Information Integration

Data Quality

23.1.2020

Felix Naumann

Overview

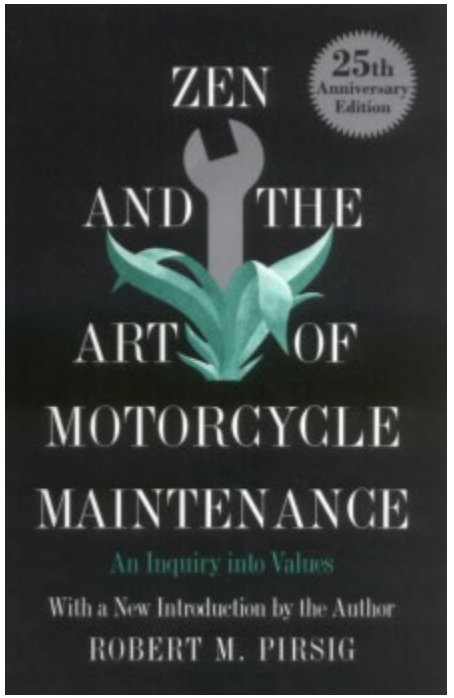
1. **Information quality**
2. IQ criteria
3. IQ assessment
4. Cleansing tasks
5. IQ anecdotes

Felix Naumann
Information Integration
Winter 2019/20

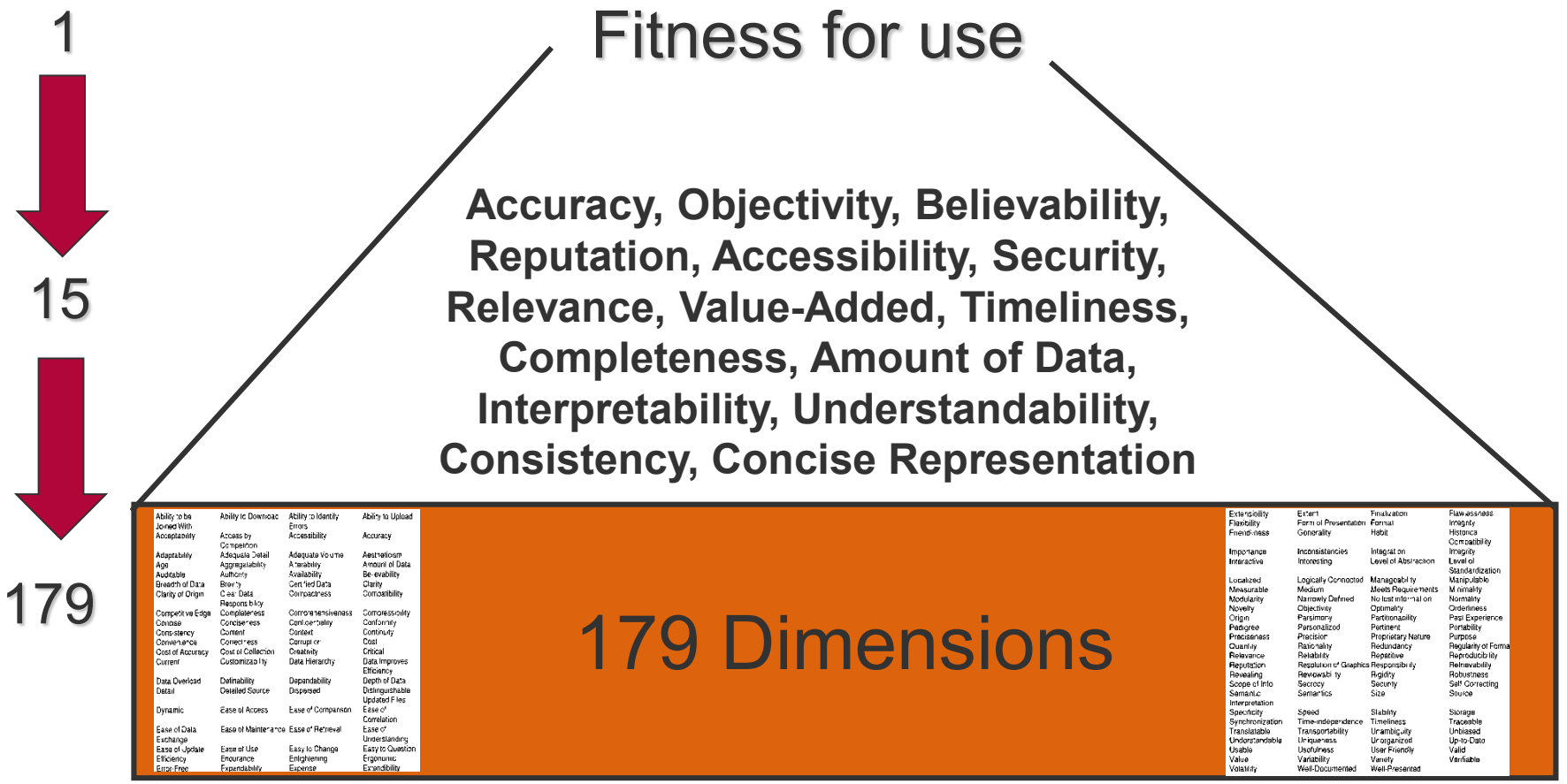
Quality

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

Robert Pirsig

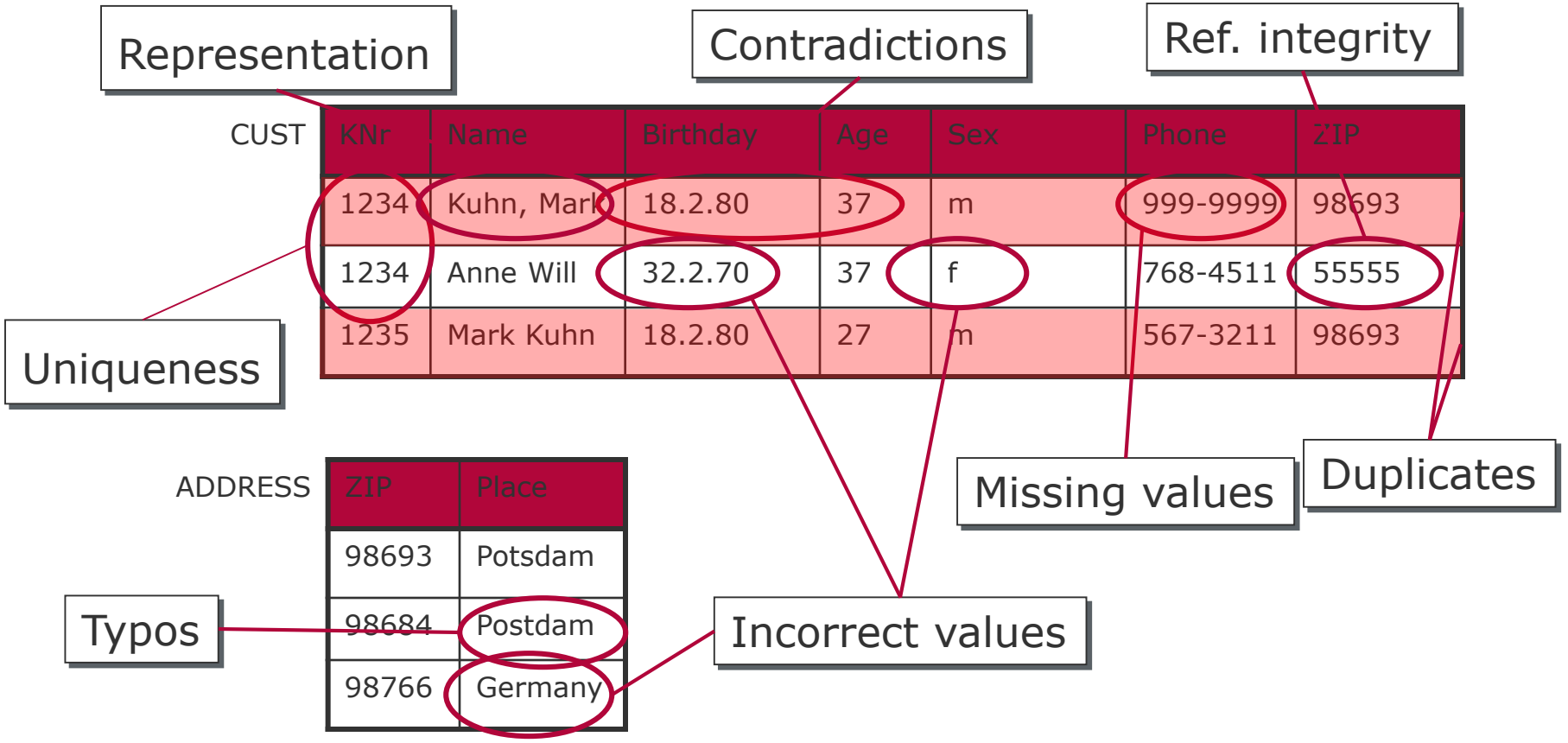


Zooming into Information Quality



Felix Naumann
Information Integration
Winter 2019/20

Data Quality: Problems



CUST

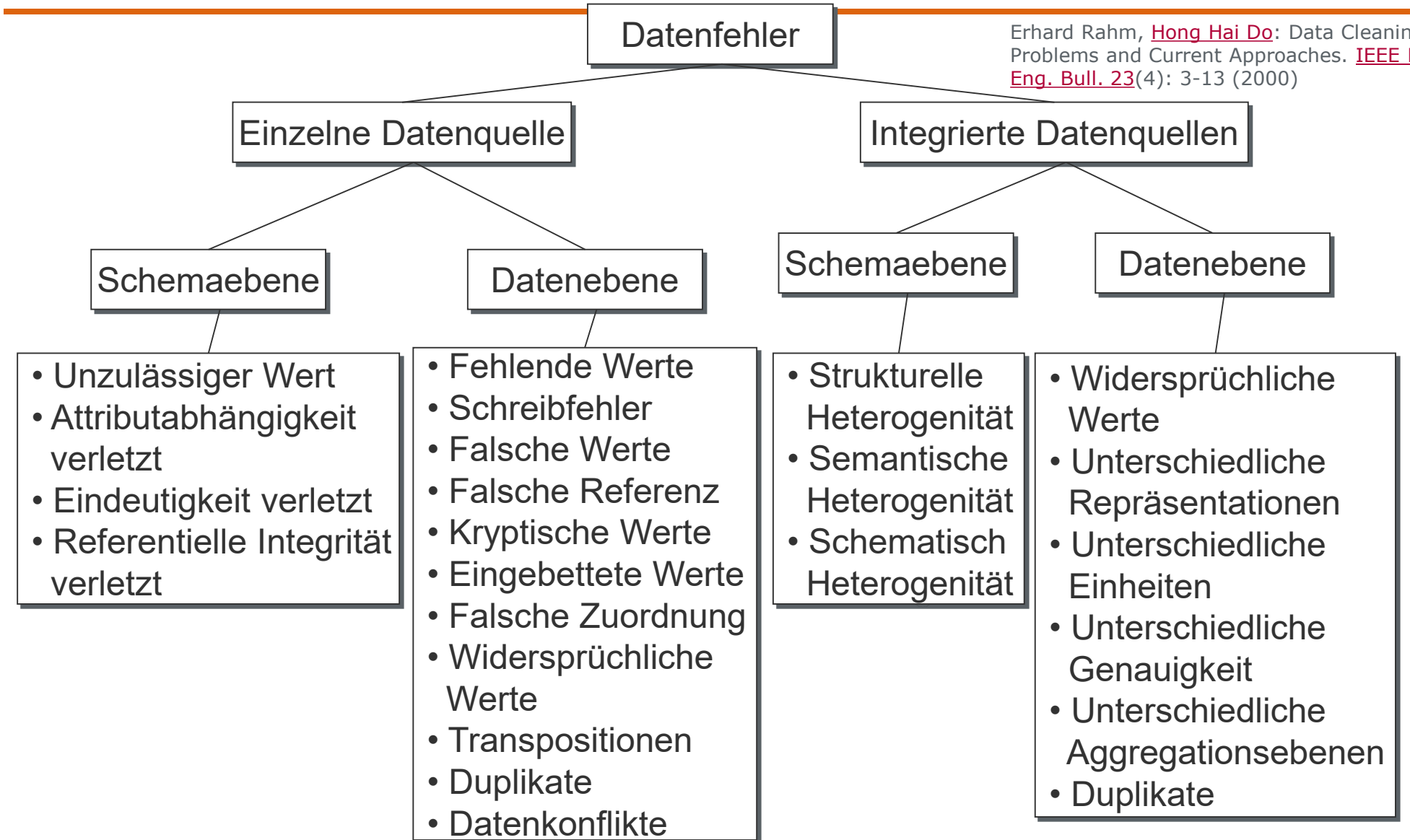
KNr	Name	Birthday	Age	Sex	Phone	ZIP
1234	Kuhn, Mark	18.2.80	37	m	999-9999	98693
1234	Anne Will	32.2.70	37	f	768-4511	55555
1235	Mark Kuhn	18.2.80	27	m	567-3211	98693

ADDRESS

ZIP	Place
98693	Potsdam
98684	Postdam
98766	Germany

Felix Naumann
Information Integration
Winter 2019/20

Classification of Errors



Erhard Rahm, Hong Hai Do: Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull. 23(4): 3-13 (2000)

DQ-Problems: Effects

- Incorrect prices in inventory retail databases [English 1999]
  - Costs for consumers 2.5 billion $
  - 80% of barcode-scan-errors to the disadvantage of consumer
- IRS 1992: almost 100,000 tax refunds not deliverable [English 1999]
- 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. [Strong et al. 1997a]
- US-Postal Service: of 100,000 mass-mailings up to 7,000 undeliverable due to incorrect addresses [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.

Hidden values / hidden value

	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	

Source: Joachim Schmid, FUZZY! Informatik AG

Felix Naumann
Information Integration
Winter 2019/20

Cost of Dirty Data

- A.T. Kearny: 25%-40% of operative costs due to poor DQ.
- Data Warehouse Institute: Industry and administration in US lose 600 billion USD annually.
- SAS study: Only 18% of German companies trust their own data.
- AT&T (70s): 20-30% of all phone lines unused due to poor data.
- 80% of all hospital records contain errors.
- ...

Overview

1. Information quality
2. **IQ criteria**
3. IQ assessment
4. Cleansing tasks
5. IQ anecdotes

Information Quality (IQ)

- What is information quality ?
  - „Fitness for use“
  - “User satisfaction”
  - Application-dependent

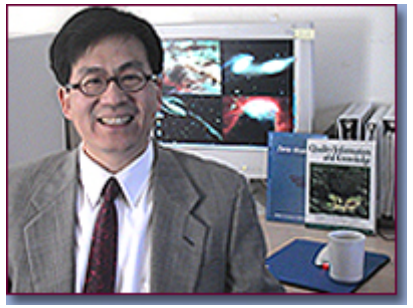
IQ := {Understandability, Reputation, Reliability, Timeliness, Availability, Price, Consistency, Coverage, Response time, Density, Completeness, Amount, Accuracy, Relevancy, ... **}**

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	

IQ Classification of Wang and Strong

- Intrinsic IQ
  - Believability, Accuracy, Objectivity, Reputation
- Contextual IQ
  - Value-added, Relevancy, Timeliness, Completeness, Amount
- Representational IQ
  - Interpretability, Understandability, Repr. Consistency, Repr. conciseness
- Accessibility IQ
  - Accessibility, Security

Wang, R. Y. & Strong, D. M.
Beyond Accuracy: What data quality means to data consumers
Management of Information Systems, 1996*, 12(4)*, 5-34



Content-based IQ Criteria

- …concern the actual data.
- Accuracy
  - is the extent to which data is correct, reliable, and certified free of error. [WS96]
- Completeness
  - is the extent to which data is not missing and is of sufficient breadth, depth, and scope for the task at hand. [WS96]
- Customer support
  - is the amount and usefulness of human help via email or telephone.
- Documentation
  - is the amount and usefulness of documents with metadata.
- Interpretability
  - is the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear. [WS96]
- Relevancy (or relevance)
  - is the extent to which data is applicable and helpful for the task at hand. [WS96]
- Reliability
  - is the degree to which the user can trust the information
- Value-Added
  - is the extent to which data is beneficial and provides advantages from its use. [WS96]

Technical IQ Criteria

- …concern software and hardware.
- Accessibility (or availability)
  - of a DBMS is the probability that a feasible query is correctly answered in a given time range.
  - Is the extent to which data are available or easily and quickly receivable [WS96].
- Latency
  - is the amount of time in seconds from issuing the query until the first data item reaches the user
- Price (cost effectiveness)
  - is the amount of money a user has to pay for a query.
  - is the extent to which the cost of collecting appropriate data is reasonable [WS96].
- Response time
  - measures the delay in seconds between submission of a query by the user and reception of the complete response from the IS.
- Security
  - is the extent to which access to data is restricted appropriately to maintain its security [WS96].
- Timeliness
  - is the extent to which the age of the data is appropriate for the task at hand [WS96].

Intellectual IQ Criteria

- …concern subjective aspects.
- Believability
  - is the extent to which data is regarded as true, real, and credible [WS96].
- Objectivity
  - is the extent to which data is unbiased, unprejudiced, and impartial [WS96].
- Reputation
  - is the extent to which data is trusted or highly regarded in terms of its source or content [WS96].

Instantiation-related IQ Criteria

- …concern the presentation of retrieved data.
- Amount of data
  - is the extent to which the quantity or volume of available data is appropriate [WS96].
- Representational conciseness
  - is the extent to which data is compactly represented without being overwhelming [WS96].
- Representational consistency
  - is the extent to which data is always represented in the same format and are compatible with previous data [WS96].
- Understandability (ease of understanding)
  - is the extent to which data are clear without ambiguity and easily comprehended [WS96].
- Verifiability (traceability, lineage)
  - Is the extent to which data are well documented, verifiable, and easily attributed to a source [WS96].

Overview

1. Information quality
2. IQ criteria
3. **IQ assessment**
4. Cleansing tasks
5. IQ anecdotes

Questionnaires

All items are measured on a 0 to 10 scale where 0 is not at all and 10 is completely. Items labels with "(R)" are reverse coded.

Accessibility. (4 items, Cronbach's Alpha=.92)
- This information is easily retrievable.
- This information is easily accessible.
- This information is easily obtainable.
- This information is quickly accessible when needed.

Appropriate Amount. (4 items, Cronbach's Alpha=.76)
- This information is of sufficient volume for our needs.
- The amount of information does not match our needs. (R)
- The amount of information is not sufficient for our needs. (R)
- The amount of information is neither too much nor too little.

Believability. (4 items, Cronbach's Alpha=.89)
- This information is believable.
- This information is of doubtful credibility. (R)
- This information is trustworthy.
- This information is credible.

Completeness. (6 items, Cronbach's Alpha=.87)
- This information includes all necessary values.
- This information is incomplete. (R)
- This information is complete.
- This information is sufficiently complete for our needs.
- This information covers the needs of our tasks.
- This information has sufficient breadth and depth for our tasks.

http://web.mit.edu/tdqm/www/tdqmpub/AIMQJun02.pdf

IQ-Assessment – New classification of IQ criteria

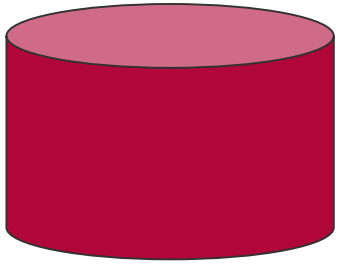
Subject

- Relevance
- Trustworthiness
- Understandability
- ...

Process

- Availability
- Response time
- ...

Object



- Completeness
- Timeliness
- ...

Felix Naumann
Information Integration
Winter 2019/20

Assessment Class	IQ Criterion	Assessment Method
Subject Criteria	Believability	User experience
	Concise representation	User sampling
	Interpretability	User sampling
	Relevancy	Continuous user assessment
	Reputation	User experience
	Understandability	User sampling
	Value-Added	Continuous user assessment
Object Criteria	Completeness	Parsing, sampling
	Customer Support	Parsing, contract
	Documentation	Parsing
	Objectivity	Expert input
	Price	Contract
	Reliability	Continuous assessment
	Security	Parsing
	Timeliness	Parsing
	Verifiability	Expert input
Process Criteria	Accuracy	Sampling, cleansing techniques
	Amount of data	Continuous assessment
	Availability	Continuous assessment
	Consistent representation	Parsing
	Latency	Continuous assessment
	Response time	Continuous assessment

From [NR00]

Overview

1. Information quality
2. IQ criteria
3. IQ assessment
4. **Cleansing tasks**
5. IQ anecdotes

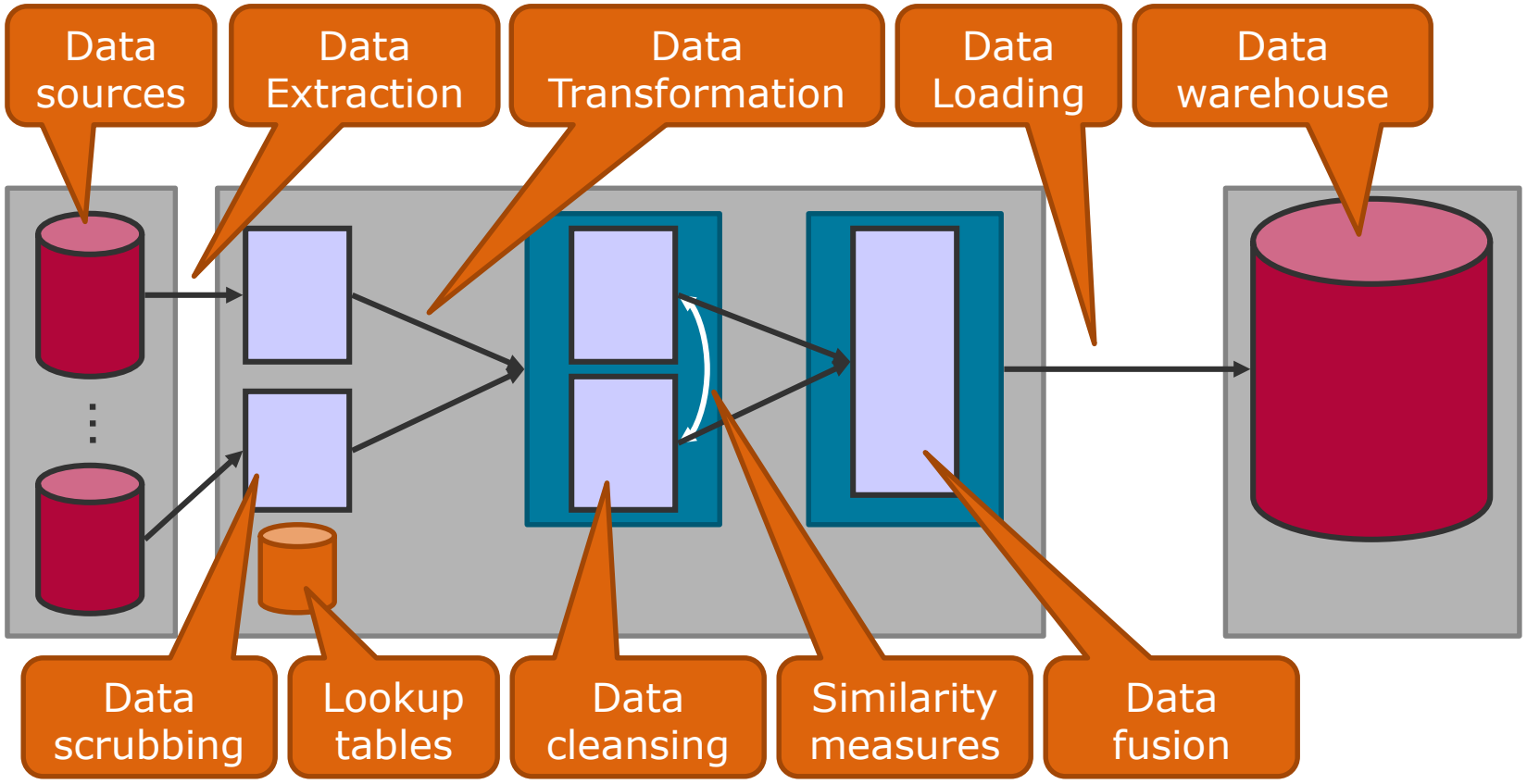
Data Cleansing Tasks – Profiling

- Data source discovery
  - Metadata
  - UDDI / matchmaking
- Schema discovery
  - Schema matching and mapping
  - Profiling for metadata (keys, foreign keys, data types, …)
- Data discovery
  - Column-level: Null-values, domains, patterns, value distributions / histograms
  - Table-level: Data mining, rules

Data Cleansing Tasks – Cleaning

- Extraction from sources
  - Technical and syntactic obstacles
- Transformation
  - Schematic obstacles
- Standardization
  - Syntactic and semantic obstacles
- Duplicate detection
  - Similarity functions
  - Algorithms
- Data fusion / consolidation
  - Semantic obstacles
- Loading into warehouse / presenting to user

Integration and Cleansing: ETL



Felix Naumann
Information Integration
Winter 2019/20

Human Interaction is Needed

- Components to implement
  - Wrappers for technical heterogeneity
  - Schema integration based on correspondences
  - Similarity measure for schema elements
  - Similarity measure for records
- Knobs to turn
  - Thresholds for similarity measures
  - Partition size / window size
- Expert guidance
  - Rule selection / rule specification
  - Schema matching
  - Duplicate detection
  - Data fusion

Overview

1. Information quality
2. IQ criteria
3. IQ assessment
4. Cleansing tasks
5. **IQ anecdotes**

Death by Typo

‘Resurrected,’ but still wallowing in red tape

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

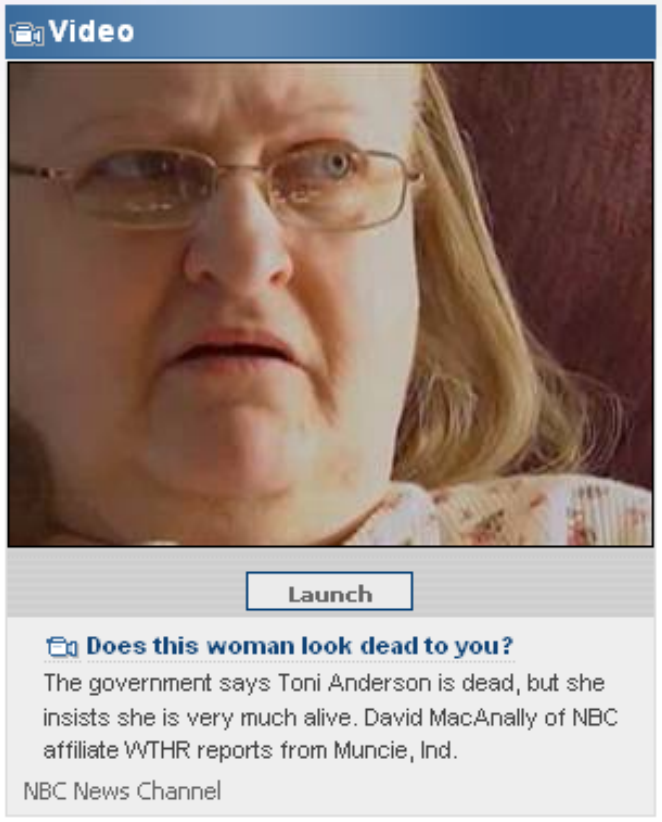
By Alex Johnson and Nancy Amons
Reporters
MSNBC and NBC News
updated 6:21 p.m. ET Feb. 29, 2008

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 ↓



Felix Naumann
Information Integration
Winter 2019/20

Decimals

SPIEGEL ONLINE

28. Januar 2008, 11:27 Uhr

FRANKREICH

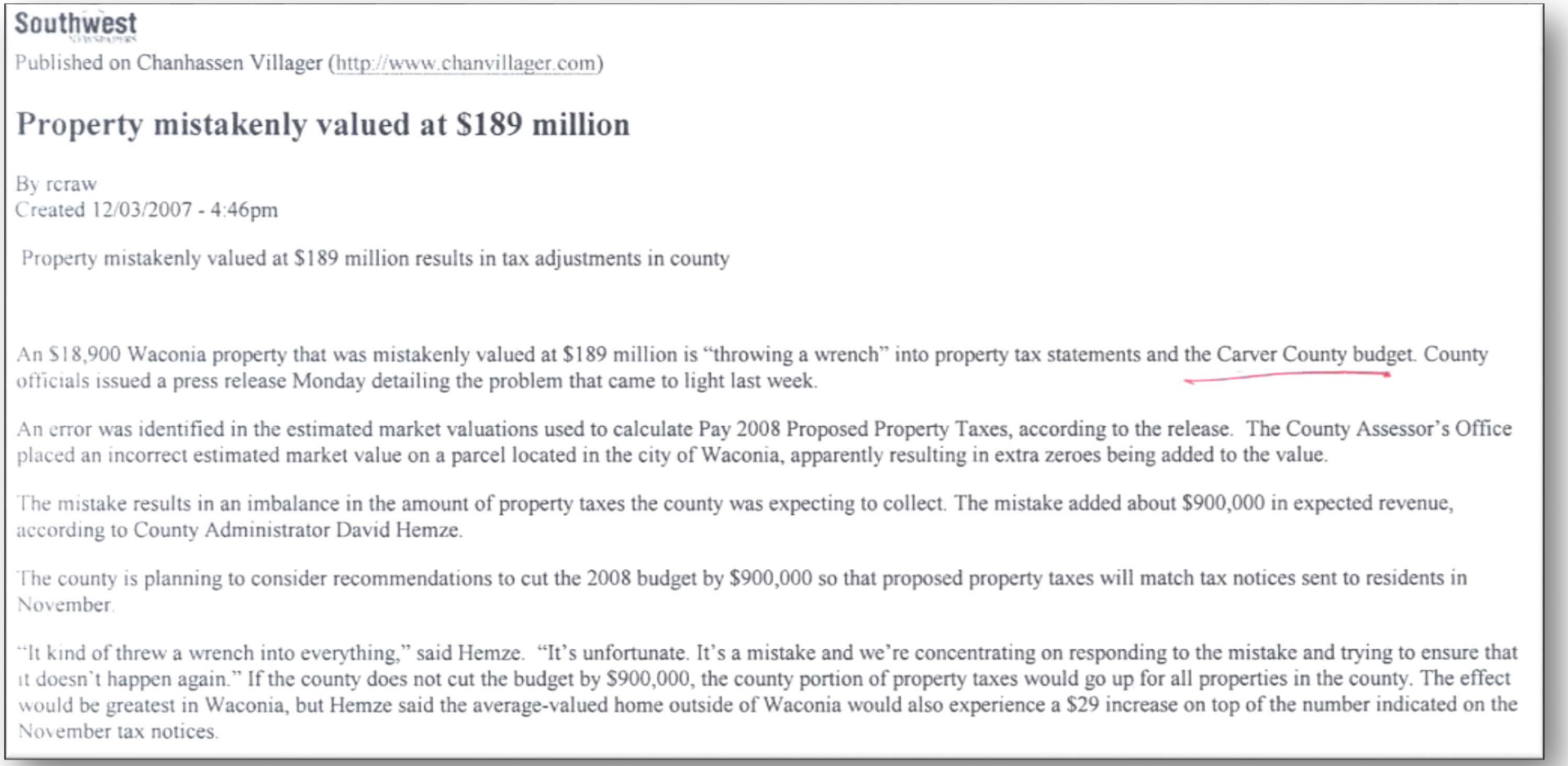
Telefonkundin erhält Rechnung über 63 Millionen Euro

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

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.

Felix Naumann
Information Integration
Winter 2019/20

Google searches for Britney Spears

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

488941 britney spears
40134 brittany spears
36315 brittney spears
24342 britany spears
7331 britny spears
6633 briteny spears
2696 britteny spears
1807 briney spears
1635 brittny spears
1479 brintey spears
1479 britanny spears
1338 britiny spears
1211 britnet spears
1096 britiney spears
991 britaney spears
991 britnay spears
811 brithney spears
811 brtiney spears
664 birtney spears
664 brintney spears
664 briteney spears
601 bitney spears
601 brinty spears
544 brittaney spears
544 brittnay spears
364 britey spears
364 brittiny spears
329 brtney spears
269 bretney spears
269 britneys spears
244 britne spears
244 brytney spears
220 breatney spears
220 britiany spears
199 britnney spears
163 britnry spears
147 breatny spears
147 brittiney spears
147 britty spears

29 britent spears
29 brittnany spears
29 britttany spears
29 btiney spears
26 birttney spears
26 breitney spears
26 brinity spears
26 britenay spears
26 britneyt spears
26 brittan spears
26 brittne spears
26 btittany spears
24 beitney spears
24 birteny spears
24 brightney spears
24 brintiny spears
24 britanty spears
24 britenny spears
24 britini spears
24 britnwy spears
24 brittni spears
24 brittnie spears
21 biritney spears
21 birtany spears
21 biteny spears
21 bratney spears
21 britani spears
21 britanie spears
21 briteany spears
21 brittay spears
21 brittinay spears
21 brtany spears
21 brtiany spears
19 birney spears
19 brirtney spears
19 britnaey spears
19 britnee spears
19 britony spears
19 brittanty spears

9 brinttany spears
9 britanay spears
9 britinany spears
9 britn spears
9 britnew spears
9 britneyn spears
9 britrney spears
9 brtiny spears
9 brtittney spears
9 brtny spears
9 brytny spears
9 rbitney spears
8 birtiny spears
8 bithney spears
8 brattany spears
8 breitny spears
8 breteny spears
8 brightny spears
8 brintay spears
8 brinttey spears
8 briotney spears
8 britanys spears
8 britley spears
8 britneyb spears
8 britnrey spears
8 britnty spears
8 brittner spears
8 brottany spears
7 baritney spears
7 birntey spears
7 biteney spears
7 bitiny spears
7 breateny spears
7 brianty spears
7 brintye spears
7 britianny spears
7 britly spears
7 britnej spears
7 britneyu spears

5 brney spears
5 broitney spears
5 brotny spears
5 bruteny spears
5 btiyney spears
5 btrittney spears
5 gritney spears
5 spritney spears
4 bittny spears
4 bnritney spears
4 brandy spears
4 brbritney spears
4 breatiny spears
4 breetney spears
4 bretiney spears
4 brfitney spears
4 briattany spears
4 brieteny spears
4 briety spears
4 briitny spears
4 briittany spears
4 brinie spears
4 brinteney spears
4 brintne spears
4 britaby spears
4 britaey spears
4 britainey spears
4 britinie spears
4 britinney spears
4 britmney spears
4 britnear spears
4 britnel spears
4 britneuy spears
4 britnewy spears
4 britnmey spears
4 brittaby spears
4 brittery spears
4 britthey spears
4 brittnaey spears

3 britiy spears
3 britmeny spears
3 britneeey spears
3 britnehy spears
3 britnely spears
3 britnesy spears
3 britnetty spears
3 britnex spears
3 britneyxxx spears
3 britnity spears
3 britntey spears
3 britnyey spears
3 britterny spears
3 brittneey spears
3 brittnney spears
3 brittnyey spears
3 brityen spears
3 briytney spears
3 brltney spears
3 broteny spears
3 brtaney spears
3 brtiiany spears
3 brtinay spears
3 brtinney spears
3 brtitany spears
3 brtiteny spears
3 brtnet spears
3 brytiny spears
3 btney spears
3 drittney spears
3 pretney spears
3 rbritney spears
2 barittany spears
2 bbbritney spears
2 bbitney spears
2 bbritny spears
2 bbrittany spears
2 beitany spears
2 beitny spears

2 brirreny spears
2 brirtany spears
2 brirttany spears
2 brirttney spears
2 britain spears
2 britane spears
2 britaneny spears
2 britania spears
2 britann spears
2 britanna spears
2 britannie spears
2 britannt spears
2 britannu spears
2 britanyl spears
2 britanyt spears
2 briteeny spears
2 britenany spears
2 britenet spears
2 briteniy spears
2 britenys spears
2 britianey spears
2 britin spears
2 britinary spears
2 britmy spears
2 britnaney spears
2 britnat spears
2 britnbey spears
2 britndy spears
2 britneh spears
2 britneney spears
2 britney6 spears
2 britneye spears
2 britneyh spears
2 britneym spears
2 britneyyy spears
2 britnhey spears
2 britnjey spears
2 britnne spears
2 britnu spears

Felix Naumann
Information Integration
Winter 2019/20

Duplicate announcement in same newspaper on same day

KURZ & KNAPP

Musikschulfest in Kleinmachnow

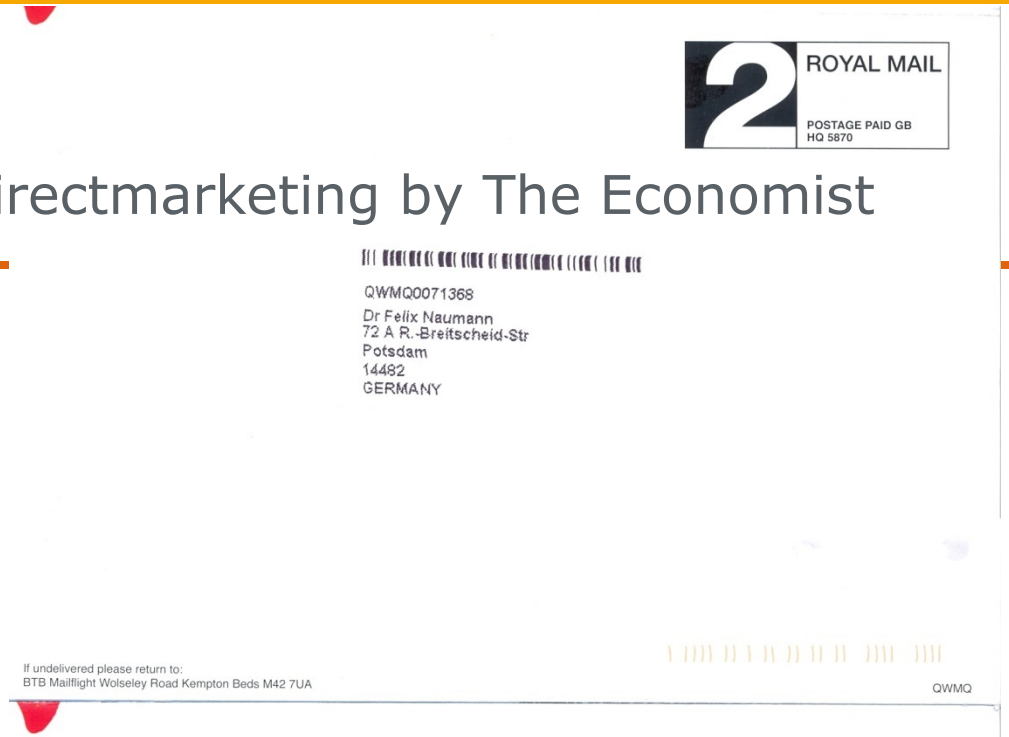
KLEINMACHNOW | Ein Musikschulfest veranstaltet die Kreismusikschule Potsdam-Mittelmark am Samstag, dem 19. Juni, in ihrem Haus in Kleinmachnow, Am Weinberg 20. Von 14 bis 20 Uhr werden Konzerte der Fachbereiche auf der Freilichtbühne, „Schnupperstunden" der Schule, das Quiz „Musikalischer Irrgarten" und vieles mehr geboten.

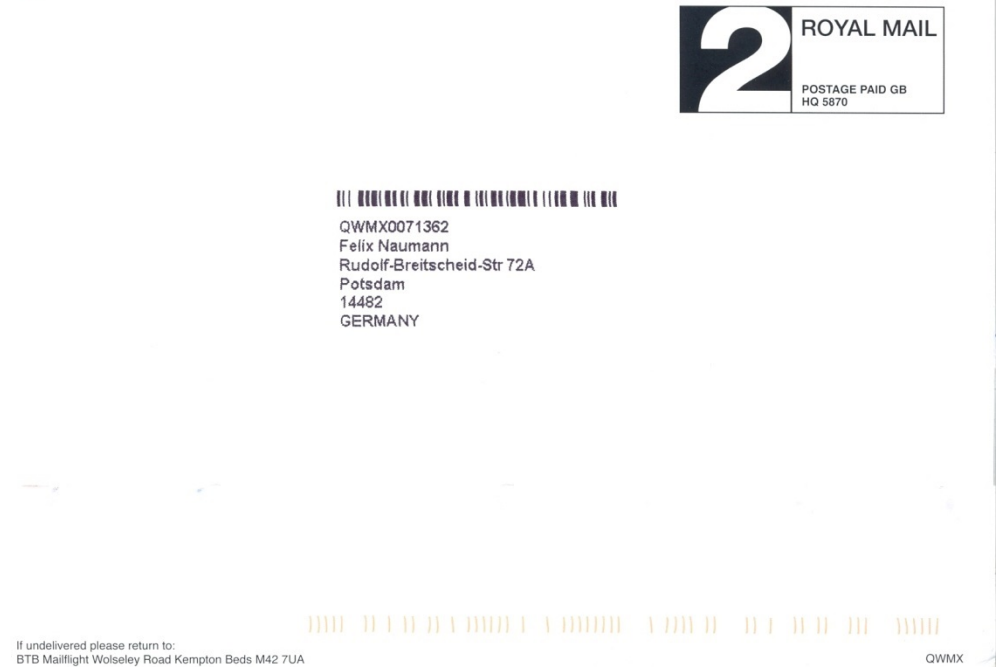
KURZ & KNAPP

Musikschule lädt ein

KLEINMACHNOW | Die Kreismusikschule „Engelbert Humperdinck" lädt für den kommenden Samstag, 19. Juni, zum diesjährigen Musikschulfest nach Kleinmachnow ein. Auf einer Freilichtbühne, in der Musikschule (Am Weinberg) und in der Aula des Weinberggymnasiums erwartet die Besucher von 14 bis 18 Uhr ein vielseitiges musikalisches Programm. Orchester, Ensembles, Bands und Solisten zeigen ihr Können. Um 15 Uhr wird „Alte Musik" vorgestellt, ab 15.30 Uhr geht es um die „Faszination Schlagzeug". Für das leibliche Wohl ist gesorgt.

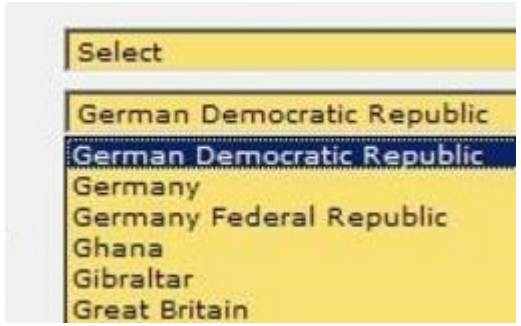
Directmarketing by The Economist





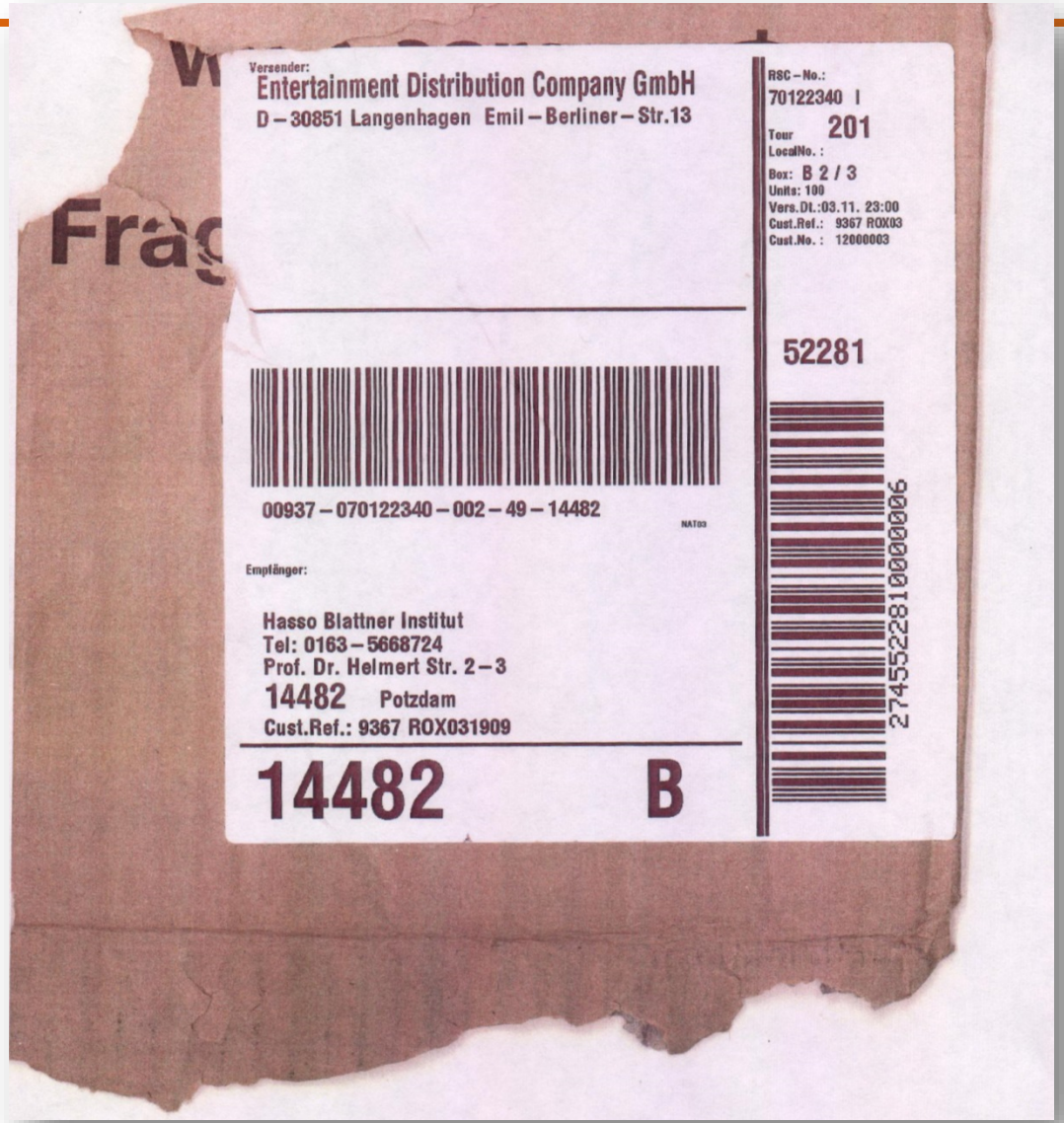
Felix Naumann
Information Integration
Winter 2019/20

FIFA registration form (2010)



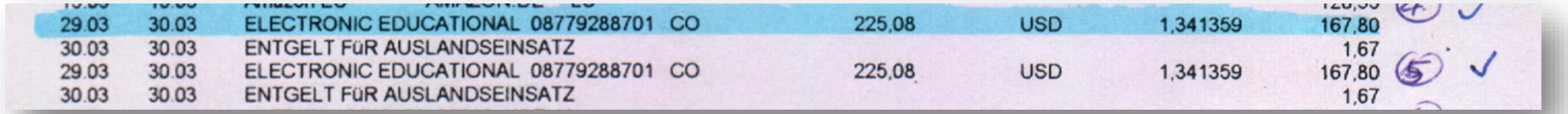
Felix Naumann
Information Integration
Winter 2019/20

Hasso Blattner Institut, Potzdam



Felix Naumann
Information Integration
Winter 2019/20

Fraudulent Transaction Duplicate



29.03	30.03	ELECTRONIC EDUCATIONAL 08779288701	CO	225,08	USD	1,341359	167,80
30.03	30.03	ENTGELT FüR AUSLANDSEINSATZ					1,67
29.03	30.03	ELECTRONIC EDUCATIONAL 08779288701	CO	225,08	USD	1,341359	167,80
30.03	30.03	ENTGELT FüR AUSLANDSEINSATZ					1,67

Felix Naumann
Information Integration
Winter 2019/20

Product Duplicate

1.

Acer P 5370 W DLP-Projektor HD-Ready (Kontrast 2000:1, 3000 ANSI Lumen, WXGA 1280 x 800 Pixel) schwarz von Acer (9. Januar 2009)

Neu kaufen: **EUR 729,00** 23 Angebote ab EUR 729,00

Lieferung bis **Dienstag, 5. Januar**: Bestellen Sie innerhalb der nächsten **1 Stunde** per Overnight-Express.

(1) Prime

Elektronik & Foto: Alle 6 Artikel ansehen

2.

Acer P5370W HD-Ready Projektor HDMI (WXGA, 1280x800, 3000 ANSI Lumen, 2400 ANSI Lumen Eco-Mode, Kontrast: 2000:1) schwarz von Acer (9. Januar 2009)

Neu kaufen: **EUR 729,00** 2 Angebote ab EUR 666,40

Auf Lager.

(2)

Elektronik & Foto: Alle 6 Artikel ansehen

Felix Naumann
Information Integration
Winter 2019/20

CD Duplicate

1.
Dile Al Sol von La Oreja de Van Gogh (**Audio CD** - 2007)
Neu kaufen: **EUR 25,99** 20 Angebote ab EUR 4,95
Nicht auf Lager. Bestellen Sie jetzt und wir liefern, sobald der Artikel verfügbar ist
Prime

2.
Dile Al Sol von La Oreja de Van Gogh (**Audio CD** - 1999)
1 Angebote ab EUR 4,99

3.
Dile Al Sol von La Oreja de Van Gogh (**Audio CD** - 2007)
Neu kaufen: ~~EUR 16,99~~ **EUR 15,97** 8 Angebote ab EUR 6,46
Lieferung bis **Dienstag, 3. Juni**: Bestellen Sie innerhalb der nächsten **7 Stunden** per Overnight-Express.
Prime

Felix Naumann
Information Integration
Winter 2019/20

Reviewer Complaint (german)

VERGRÖSSERN
Zoom und weitere Ansichten

Für Kunden: Stellen Sie Ihre eigenen Bilder ein.

Das neue Filmlexikon 2009 (DVD-ROM)
von United Soft Media Verlag GmbH

Plattform: Windows XP / Vista

(1 Kundenrezension)

Preis: EUR 30,95 & kostenlose Lieferung mit **Amazon Prime**.

Auf Lager.

1 von 1 Kunden fanden die folgende Rezension hilfreich:

Redaktionell lieblos erweitert, 22. Januar 2009

Von **Thomas Georg Maria Mainka** - Alle meine Rezensionen ansehen

Zuerst das positive: Gegenüber der 2006er Ausgabe ist die Software deutlich stabiler geworden. Abstürze, die man mit Affengriff und Taskmanager beseitigen muß, kommen glücklicherweise nicht mehr vor.

Ansonsten hat sich an der Software so gut wie nichts geändert.

Redaktionell mußte ich jedoch feststellen, daß die Anzahl an Doubletten (also Doppeleinträge ein und des selben Films mit geringfügigen abweichenden Daten wie Laufzeitunterschiede von 1 Minute, oder 35mm ./. Video bzw. teilweise unvollständigen Darsteller- bzw. Filmstab-Listen) deutlich zugenommen haben. (Zumindestens ist mir dies bei den Märchenfilmen - meinem privaten Sondersammelgebiet - besonders aufgefallen)

German Umlaute

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

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

Real-world false positive

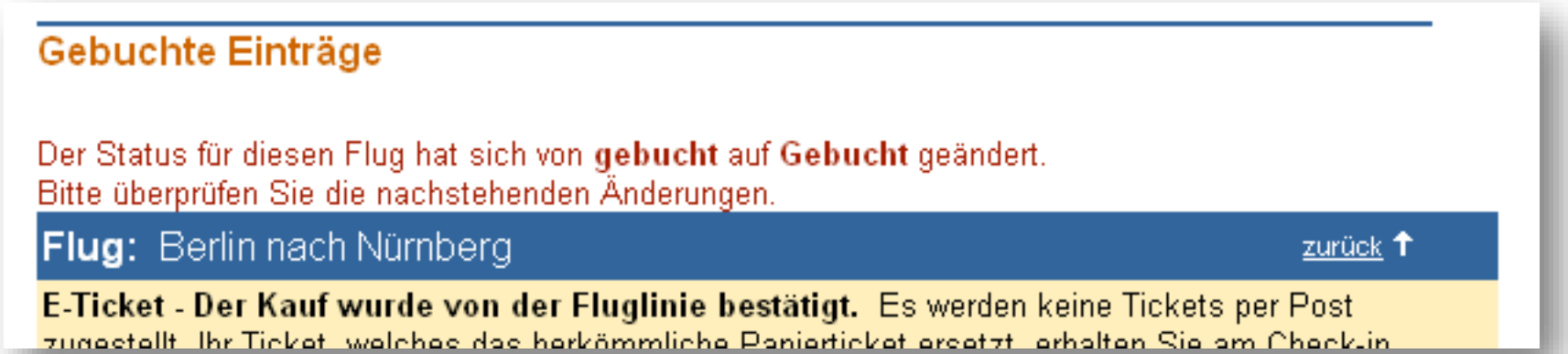
Die doppelte Gabi!

Links Gabi, geboren am 10. August in Osterburg 1957, wohnt jetzt in Dresden, verheiratet, 2 Söhne (25, 27) und rechts Gabi, geboren am 10. August 1957, wohnt in Dresden, verwitwet, 2 Töchter (33, 34)

Foto: Dirk Sukow

Felix Naumann
Information Integration
Winter 2019/20

Status change
from „booked“ to „Booked“

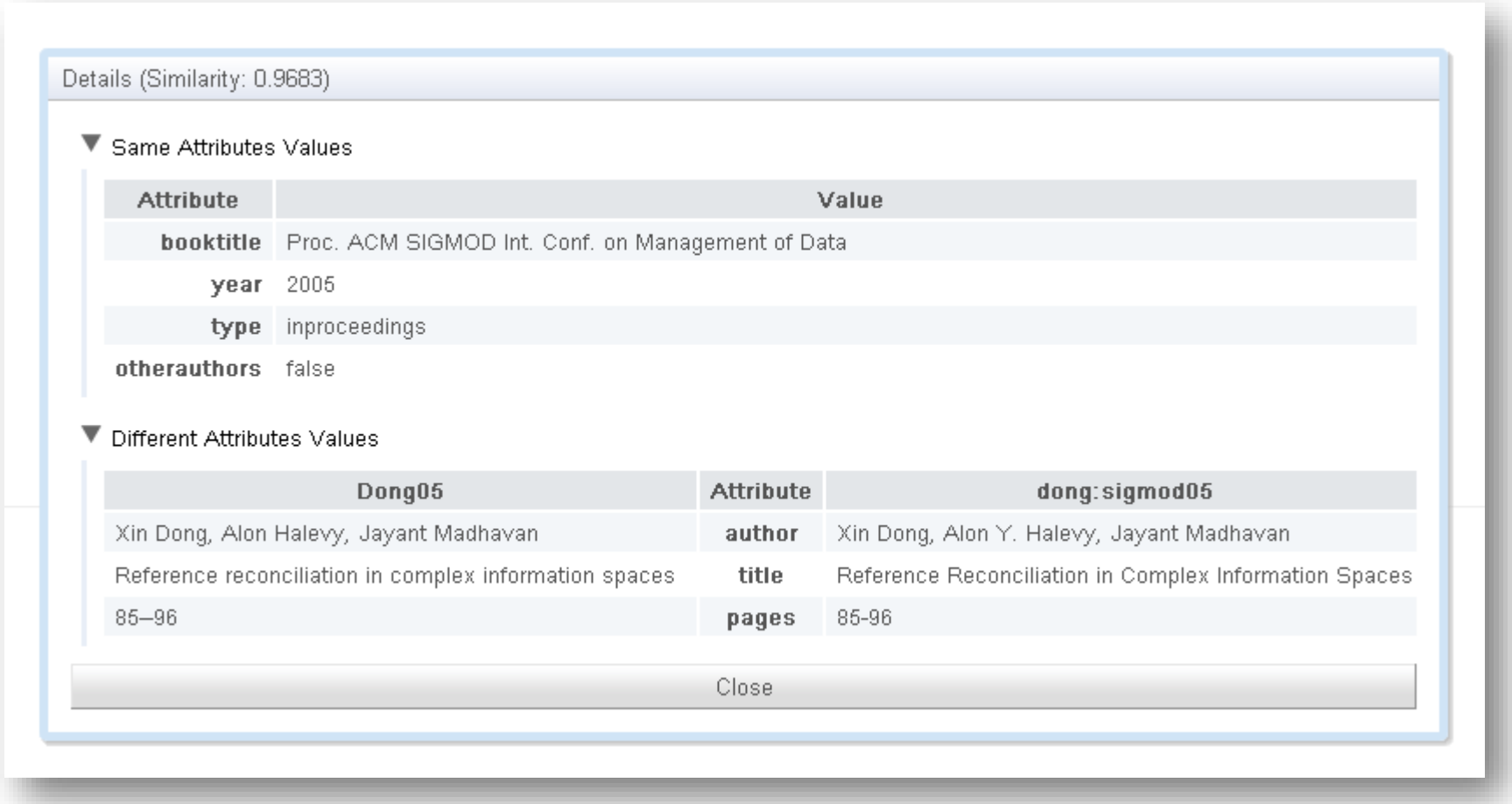


Felix Naumann
Information Integration
Winter 2019/20

Manual Data Fusion

Mary: Hi, my name is Mary. How may I help you?
Felix Naumann: Hi, I am Felix
Felix Naumann: I am area editor of "Information System".
Mary: Hello Felix, how can I help you today?
Felix Naumann: I would like to assign a reviewer to a submission. The reviewer is already registered twice in the system. Once with an old email adress, and once with an incorrect name. The person is Barbara Pernici.
Mary: Ok, did you want me to merge her accounts?
Felix Naumann: Sure, merging her accounts would be a first step. The correct email address is barbara.pernici@polimi.it and of course her name is "Barbara Pernici" and not "Barbara Barbara Pernici".
Felix Naumann: Can you let me know, once the files are corrected. Thansk for your help!
Mary: Her name has been corrected and the accounts merged.
Mary: Is there anything else I can help you with?
Felix Naumann: No, perfect. Thanks, Mary.
Mary: You are welcome Felix.
Mary: It was a pleasure assisting you today. If you require any further assistance please contact us again. To close this chat session, please click the Close button in the top corner of this chat window.

An Ironic Duplicate



Felix Naumann
Information Integration
Winter 2019/20

Translation problems

LIVE BBC NEWS CHANNEL

Page last updated at 11:45 GMT, Thursday, 19 February 2009

E-mail this to a friend Printable version

The mystery of Ireland's worst driver

Details of how police in the Irish Republic finally caught up with the country's most reckless driver have emerged, the Irish Times reports.

Poles are Ireland's largest immigrant population

He had been wanted from counties Cork to Cavan after racking up scores of speeding tickets and parking fines.

However, each time the serial offender was stopped he managed to evade justice by giving a different address.

But then his cover was blown.

It was discovered that the man every member of the Irish police's rank and file had been looking for - a Mr Prawo Jazdy - wasn't exactly the sort of prized villain whose apprehension leads to an officer winning an award.

In fact he wasn't even human.

"*Prawo Jazdy* is actually the Polish for driving licence and not the first and surname on the licence," read a letter from June 2007 from an officer working within the Garda's traffic division.

"Having noticed this, I decided to check and see how many times officers have made this mistake.

"It is quite embarrassing to see that the system has created Prawo Jazdy as a person with over 50 identities."

The officer added that the "mistake" needed to be rectified immediately and asked that a memo be circulated throughout the force.

In a bid to avoid similar mistakes being made in future relevant guidelines were also amended.

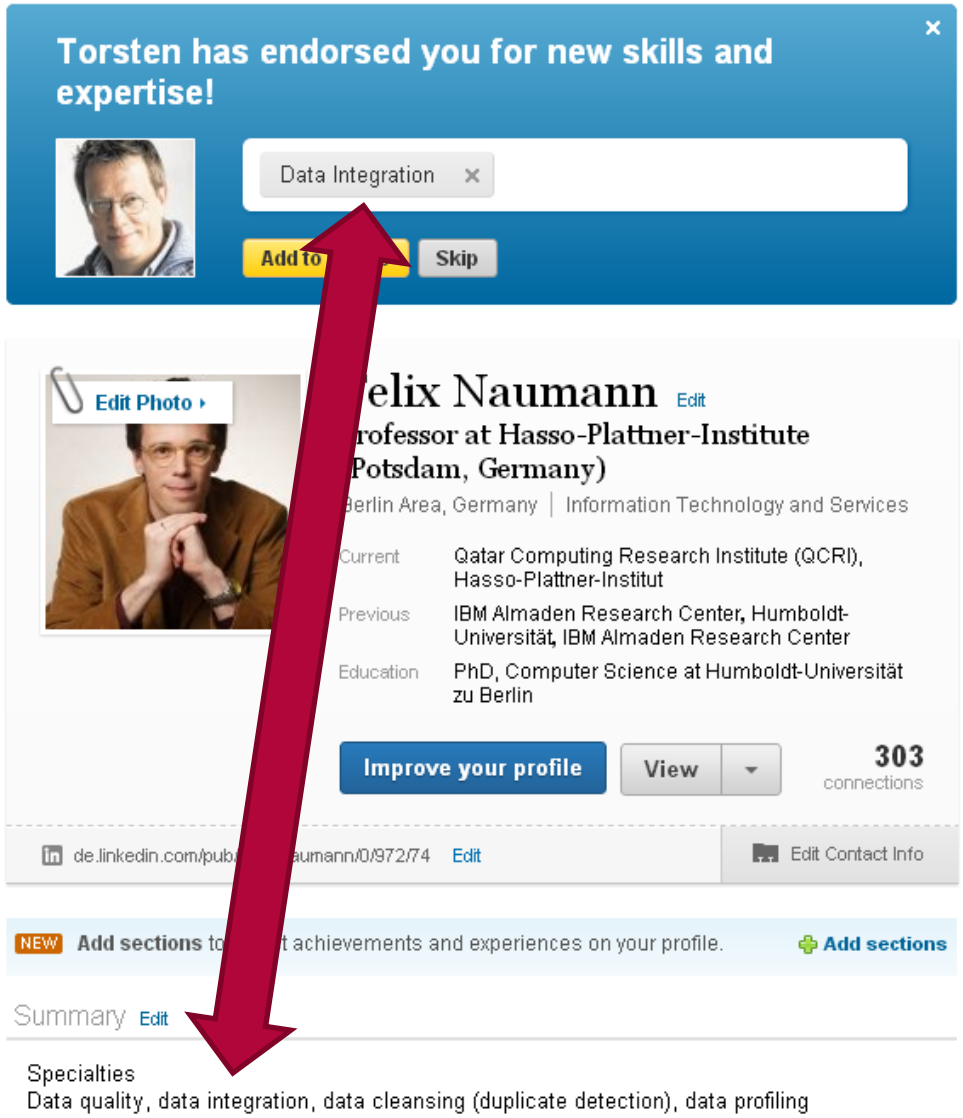
And if nothing else is learnt from this driving-related debacle, Irish police officers should now know at least two words of Polish.

As for the seemingly elusive Mr Prawo Jazdy, he has presumably become a cult hero among Ireland's largest immigrant population.

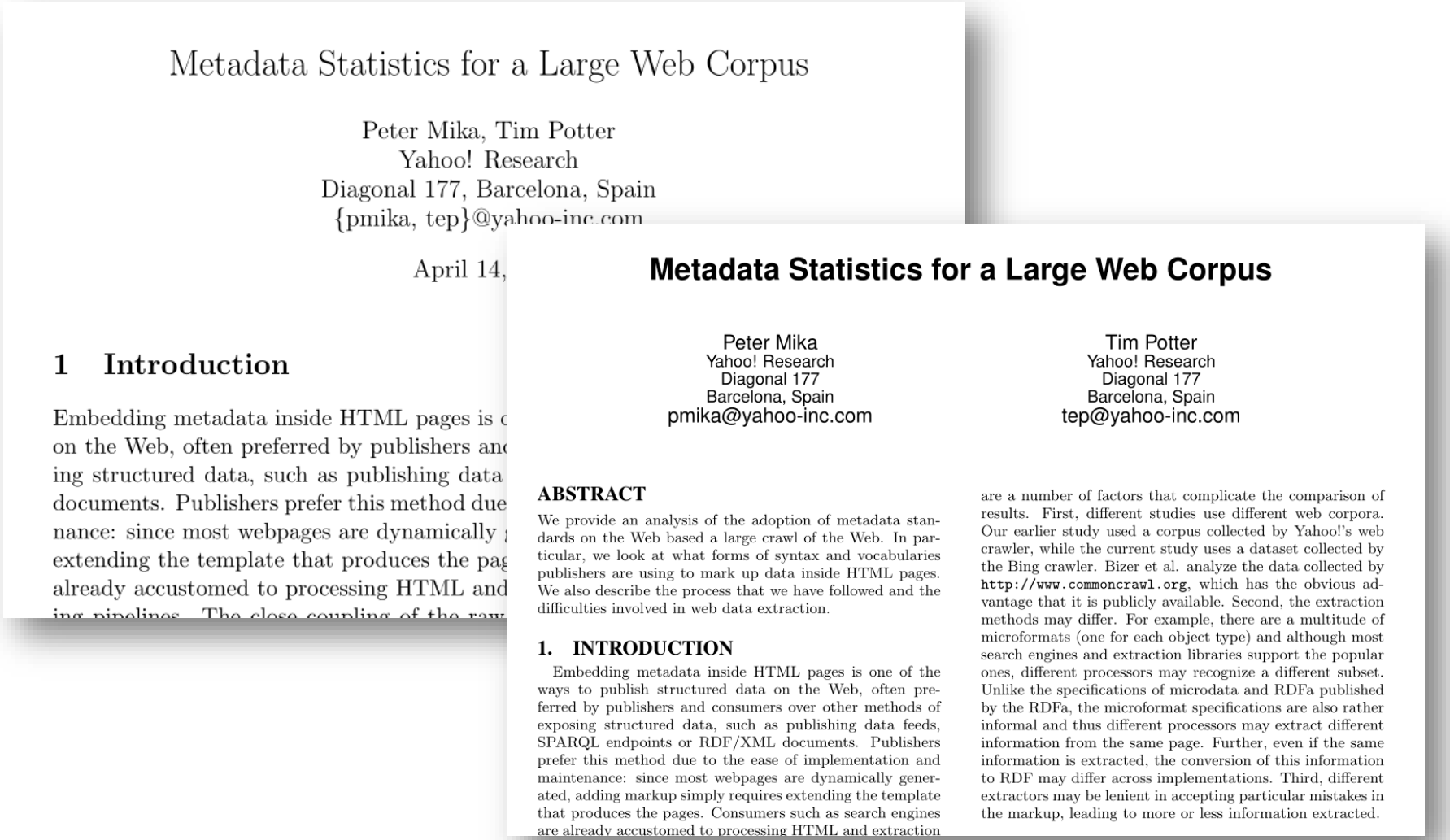
Felix Naumann
Information Integration
Winter 2019/20

5 Participants =>
5 Affiliations

Skill duplicate



Two different(!) LDOW 2012 papers



Metadata Statistics for a Large Web Corpus

Peter Mika, Tim Potter
Yahoo! Research
Diagonal 177, Barcelona, Spain
{pmika, tep}@yahoo-inc.com

April 14,

1 Introduction

Embedding metadata inside HTML pages is c
on the Web, often preferred by publishers and
ing structured data, such as publishing data
documents. Publishers prefer this method due
nance: since most webpages are dynamically
extending the template that produces the pag
already accustomed to processing HTML and
ing pipelines. The close coupling of the raw

Metadata Statistics for a Large Web Corpus

Peter Mika
Yahoo! Research
Diagonal 177
Barcelona, Spain
pmika@yahoo-inc.com

Tim Potter
Yahoo! Research
Diagonal 177
Barcelona, Spain
tep@yahoo-inc.com

ABSTRACT

We provide an analysis of the adoption of metadata standards on the Web based a large crawl of the Web. In particular, we look at what forms of syntax and vocabularies publishers are using to mark up data inside HTML pages. We also describe the process that we have followed and the difficulties involved in web data extraction.

1. INTRODUCTION

Embedding metadata inside HTML pages is one of the ways to publish structured data on the Web, often preferred by publishers and consumers over other methods of exposing structured data, such as publishing data feeds, SPARQL endpoints or RDF/XML documents. Publishers prefer this method due to the ease of implementation and maintenance: since most webpages are dynamically generated, adding markup simply requires extending the template that produces the pages. Consumers such as search engines are already accustomed to processing HTML and extraction

are a number of factors that complicate the comparison of results. First, different studies use different web corpora. Our earlier study used a corpus collected by Yahoo!'s web crawler, while the current study uses a dataset collected by the Bing crawler. Bizer et al. analyze the data collected by `http://www.commoncrawl.org`, which has the obvious advantage that it is publicly available. Second, the extraction methods may differ. For example, there are a multitude of microformats (one for each object type) and although most search engines and extraction libraries support the popular ones, different processors may recognize a different subset. Unlike the specifications of microdata and RDFa published by the RDFa, the microformat specifications are also rather informal and thus different processors may extract different information from the same page. Further, even if the same information is extracted, the conversion of this information to RDF may differ across implementations. Third, different extractors may be lenient in accepting particular mistakes in the markup, leading to more or less information extracted.

Felix Naumann
Information Integration
Winter 2019/20

Bewerbung für das Wintersemester 09/10

Angaben zur Person

Nachname:	[redacted]
Vorname:	[redacted]
Geschlecht	männlich
Titel:	
Namenszusatz:	
Geburtsdatum:	24.07.1991

Hinweis: Ihr Online-Bewerbungssystem erfordert es, dass ich mich als ein Jahr älter ausgebe, also ich bin ("Das Mindestalter für den Hochschulzugang beträgt: 17"). Mein eigentliches Geburtsdatum ist der 24.7.1992, sodass ich zu Studienbeginn bereits 17 Jahre alt sein werde. Diese Änderung wurde mir durch das Studierendensekretariat empfohlen

Geburtsort:	Erlangen
Geburtsname:	
Staatsangehörigkeit:	Deutschland

Me – a duplicate

Dear CMT Support,

I noticed that the April/May Revision tracks have two records for meta reviewer, Felix Naumann. I tried to delete them and add him again, but each time the system produced two entries for him. This problem doesn't exist for other tracks. Can you look into the reason? Thank you.

Chen

Felix	Naumann	Felix.Naumann@hpi.uni-potsdam.de	Hasso Plattner Institut Potsdam	**Primary:** Information Integration\Information Integration **Secondary:** Information Integration\Data Cleansing and Data Profiling; Information Integration\Heterogeneous and Federated DBMS; Information Integration\Information Filtering and Dissemination; Information Integration\Meta-data Management; Information Integration\Other Integration or Web; Information Integration\Social Systems and Recommendations; Information Integration\Web Data Management; Information Integration\Web Services and Web Service Composition;	No	No	0	[X]	[X]
Felix	Naumann	Felix.Naumann@hpi.uni-potsdam.de	Hasso Plattner Institut Potsdam	**Primary:** Information Integration **Secondary:** Information Integration\Data Cleansing and Data Profiling; Information Integration\Heterogeneous and Federated DBMS; Information Integration\Information Filtering and Dissemination; Information Integration\Meta-data Management; Information Integration\Other Integration or	No	No	0	[X]	

Weitere Informationen zu VLDB2015 Chairs anzeigen.

JDIQ author list by ACM

real-world applications, datasets and other experiences in handling poor

Ramamurthy, Karthikeyan ; Mojsilovic, Aleksandra

Top Cited Articles | Metrics | Award Winners | Authors | Affiliations | JDIQ Archive

Filter By:

First Name	Last Name	Paper Counts
Maurice	Mulvenna	1
Heiko	Müller	1
Felix	Naumann	1
Felix	Naumann	1
H	Nehemiah	1
Eric	Nelson	1
Benjamin	Ngugi	1
Jay	Nunamaker,	1
Rabia	Nuray-Turan	1
Boris	Otto	1
John	O'Donoghue	1
Fabian	Panse	1
Jeffrey	Parsons	1