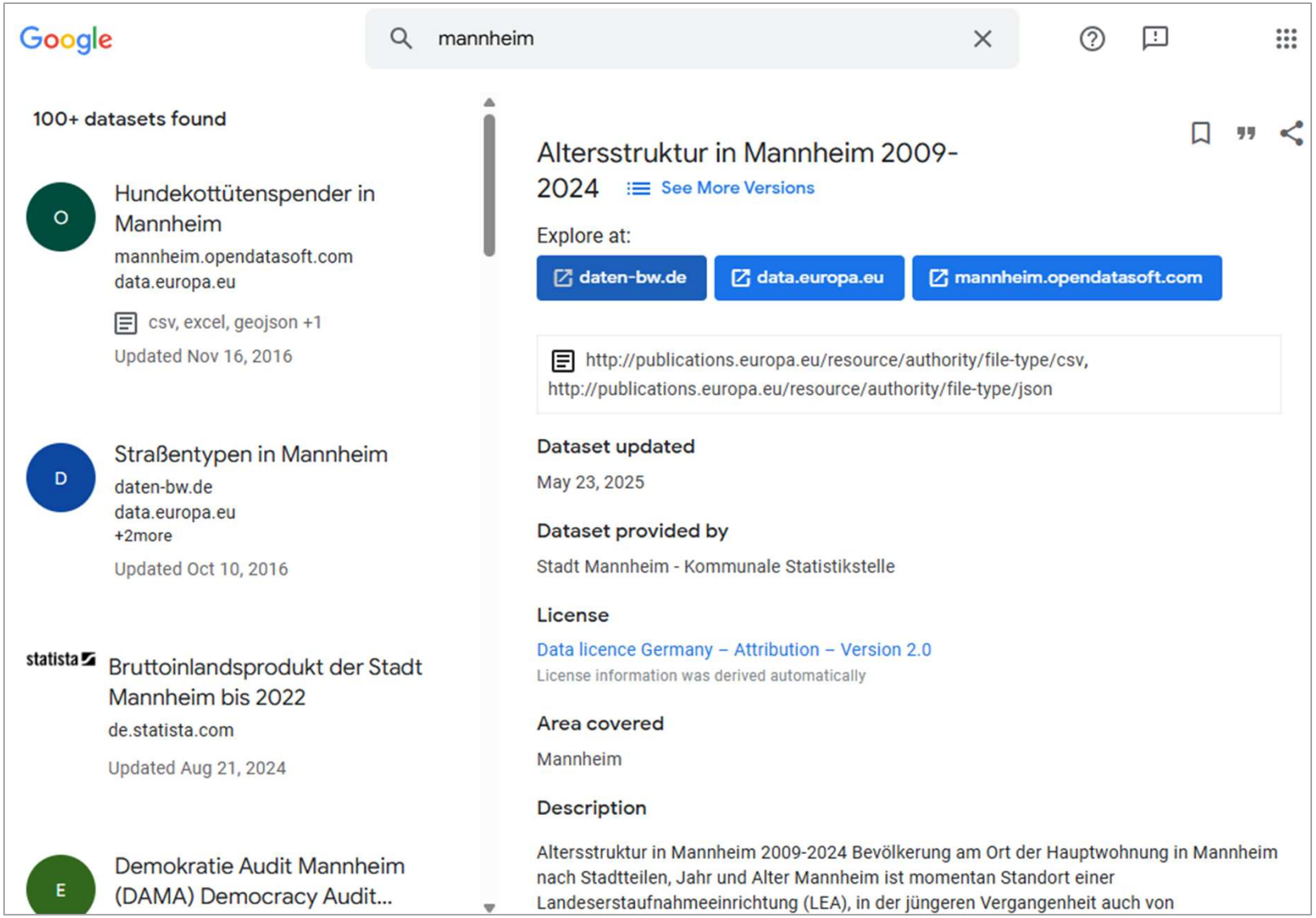
1. Data Discovery
2. Dataset Metadata
3. Data Profiling
4. Data Quality
5. Data Quality Assessment

1. Data Discovery

Goal: Given a query describing the user's information need, generate a ranked list of relevant datasets.

- Queries may consist of:
  1. **keywords** being matched (separately) against
     - dataset metadata
     - schema elements
     - dataset content
  2. **tables** (query-by-table) for finding
     - join-able tables
     - union-able tables
  3. additional **constraints**, e.g.
     - license requirements
     - timeliness requirements
- Retrieval methods may eighter index data and metadata as text (traditional retrieval) or rely on embeddings (neighborhood search)
     - see course: IE663 Information Retrieval and Web Search

Example: Dataset Search Engine



Crawls dataset metadata from the Web.

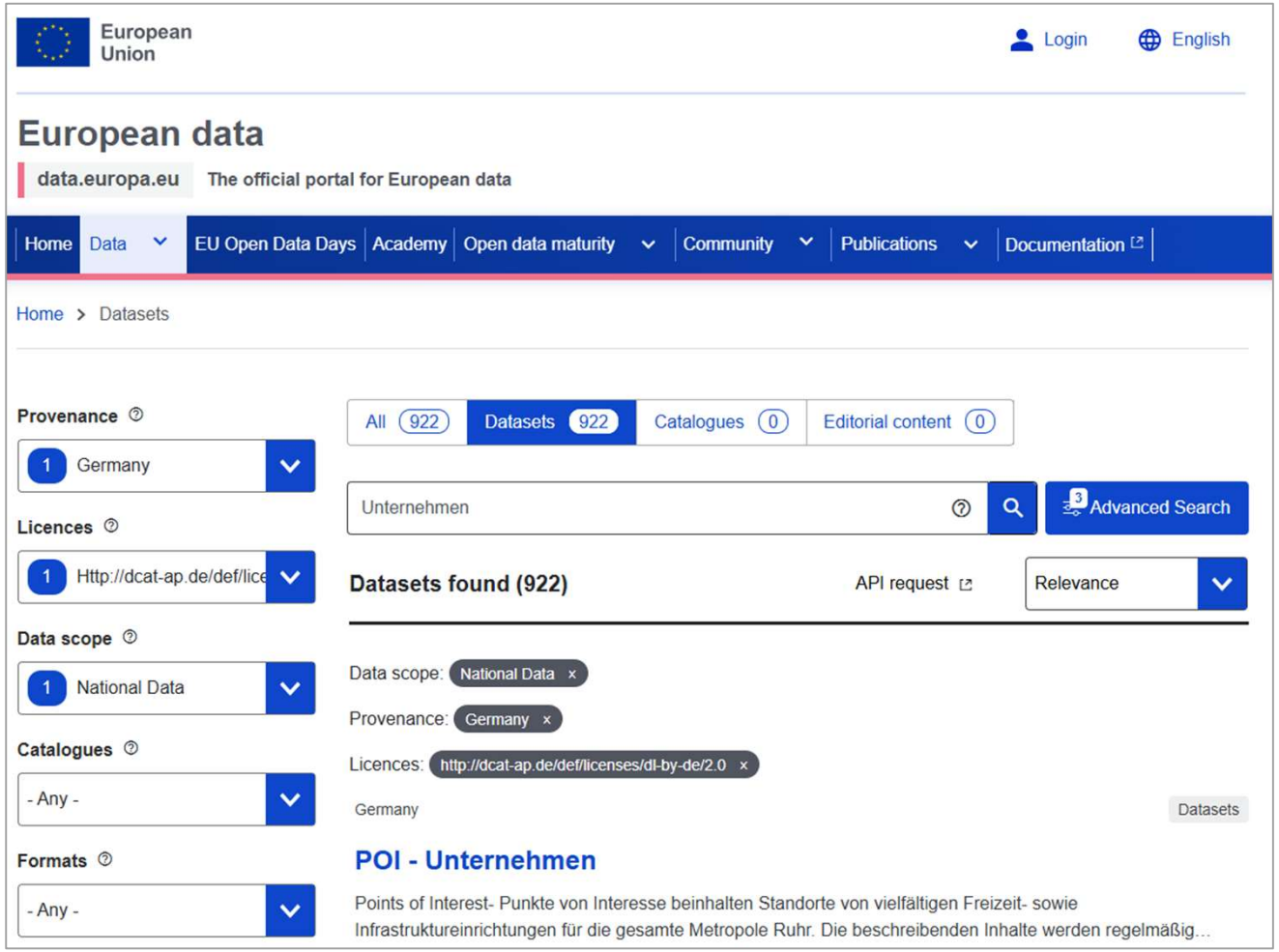
schema.org

https://datasetsearch.research.google.com/

Faceted Dataset Search

Faceted search combines keywords and filter constrains (facets)

- on license,
- on format,
- or source,
- …



https://data.europa.eu/

Query-by-Table: Joinable Table Search

Goal: Given a table, find tables that can be joined with the query table to augment it with additional attributes.

Query table

Region	Unemp. Rate
Lorraine	8 %
Guadeloupe	22 %
Centre	10 %

Ranked list of joinable tables

Region	GDP per Capita
Lorraine	
Guadeloup	
Centre	

Region	Population Growth
Martinique	1,64 %
Lorraine	-0,05 %
Guadeloupe	1,34 %
Centre	0,16 %

Extended query table

Region	Unemp. Rate	Population Growth
Lorraine	8 %	-0,05 %
Guadeloupe	22 %	1,34 %
Centre	10 %	0,16 %

Label column

New attribute

- Approach:
  1. Index label columns (and attribute names)
  2. Rank tables by label column value (and attribute) overlap
- Use cases:
  - enrich data for analysis, e.g. discover novel correlations
  - extend training dataset with additional features

Zhu, et al.: JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. SIGMOD, 2019.

Query-by-Table: Table Union Search

Goal: Given a table, find tables that can be unioned with the query table to augment it with additional tuples.

Query table

Region	Unemp. Rate
Lorraine	8 %
Guadeloupe	22 %

Ranked list of union-able tables

Region	Unemp. Rate
Centre	
Lorraine	

Region	Unemp. Rate
Berlin	9 %
Bayern	5 %

Extended query table

Region	Unemp. Rate
Lorraine	8 %
Guadeloupe	22 %
Centre	10 %
Berlin	9%
Bayern	5%

Additional tuples

- Approach:
  1. Semantically index schema attributes
  2. Rank tables by attribute overlap
- Use cases:
  - extend training data with additional examples
  - extend dataset to cover additional geographic regions or time periods

Nargesian, et al.: Table Union Search on Open Data. PVLDB, 2018.
Li, et al.: A Survey on Open Dataset Search in the LLM Era. arXiv:2509.00728, 2025.

2. Dataset Metadata

Data discovery may rely on different types of dataset metadata (data describing data) in addition to the dataset content.

Relevant types of metadata include:

- **Basic metadata:** Title, description, publication date, tags
- **Business metadata:** License, price, provenance information, contact information, results of quality checks,
- **Technical metadata**: schemata and value format, content fingerprints / examples, mappings to other schemata or ontologies (declarative or code), change history
- **Usage metadata**, access logs, user ratings, user comments

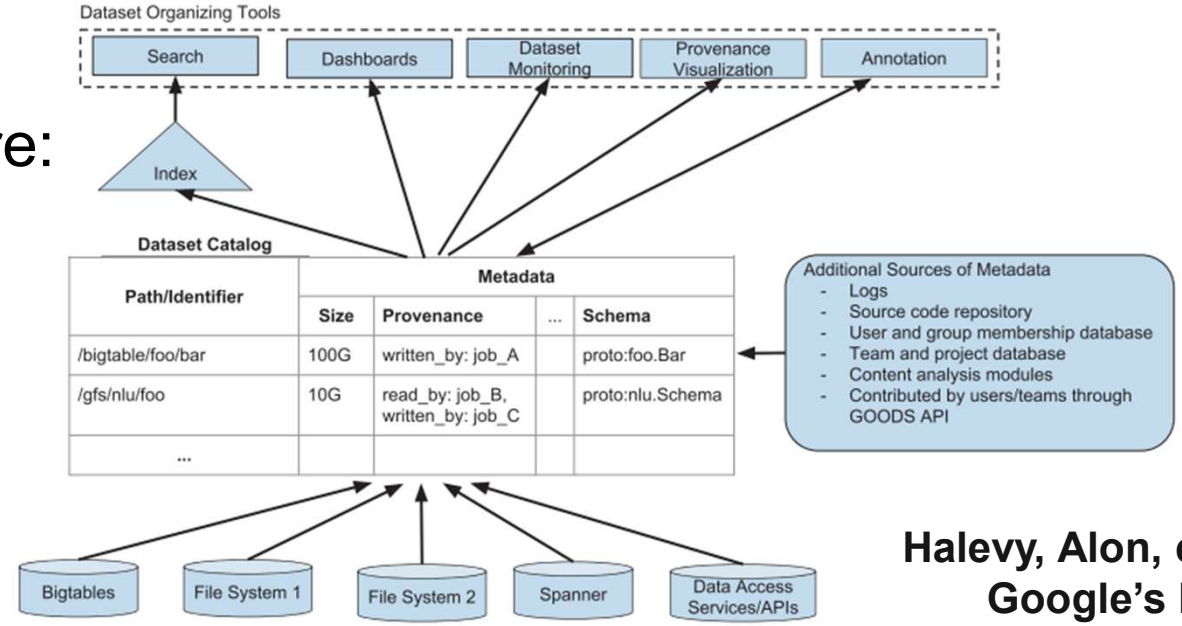
Problem: Metadata that needs to be manually created is often scarce or superficial ➔ Data profiling to partly automate creation

Example: GOODS Data Catalog

- Data catalog used internally by Google
  - number of datasets (2016): 16 billion
- Types of metadata stored

Metadata Groups	Metadata
Basic	size, format, aliases, last modified time, access control lists
Content-based	schema, number of records, data fingerprint, key field, frequent tokens, similar datasets
Provenance	reading jobs, writing jobs, downstream datasets, upstream datasets
User-supplied	description, annotations
Team and Project	project description, owner team name
Temporal	change history

- System architecture:



Halevy, Alon, et al.: Goods: Organizing Google's Datasets. SIGMOD, 2016.

2.1 Publishing Dataset Metadata on the Web

To support data discovery, datasets should be published together with rich and easy-to-understand metadata.

1. Vocabularies for representing dataset metadata
2. The FAIR data principles
3. Representing data together with metadata

Dublin Core Vocabulary (DC)

- The Dublin Core vocabulary defines terms for representing basic metadata
  - creator, contributor, publisher, date, rights, format, language, …
- The terms are used in different contexts
  - HTML document metadata, dataset catalogs, library catalogs
  - Example of a Linked Data document:

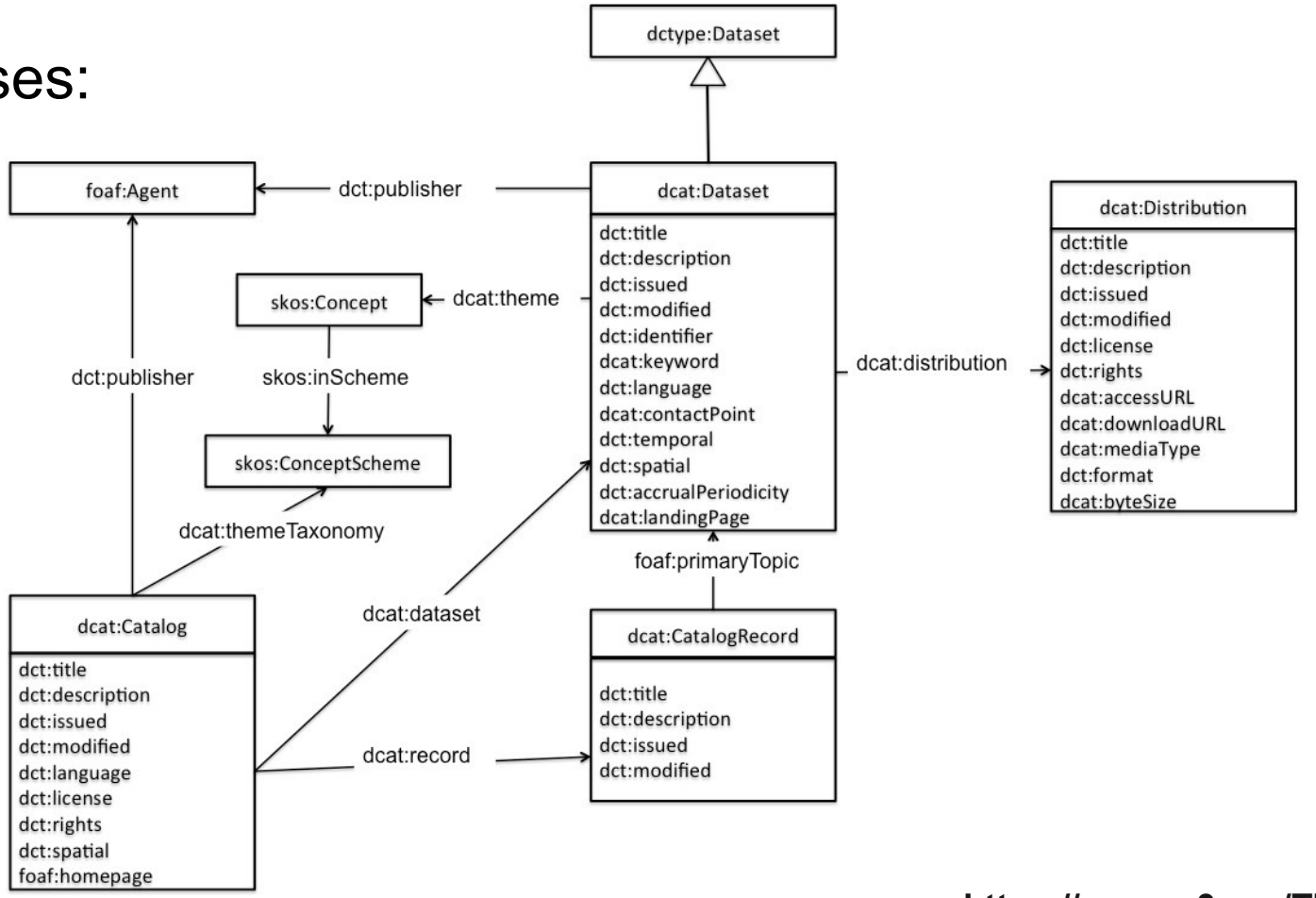
http://dbpedia.org/data/Alec_Empire

```
# 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 Data Catalog Vocabulary (DCAT)

- DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web
  - reuses Dublin Core terms
- Main classes:



https://www.w3.org/TR/vocab-dcat-3/

Example: DCAT Dataset Descriptions

Example uses Turtle syntax for RDF (see Data Exchange Formats – Part 2)

```
ex:dataset-001
  rdf:type dcat:Dataset ;
  dcterms:title "Imaginary dataset"@en ;
  dcat:keyword "accountability"@en, "transparency"@en, "payments"@en ;
  dcterms:creator ex:finance-employee-001 ;
  dcterms:issued "2011-12-05"^^xsd:date ;
  dcterms:modified "2011-12-15"^^xsd:date ;
  dcat:contactPoint <http://dcat.example.org/transparency-office/contact> ;
  dcterms:temporal [ a dcterms:PeriodOfTime ;
    dcat:startDate "2011-07-01"^^xsd:date ;
    dcat:endDate   "2011-09-30"^^xsd:date ;
  ];
  dcat:temporalResolution "P1D"^^xsd:duration ;
  dcterms:spatial <http://sws.geonames.org/6695072/> ;
  dcat:spatialResolutionInMeters "30.0"^^xsd:decimal ;
  dcterms:publisher ex:finance-ministry ;
  dcat:distribution ex:dataset-001-csv ;
```

Schema.org Dataset Vocabulary

- recommended by Google and used by Google dataset search
- Terms overlap with Dublin Core and DCAT vocabularies

Example uses JSON-LD syntax

```
<script type="application/ld+json">
    { "@context":"https://schema.org/",
      "@type":"Dataset",
      "name":"NCDC Storm Events Database",
      "description":"Storm Data is provided by the National Weather Service...",
      "url":"https://catalog.data.gov/dataset/ncdc-storm-events-database",
      "keywords":[ "CYCLONES", "DROUGHT","FOG","FREEZE"],
      "license" : "https://creativecommons.org/publicdomain/zero/1.0/",
      "creator":{
         "@type":"Organization",
         "url": "https://www.ncei.noaa.gov/",
         "name":"National Centers for Environmental Information",
      "temporalCoverage":"1950-01-01/2013-12-18",
      "spatialCoverage":{
         "@type":"Place",
         "geo":{
            "@type":"GeoShape",
            "box":"18.0 -65.0 72.0 172.0" }  }
      }
</script>
```

https://schema.org/Dataset
https://developers.google.com/search/docs/appearance/structured-data/dataset?hl=en

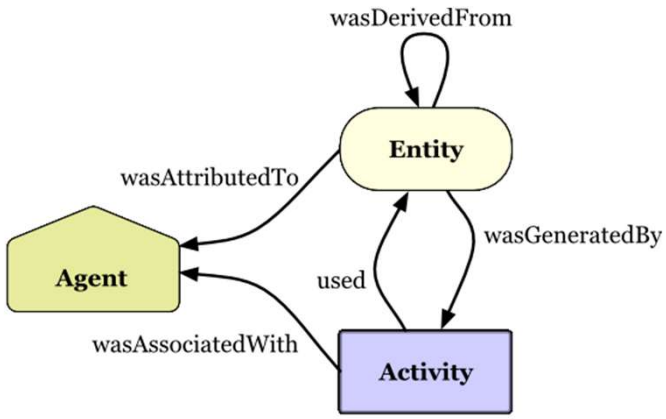
W3C Provenance Vocabulary (PROV)

- Provenance: Information about the data creation process.
- The W3C PROV vocabulary defines terms for representing detailed provenance chains.

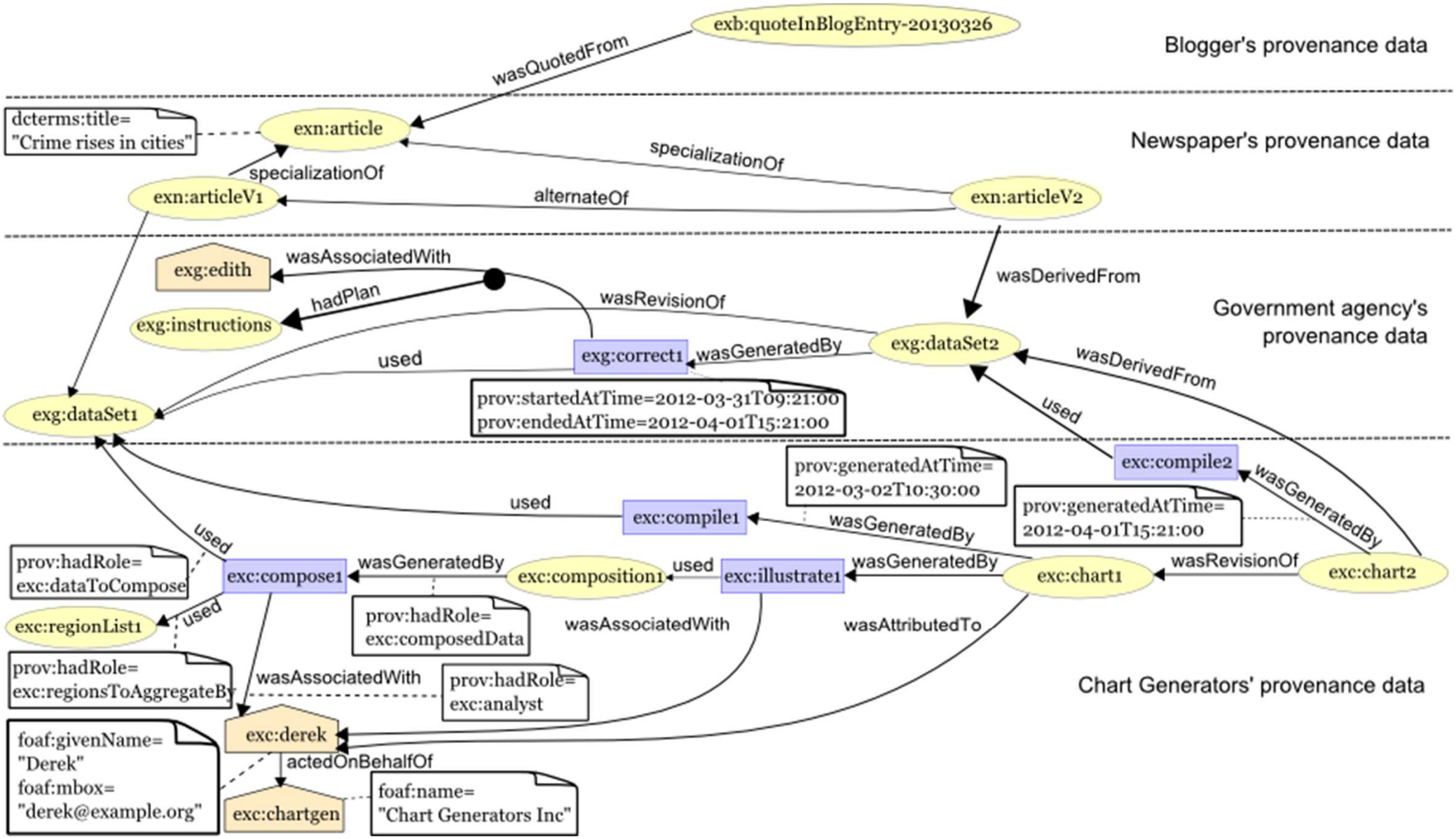
PROV

Example uses XML syntax

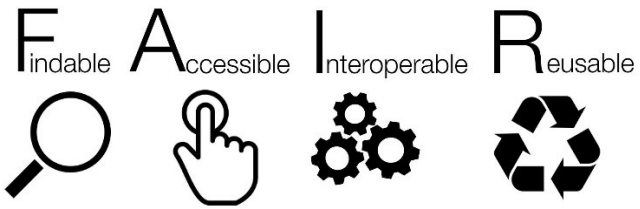
```
<prov:document>
  <!-- Entities -->
  <prov:entity prov:id="exn:article">
        <dct:title>Crime rises in cities</dct:title>
  </prov:entity>
  <!-- Agents -->
  <prov:agent prov:id="exc:derek">
      <prov:type>prov:Person</prov:type>
      <foaf:givenName>Derek Smith</foaf:givenName>
      <foaf:mbox>mailto:derek@example.org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



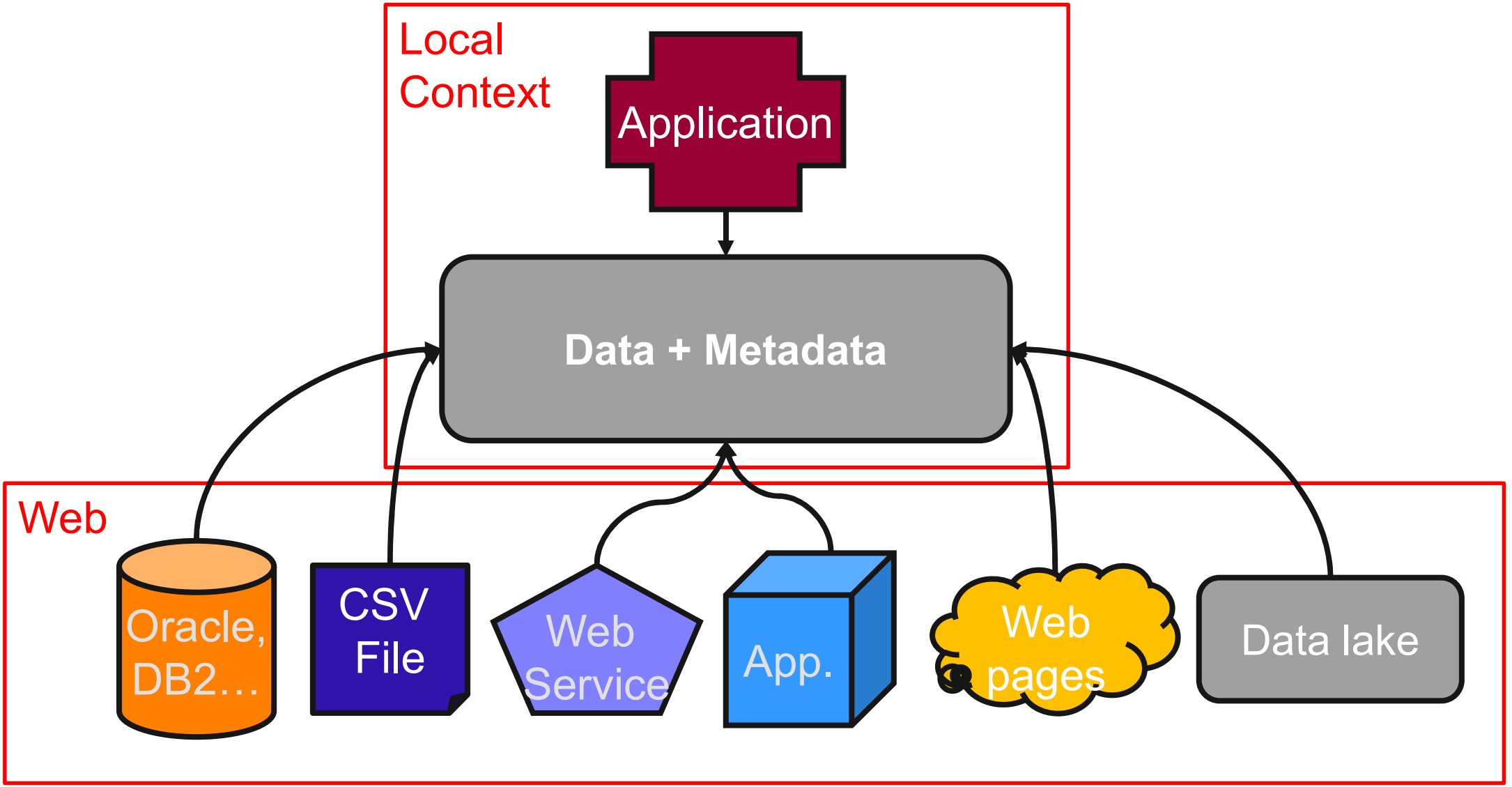
The FAIR Data Principles



- **F**indable
  1. (Meta)data are assigned a globally unique identifier
  2. Data are described with rich metadata
  3. (Meta)data are registered or indexed in a searchable resource
- **A**ccessible
  1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  2. Metadata are accessible, even when the data are no longer available
- **I**nteroperable
  1. (Meta)data use a formal, broadly applicable language for knowledge representation
  2. (Meta)data use vocabularies that follow FAIR principles
  3. (Meta)data include qualified references to other (meta)data
- **R**eusable
  1. (Meta)data are released with a clear data usage license
  2. (Meta)data are associated with detailed provenance
  3. (Meta)data meet domain-relevant community standards

https://www.go-fair.org/fair-principles/

2.3 Representing Data together with Metadata



Relational Data Model

- Alternative 1: Dataset-level metadata (coase grained, supported by Pandas)
- Alternative 2: Record-level metadata (finer grained, fast queries)
- Alternative 3: Value-level metadata (very fine grained, but slow queries)
- Alternative 4: Use specific database engine that support extended relational model.

Physicians with record level metadata

Key	Name	Street	ProvID
1425	Dr. Mark Smith	14 Main Street	001
1425	Mark Smith	12 Main St.	002
…	…	…	…

Physicians with value-level metadata

Key	Attribute	Value	ProvID
1425	Name	Dr. Mark Smith	001
1425	Name	Mark Smith	002
1425	Street	14 Main Street	001
…	…	…	…

Provenance metadata table

ProvID	Source	Date
001	www.mark-smith.com	12/6/2023 18:42:12
002	www.doc-find.com	12/1/2023 12:21:54
…	…	…

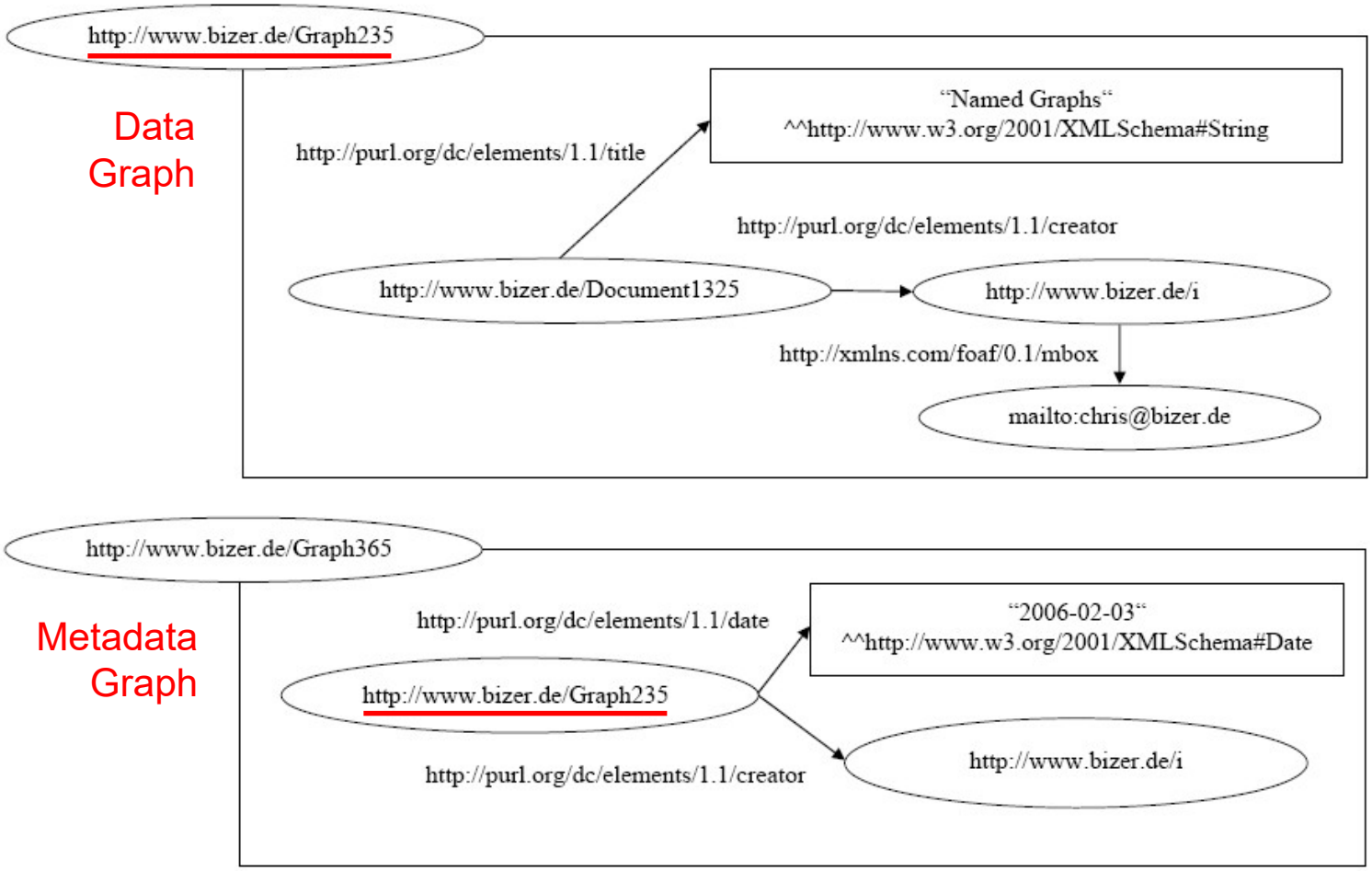
XML Data Model

Represent provenance using multiple value elements and XML 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 2023 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 metadata by talking about a graph in another graph
- Named Graphs can be queried using the SPARQL keyword GRAPH

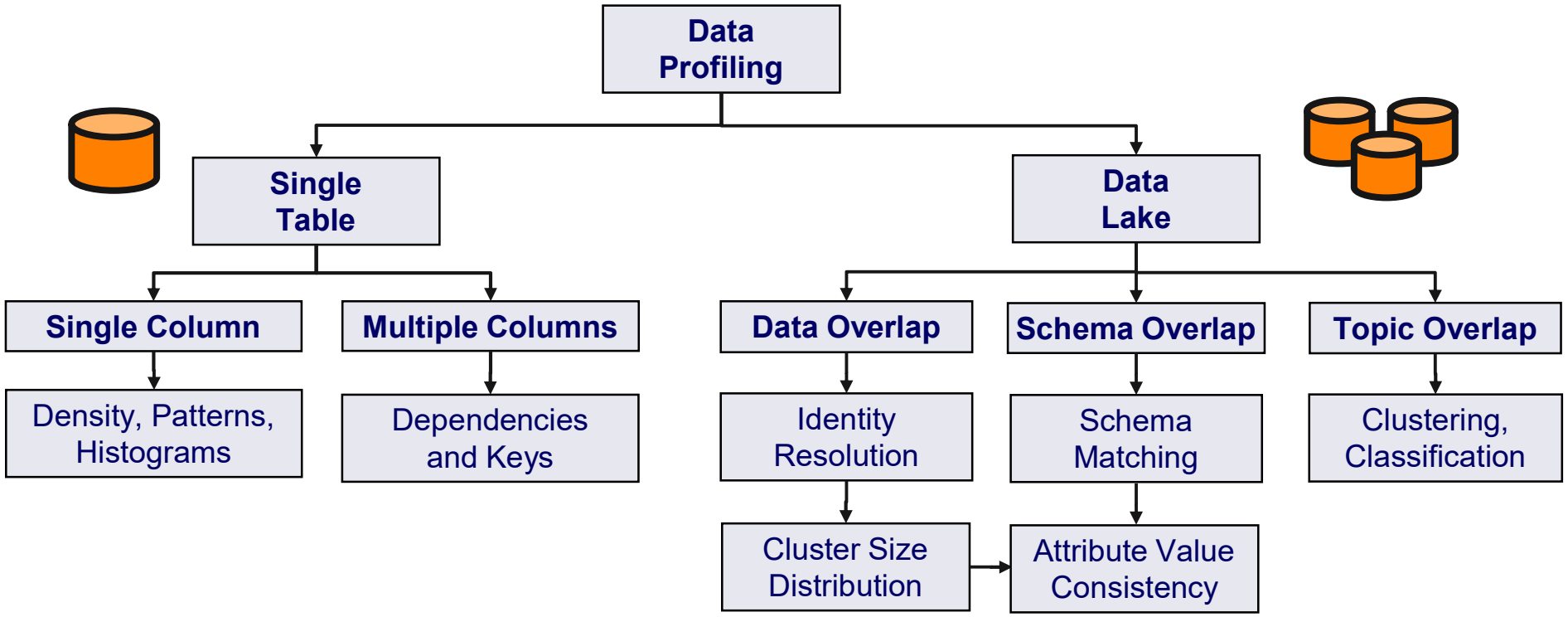


Carroll, Bizer, Hayes, Stickler: Named Graphs. Journal of Web Semantics, 2005.

3. 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 by profiling results



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

3.1 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, identier
	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

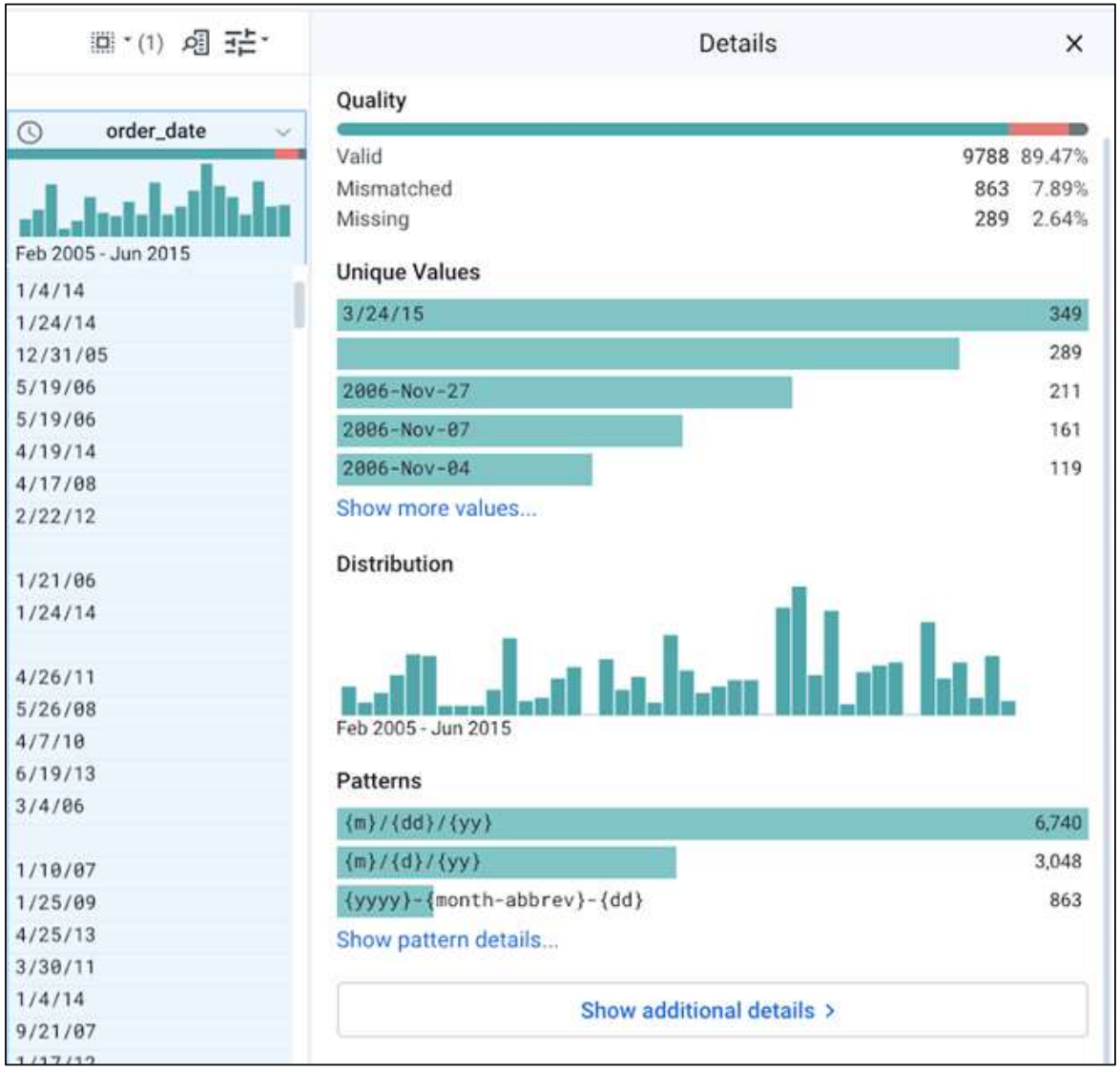
- outliers
- skewed distributions

Data types and lengths should always be inspected

Good summarization of an attibute

Example: Frequent Values and Frequent Value Patterns

Goolge Cloud Dataprep by Trifacta



Data type mismatch

Most frequent values

Most frequent value patterns

Dataset Profiling in Python

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

```
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)
```

Overview | Variables | Correlations | Missing values | Sample

Dataset info

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

Variables types

CAT	6
NUM	2

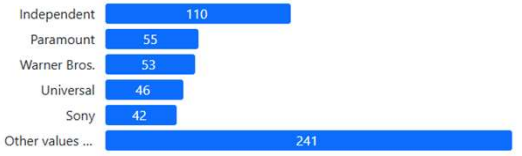
Data types

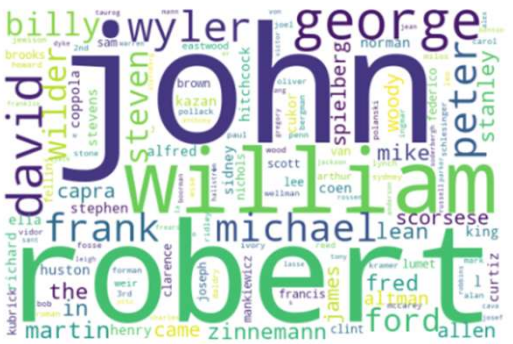
Toggle Reproduction Information

Toggle Warnings

Warnings

`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





Attribute summaries
Missing values
Correlations

Warnings

Dataset Profiling using LLMs

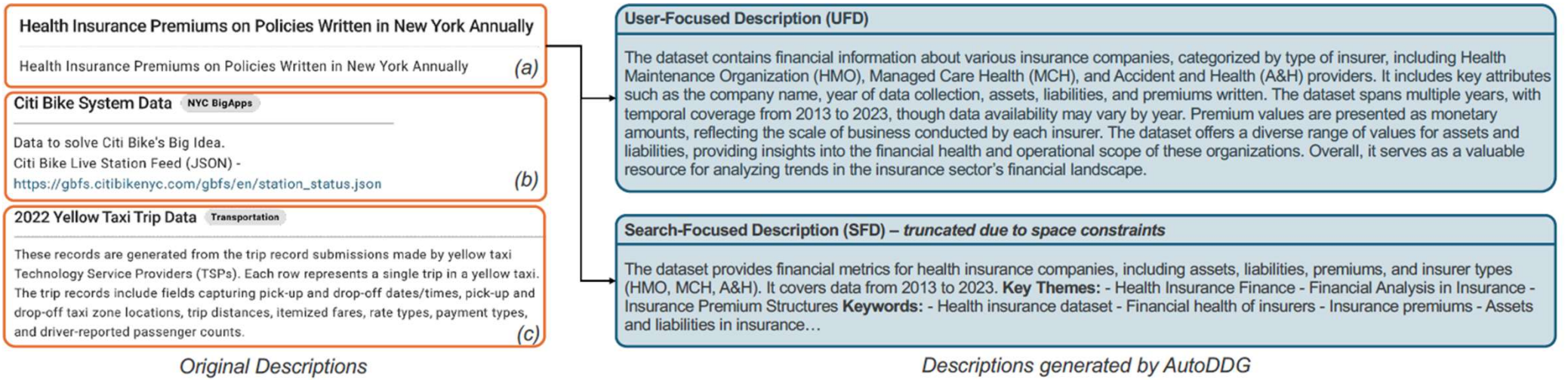
- LLMs can be used to generate textual dataset metadata.
  - descriptions, topics, keywords, use cases
- AutoDDG prototype creates search-focused and user-focused dataset descriptions
  - Example prompts:

Dataset Topic Generation Prompt

```
Using the dataset information provided, generate a concise
topic in 2-3 words that best describes the dataset's
primary theme:
- Title: {title}
- Original Description: {original_description} (optional)
- Dataset Sample: {dataset_sample}
- Topic (2-3 words):
```

Search-Focused Description

Prompt: You are given a dataset about the topic *D_topic*, with the following initial description: *D_initial_description*. Please expand the description by including the exact topic. Additionally, add as many related concepts, synonyms, and relevant terms as possible based on the initial description and the topic. Unlike the initial description, which is focused on presentation and readability, the expanded description is intended to be indexed at backend of a dataset search engine to improve searchability. Therefore, focus less on readability and more on including all relevant terms related to the topic. Make sure to include any variations of the key terms and concepts that could help improve retrieval in search results. Please follow the structure of following example template: *Template*.



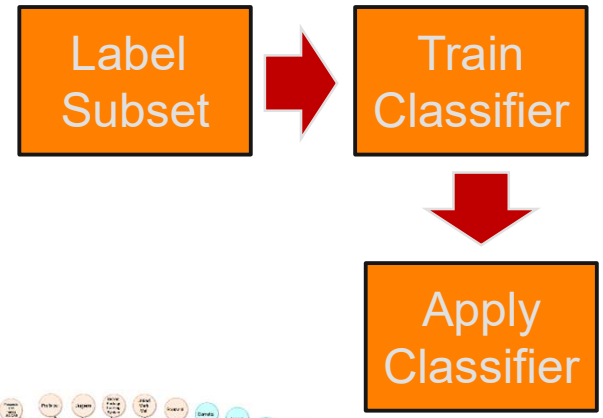
Original Descriptions

Descriptions generated by AutoDDG

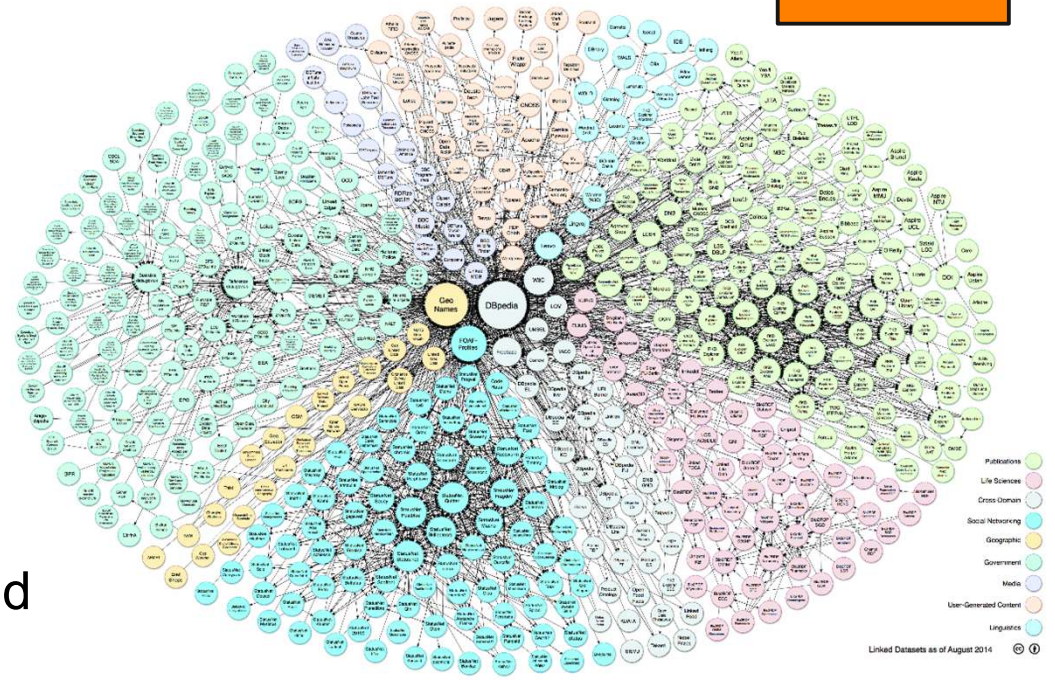
Haoxiang et al.: AutoDDG: Automated dataset description generation using LLMs. arXiv:2502.01050, 2025.

3.2 Data Lake Profiling: Topical 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.

Data Lake Profiling: Table Annotation

- Annotate the columns of tables in a large table corpus with terms from a knowledge graph or shared vocabulary
  - use case: column indexing for data search
- Subtasks:
  - Column Type annotation (CTA): distance, weight, location, or person
  - Column Property annotation (CPA): proteinContent, fatContent, director, producer

film	director	producer	country
???	???	???	???
Happy Feet	George Miller, Warren Coleman, Judy Morris	Bill Miller, George Miller, Doug Mitchell	USA
Cars	John Lasseter, Joe Ranft	Darla K. Anderson	UK
Flushed Away	David Bowers, Sam Fell	Dick Clement, Ian La Frenais, Simon Nye	France

SemTab evaluation campaign: https://www.cs.ox.ac.uk/isg/challenges/sem-tab/
Survey: Liu, et al.: From tabular data to knowledge graphs. Journal of Web Semantics, 2023.

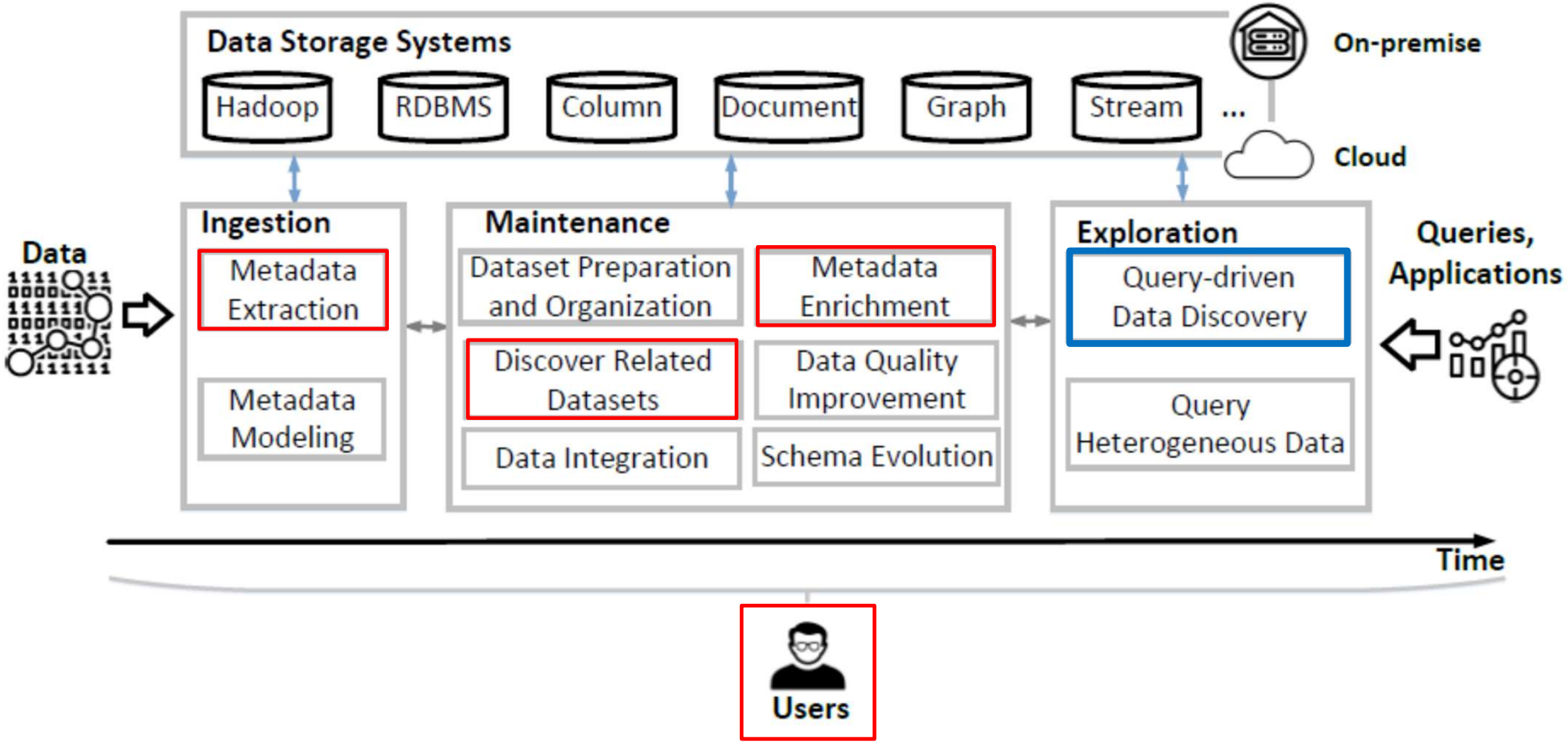
Data Lake Profiling: Data and Schema Overlap

- 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			
	T_0	T_c	V_c	Numeric	Date	String	Reference
+ 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 Institution	25 737	14 415	238 365	38 056	64 578	13 334	122 397
|- 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 Profiling within a Data Lake Management System



Hai, Rihan, Christoph Quix, and Matthias Jarke: Data Lake Concept and Systems: A Survey. arXiv:2106.09592, 2021.

4. 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, relevancy, …
   - which quality dimensions matter depends on the task at hand
2. is task-dependent
   - high quality requirements when you invest one million €
   - lower requirements when you choose which movie to watch
3. is subjective
   - some people are more paranoid than others

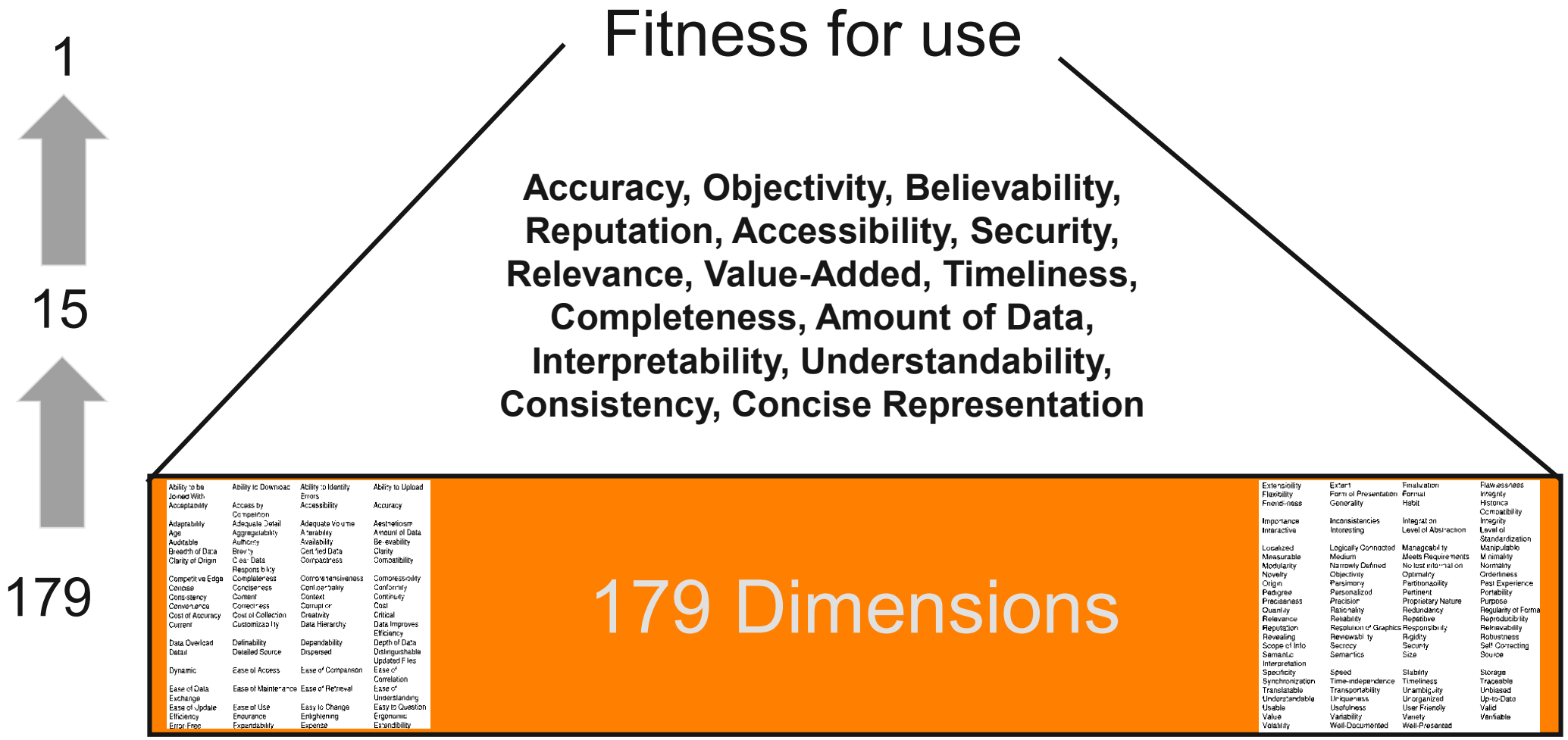
Data Quality in the Enterprise and Web Context

- 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:

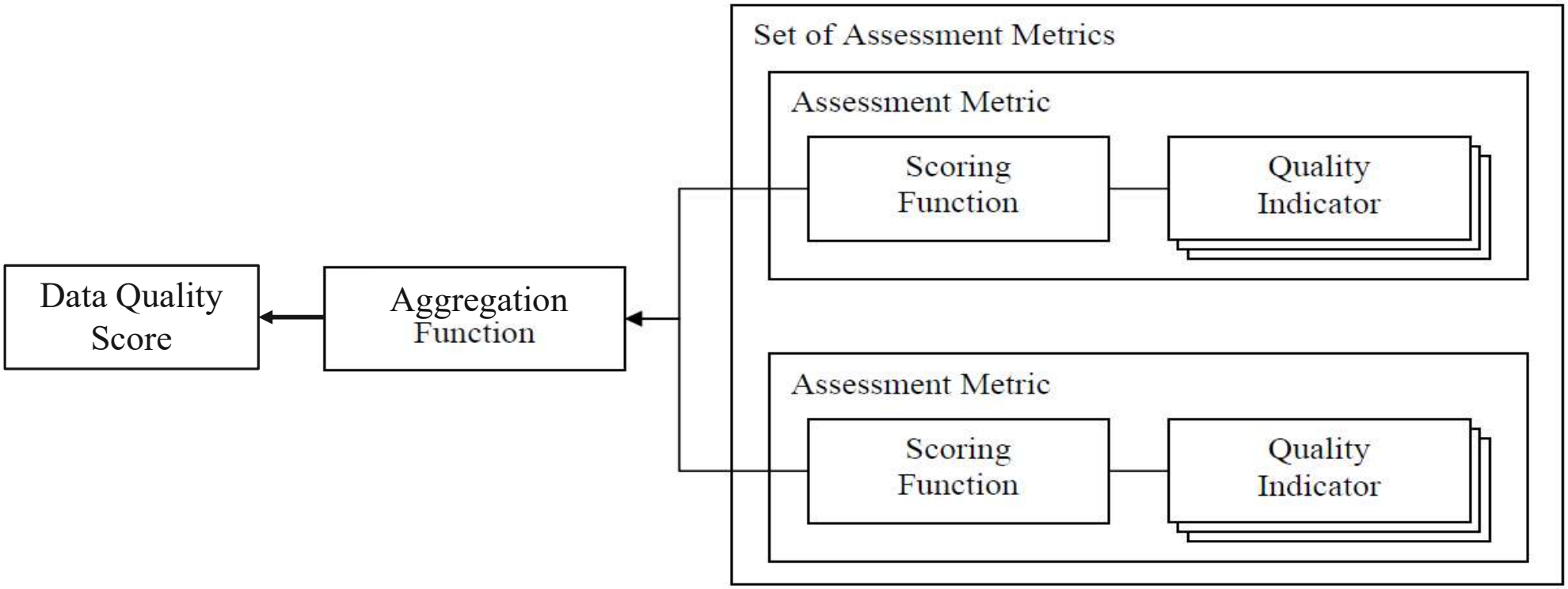


Category	IQ Criteria	TDQM	MBIS	Weikum	DWQ	SCOUG	Chen
Content-related Criteria	Accuracy	Yes	Yes	Yes	Yes	Yes	Yes
	Documentation					Yes	
	Relevancy	Yes	Yes		Yes		Yes
	Value-Added	Yes				Yes	
	Completeness	Yes	Yes	Yes	Yes	Yes	Yes
	Interpretability	Yes			Yes		
Technical Criteria	Timeliness	Yes	Yes	Yes	Yes	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	

Source: Felix Naumann

4.2. Data Quality Assessment

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



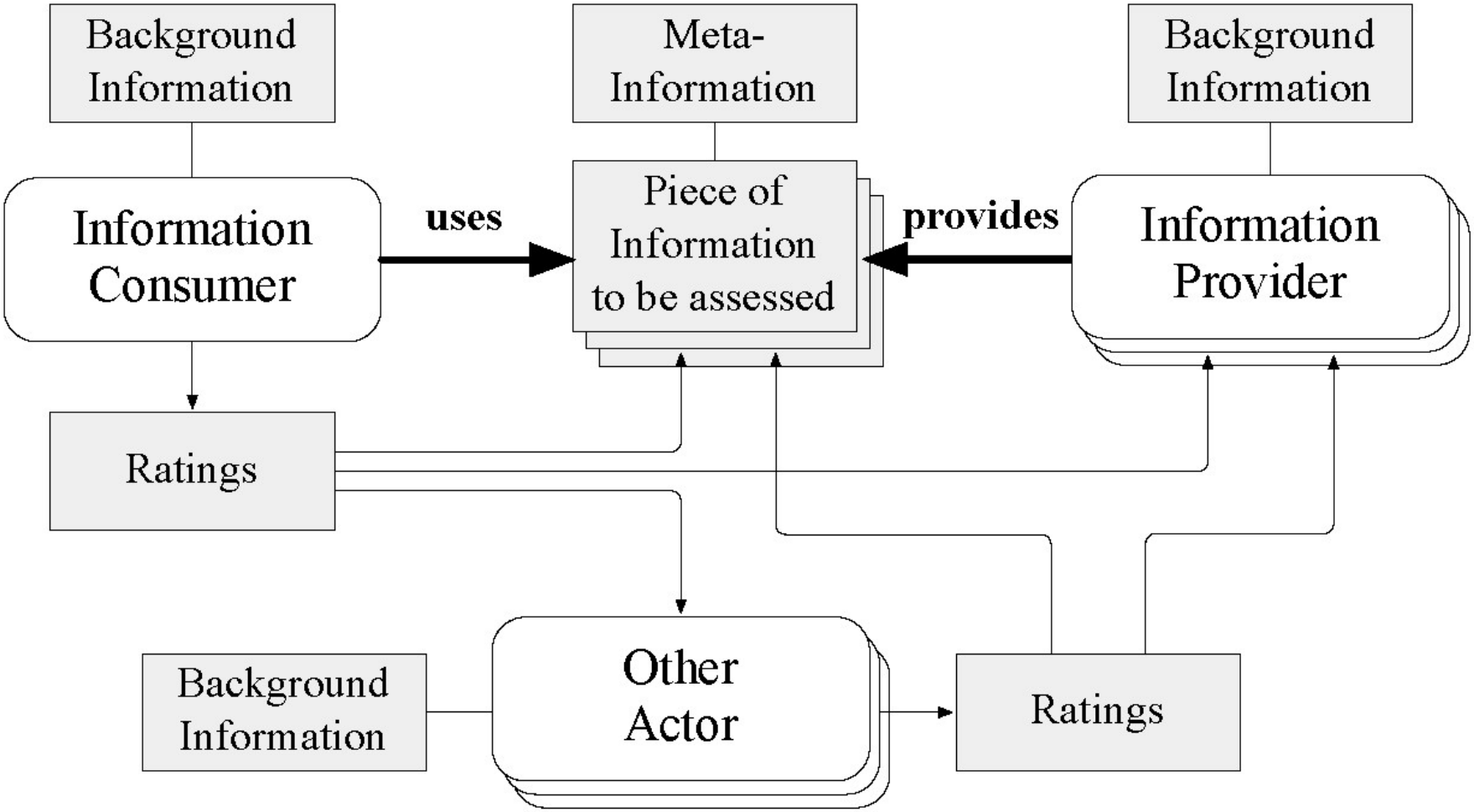
The applicability of specific heuristics depends on

1. Availability of quality indicators (like provenance information or ratings)
2. Quality of quality indicators (fake ratings, sparse provenance information)

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 positive ratings”, “Accept recommendations from a friend on restaurants, but distrust him on computers”, “Prefer content from websites having a high PageRank”

Quality Indicators in the Web Context

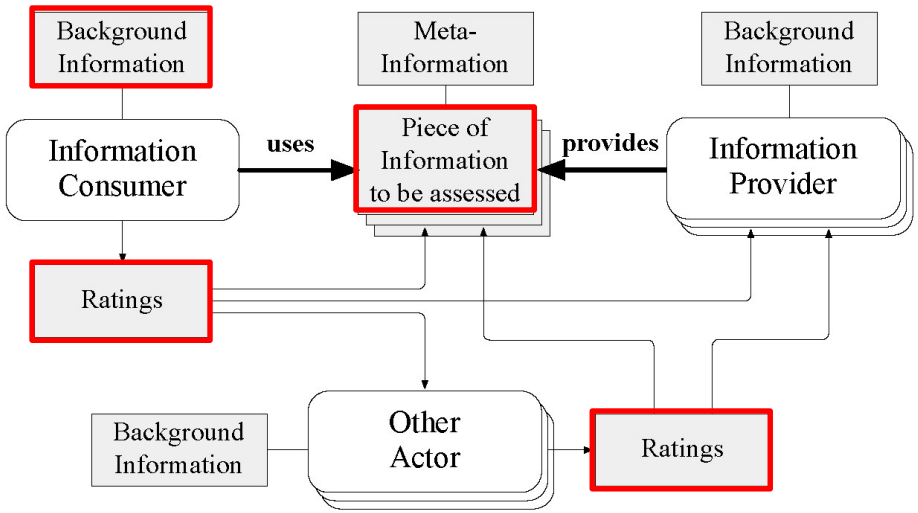


4.2.1 Assessing Data Accuracy

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

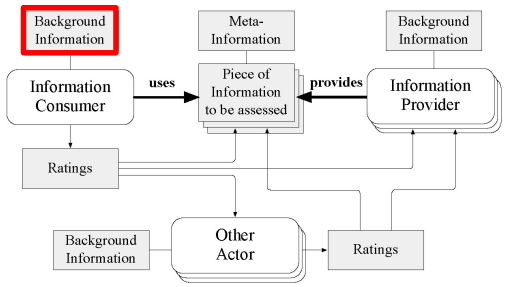
- also called: Truth Discovery, Fake News Detection

- 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.



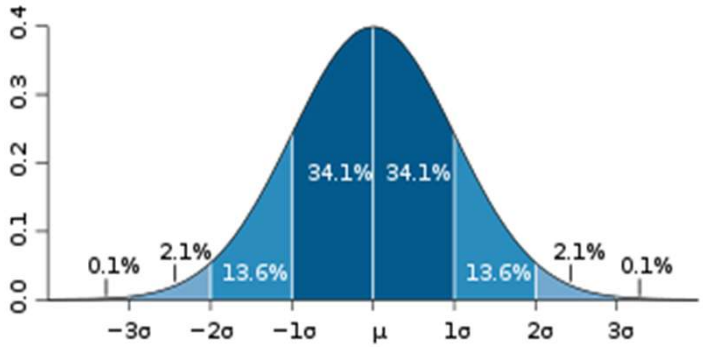
- Examples of constraints
  - the age of humans should be between 0 and 130
  - books must have at least one author
- Examples of consistency rules
  - if person is in middle school, then age is (likely) below 25
  - if area code is 131, then the city should be Edinburgh
- Rule and constraint acquisition
  - define rules and constraints manually
  - or learn from examples e.g. using association analysis (see lecture Data Mining)

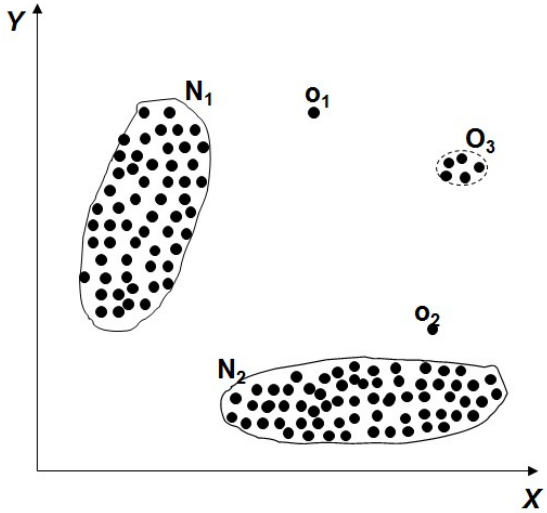
Fan, Geerts: Foundations of Data Quality Management. Morgan & Claypool, 2012.

Outlier Detection

An outlier is 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

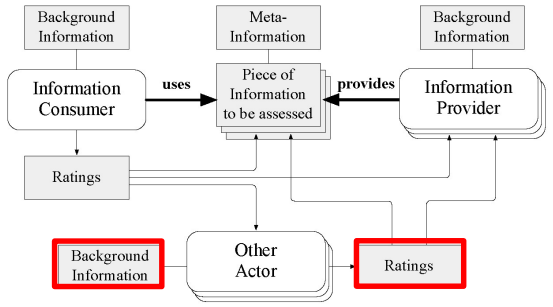




Chandola, et al.: Anomaly Detection: A Survey. ACM Computing Surveys, 2009.

Ratings

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



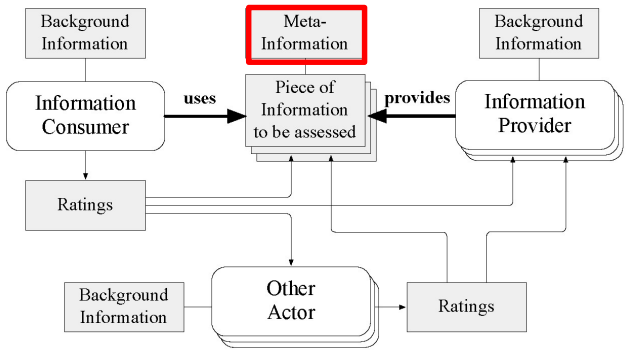
- Various scoring functions exist
  - practical systems often use simple, easily understandable functions
- Challenges:
  1. Motivate users to rate

  2. Quality of the ratings
     - fake ratings
     - clueless raters
- Events interpretable as positive ratings
  - clicks, page views
  - time spent on some page
  - hyperlinks between pages (PageRank)

4.2.2 Assessing Data Timeliness

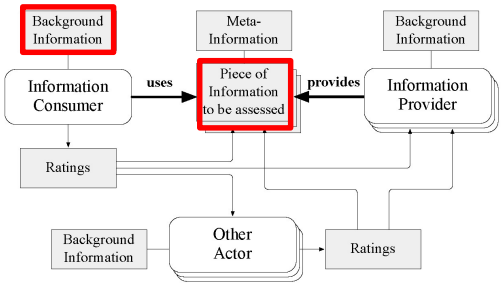
Definition Timeliness: The extent to which the age of the data is appropriate for the task at hand.

- The assessment of the timeliness of data usually requires provenance metadata.
- Provenance metadata
  - HTTP Last-Modified
  - dc:date
- Fallbacks if no timestamps are available
  - propagate timestamps to data without timestamps
    - e.g. two tables state same profit for a company, only one table has a timestamp
    - Zhang, Chakrabarti: InfoGather+, SIGMOD 2013.
  - use rules instead of timestamps
    - number of inhabitants: Prefer higher value as world population 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.

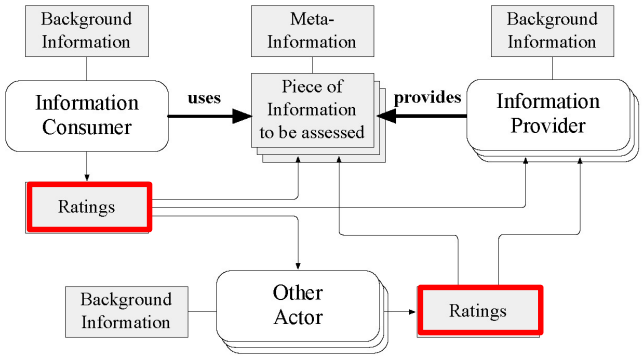


- Two perspectives on completeness:
  - Density: Fraction of attributes filled
  - Coverage: Fraction of real-world objects represented
- Assessment:
  - Density
    - sample data source and calculate density from sample
  - Coverage
    - hard to calculate as overall number of real-world objects is unknown in many cases: countries fine; products or people problematic
    - fallback: prefer data sources that describe more entities

4.2.4 Assessing Believability / Trustworthiness

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

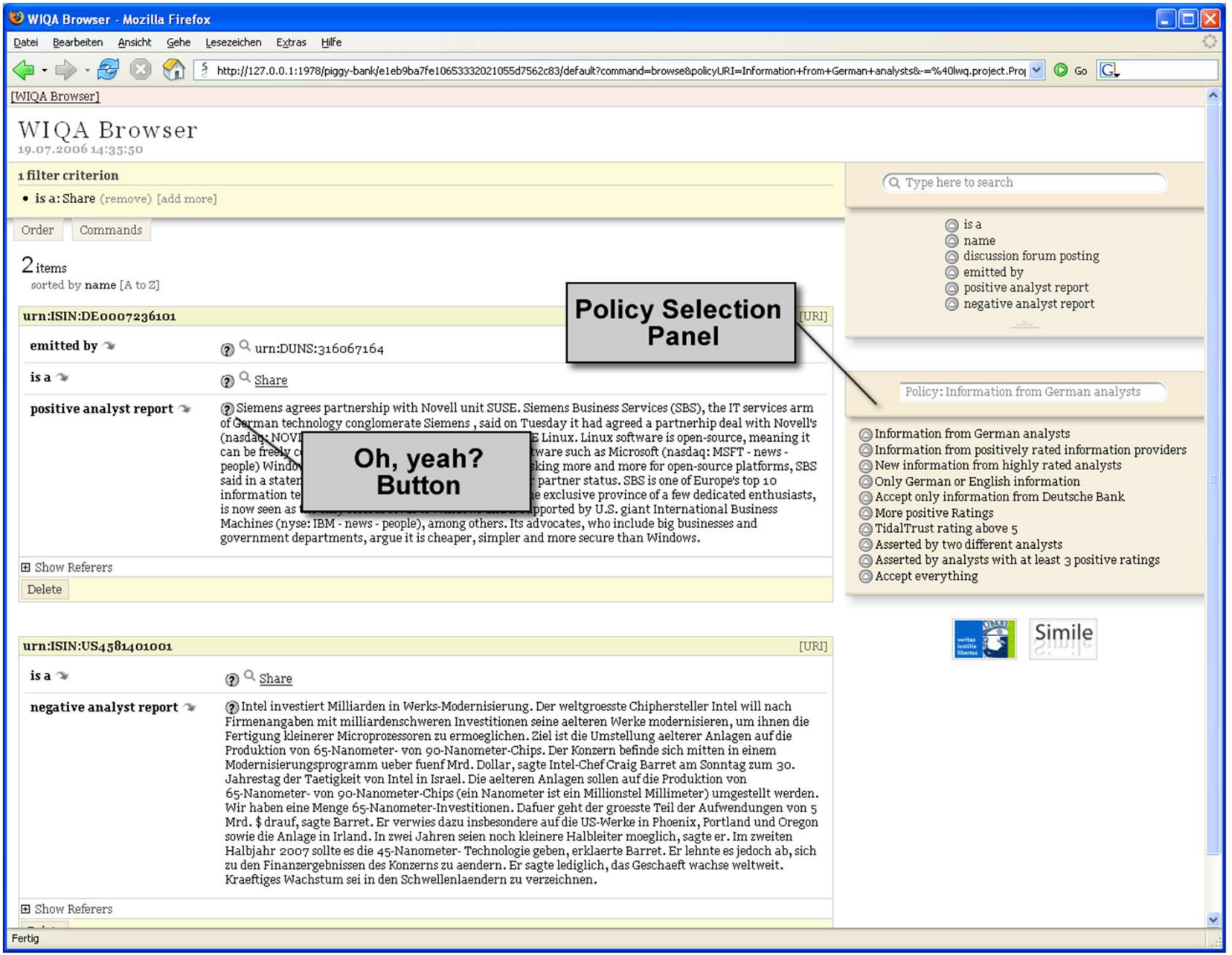
- Subjective dimension which depends on the individual user
- Assessment:
  - individual experience with the data
  - fallbacks:
    - corporate guidance about sources
    - trust networks



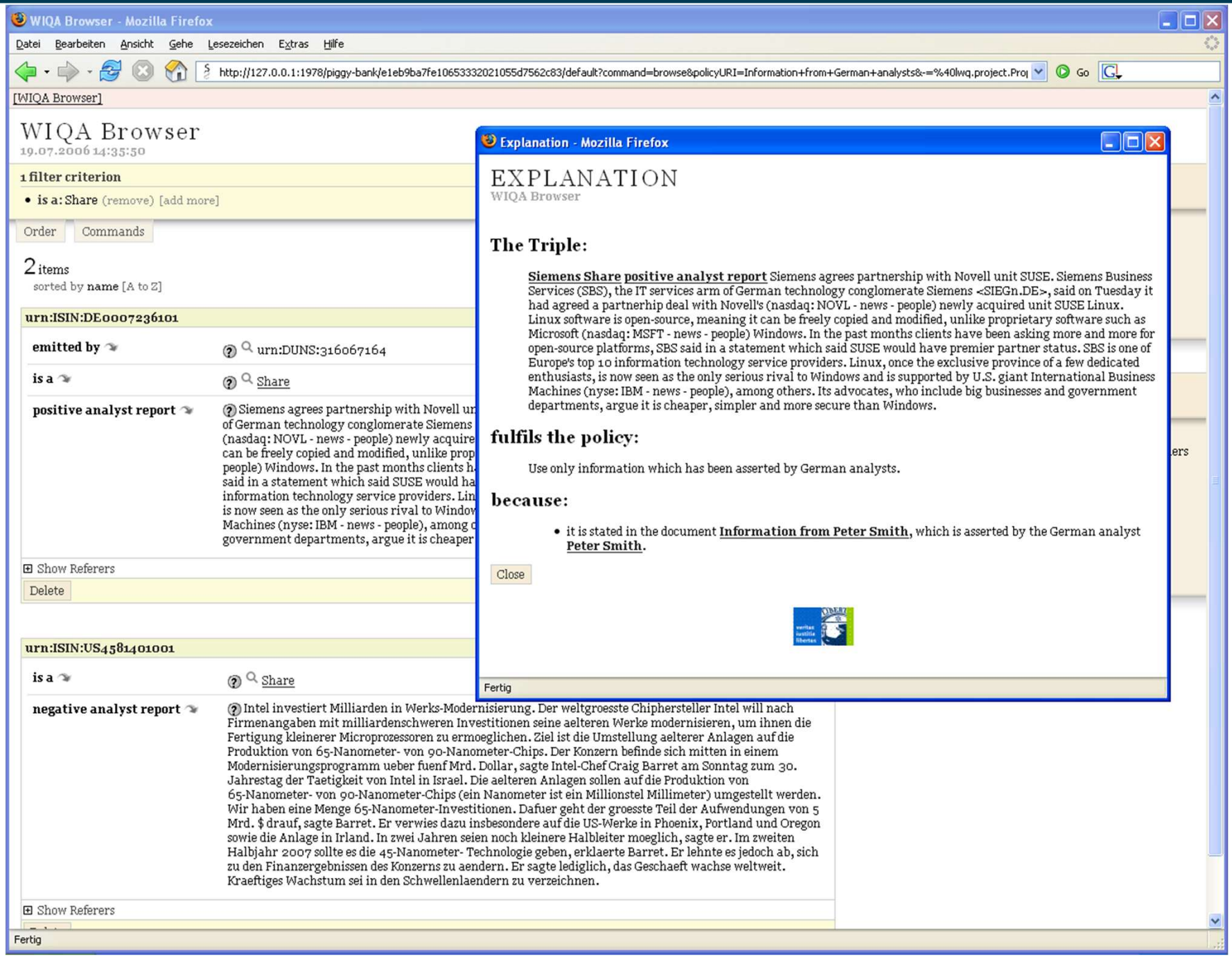
- Explanations about the data quality assessment process
  - in order to trust data, users should understand why the system regards data to be high quality
  - Tim Berners-Lee's "Oh, yeah?"-button

Prototype: The WIQA - Browser

- Enables users to employ different quality assess-ment policies
- Can explain assessment results



Explanation about an Assessment Decision



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

- Data quality assessment is essential for 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 situations, data scientists only perform shallow, ad hoc data quality assessment before selecting data for further use.
  - Read metadata and gaze over some example records
  - Reduces the effort now, but introduces risk of problems later

6. References

- Data Discovery
  - Li, Pengyue, et al.: A Survey on Open Dataset Search in the LLM Era. arXiv:2509.00728, 2025.
- 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
  - Naumann, Rolker: Assessment Methods for Information Quality Criteria. Conference on Information Quality, 2000.
  - Rekatsinas, et al.: HoloClean: Holistic Data Repairs with Probabilistic Inference. VLDB 2017.
  - Chandola, et al.: Anomaly Detection: A Survey. ACM Computing Surveys, 2009.