Duplikaterkennung - Motivation

Masterseminar
16.4.2008

Felix Naumann
Hasso-Plattner-Institut
Fachgebiet Informationssysteme

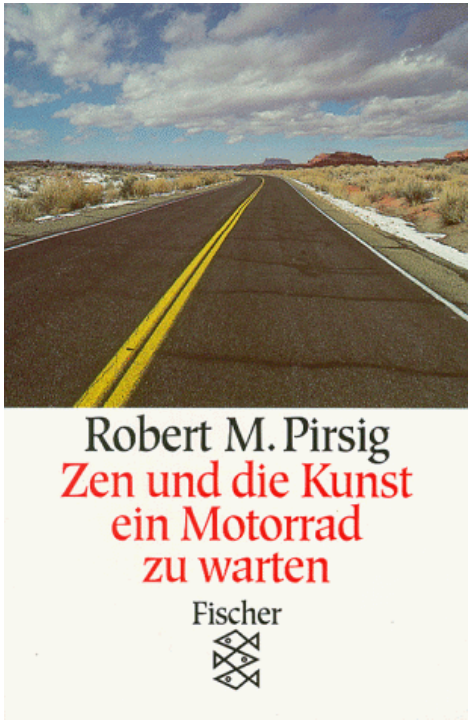
Überblick

- Datenqualität
- Datenfehler und ihre Ursachen
- Datenreinigung
- Duplikaterkennung

Qualität

"Even though quality cannot be defined, you know what it is."

Robert Pirsig



Zoom in die Datenqualität

1

15

179

Fitness for use

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

Ability to be Joined With	Ability to Download	Ability to Identify Errors	Ability to Upload
Acceptability	Access by Competitors	Accessibility	Accuracy
Adaptability	Adequate Detail	Adequate Volume	Aesthetic
Age	Aggregatability	Alterability	Amount of Data
Auditable	Authority	Availability	Believability
Breadth of Data	Brevity	Certified Data	Clarity
Clarity of Origin	Clear Data	Compactness	Comparability
	Responsibility		
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	Expert	Familiarity	Flawlessness
Flexibility	Form of Presentation	Formal	Integrity
Friendliness	Generality	Habit	Historical
			Compatibility
Importance	Inconsistencies	Integration	Integrity
Interactive	Interesting	Level of Abstraction	Level of Standardization
			Manipulable
Localized	Logically Connected	Manageability	Minimality
Measurable	Medium	Meets Requirements	Normality
Modularity	Narrowly Defined	No least information	Orderliness
Novelty	Objectivity	Optimality	Past Experience
Origin	Parsimony	Partitionability	Portability
Pedigree	Personalized	Pertinent	Purpose
Preciseness	Precision	Proprietary Nature	Regularity of Format
Quantity	Rationality	Redundancy	Reproducibility
Relevance	Reliability	Repetitive	Relevability
Reputation	Resolution of Graphics	Responsibility	Robustness
Revealing	Revocability	Rigidity	Self Correcting
Scope of Info	Secrecy	Security	Source
Semantic Interpretation	Semantics	Size	
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	

179 Dimensionen

DQ-Probleme: Auswirkungen

- Fehlerhafte Warenpreise in Artikel-DB des US-Einzelhandels
  - Kosten für Konsumenten 2.5 Mrd $
  - 80% der Barcode-Scan-Fehler zulasten der Konsumenten
- US-Finanzbehörde 1992: knapp 100.000 Steuererstattungsbescheide unzustellbar
- 50-80% der Einträge im US-Vorstrafenregister ungenau, unvollständig oder fehlerhaft
- US-Post: von 100.000 Massen-Postsendungen bis zu 7.000 aufgrund von Adressfehlern unzustellbar

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.

Überblick

- Datenqualität
- Datenfehler und ihre Ursachen
- Datenreinigung
- Duplikaterkennung

Nachkommastellen

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.

Deutsche Namen

Schwierige Namen

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	breatiny 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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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

http://www.google.com/jobs/britney.html

Falsche Duplikate

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

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

Add this position or Skip this

Positions already in your profile:

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

Datenqualität: Probleme

Repräsentation

Widersprüche

Ref. Integrität

KUNDE

KNr	Name	Geb.datum	Alter	Geschlecht	Telefon	PLZ
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

Eindeutigkeit

Fehlende Werte

Duplikate

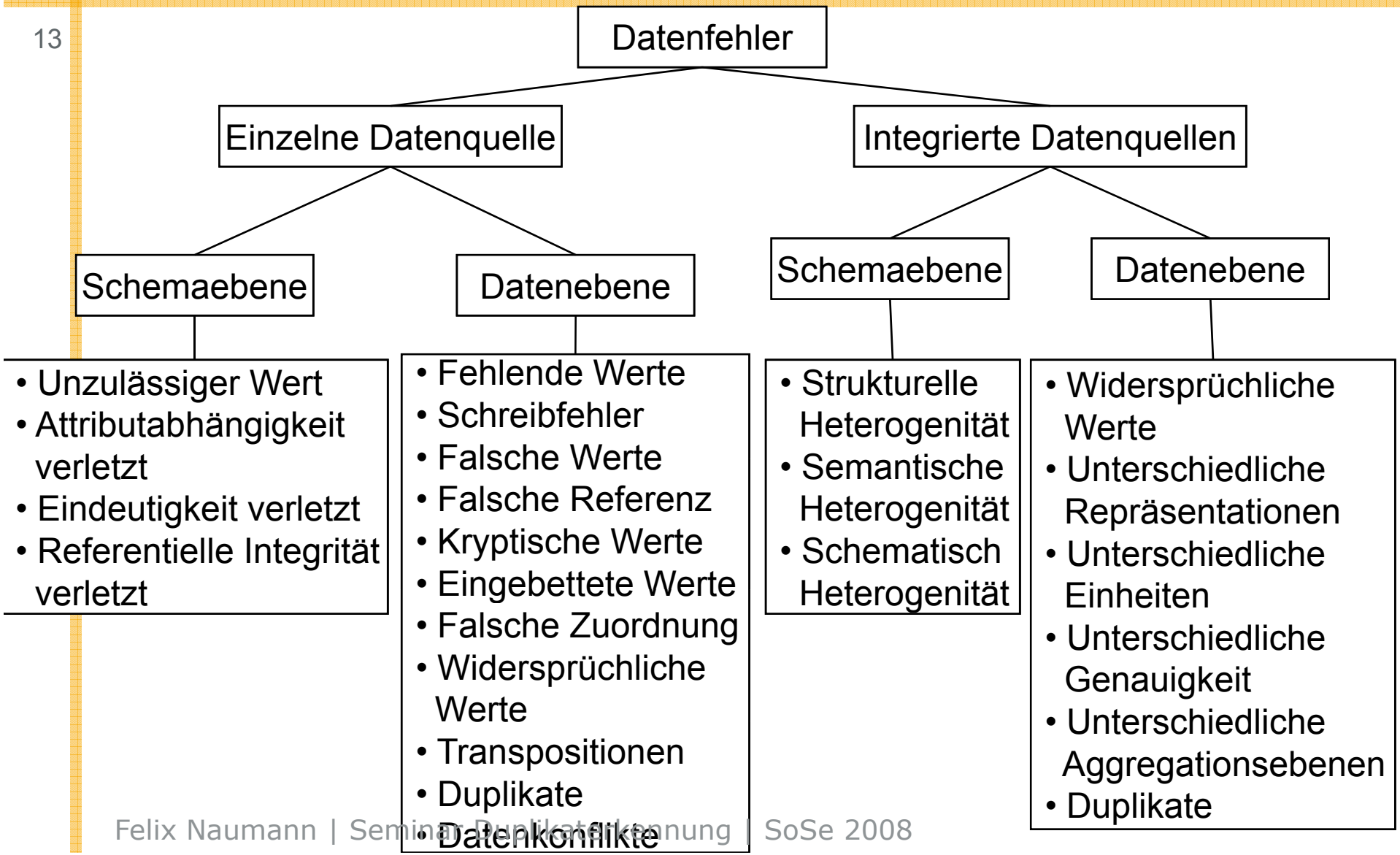
ADRESSE

PLZ	Ort
98693	Potsdam
98684	Postdam
98766	BRD

Schreibfehler

Falsche Werte

Klassifikation von Datenfehlern



Überblick

- Datenqualität
- Datenfehler und ihre Ursachen
- Datenreinigung
- Duplikaterkennung

Hilfsmittel

- Gesunder Menschenverstand
- Text-Editor
- Datenbanken und SQL („Bordmittel")
- Programmiesprachen
- Selbstentwickelte Tools
- Kommerzielle Tools
- Referenzdaten
  - Orts- und PLZ-Verzeichnisse, Straßenverzeichnisse
  - Namensverzeichnisse
    - z.B. IBMs Global Name Recognition mit 1 Milliarde Namen

Gesunder Menschenverstand

Southwest

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.

Standardisierung

- Adressen
  - Str. → Straße, Ch. → Chaussee, usw.
  - R.-Breitscheid-Str. 72 a → Rudolf-Breitscheid.-Str. 72A
  - 128 Schreibweisen für Frankfurt am Main
    - Frankfurt a.M., Frankfurt/M Frankfurt, Frankfurt a. Main, …
- Namen
  - Dr. Ing. h.c. F. Porsche AG
  - Hewlett-Packard Development Company, L.P.
- Numerische Daten
  - 10.000 € = 10T EURO = 10k EUR = 10.000,00€ = 10,000.- €
- Telefonnummern
- Geburtsdaten

Strukturierung – Personen

Anrede	Titel	Vorname	Nachname	...
		Herr Prof. Dr. Felix	Naumann	

Anrede	Titel	Vorname	Nachname	...
Herr	Prof. Dr.	Felix	Naumann	

Anrede	Titel	Vorname	Nachname	...
		Herr Prof. Dr. Felix und Marie	Naumann	

Anrede	Titel	Vorname	Nachname	...
Herr	Prof. Dr.	Felix	Naumann	
Frau		Marie	Naumann	

Strukturierung – Straßen

...	Straße	HNR	...
	Hauptstr. 7		

Ziele der Datenreinigung

- Verbesserung der Datenqualität einzelner Datensätze
- Gewinnung neuer Informationen
- Verbesserung der Ergebnisse der Duplikaterkennung
  - Standardisierte Datensätze sind besser vergleichbar.

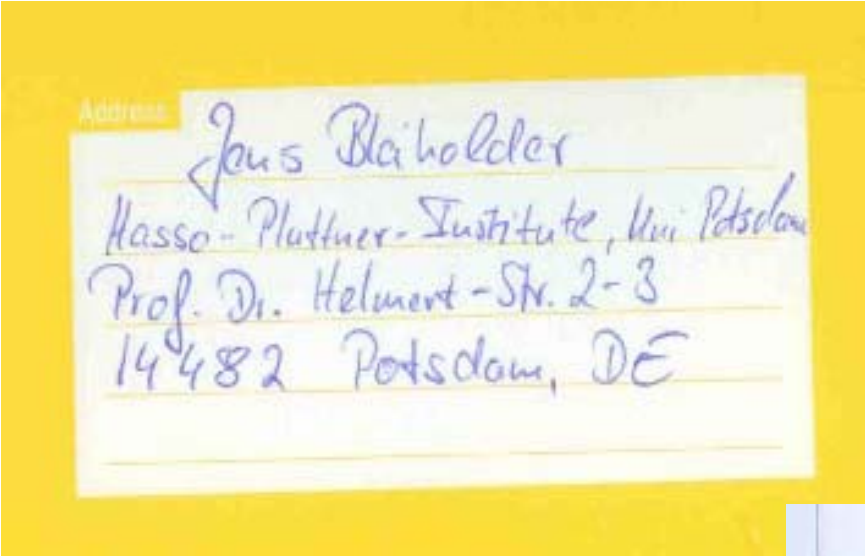
Praxisbeispiel – Versteckter Mehrwert

	Feld						
Datenelement	**Name1**	**Name2**	**Name3**	**Ortsname**	**Ortsteil**	**Straße**	**Summe**
Handy-Nummer	41	501	10	0	**2677**	297	3526
Festnetznummer	15	98	6	0	**221**	9579	9919
Kostenstelle	283	1112	73	2	**87**	16	1573
Registriernummer	11	583	1	1	**0**	3	599
Lieferungsnummer	55	390	9	0	**212**	15	681
Abteilung	3711	**9997**	115	60	**439**	175	14497
Sperrkennzeichen	129	143	2	0	**66**	9	349
Löschkennzeichen	1028	442	5	36	**113**	10	1634
Rechtsform	**131700**	66136	187	6	**64**	57	198150
Kreditoreninfo	0	100	11	0	**18**	0	129
Kommissionsinfo	216	352	1	2	**36**	10	617
Baustelle	2013	3452	42	5	**124**	222	5858
Abladestelle	2923	3808	94	1503	**958**	3065	12351
Behörde	13410	12461	172	19	**295**	7075	33432
Summe	155535	99575	728	1634	**5310**	20533	

Überblick

- Datenqualität
- Datenfehler und ihre Ursachen
- Datenreinigung
- Duplikaterkennung

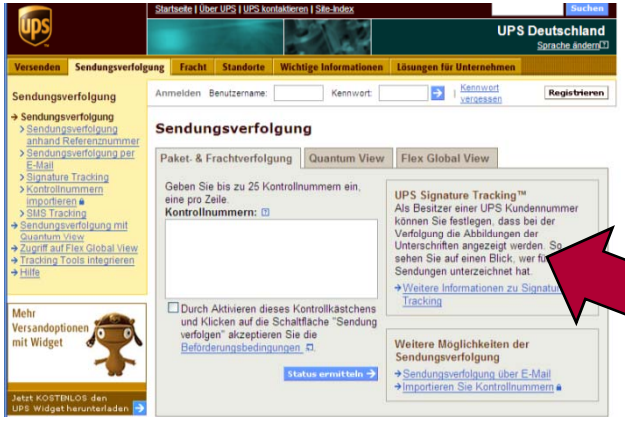
Wie entstehen Duplikate?



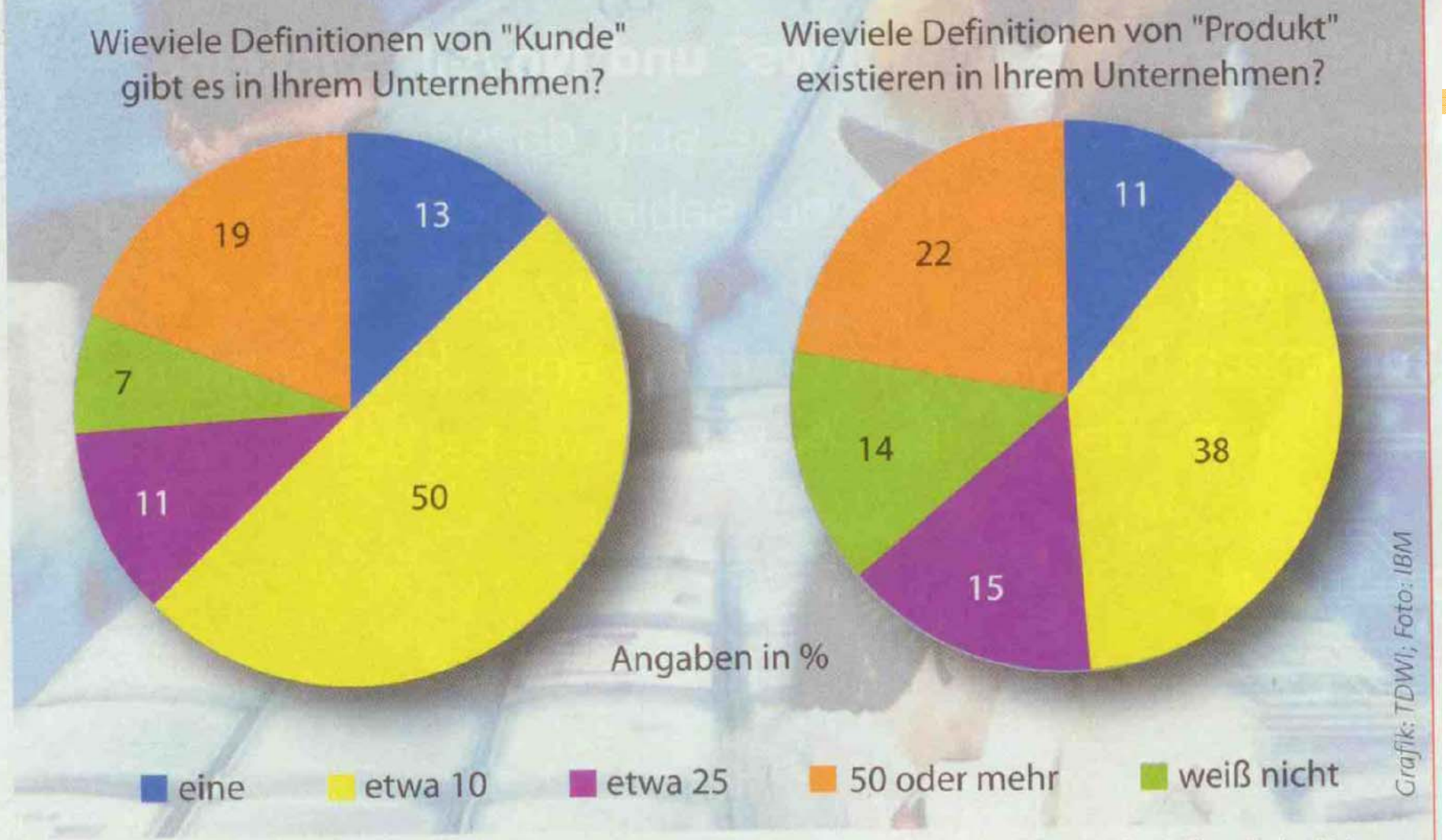
Original

Zugestellt

Wie entstehen Duplikate?

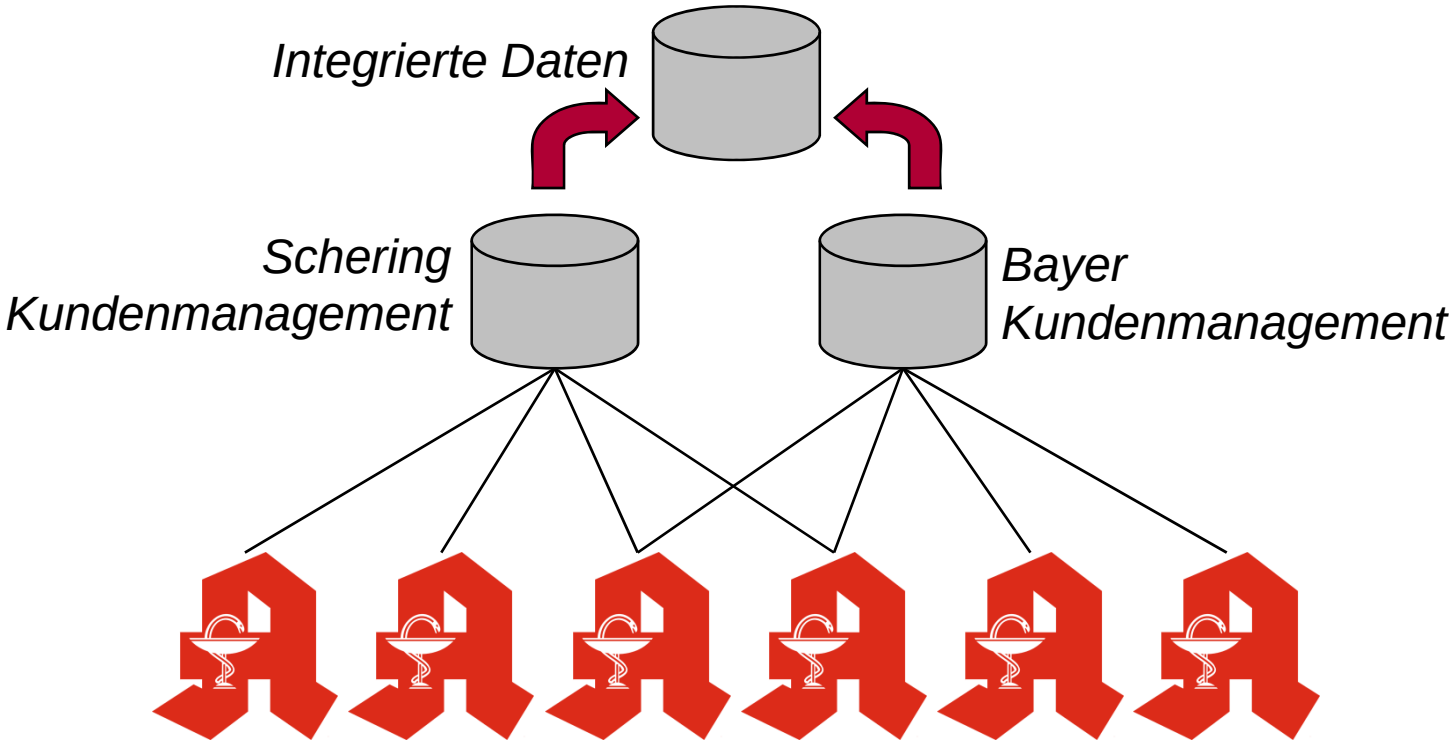


Mehrere Definitionen eines Kernobjekts sind die Regel



Der Kunde ist das am häufigsten in Stammdaten definierte Objekt, hat The Data Warehouse Institute in einer Umfrage herausgefunden. Einheitliche Daten gibt es trotzdem nicht: In 87 Prozent der Unternehmen existieren mehrere Kundendefinitionen. sr

Wie entstehen Duplikate?



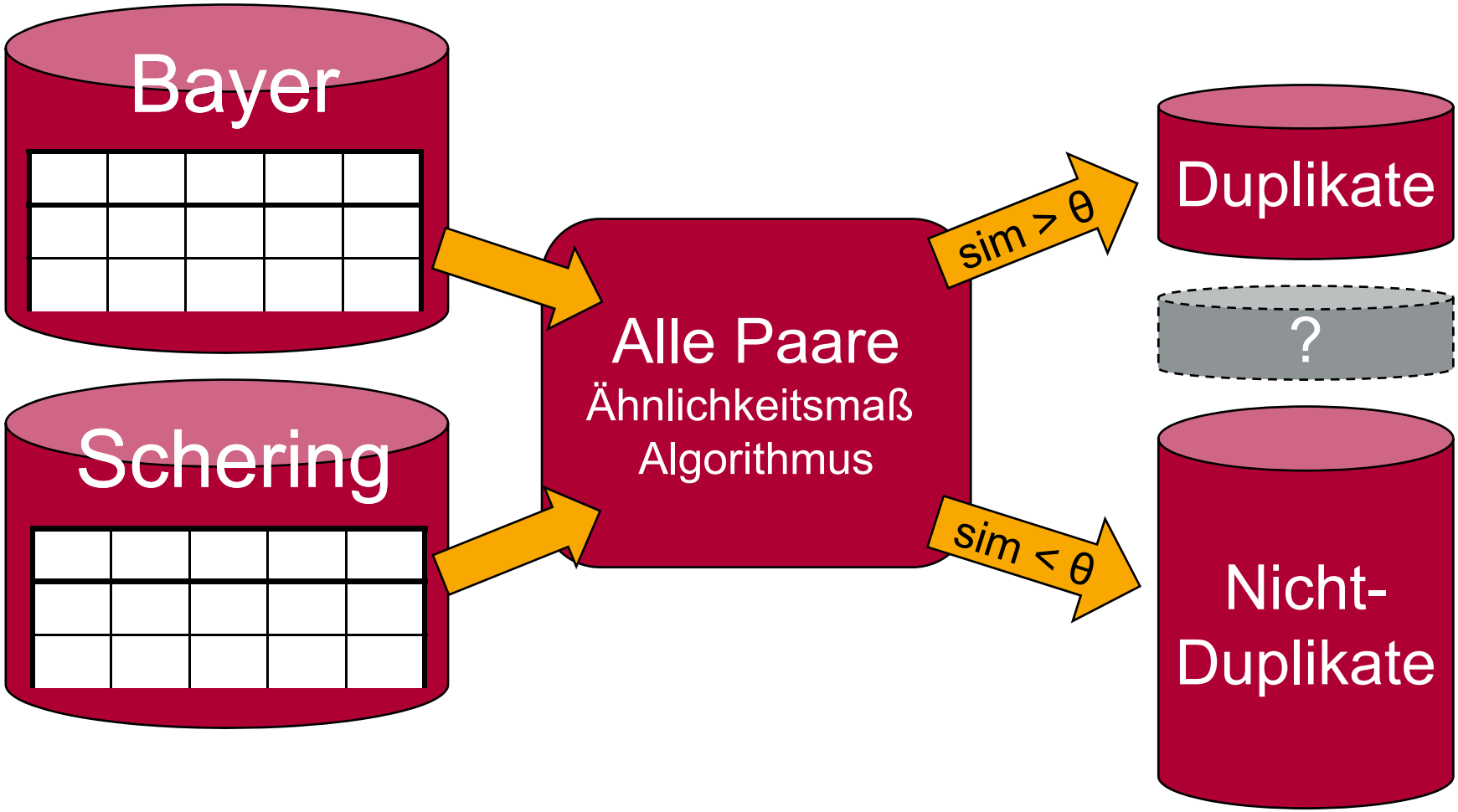
Duplikaterkennung

Duplikaterkennung ist das Finden mehrerer Repräsentationen desselben Realweltobjekts.

- Problem 1: Repräsentationen sind nicht identisch.
  - *Fuzzy duplicates*
- Lösung: Ähnlichkeitsmaße
  - Wert- und Datensatzvergleiche
  - Domänenunabhängig oder -abhängig

- Problem 2: Die Datenmenge ist groß.
  - Quadratischer Aufwand: Jedes Paar muss verglichen werden.
- Lösung: Algorithmen
  - Z.B. Vergleiche durch Partitionierung vermeiden

Duplikaterkennung



Wirkungen von Duplikaten

- Mehrfache Zusendung von Katalogen
- Rechnungen werden doppelt bezahlt
- Banken
  - Überschreiten des Kreditlimits wird nicht erkannt
- Lagerhaltung / Einkauf
  - Zu niedriger Lagerbestand einzelner Waren wird ausgewiesen.
  - Kein Ausnutzen von Mengenrabatten bei Bestellungen
- Gesamtumsatz eines Kunden bleibt unbekannt.

Kunde	Umsatz
BMW	20.000
BaMoWe	5.000.000
Bayerische Motorenwerke	300.000
...	...

- Mehraufwand in der IT
- Sinkende Kundenzufriedenheit
- Potenziale und Gefahren nicht erkannt
- Inkorrekte Kennzahlen

“Duplikaterkennung” hat viele Duplikate

Household matching

Fuzzy match

Duplicate detection

Mixed and split citation problem

Record linkage

Object identification

Match

Deduplication

Dubletten

Object consolidation

Entity resolution

Entity clustering

Approximate match

Identity uncertainty

Reference reconciliation

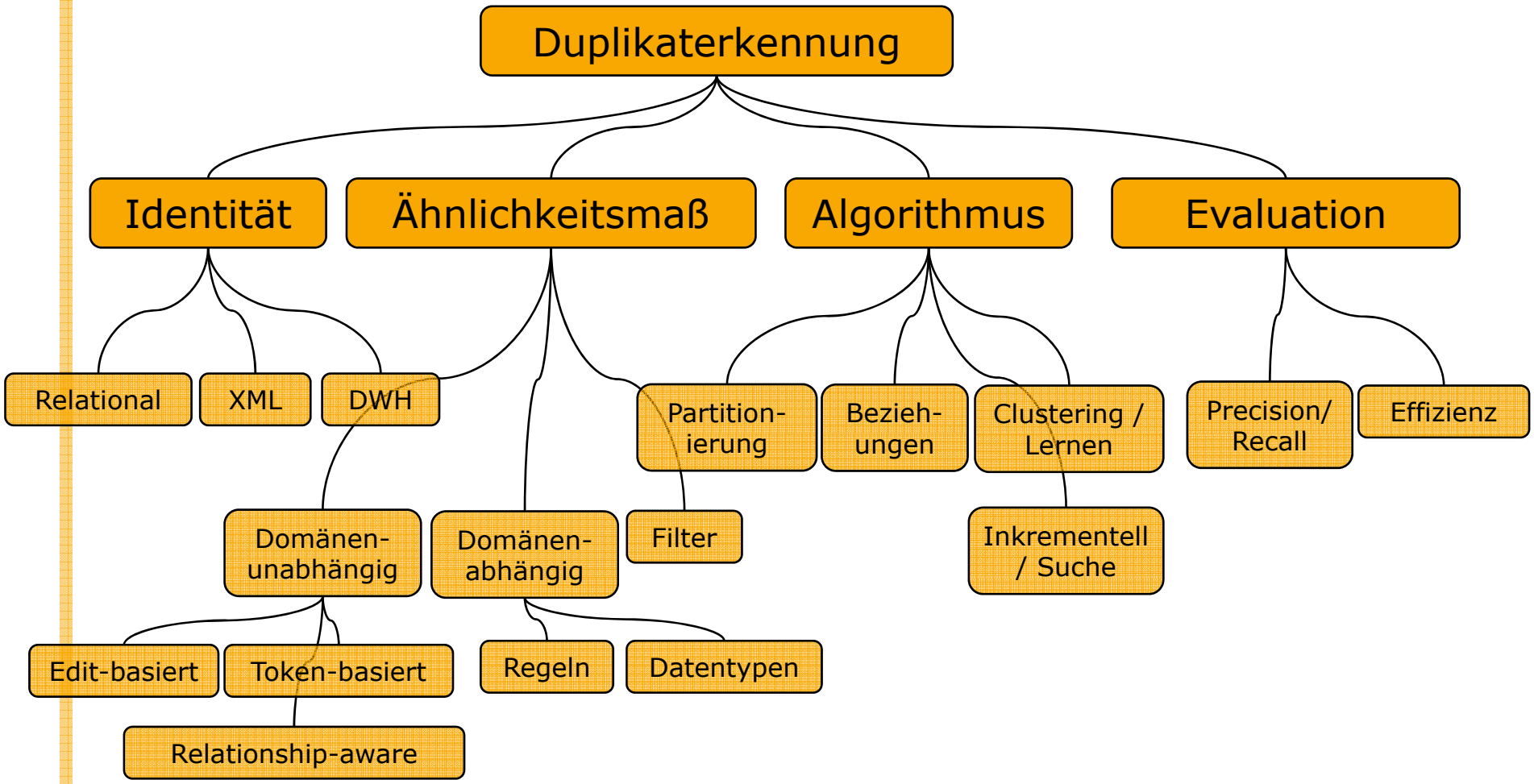
Merge/purge

Hardening soft databases

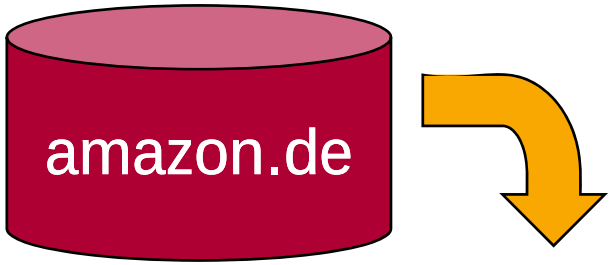
Householding

Reference matching

Duplikaterkennung

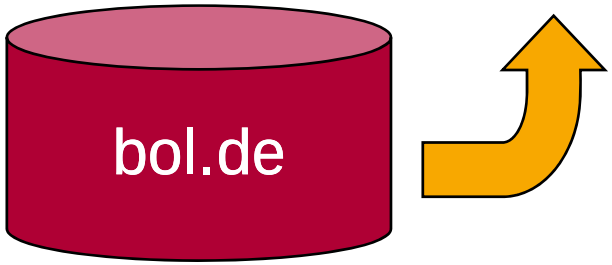


Datenfusion



0766607194	H. Melville		$3.98	

0766607194	Herman Melville	Moby Dick	$5.99	



Datenfusion in Outlook

Mehrfach vorhandener Kontakt

Der Name oder die E-Mail-Adresse dieses Kontakts ist bereits im Ordner 'Kontakte' vorhanden. Möchten Sie:

- ○ Neuen Kontakt hinzufügen
- ◉ Informationen des ausgewählten Kontakts aktualisieren. Eine Sicherungskopie wird im Ordner 'Gelöschte Objekte' gespeichert.

Name	Position	Firma	E-Mail
Loeser, Alexander			alexander.loeser@sap.com

Vorschau der aktualisierten Visitenkarte:

Alexander Löser
SAP Research CEC Dresden
Senior Researcher
+49 (0) 351 - 48 11 - 6136 Geschäftlich
0174 - 61 96 808 Mobiltelefon
alexander.loeser@sap.com
SAP AG
Chemnitzer Straße 48
Dresden 01187
Deutschland

Änderungen am ausgewählten Kontakt:

Name:	Alexander Löser ~~Loeser, Alexander~~
Position:	Senior Researcher
Firma:	SAP Research CEC Dresden
E-Mail:	alexander.loeser@sap.com
Adresse geschäftlich:	SAP AG Chemnitzer Straße 48 Dresden 01187 Deutschland
Adresse privat:	Max Steinke Strasse 25 13086 Berlin (Inga Herdrich)
Telefon geschäftlich:	+49 (0) 351 - 48 11 - 6136 ~~+49 351 4811 6136~~
Fax geschäftl.:	+49 (0) 6227 - 78-4 84 79 ~~+49 6227 7848479~~
Mobiltelefon:	0174 - 61 96 808 ~~0174 619 6808~~
Kontaktbild:	Keine Änderung
Notizen:	Keine Änderung

Aktualisieren | Abbrechen