Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Data Quality in Databases

OpEN.SC Symposium 8.5.2009

Felix Naumann
Hasso-Plattner-Institut
Fachgebiet Informationssysteme

The HPI – Hasso Plattner Institut

- Founded in 1998 as a Public Private Partnership
- Hasso Plattner, co-founder of SAP, endowed over 200 Mio. Euro.
- Adjoined with the University of Potsdam
  - Capital of Brandenburg, bordering Berlin
- 400 students – Bachelor, Master, and PhD

Information systems team

Prof. Felix **Naumann**

Katrin **Heinrich**

Jens **Bleiholder**

project **fusem**

project **ViQTOR**

Paul **Führing**

DQ Annotation & Assessment

Data Fusion

project **HumMer**

Christoph Böhm

Data Profiling & Cleaning

Information Integration

Information Quality

Armin **Roth**

project **System P**

Peer Data Management Systems

Matching

Data Integration for Life Science Data Sources

project **Aladin**

Alexander **Albrecht**

ETL Management

Service-Oriented Systems

Ontologies, Profiling

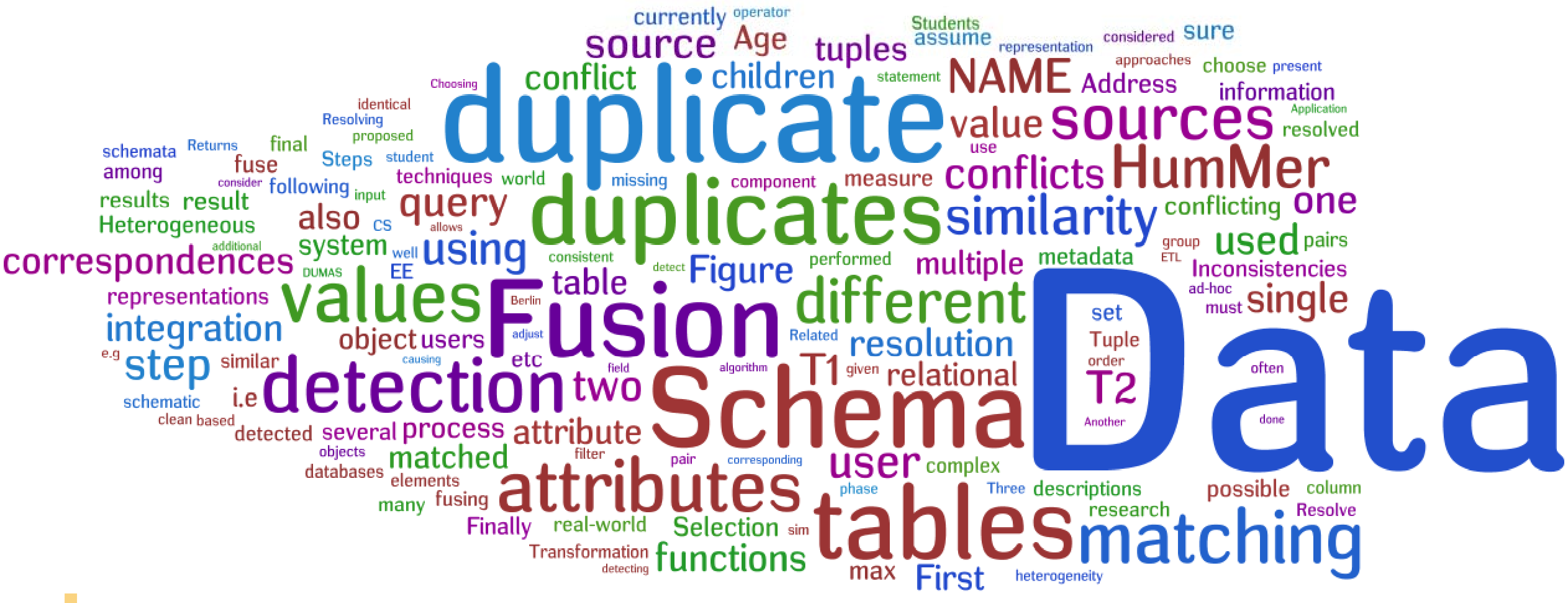
Mohammed **AbuJarour**

Frank **Kaufer**

Jana **Bauckmann**

Data Profiling for Schema Management

“Data Fusion in Three Steps” DE Bulletin, 2006



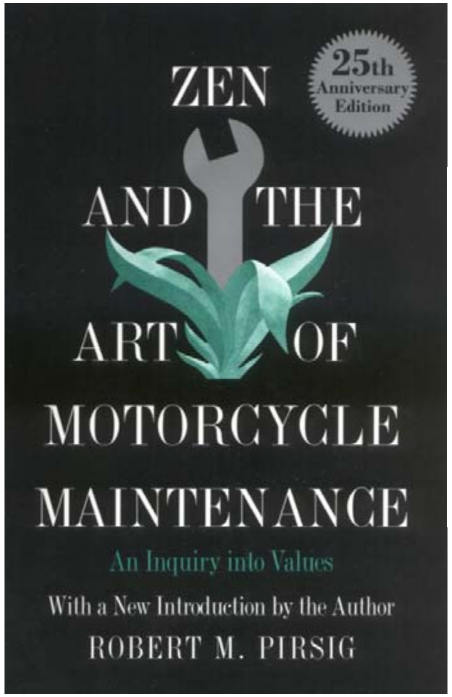
Wordle.net

Overview

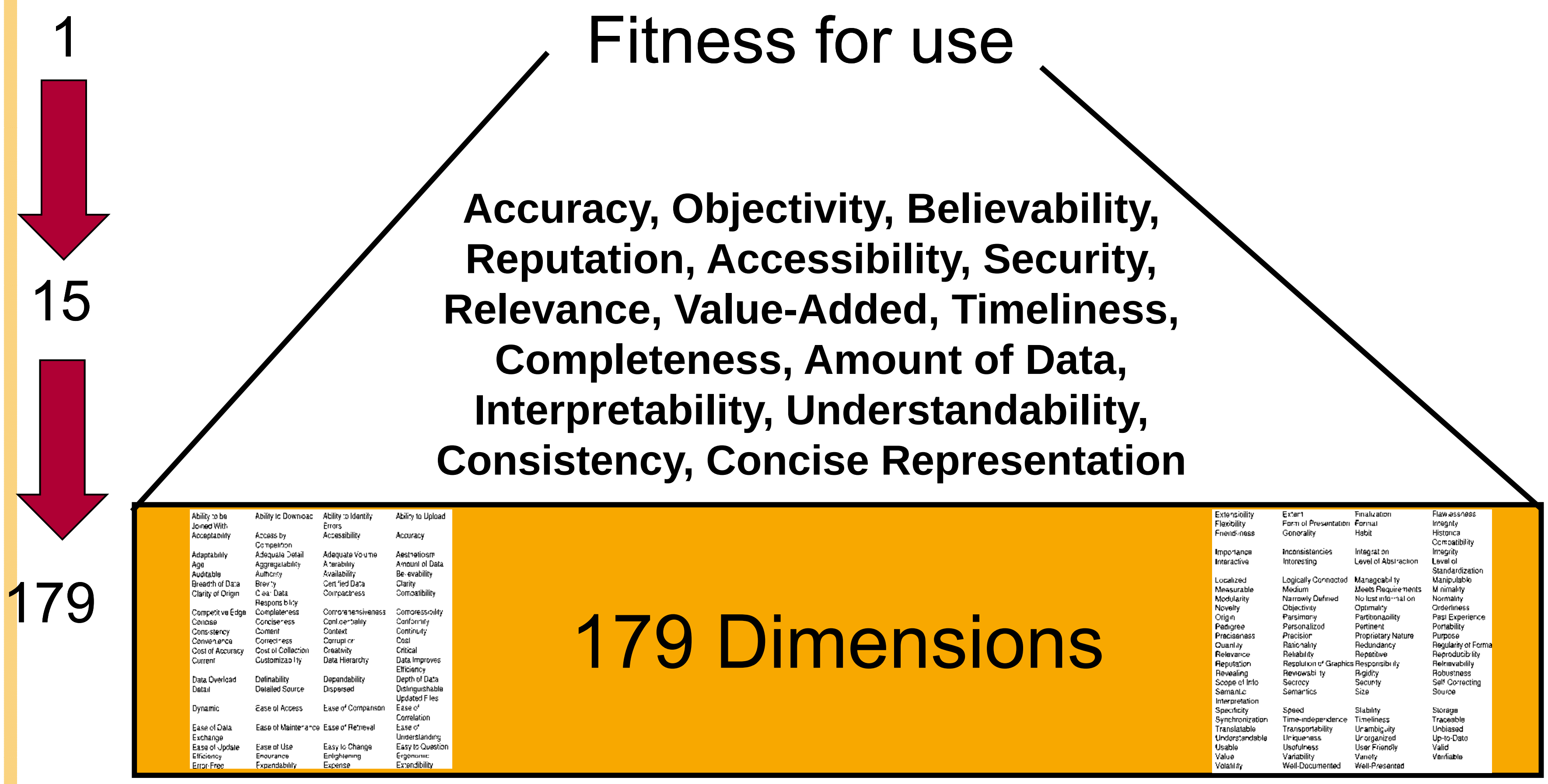
- Information Quality
- Step 1: Schema Matching
- Step 2: Duplicate detection
- Step 3: Data fusion
- Summary

Quality

> “*Even though quality cannot be defined, you know what it is.*”
>
> **Robert Pirsig**


ZEN AND THE ART OF MOTORCYCLE MAINTENANCE
25th Anniversary Edition
An Inquiry into Values
With a New Introduction by the Author
ROBERT M. PIRSIG

Zooming into Information Quality



Fitness for use

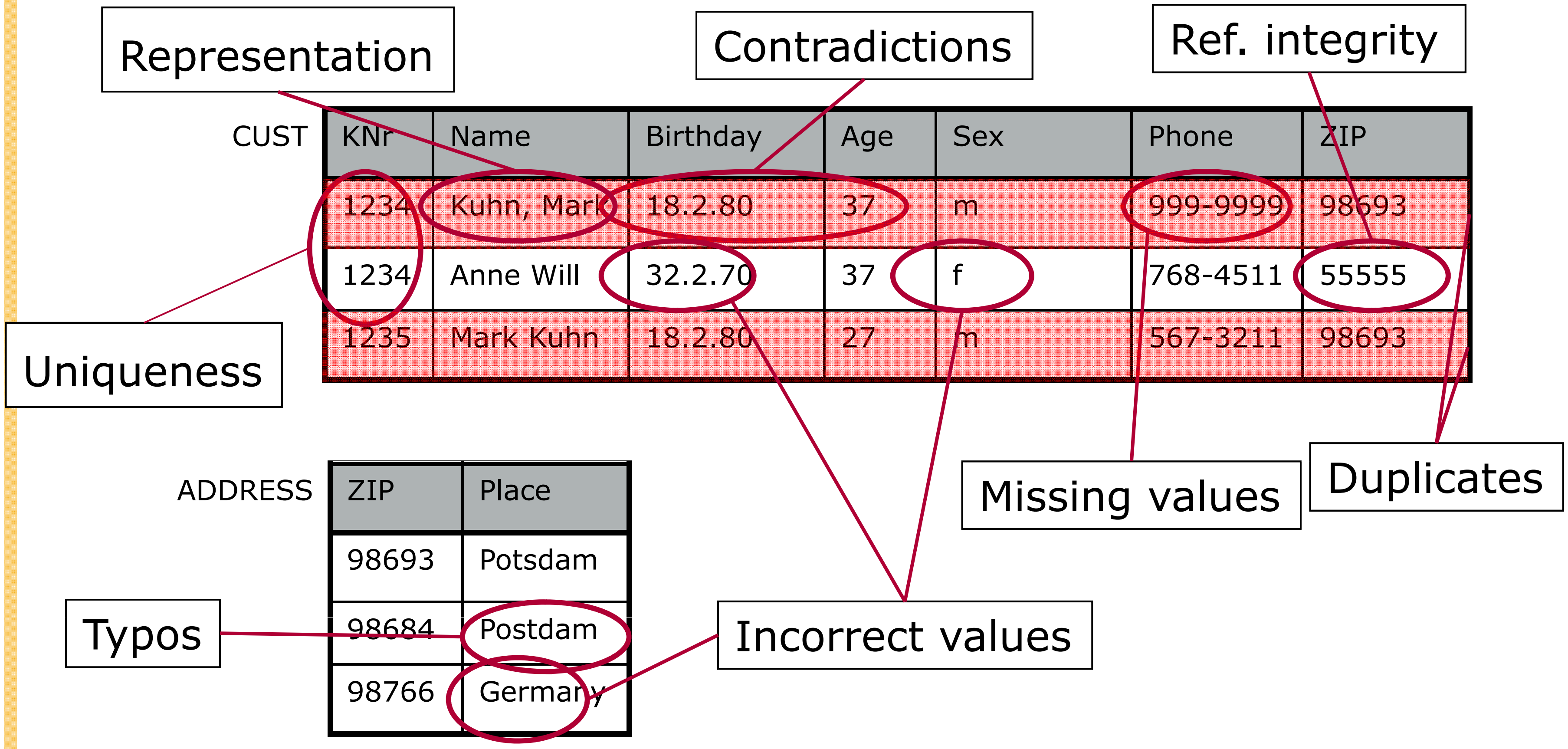
Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Understandability, Consistency, Concise Representation

179 Dimensions

Ability to be Joined With	Ability to Download	Ability to Identify Errors	Ability to Upload
Acceptability	Access by Competition	Accessibility	Accuracy
Adaptability	Adequate Detail	Adequate Volume	Aestheticism
Age	Aggregatability	Alterability	Amount of Data
Auditable	Authority	Availability	Believability
Breadth of Data	Brevity	Certified Data	Clarity
Clarity of Origin	Clear Data Responsibility	Compactness	Compatibility
Competitive Edge	Completeness	Comprehensiveness	Compressibility
Concise	Conciseness	Confidentiality	Conformity
Consistency	Content	Context	Continuity
Convenience	Correctness	Corruption	Cost
Cost of Accuracy	Cost of Collection	Creativity	Critical
Current	Customizability	Data Hierarchy	Data Improves Efficiency
Data Overload	Definability	Dependability	Depth of Data
Detail	Detailed Source	Dispersed	Distinguishable Updated Files
Dynamic	Ease of Access	Ease of Comparison	Ease of Correlation
Ease of Data Exchange	Ease of Maintenance	Ease of Retrieval	Ease of Understanding
Ease of Update	Ease of Use	Easy to Change	Easy to Question
Efficiency	Endurance	Enlightening	Ergonomic
Error-Free	Expandability	Expense	Extendibility

Extensibility	Extent	Finalization	Flawlessness
Flexibility	Form of Presentation	Format	Integrity
Friendliness	Generality	Habit	Historical Compatibility
Importance	Inconsistencies	Integration	Integrity
Interactive	Interesting	Level of Abstraction	Level of Standardization
Localized	Logically Connected	Manageability	Manipulable
Measurable	Medium	Meets Requirements	Minimality
Modularity	Narrowly Defined	No lost information	Normality
Novelty	Objectivity	Optimality	Orderliness
Origin	Parsimony	Partitionability	Past Experience
Pedigree	Personalized	Pertinent	Portability
Preciseness	Precision	Proprietary Nature	Purpose
Quantity	Rationality	Redundancy	Regularity of Forma
Relevance	Reliability	Repetitive	Reproducibility
Reputation	Resolution of Graphics	Responsibility	Retrievability
Revealing	Reviewability	Rigidity	Robustness
Scope of Info	Secrecy	Security	Self Correcting
Semantic Interpretation	Semantics	Size	Source
Specificity	Speed	Stability	Storage
Synchronization	Time-independence	Timeliness	Traceable
Translatable	Transportability	Unambiguity	Unbiased
Understandable	Uniqueness	Unorganized	Up-to-Date
Usable	Usefulness	User Friendly	Valid
Value	Variability	Variety	Verifiable
Volatility	Well-Documented	Well-Presented	

Data Quality: Problems



CUST

KNr	Name	Birthday	Age	Sex	Phone	ZIP
1234	Kuhn, Mark	18.2.80	37	m	999-9999	98693
1234	Anne Will	32.2.70	37	f	768-4511	55555
1235	Mark Kuhn	18.2.80	27	m	567-3211	98693

ADDRESS

ZIP	Place
98693	Potsdam
98684	Postdam
98766	Germany

Representation

Contradictions

Ref. integrity

Uniqueness

Missing values

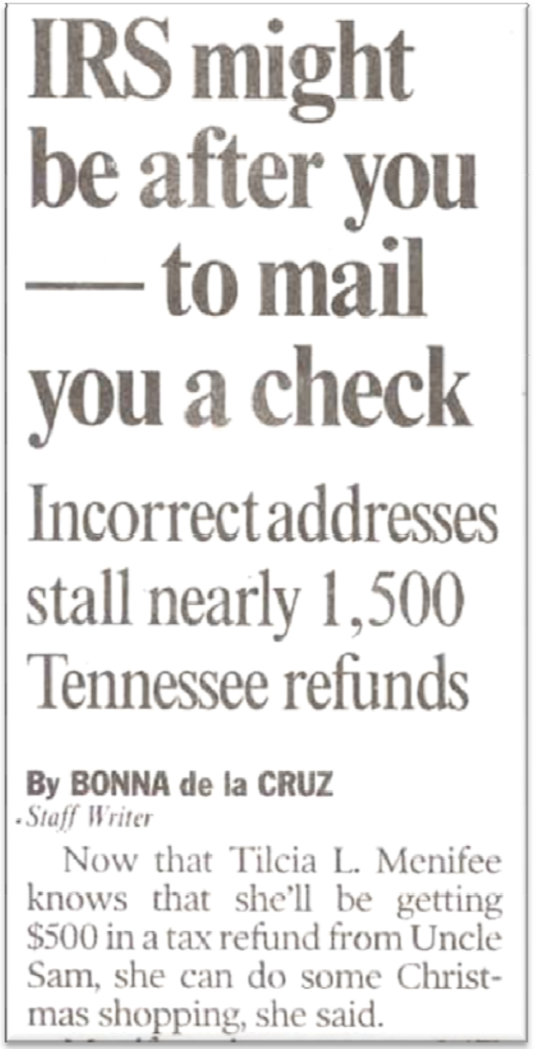
Duplicates

Typos

Incorrect values

DQ-Problems: Effects

- Fehlerhafte Warenpreise in Artikel-DB des US-Einzelhandels [English 1999]
  - □ Kosten für Konsumenten 2.5 Mrd $
  - □ 80% der Barcode-Scan-Fehler zulasten der Konsumenten
- US-Finanzbehörde 1992: knapp 100.000 Steuererstattungsbescheide unzustellbar [English 1999]
- 50-80% der Einträge im US-Vorstrafenregister ungenau, unvollständig oder fehlerhaft [Strong et al. 1997a]
- US-Post: von 100.000 Massen-Postsendungen bis zu 7.000 aufgrund von Adressfehlern unzustellbar [Pierce 2004]



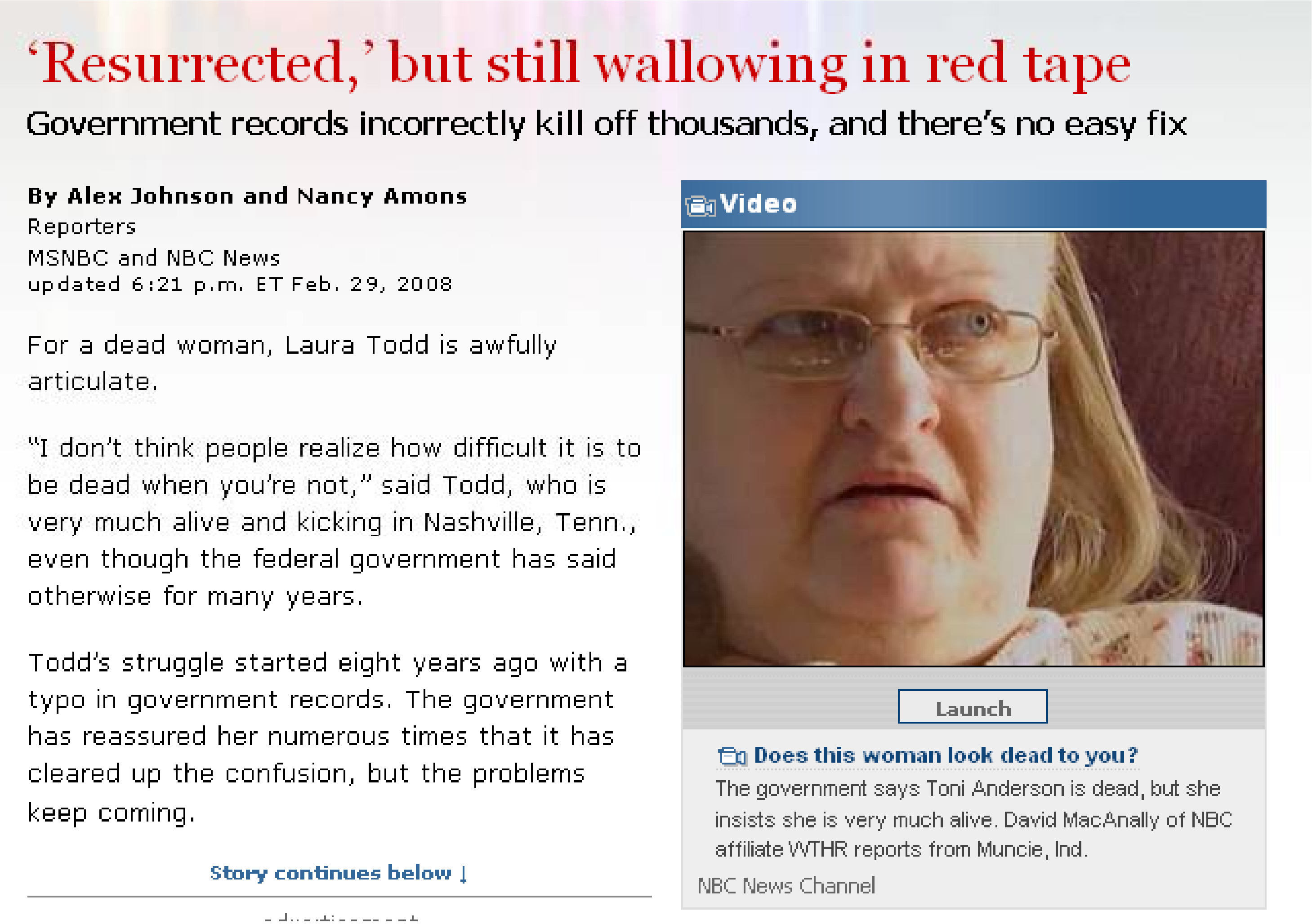
IRS might be after you — to mail you a check

Incorrect addresses stall nearly 1,500 Tennessee refunds

By BONNA de la CRUZ
Staff Writer

Now that Tilcia L. Menifee knows that she'll be getting $500 in a tax refund from Uncle Sam, she can do some Christmas shopping, she said.

Death by Typo



‘Resurrected,’ but still wallowing in red tape

Government records incorrectly kill off thousands, and there’s no easy fix

By Alex Johnson and Nancy Amons
Reporters
MSNBC and NBC News
updated 6:21 p.m. ET Feb. 29, 2008

For a dead woman, Laura Todd is awfully articulate.

“I don’t think people realize how difficult it is to be dead when you’re not,” said Todd, who is very much alive and kicking in Nashville, Tenn., even though the federal government has said otherwise for many years.

Todd’s struggle started eight years ago with a typo in government records. The government has reassured her numerous times that it has cleared up the confusion, but the problems keep coming.

Story continues below ↓

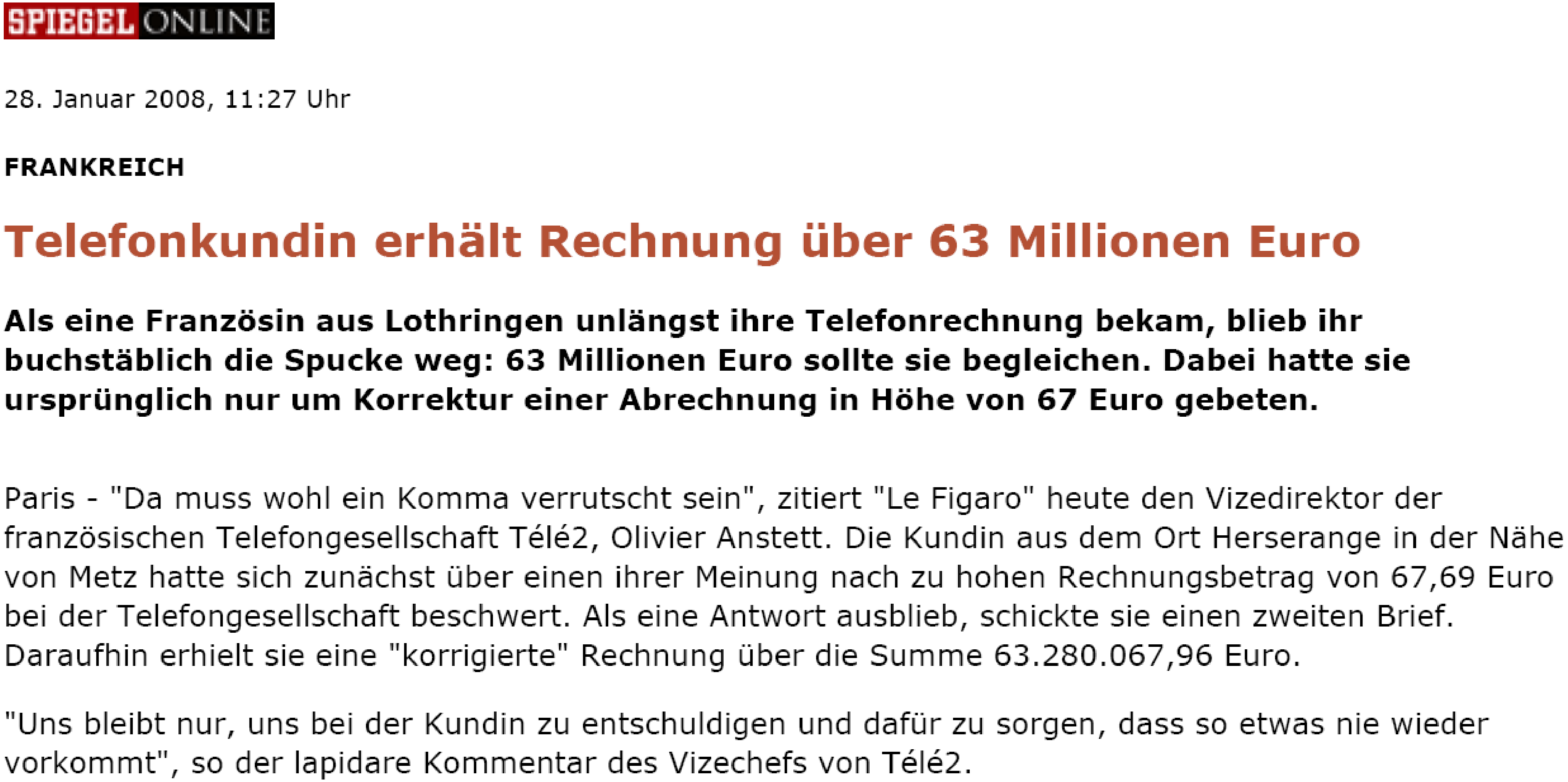
advertisement

Does this woman look dead to you?

The government says Toni Anderson is dead, but she insists she is very much alive. David MacAnally of NBC affiliate WTHR reports from Muncie, Ind.

NBC News Channel

Decimals



SPIEGEL ONLINE

28. Januar 2008, 11:27 Uhr

FRANKREICH

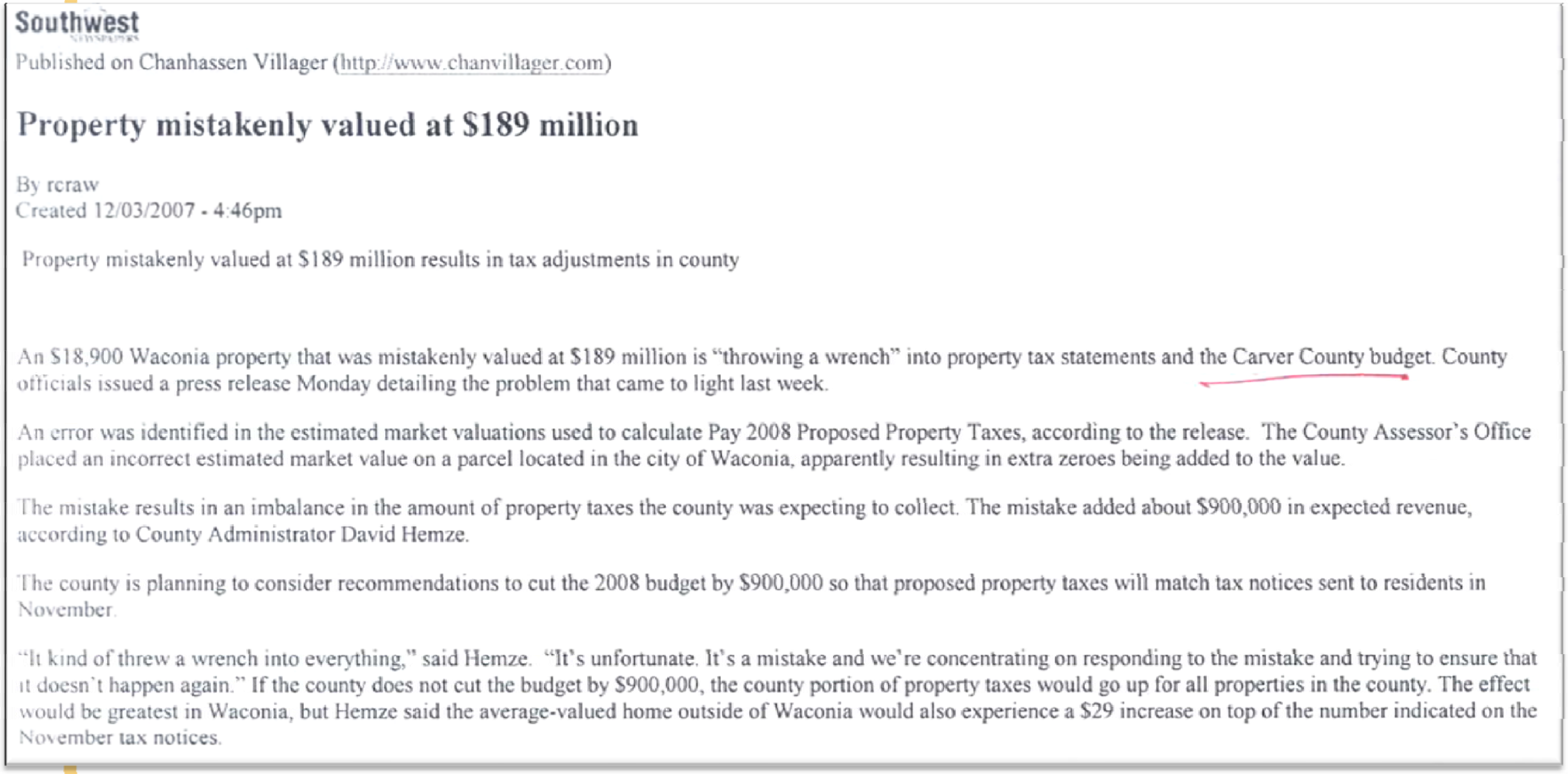
Telefonkundin erhält Rechnung über 63 Millionen Euro

Als eine Französin aus Lothringen unlängst ihre Telefonrechnung bekam, blieb ihr buchstäblich die Spucke weg: 63 Millionen Euro sollte sie begleichen. Dabei hatte sie ursprünglich nur um Korrektur einer Abrechnung in Höhe von 67 Euro gebeten.

Paris - "Da muss wohl ein Komma verrutscht sein", zitiert "Le Figaro" heute den Vizedirektor der französischen Telefongesellschaft Télé2, Olivier Anstett. Die Kundin aus dem Ort Herserange in der Nähe von Metz hatte sich zunächst über einen ihrer Meinung nach zu hohen Rechnungsbetrag von 67,69 Euro bei der Telefongesellschaft beschwert. Als eine Antwort ausblieb, schickte sie einen zweiten Brief. Daraufhin erhielt sie eine "korrigierte" Rechnung über die Summe 63.280.067,96 Euro.

"Uns bleibt nur, uns bei der Kundin zu entschuldigen und dafür zu sorgen, dass so etwas nie wieder vorkommt", so der lapidare Kommentar des Vizechefs von Télé2.

Common Sense



Southwest
NEWSPAPERS
Published on Chanhassen Villager (http://www.chanvillager.com)

Property mistakenly valued at $189 million

By rcraw
Created 12/03/2007 - 4:46pm

Property mistakenly valued at $189 million results in tax adjustments in county

An $18,900 Waconia property that was mistakenly valued at $189 million is "throwing a wrench" into property tax statements and the Carver County budget. County officials issued a press release Monday detailing the problem that came to light last week.

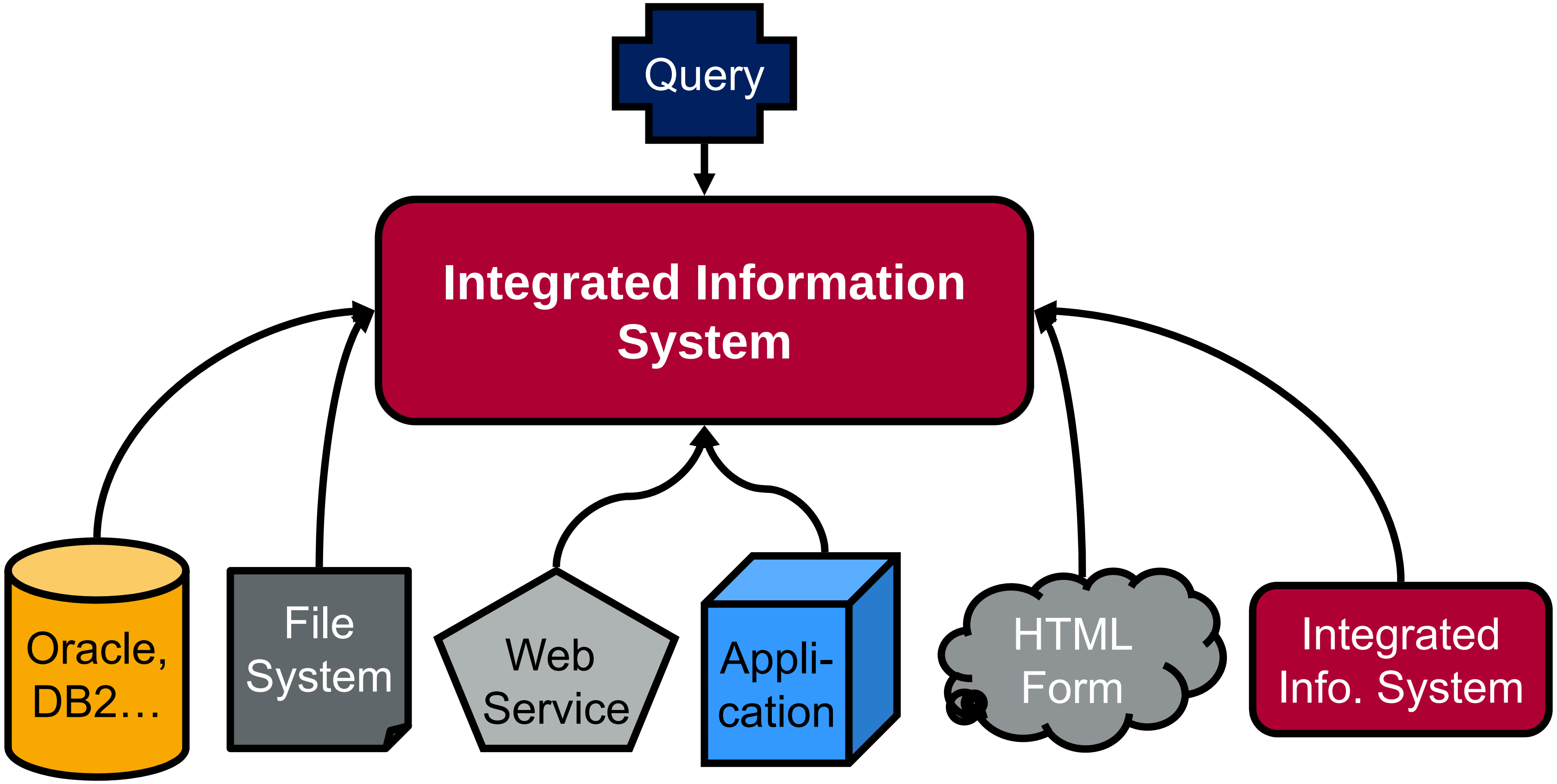
An error was identified in the estimated market valuations used to calculate Pay 2008 Proposed Property Taxes, according to the release. The County Assessor's Office placed an incorrect estimated market value on a parcel located in the city of Waconia, apparently resulting in extra zeroes being added to the value.

The mistake results in an imbalance in the amount of property taxes the county was expecting to collect. The mistake added about $900,000 in expected revenue, according to County Administrator David Hemze.

The county is planning to consider recommendations to cut the 2008 budget by $900,000 so that proposed property taxes will match tax notices sent to residents in November.

"It kind of threw a wrench into everything," said Hemze. "It's unfortunate. It's a mistake and we're concentrating on responding to the mistake and trying to ensure that it doesn't happen again." If the county does not cut the budget by $900,000, the county portion of property taxes would go up for all properties in the county. The effect would be greatest in Waconia, but Hemze said the average-valued home outside of Waconia would also experience a $29 increase on top of the number indicated on the November tax notices.

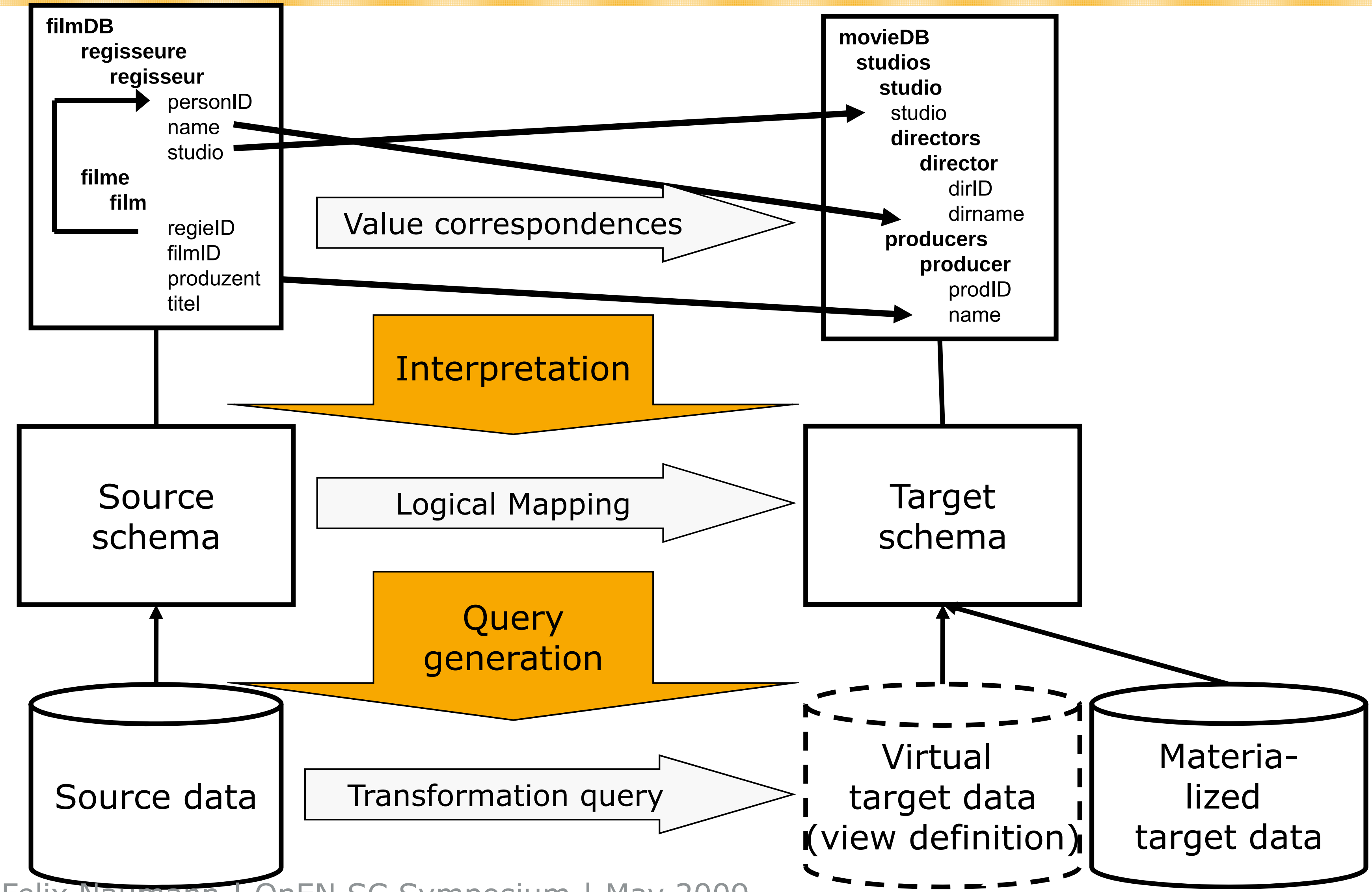
Integrated Information Systems



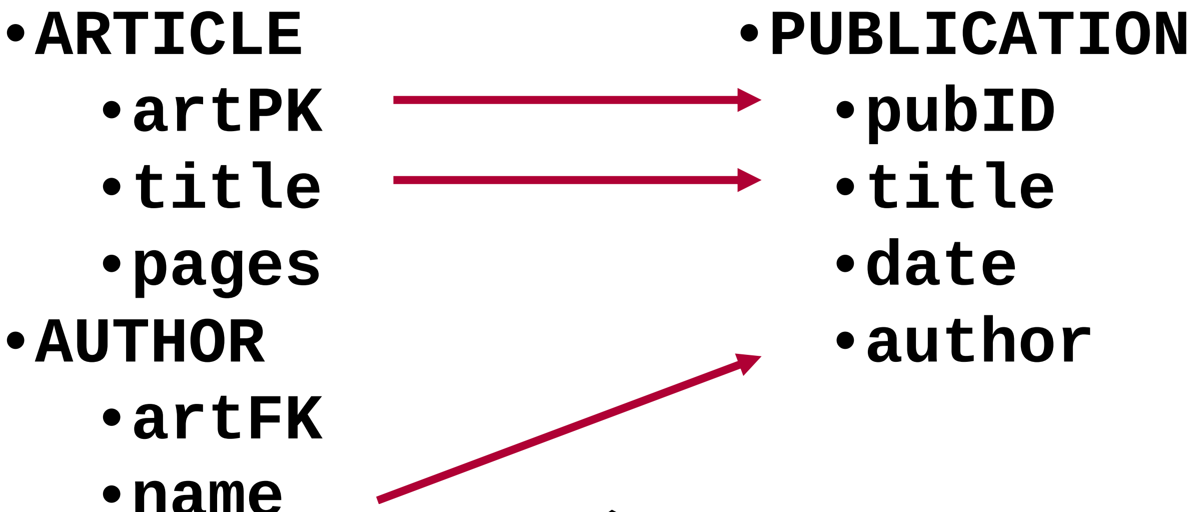
Overview

- Information Quality
- Step 1: Schema Matching
- Step 2: Duplicate detection
- Step 3: Data fusion
- Summary

Schema Mapping in Context



Schema Mapping Example

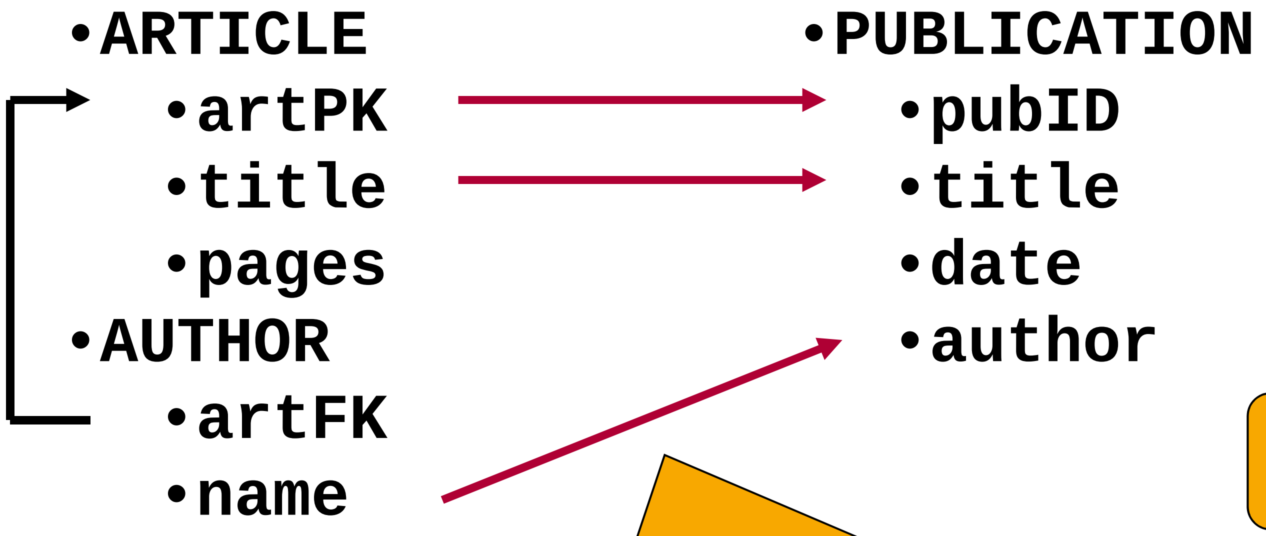


- ARTICLE
  - artPK
  - title
  - pages
- AUTHOR
  - artFK
  - name

- PUBLICATION
  - pubID
  - title
  - date
  - author

```
SELECT artPK AS pubID   UNION  SELECT null AS pubID
       title AS title                 null AS title
       null AS date                   null AS date
       null AS author                 name AS author
FROM   ARTICLE                 FROM   AUTHOR
```

Schema Mapping Example



- ARTICLE
  - artPK
  - title
  - pages
- AUTHOR
  - artFK
  - name

- PUBLICATION
  - pubID
  - title
  - date
  - author

```
SELECT      artPK AS pubID
            title AS title
            null AS date
            name AS author
FROM        ARTICLE, AUTHOR
WHERE       ARTICLE.artPK = AUTHOR.artFK
```

Further interpretations?

Schema Matching – Motivation

Schemata are

- large
- complex
- foreign
- confusing
- different language
- cryptic

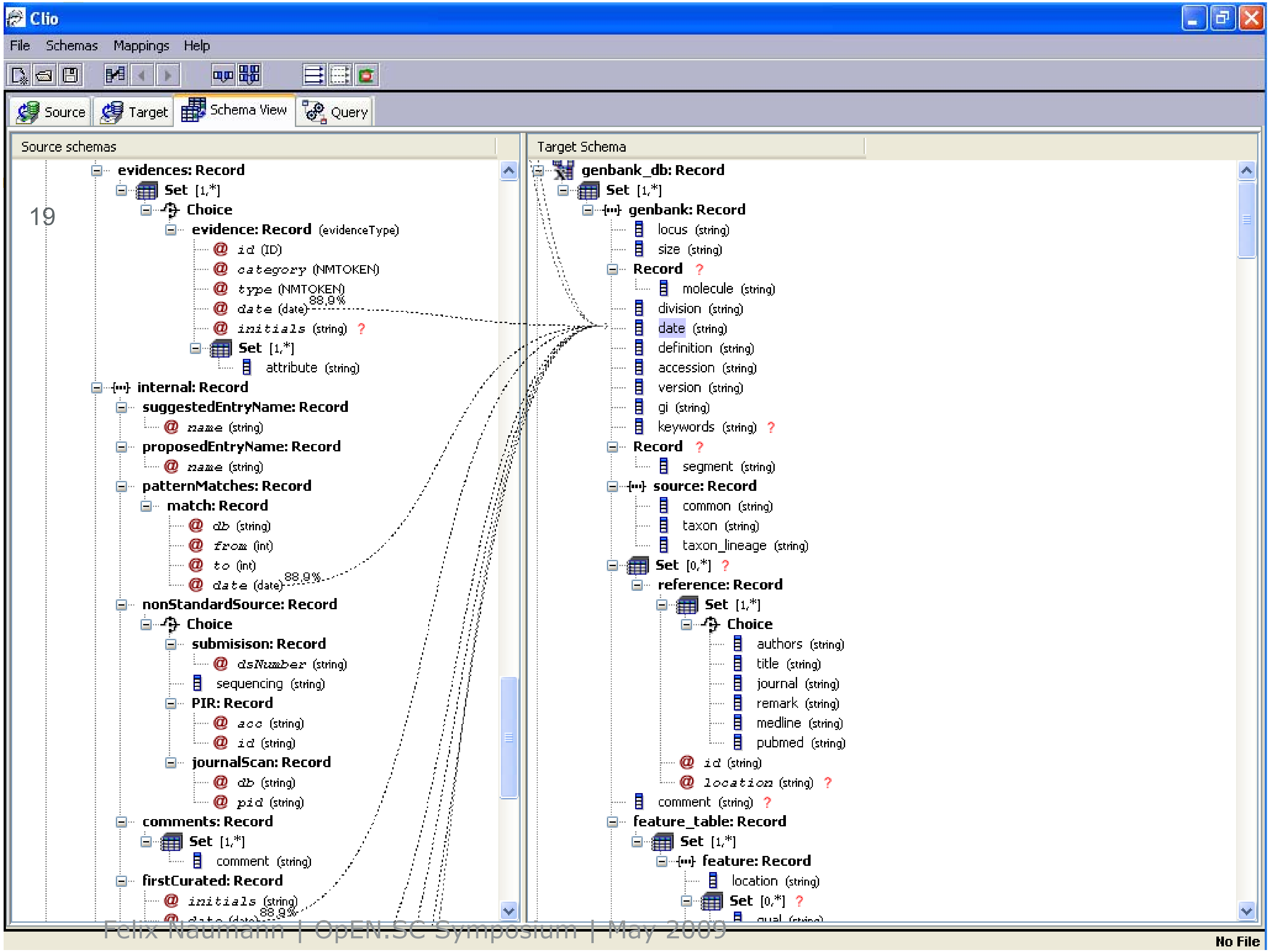
> 100 tables, many attributes

Deep Nesting
Foreign keys
XML Schema

Unknown synonyms

Unknown homonyms

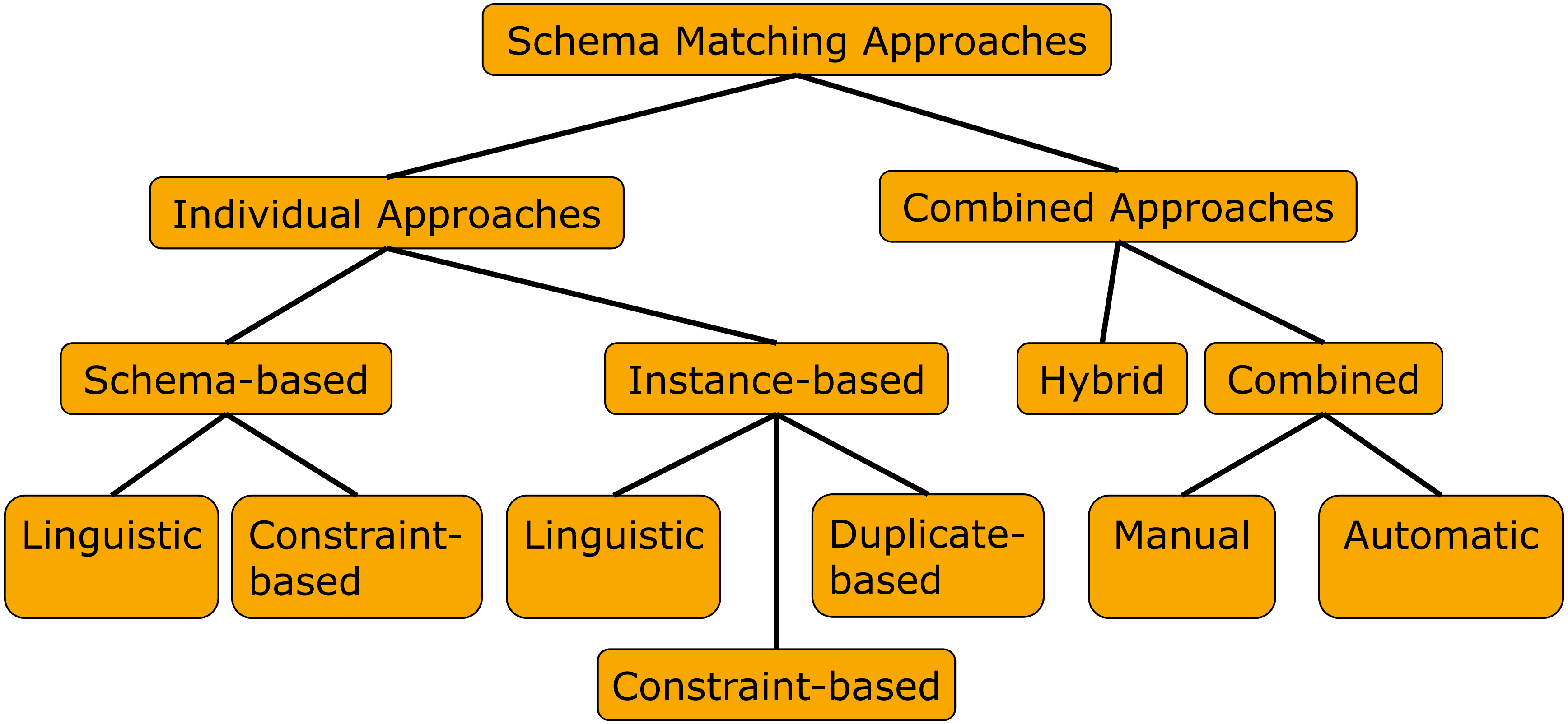
|attribute name| ≤ 8
|table name| ≤ 8



```
Source schemas
evidences: Record
  Set [1,*]
    Choice
      evidence: Record (evidenceType)
        @ id (ID)
        @ category (NMTOKEN)
        @ type (NMTOKEN)
        @ date (date)
        @ initials (string) ?
        Set [1,*]
          attribute (string)
internal: Record
  suggestedEntryName: Record
    @ name (string)
  proposedEntryName: Record
    @ name (string)
  patternMatches: Record
    match: Record
      @ db (string)
      @ from (int)
      @ to (int)
      @ date (date)
  nonStandardSource: Record
    Choice
      submisison: Record
        @ dsNumber (string)
        sequencing (string)
      PIR: Record
        @ acc (string)
        @ id (string)
      journalScan: Record
        @ db (string)
        @ pid (string)
  comments: Record
    Set [1,*]
      comment (string)
  firstCurated: Record
    @ initials (string)
    @ date (date)
```

```
Target Schema
genbank_db: Record
  Set [1,*]
    genbank: Record
      locus (string)
      size (string)
      Record ?
        molecule (string)
      division (string)
      date (string)
      definition (string)
      accession (string)
      version (string)
      gi (string)
      keywords (string) ?
      Record ?
        segment (string)
      source: Record
        common (string)
        taxon (string)
        taxon_lineage (string)
      Set [0,*] ?
        reference: Record
          Set [1,*]
            Choice
              authors (string)
              title (string)
              journal (string)
              remark (string)
              medline (string)
              pubmed (string)
          @ id (string)
          @ location (string) ?
      comment (string) ?
      feature_table: Record
        Set [1,*]
          feature: Record
            location (string)
            Set [0,*] ?
              qual (string)
```

Schema Matching Classification [RB01]



Overview

- Information Quality
- Step 1: Schema Matching
- Step 2: Duplicate detection
- Step 3: Data fusion
- Summary

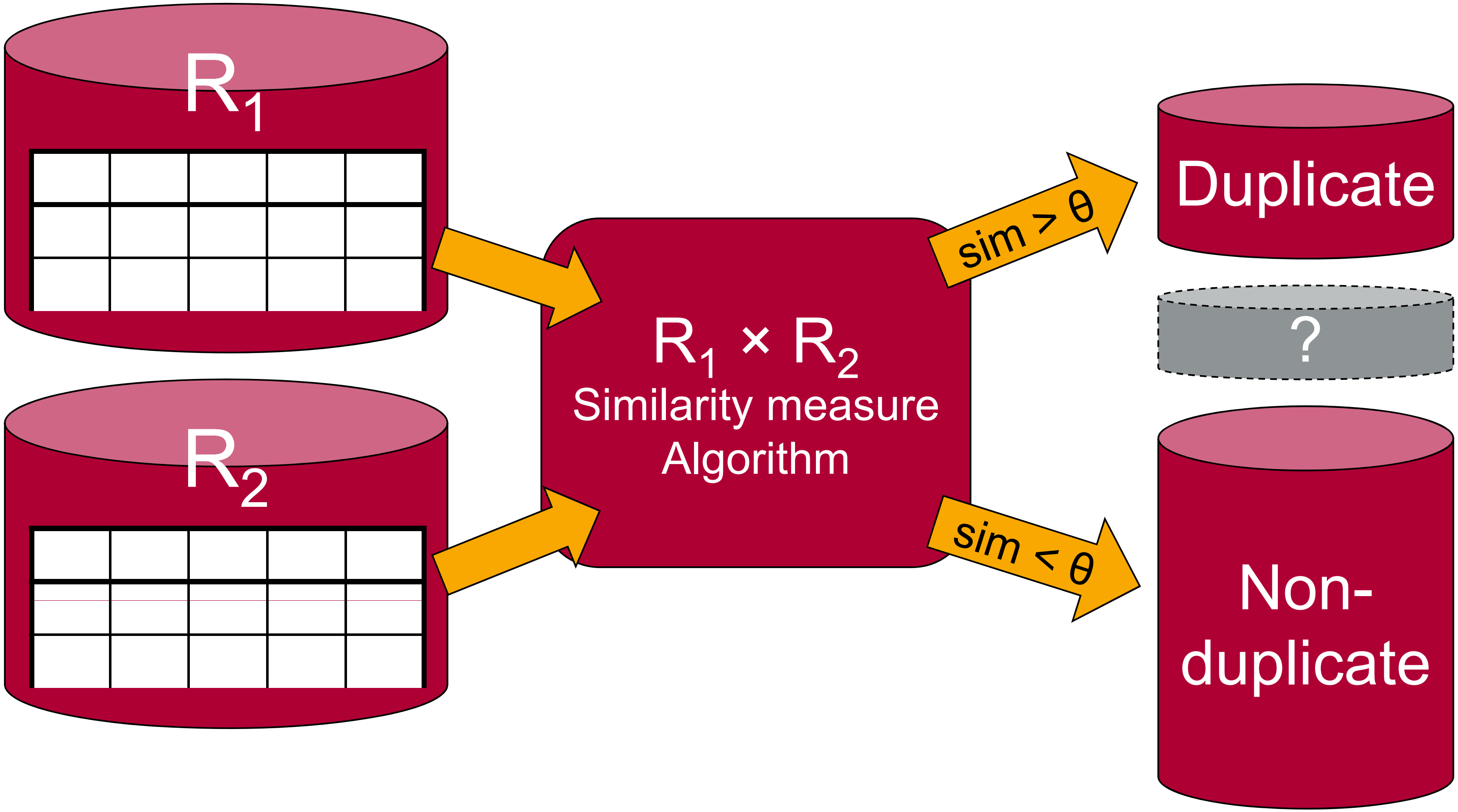
Duplicate Detection

Duplicate detection is the discovery of multiple representations of the same real-world object.

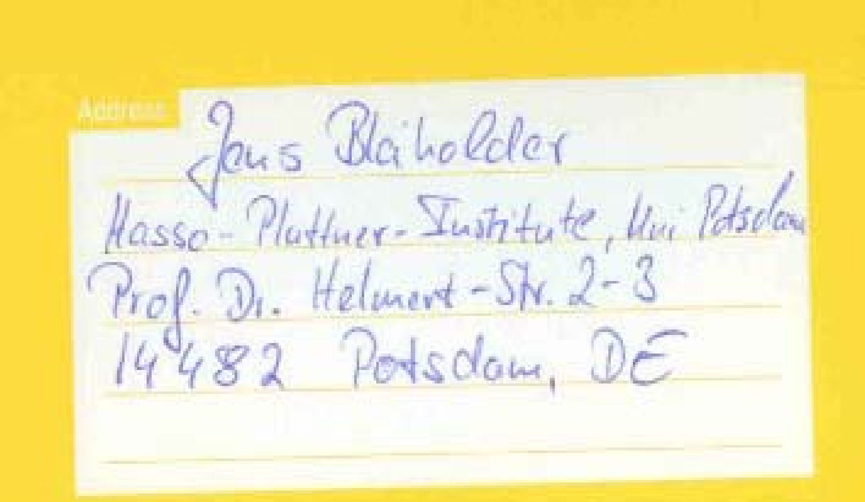
- Problem 1: Representations are not identical.
  - *Fuzzy duplicates*
- Solution: Similarity measures
  - Value- and record-comparisons
  - Domain-dependent or domain-independent

- Problem 2: Data sets are large.
  - Quadratic complexity: Comparison of every pair of records.
- Solution: Algorithms
  - E.g., avoid comparisons by partitioning.

Duplicate Detection



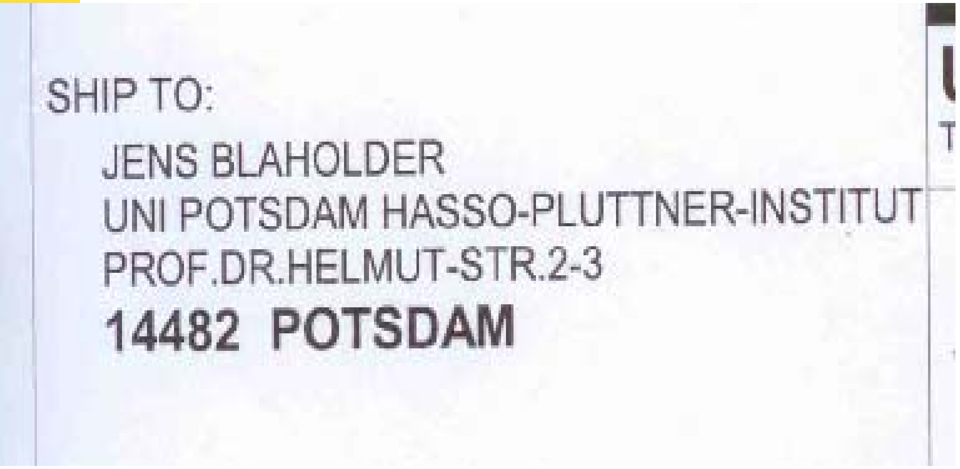
Origins of duplicates



Address

Jens Blaholder
Hasso-Plattner-Institute, Uni Potsdam
Prof. Dr. Helmert-Str. 2-3
14482 Potsdam, DE

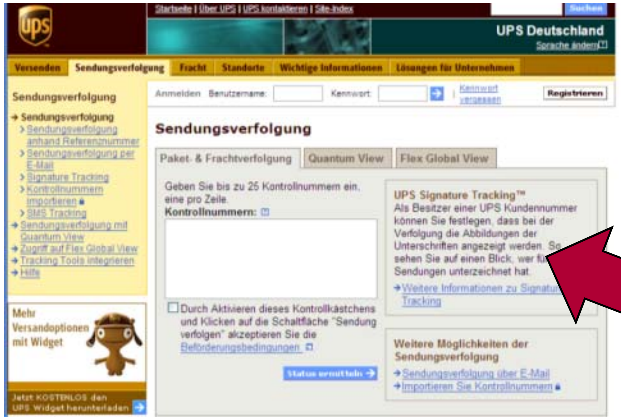
Original



SHIP TO:
JENS BLAHOLDER
UNI POTSDAM HASSO-PLUTTNER-INSTITUT
PROF.DR.HELMUT-STR.2-3
14482 POTSDAM

Scanned

Origins of duplicates



Startseite | Über UPS | UPS kontaktieren | Site-Index

UPS Deutschland

Sprache ändern

Sendungsverfolgung

- Sendungsverfolgung
  - Sendungsverfolgung anhand Referenznummer
  - Sendungsverfolgung per E-Mail
  - Signature Tracking
  - Kontrollnummern importieren
  - SMS Tracking
- Sendungsverfolgung mit Quantum View
- Zugriff auf Flex Global View
- Tracking Tools integrieren
- Hilfe

Mehr Versandoptionen mit Widget

Jetzt KOSTENLOS den UPS Widget herunterladen

Anmelden Benutzername: Kennwort: | Kennwort vergessen

Sendungsverfolgung

Geben Sie bis zu 25 Kontrollnummern ein, eine pro Zeile.

Kontrollnummern:

□ Durch Aktivieren dieses Kontrollkästchens und Klicken auf die Schaltfläche "Sendung verfolgen" akzeptieren Sie die Beförderungsbedingungen

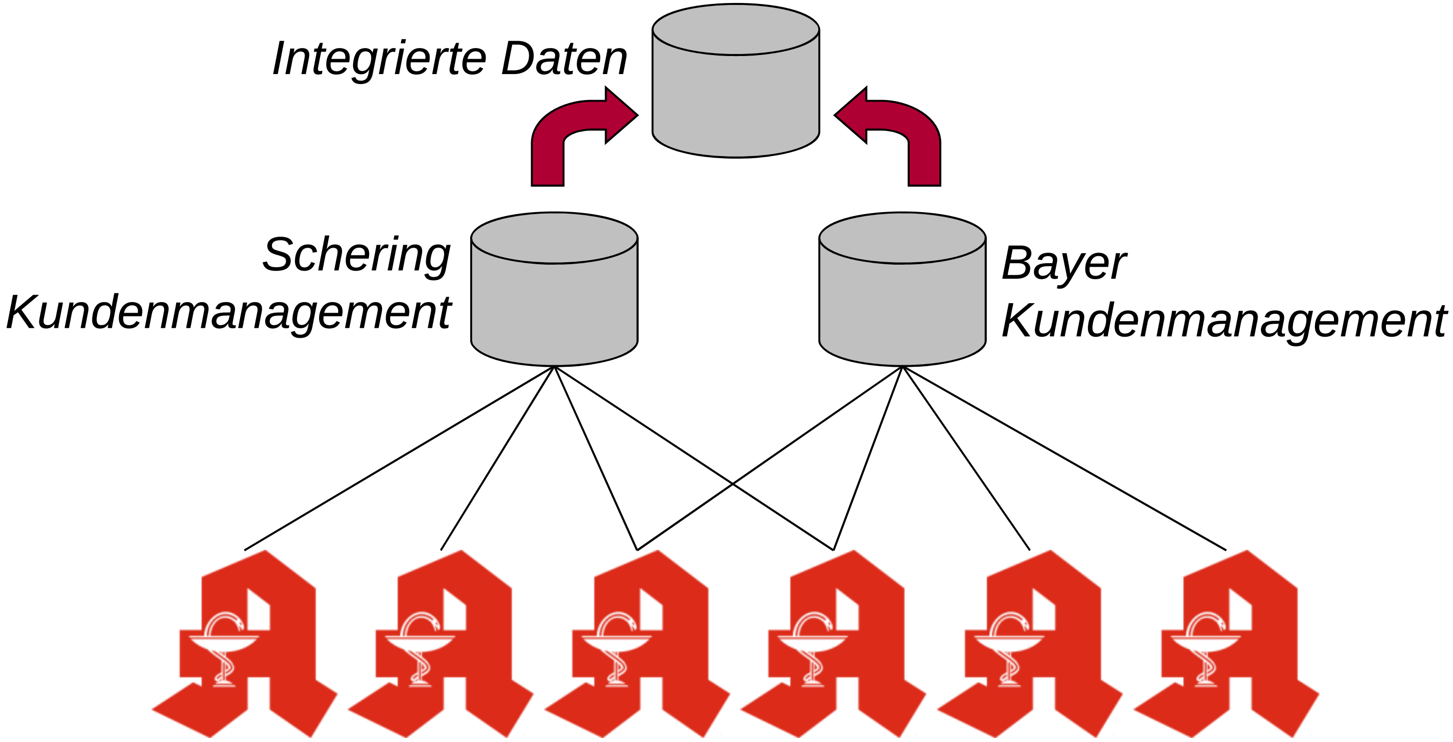
UPS Signature Tracking™
Als Besitzer einer UPS Kundennummer können Sie festlegen, dass bei der Verfolgung die Abbildungen der Unterschriften angezeigt werden. So sehen Sie auf einen Blick, wer fü Sendungen unterzeichnet hat.

→ Weitere Informationen zu Signatur Tracking

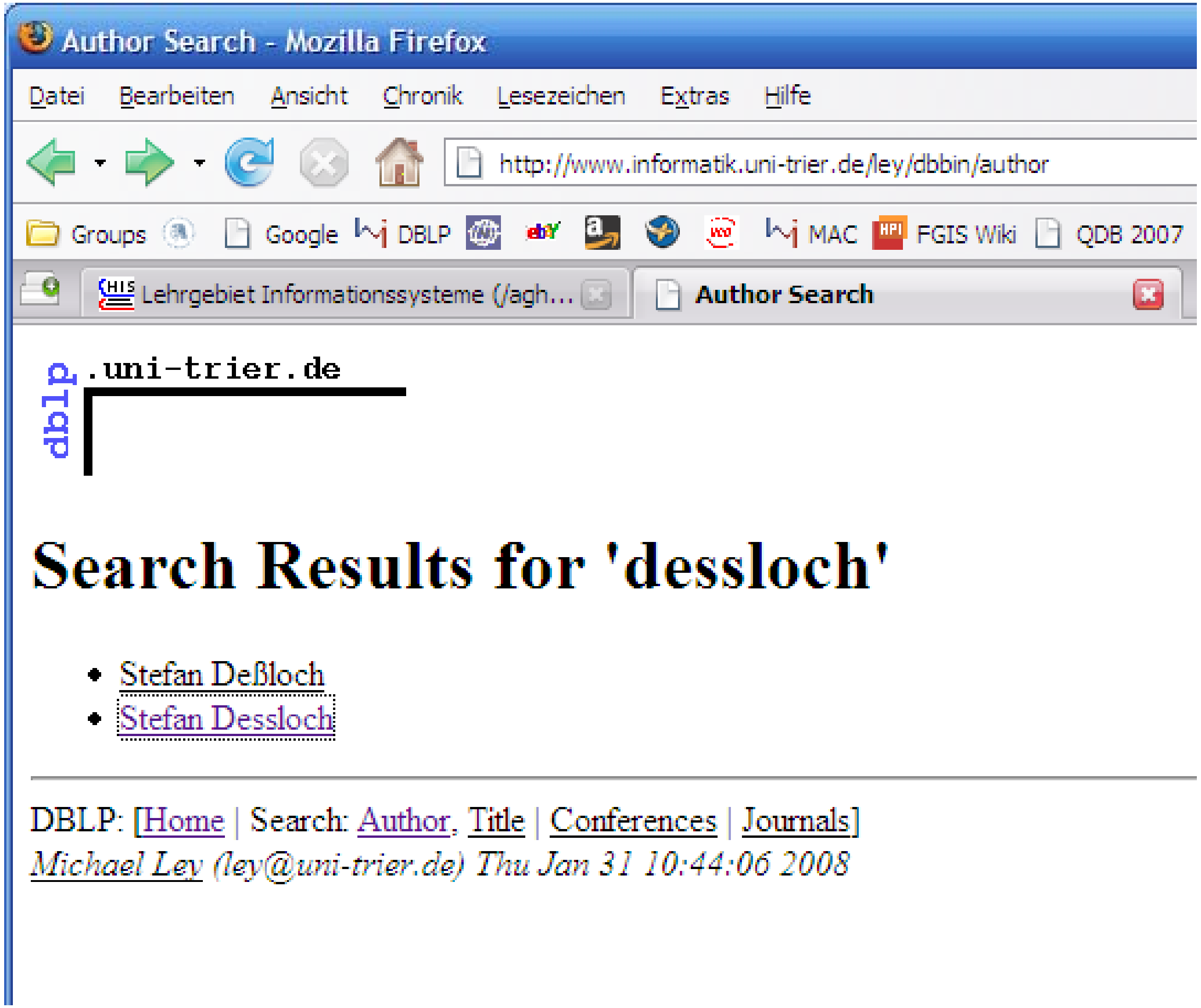
Weitere Möglichkeiten der Sendungsverfolgung

→ Sendungsverfolgung über E-Mail
→ Importieren Sie Kontrollnummern

Origins of duplicates



German names


Author Search - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
http://www.informatik.uni-trier.de/ley/dbbin/author
Groups Google DBLP MAC FGIS Wiki QDB 2007
Lehrgebiet Informationssysteme (/agh...
Author Search
dblp.uni-trier.de

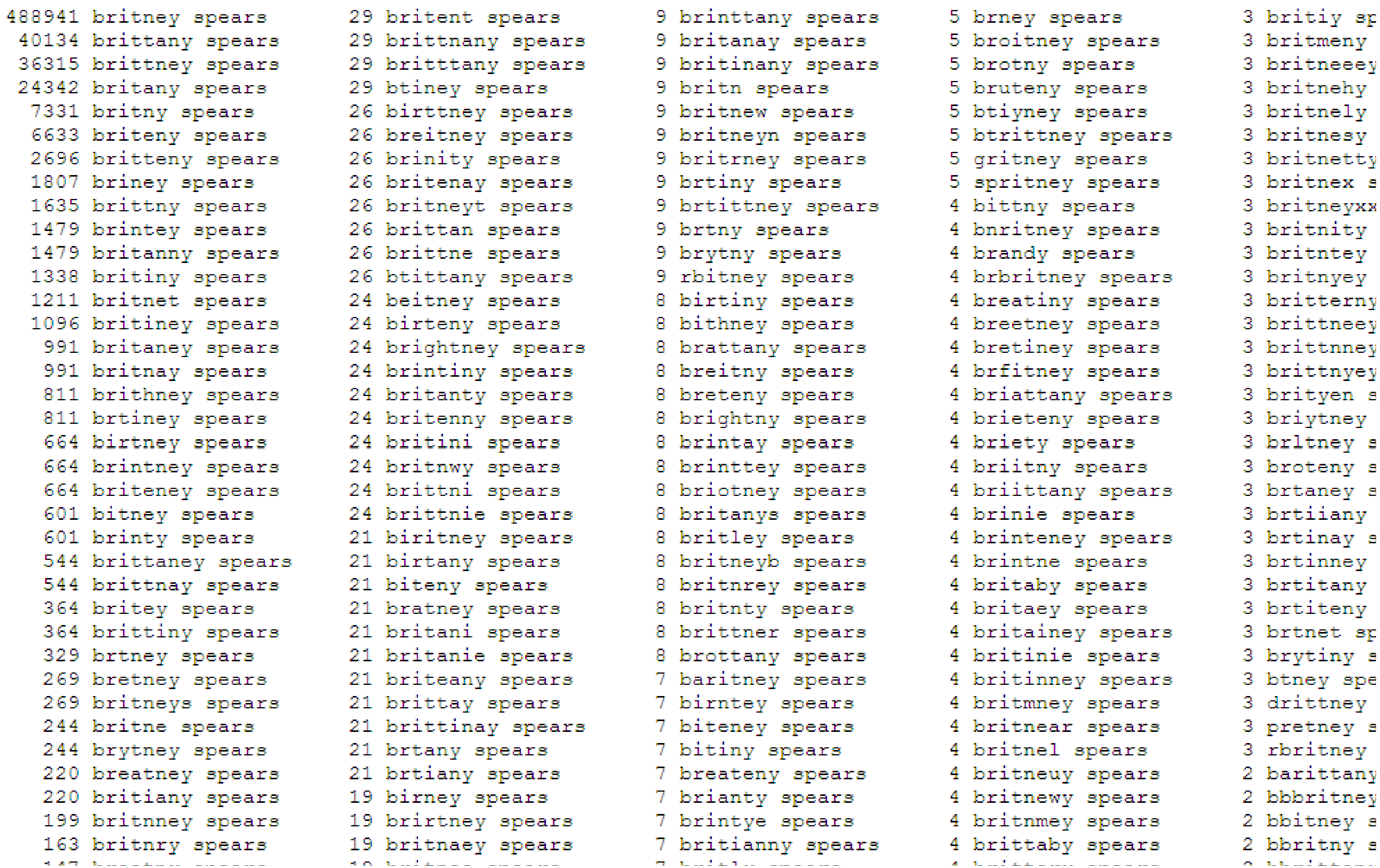
Search Results for 'dessloch'

- Stefan Deßloch
- Stefan Dessloch

DBLP: [Home | Search: Author, Title | Conferences | Journals]

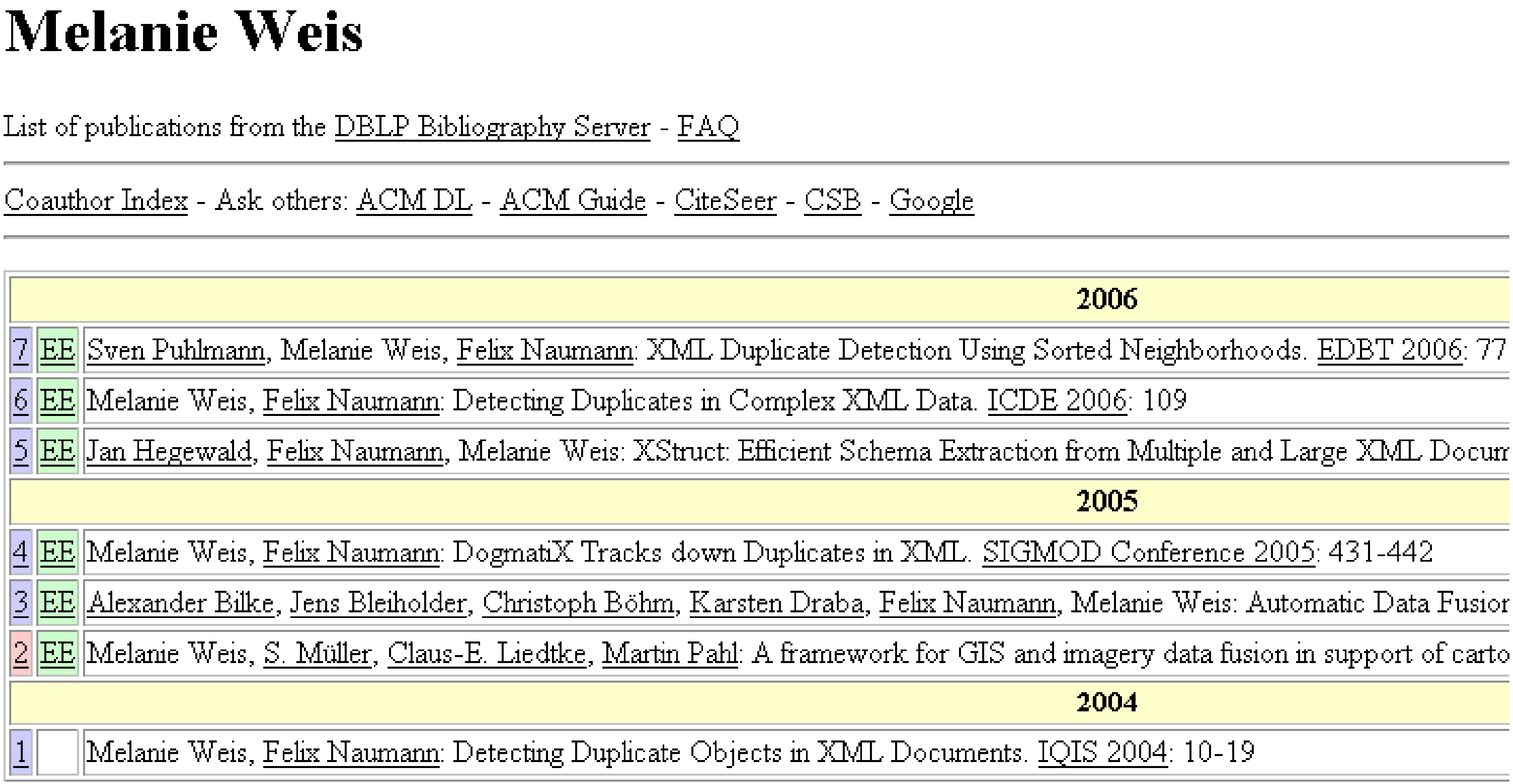
Michael Ley (ley@uni-trier.de) Thu Jan 31 10:44:06 2008

Difficult names



```
488941 britney spears      29 britent spears      9 brinttany spears    5 brney spears        3 britiy sp
 40134 brittany spears     29 brittnany spears    9 britanay spears     5 broitney spears     3 britmeny
 36315 brittney spears     29 britttany spears    9 britinany spears    5 brotny spears       3 britneeey
 24342 britany spears      29 btiney spears       9 britn spears        5 bruteny spears      3 britnehy
  7331 britny spears       26 birttney spears     9 britnew spears      5 btiyney spears      3 britnely
  6633 briteny spears      26 breitney spears     9 britneyn spears     5 btrittney spears    3 britnesy
  2696 britteny spears     26 brinity spears      9 britrney spears     5 gritney spears      3 britnetty
  1807 briney spears       26 britenay spears     9 brtiny spears       5 spritney spears     3 britnex s
  1635 brittny spears      26 britneyt spears     9 brtittney spears    4 bittny spears       3 britneyxx
  1479 brintey spears      26 brittan spears      9 brtny spears        4 bnritney spears     3 britnity
  1479 britanny spears     26 brittne spears      9 brytny spears       4 brandy spears       3 britntey
  1338 britiny spears      26 btittany spears     9 rbitney spears      4 brbritney spears    3 britnyey
  1211 britnet spears      24 beitney spears      8 birtiny spears      4 breating spears     3 britterny
  1096 britiney spears     24 birteny spears      8 bithney spears      4 breetney spears     3 brittneey
   991 britaney spears     24 brightney spears    8 brattany spears     4 bretiney spears     3 brittnney
   991 britnay spears      24 brintiny spears     8 breitny spears      4 brfitney spears     3 brittnyey
   811 brithney spears     24 britanty spears     8 breteny spears      4 briattany spears    3 brityen s
   811 brtiney spears      24 britenny spears     8 brightny spears     4 brieteny spears     3 briytney
   664 birtney spears      24 britini spears      8 brintay spears      4 briety spears       3 brltney s
   664 brintney spears     24 britnwy spears      8 brinttey spears     4 briitny spears      3 broteny s
   664 briteney spears     24 brittni spears      8 briotney spears     4 briittany spears    3 brtaney s
   601 bitney spears       24 brittnie spears     8 britanys spears     4 brinie spears       3 brtiiany
   601 brinty spears       21 biritney spears     8 britley spears      4 brinteney spears    3 brtinay s
   544 brittaney spears    21 birtany spears      8 britneyb spears     4 brintne spears      3 brtinney
   544 brittnay spears     21 biteny spears       8 britnrey spears     4 britaby spears      3 brtitany
   364 britey spears       21 bratney spears      8 britnty spears      4 britaey spears      3 brtiteny
   364 brittiny spears     21 britani spears      8 brittner spears     4 britainey spears    3 brtnet sp
   329 brtney spears       21 britanie spears     8 brottany spears     4 britinie spears     3 brytiny s
   269 bretney spears      21 briteany spears     7 baritney spears     4 britinney spears    3 btney spe
   269 britneys spears     21 brittay spears      7 birntey spears      4 britmney spears     3 drittney
   244 britne spears       21 brittinay spears    7 biteney spears      4 britnear spears     3 pretney s
   244 brytney spears      21 brtany spears       7 bitiny spears       4 britnel spears      3 rbritney
   220 breatney spears     21 brtiany spears      7 breateny spears     4 britneuy spears     2 barittany
   220 britiany spears     19 birney spears       7 brianty spears      4 britnewy spears     2 bbbritney
   199 britnney spears     19 brirtney spears     7 brintye spears      4 britnmey spears     2 bbitney s
   163 britnry spears      19 britnaey spears     7 britianny spears    4 brittaby spears     2 bbritny s
```

False Duplicates



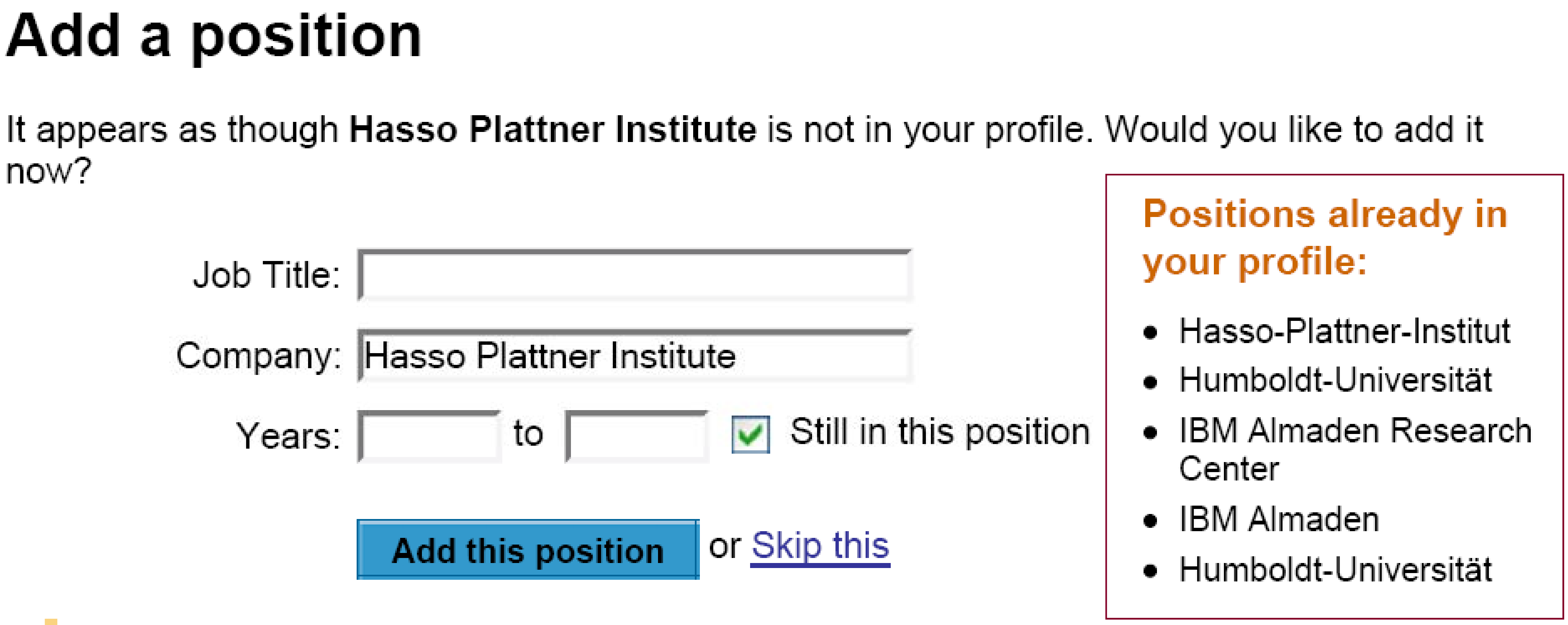
Melanie Weis

List of publications from the DBLP Bibliography Server - FAQ

Coauthor Index - Ask others: ACM DL - ACM Guide - CiteSeer - CSB - Google

		2006
7	EE	Sven Puhlmann, Melanie Weis, Felix Naumann: XML Duplicate Detection Using Sorted Neighborhoods. EDBT 2006: 77
6	EE	Melanie Weis, Felix Naumann: Detecting Duplicates in Complex XML Data. ICDE 2006: 109
5	EE	Jan Hegewald, Felix Naumann, Melanie Weis: XStruct: Efficient Schema Extraction from Multiple and Large XML Docum
		2005
4	EE	Melanie Weis, Felix Naumann: DogmatiX Tracks down Duplicates in XML. SIGMOD Conference 2005: 431-442
3	EE	Alexander Bilke, Jens Bleiholder, Christoph Böhm, Karsten Draba, Felix Naumann, Melanie Weis: Automatic Data Fusior
2	EE	Melanie Weis, S. Müller, Claus-E. Liedtke, Martin Pahl: A framework for GIS and imagery data fusion in support of carto
		2004
1		Melanie Weis, Felix Naumann: Detecting Duplicate Objects in XML Documents. IQIS 2004: 10-19

Company duplicates



Add a position

It appears as though **Hasso Plattner Institute** is not in your profile. Would you like to add it now?

Job Title:	
Company:	Hasso Plattner Institute
Years:	to ☑ Still in this position

Positions already in your profile:

- Hasso-Plattner-Institut
- Humboldt-Universität
- IBM Almaden Research Center
- IBM Almaden
- Humboldt-Universität

Motivation

Customer	Revenue
BMW	20.000
BaMoWe	5.000.000
Bayerische Motorenwerke	300.000
…	…

- Possible effects
  - Example: Portfolio Management Offers
  - Credit maximum not detected
  - Too low inventory levels
  - No quantity discount for multiple orders
  - Total revenue of preferred customers unknown
  - Multiple mailings of same catalog to same household
- General problems
  - Additional, unnecessary IT expenses
  - Low customer satisfaction
  - Potentials and dangers not detected
  - Poor quality financial data

Ironically, “Duplicate Detection” has many Duplicates

Household matching

Doubles

Duplicate detection

Mixed and split citation problem

Record linkage

Object identification

Match

Deduplication

Fuzzy match

Object consolidation

Entity resolution

Entity clustering

Approximate match

Identity uncertainty

Reference reconciliation

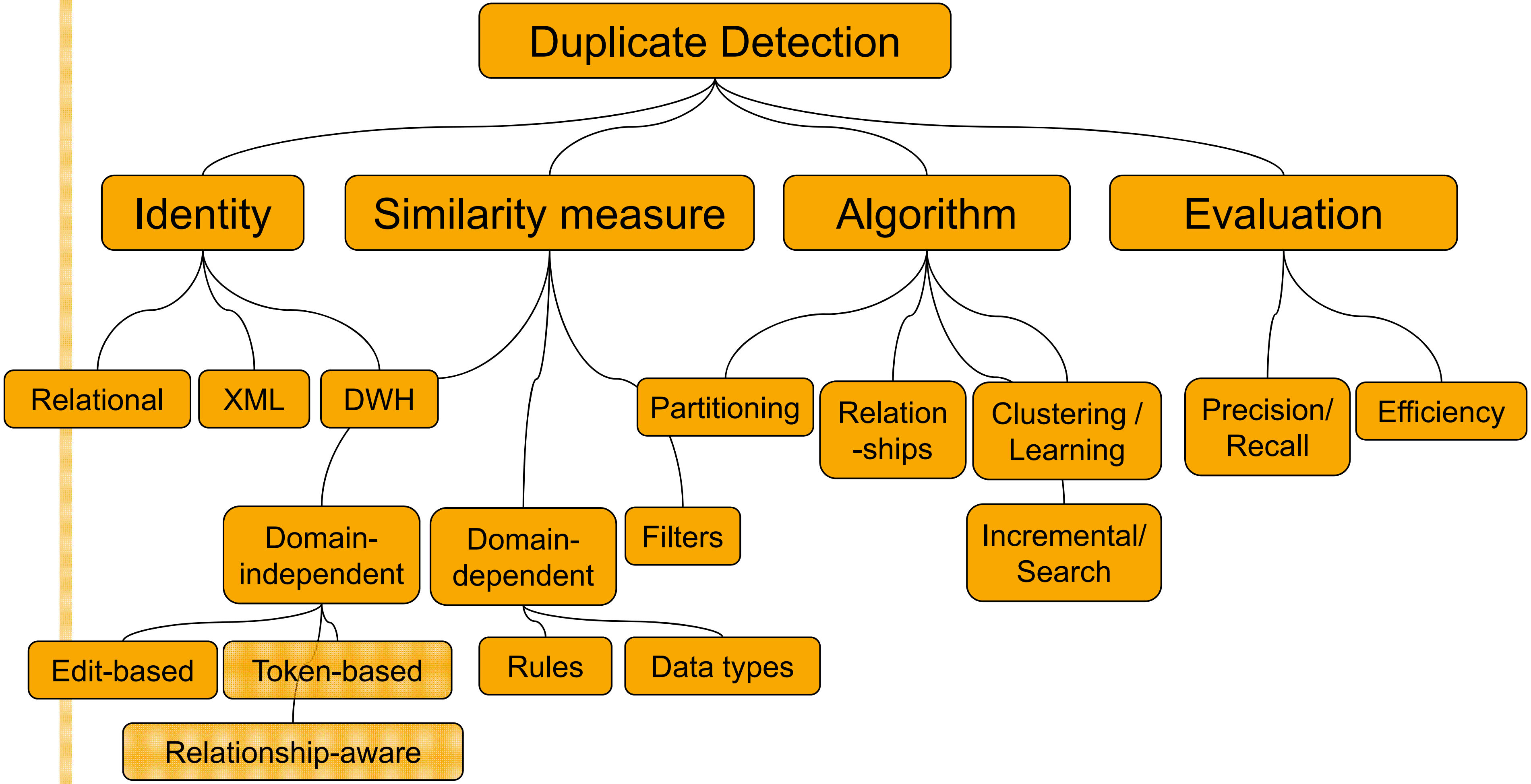
Merge/purge

Hardening soft databases

Householding

Reference matching

Duplicate Detection – Research



Token-based Similarity Measures

- Tokens
  - □ Words / Terms
  - □ n-grams
- Jaccard
  - □ |{common tokens}| / |{all tokens}|
- TFIDF [Cohen et al. 2003]
  - □ Term frequency: *tf*
  - □ Inverse document frequency: *idf*
  - □ TFIDF: log (*tf*+1) x log (*idf*)
  - □ Common words have low weight
  - □ Similarity measure: Cosine similarity of term vectors weighted by TFID
- And many more
  [Koudas Srivastavasa 2005]

Edit-based Similarity Measures

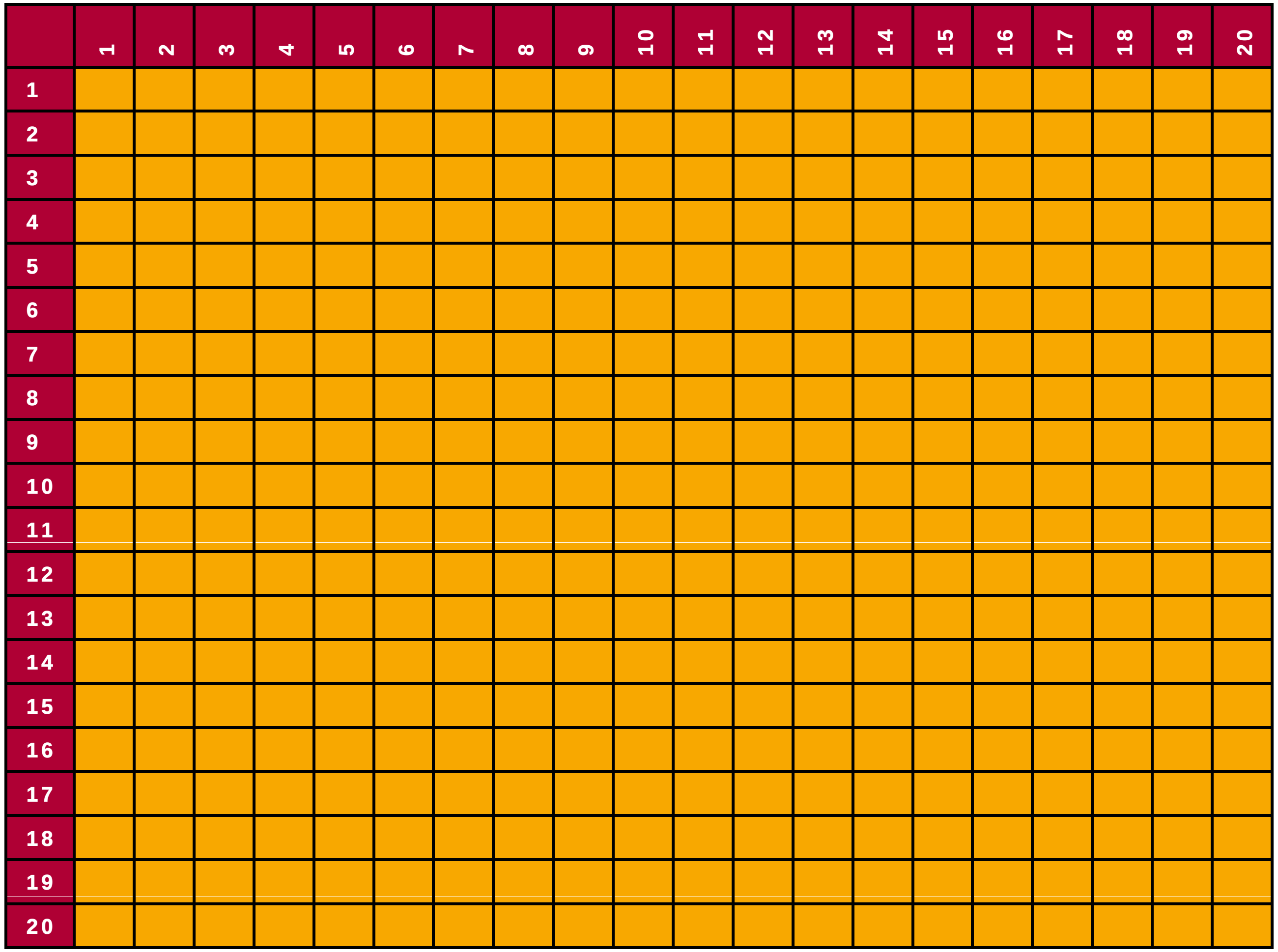
- Jaro [Jaro 1989] / Jaro-Winkler [Winkler 1999]
  - Common letters within ½ string length
  - Transposed letters
- Edit-distance / Levenshtein-distance [Levenshtein 1965]
  - Minimum number of edits from one word to the other
  - Domain-specific costing
  - Dynamic Programming
- Soundex
  - 4-letter code for each word
  - SOUNDEX('Farwick ') = F620
- …

Frass, Fricke, Fahruschi, Feuerhake

Record Pairs as Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1																				
2																				
3																				
4																				
5																				
6																				
7																				
8																				
9																				
10																				
11																				
12																				
13																				
14																				
15																				
16																				
17																				
18																				
19																				
20																				

Number of comparisons: All pairs



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
3	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
4	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
5	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
6	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
7	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
8	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
9	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
10	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
11	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
12	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
13	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
14	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
15	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
16	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
17	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
18	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
19	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
20	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧

400
comparisons

Reflexivity of Similarity

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
3	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
4	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
5	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
6	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
7	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
8	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
9	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
10	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
11	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
12	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
13	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧
14	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧	🟧
15	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧	🟧
16	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧	🟧
17	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧	🟧
18	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧	🟧
19	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜	🟧
20	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	⬜

380
comparisons

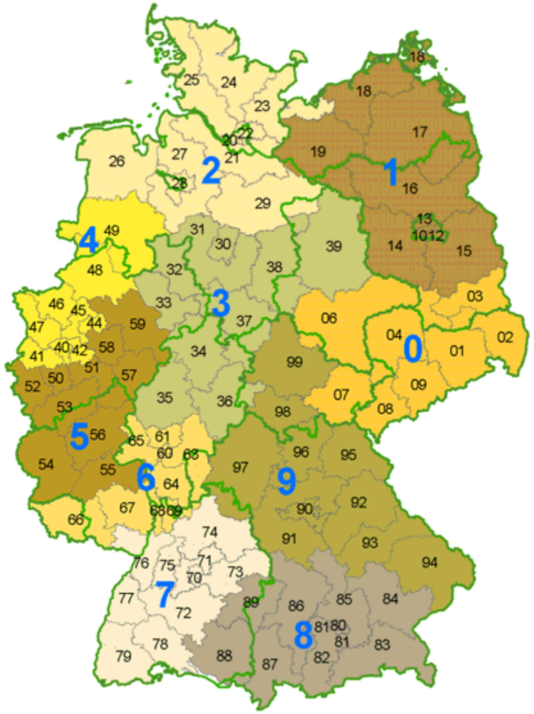
Symmetry of Similarity

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
3	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
4	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
5	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
6	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
7	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
8	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
9	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
10	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
11	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
12	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
13	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧
14	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧
15	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧
16	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧
17	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧
18	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧
19	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧
20	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

190
comparisons

Partitioning / Blocking

- Partition the records (horizontally) and compare pairs of records only within a partition.
  - □ Partitioning by first two zip-digits
    - ◇ Ca. 100 partitions in Germany
    - ◇ Ca. 100 customers per partition
    - ◇ => 495.000 comparisons
  - □ Partition by first letter of surname
  - □ …



Source: wikipedia.de

- Idea: Partition multiple times by different criteria.
  - □ Then apply transitive closure on discovered duplicates.

Complexity

Still: Too many comparisons

- 10.000 customers => 49.995.000 comparisons
  - □ $(n^2 - n) / 2$
  - □ Each comparison is expensive (complex similarity measures).

Idea: Avoid comparisons by heuristics

- Filtering of records
- Partitionierung

Records sorted by ZIP

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
2	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
3	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
4	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
5	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
6	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
7	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
8	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
9	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
10	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
11	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
12	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧
13	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧	🟧
14	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧	🟧
15	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧	🟧
16	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧	🟧
17	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧	🟧
18	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧	🟧
19	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟧
20	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜

190 comparisons

Blocking by ZIP

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		🟧	🟧	🟧	🟧															
2			🟧	🟧	🟧															
3				🟧	🟧															
4					🟧															
5						🟧	🟧	🟧												
6							🟧	🟧												
7								🟧												
8									🟧	🟧	🟧	🟧								
9										🟧	🟧	🟧								
10											🟧	🟧								
11												🟧								
12													🟧	🟧	🟧	🟧	🟧			
13														🟧	🟧	🟧	🟧			
14															🟧	🟧	🟧			
15																🟧	🟧			
16																	🟧			
17																		🟧	🟧	🟧
18																			🟧	🟧
19																				🟧
20																				

47
comparisons

Sorted Neighborhood
[Hernandez Stolfo 1998]

- Idea
  - Sort tuples so that similar tuples are close to each other.
  - Only compare tuples within a small neighborhood (window).

1. Generate key
   - E.g.: SSN+"first 3 letters of name" + ...
2. Sort by key
   - Similar tuples end up close to each other.
3. Slide window over sorted tuples
   - Compare all pairs of tuples within window.

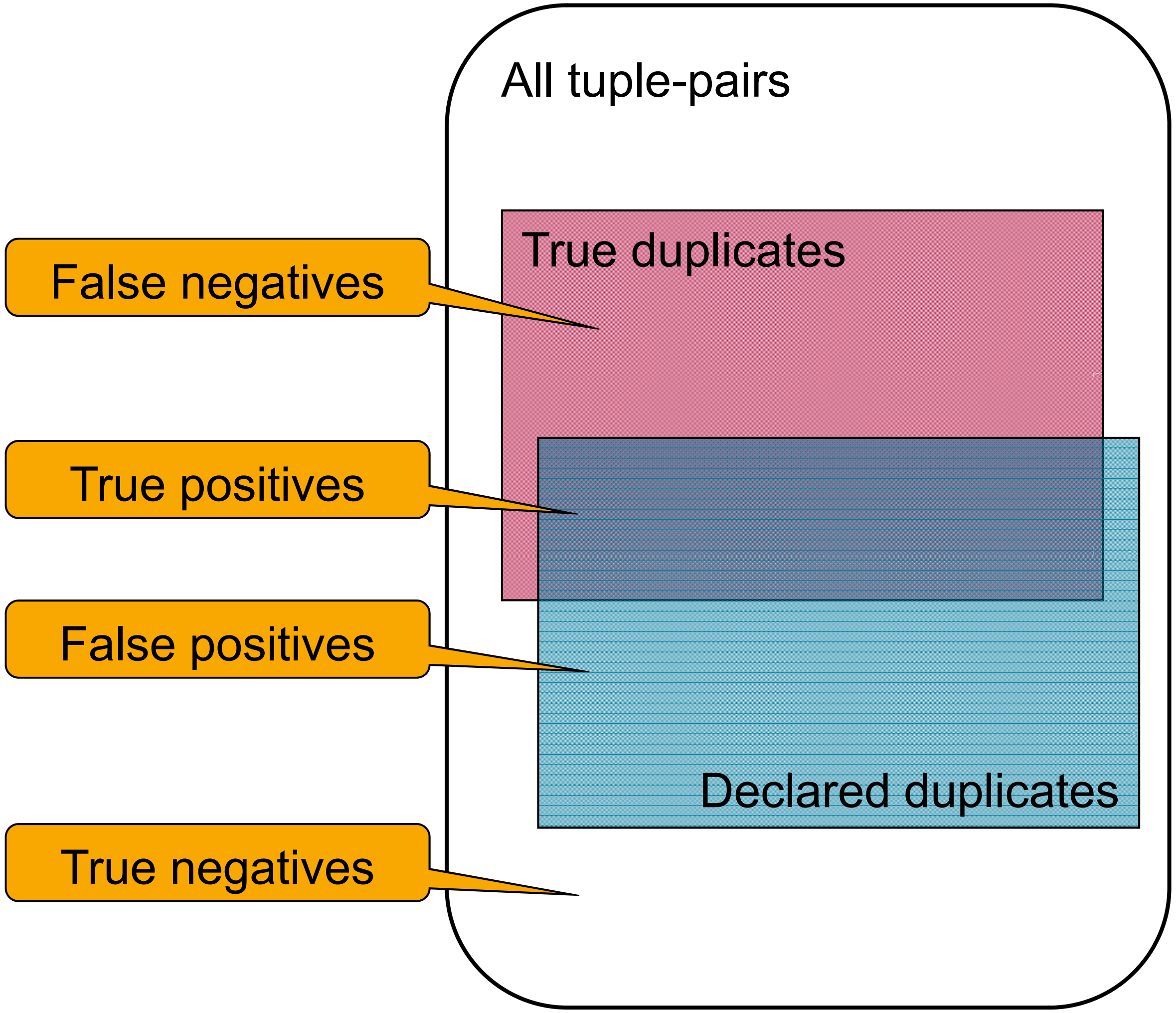
- Problems
  - Choice of key
  - Choice of window size
- Complexity: At least 3 passes over data
  - Sorting!

SNM by ZIP (window size 4)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		🟧	🟧	🟧	⬜															
2			🟧	🟧	🟧															
3				🟧	🟧	🟧														
4					🟧	🟧	🟧													
5						🟧	🟧	🟧												
6							🟧	🟧	🟧											
7								🟧	🟧	🟧										
8									🟧	🟧	🟧	⬜								
9										🟧	🟧	🟧								
10											🟧	🟧	🟧	⬜						
11												🟧	🟧	🟧						
12													🟧	🟧	🟧	⬜	⬜			
13														🟧	🟧	🟧	⬜			
14															🟧	🟧	🟧			
15																🟧	🟧	🟧	⬜	
16																	🟧	🟧	🟧	
17																		🟧	🟧	🟧
18																			🟧	🟧
19																				🟧
20																				

54
comparisons

Precision & Recall
(≈ correctness and completeness)

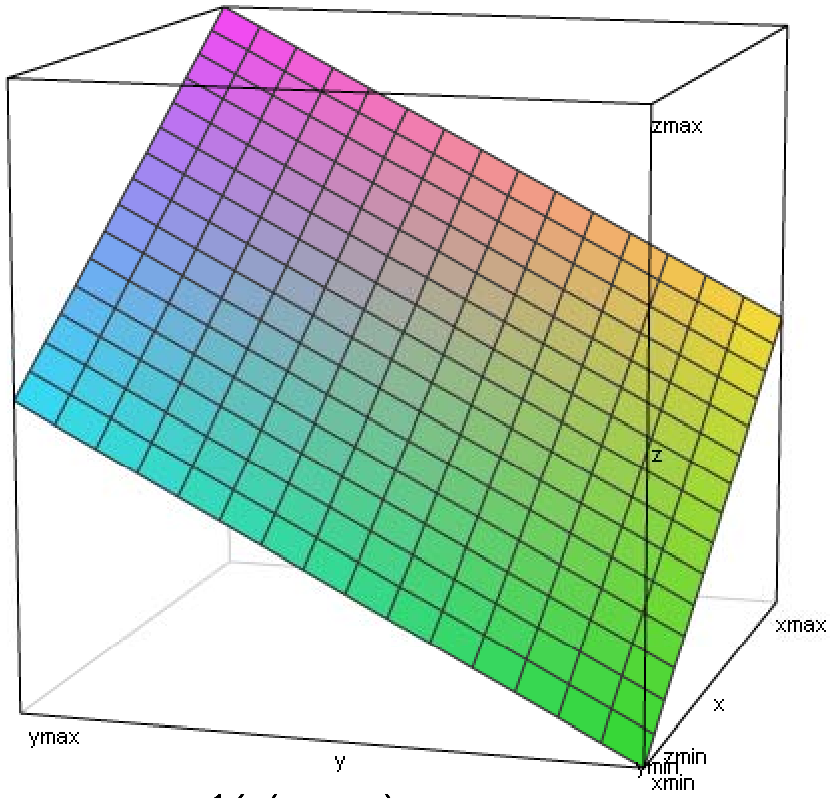


$$\text{Precision} = \frac{\text{True positives}}{\text{Declared duplicates}}$$

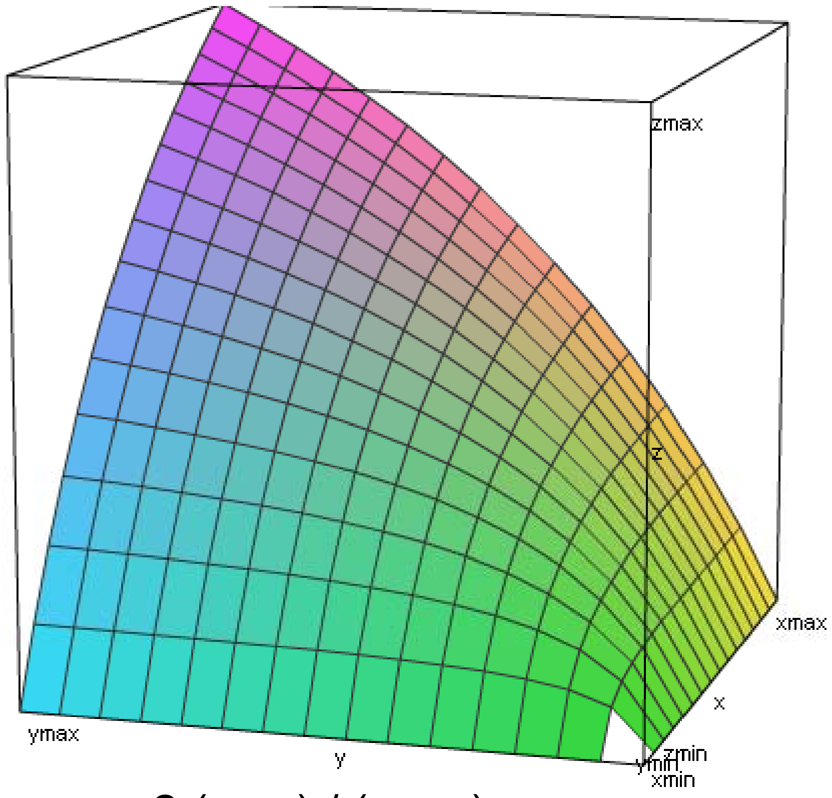
$$\text{Recall} = \frac{\text{True positives}}{\text{True duplicates}}$$

$$\text{F-Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Arithmetic mean („Average“) vs. Harmonic mean („F-Measure“)

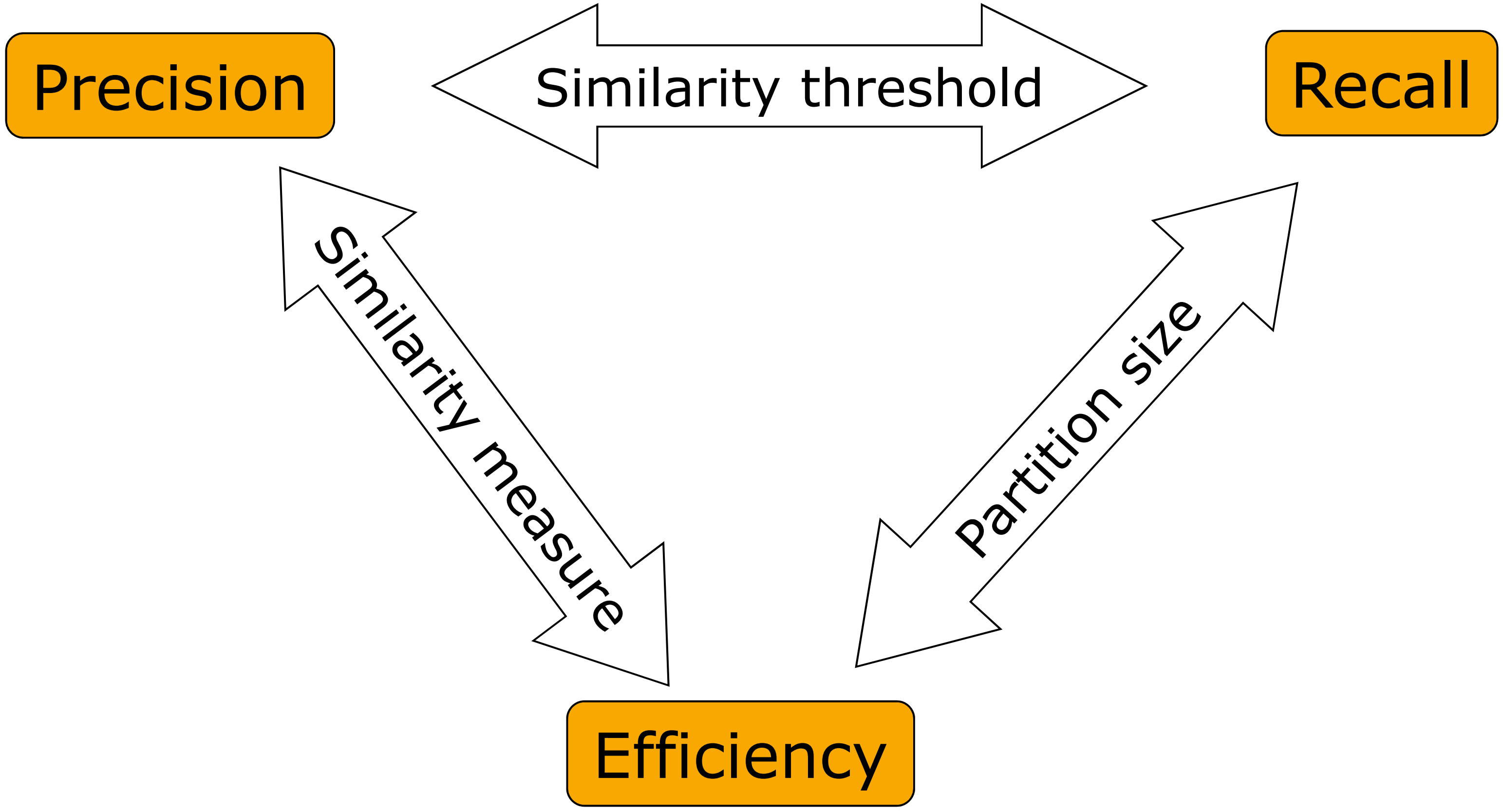


$z = \frac{1}{2}(x + y)$



$z = 2\,(x \cdot y) / (x + y)$

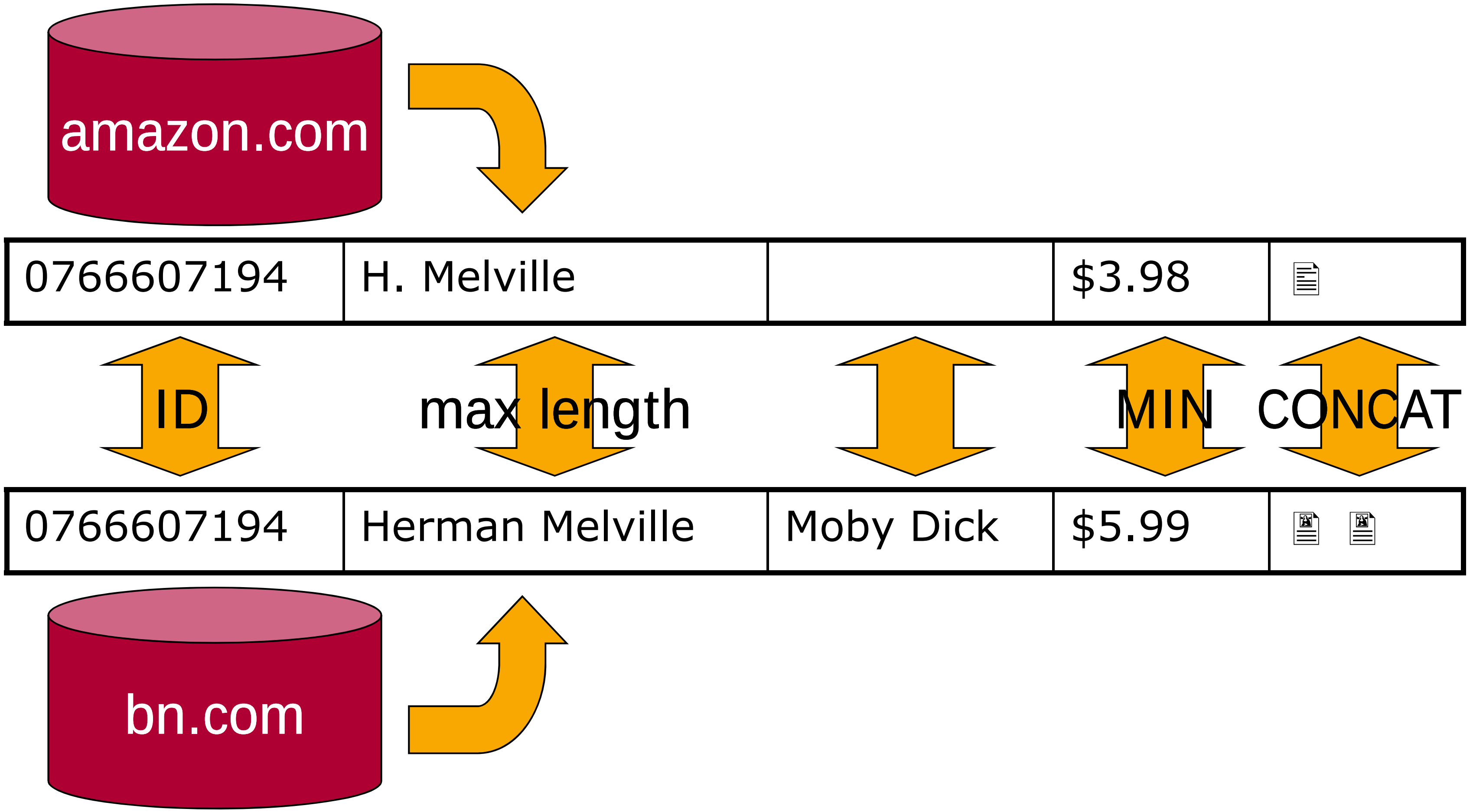
Duplikaterkennung – Zielkonflikte



Overview

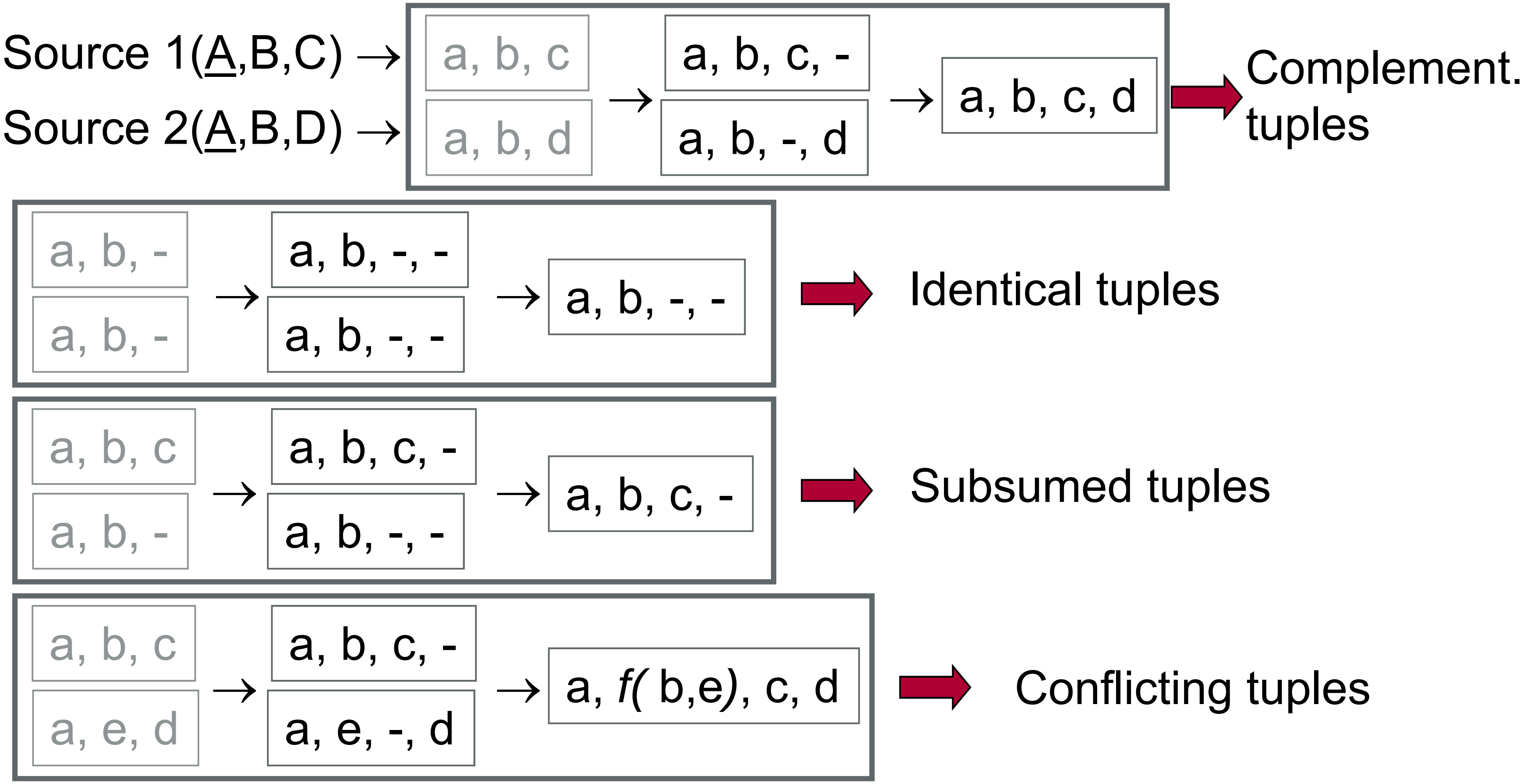
- Information Quality
- Step 1: Schema Matching
- Step 2: Duplicate detection
- ➡ Step 3: Data fusion
- Summary

Data Fusion



0766607194	H. Melville		$3.98	🗎
0766607194	Herman Melville	Moby Dick	$5.99	🗎 🗎

“Proper” Data Fusion



Source 1(A,B,C) →

Source 2(A,B,D) →

a, b, c → a, b, c, - → a, b, c, d

a, b, d → a, b, -, d

Complement. tuples

a, b, - → a, b, -, - → a, b, -, -

a, b, - → a, b, -, -

Identical tuples

a, b, c → a, b, c, - → a, b, c, -

a, b, - → a, b, -, -

Subsumed tuples

a, b, c → a, b, c, - → a, *f*(b,e), c, d

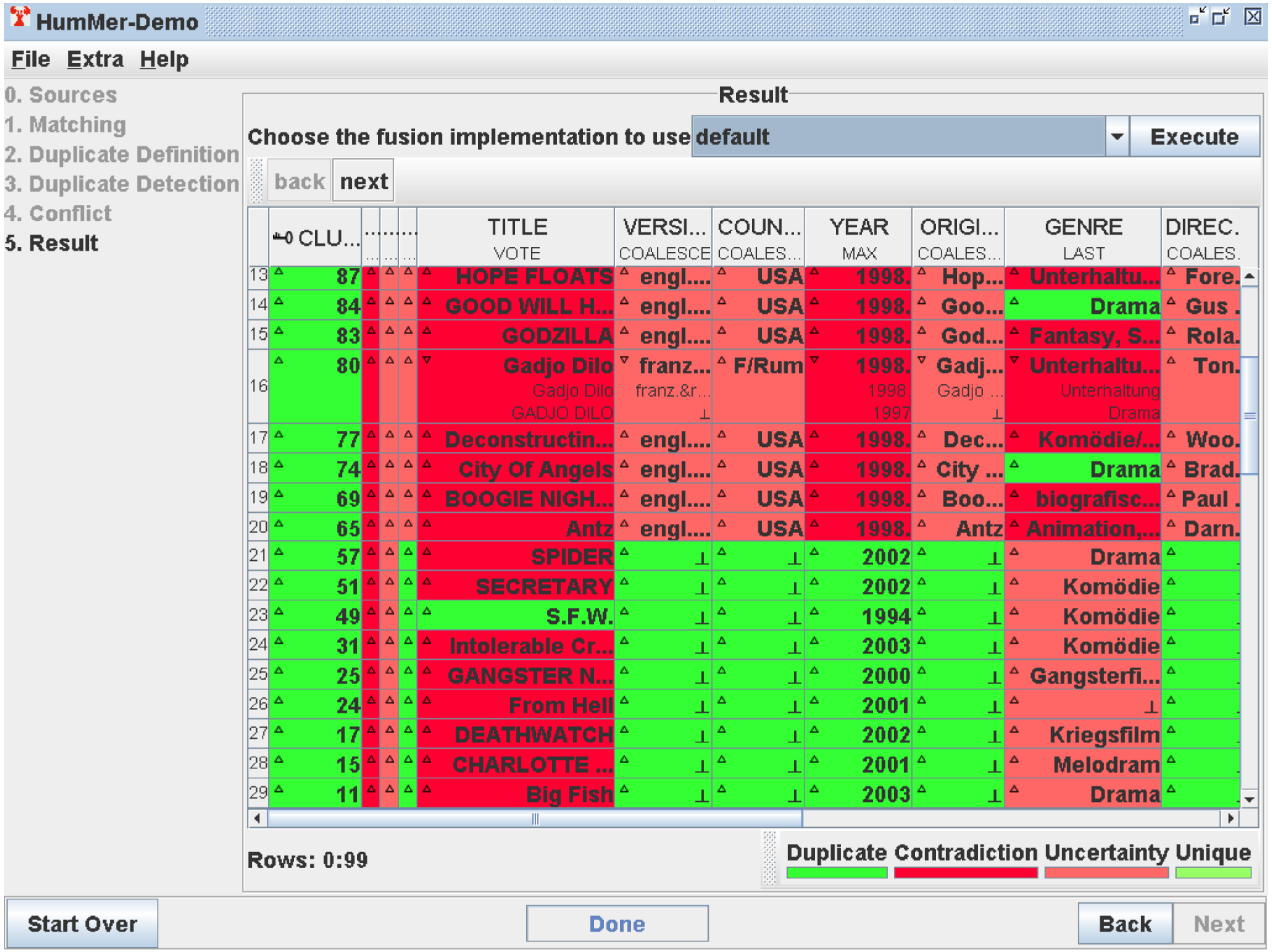
a, e, d → a, e, -, d

Conflicting tuples

Conflict Resolution Functions

Min, Max, Sum, Count, Avg, StdDev	Standard aggregation
Random	Random choice
First, Last	Choose first/last value; depends on order
Longest, Shortest	Choose longest/shortest value
Choose(*source*)	Choose value froma particular source
ChooseDepending(*col*, *val*)	Choose depending on *val* in other column *col*
Vote	Majority decision
Coalesce	Choose first non-null value
Group, Concat	Group or concatenate all values
MostRecent	Choose most recent (up-to-date) value
MostAbstract, MostSpecific	Use a taxonomy / ontology
….	….

Visualization of Integrated Data



Result

Choose the fusion implementation to use: default

	CLU...	...	...	...	TITLE	VERSI...	COUN...	YEAR	ORIGI...	GENRE	DIREC.
		...	...	...	VOTE	COALESCE	COALES...	MAX	COALES...	LAST	COALES.
13	87				HOPE FLOATS	engl....	USA	1998.	Hop...	Unterhaltu...	Fore.
14	84				GOOD WILL H...	engl....	USA	1998.	Goo...	Drama	Gus .
15	83				GODZILLA	engl....	USA	1998.	God...	Fantasy, S...	Rola.
16	80				Gadjo Dilo Gadjo Dilo GADJO DILO	franz... franz.&r... ⊥	F/Rum	1998. 1998. 1997	Gadj... Gadjo ... ⊥	Unterhaltu... Unterhaltung Drama	Ton.
17	77				Deconstructin...	engl....	USA	1998.	Dec...	Komödie/...	Woo.
18	74				City Of Angels	engl....	USA	1998.	City ...	Drama	Brad.
19	69				BOOGIE NIGH...	engl....	USA	1998.	Boo...	biografisc...	Paul .
20	65				Antz	engl....	USA	1998.	Antz	Animation,...	Darn.
21	57				SPIDER	⊥	⊥	2002	⊥	Drama	
22	51				SECRETARY	⊥	⊥	2002	⊥	Komödie	
23	49				S.F.W.	⊥	⊥	1994	⊥	Komödie	
24	31				Intolerable Cr...	⊥	⊥	2003	⊥	Komödie	
25	25				GANGSTER N...	⊥	⊥	2000	⊥	Gangsterfi...	
26	24				From Hell	⊥	⊥	2001	⊥	⊥	
27	17				DEATHWATCH	⊥	⊥	2002	⊥	Kriegsfilm	
28	15				CHARLOTTE ...	⊥	⊥	2001	⊥	Melodram	
29	11				Big Fish	⊥	⊥	2003	⊥	Drama	

Rows: 0:99

Duplicate 🟩 Contradiction 🟥 Uncertainty 🟥 Unique 🟩

Overview

- Information Quality
- Step 1: Schema Matching
- Step 2: Duplicate detection
- Step 3: Data fusion
- Summary

Summary

- Data Quality
- Step 1: Schema Matching
  - Similarity Measure
  - Combination of methods
- Step 2: Duplicate Detection
  - Similarity Measure
  - Algorithm
  - Data Model
- Step 3: Data Fusion
  - Relational Operators
  - Conflict Resolution
  - Visualization of Semantics and Overlap