# Beyond Information Integration: Content Merging

## Forschungskolloquium CIS/ISST

Felix Naumann
naumann@informatik.hu-berlin.de

## Einige Untertitel

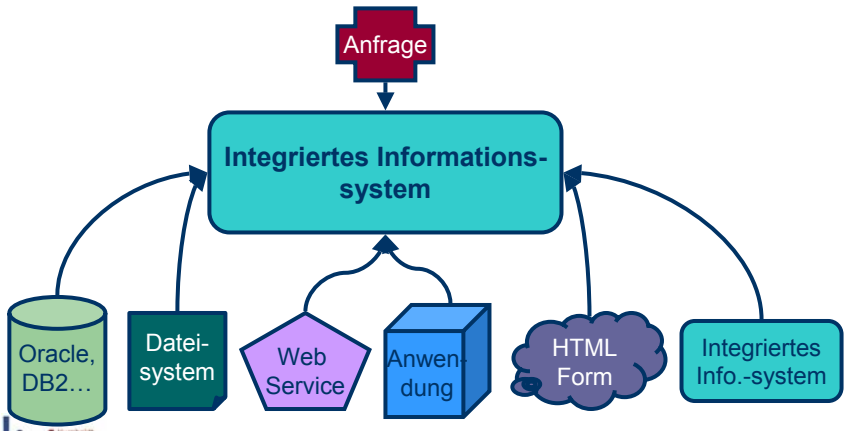
- Content Merging
- Objekt/Data Fusion
- Data Amalgamation
- Data Consolidation
- Intelligent Information Integration
- Data Cleansing
- Datenintegration
- Datenverschmelzung

# Forschungsgruppe Informationsintegration

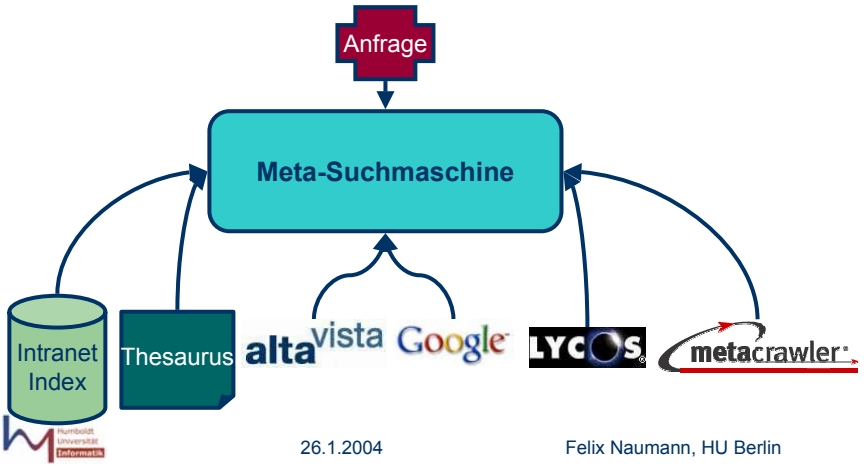
- Institut für Informatik der Humboldt-Universität zu Berlin
- Juniorprofessor: Felix Naumann
- Mitarbeiter
  - Jens Bleiholder & Melanie Weis
- Forschungsthemen
  - Objektidentifikation
  - Informationsintegration
  - Optimierung
  - Visualisierung



# Integrierte Informationssysteme



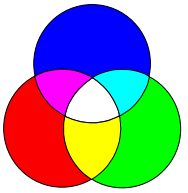
# Integrierte Suchmaschinen



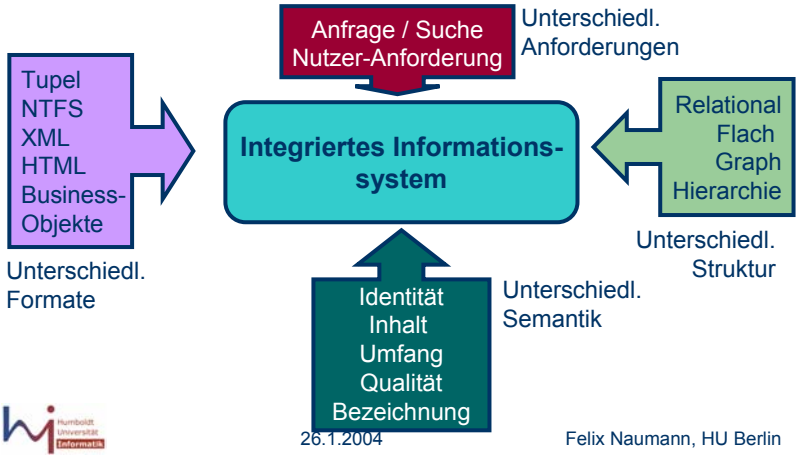



# Redundanz – pro & contra

- Ist gut, denn sie bietet
  - mehr Informationen,
  - detaillierter Informationen,
  - und verifizierbare Informationen.
  - Deshalb sollten wir integrieren!
- Ist problematisch, denn
  - Redundanz herrscht nur konzeptionell.
  - Technische und strukturelle Schwierigkeiten
  - Konflikte und Alternativen
  - Deshalb ist Informationsintegration interessant.

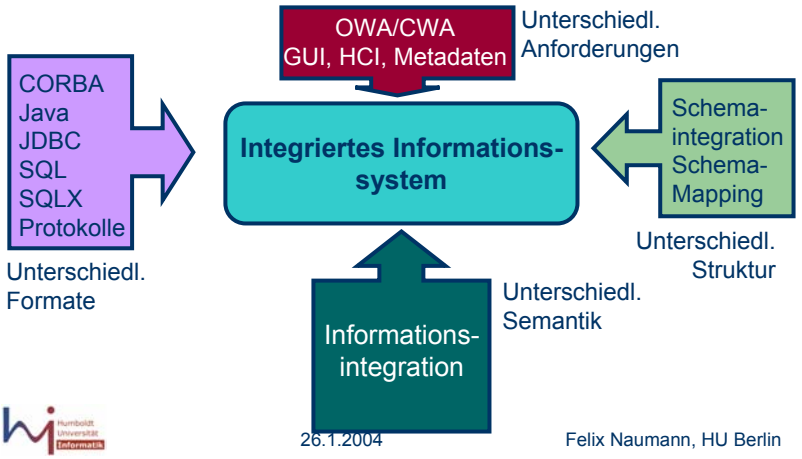


# Integrierte Informationssysteme

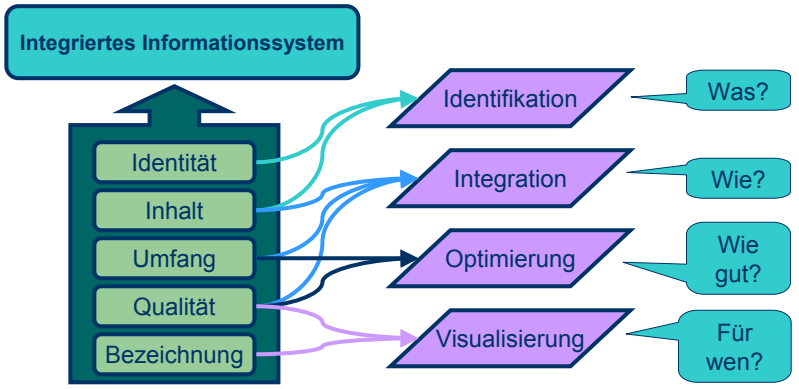





# Integrierte Informationssysteme

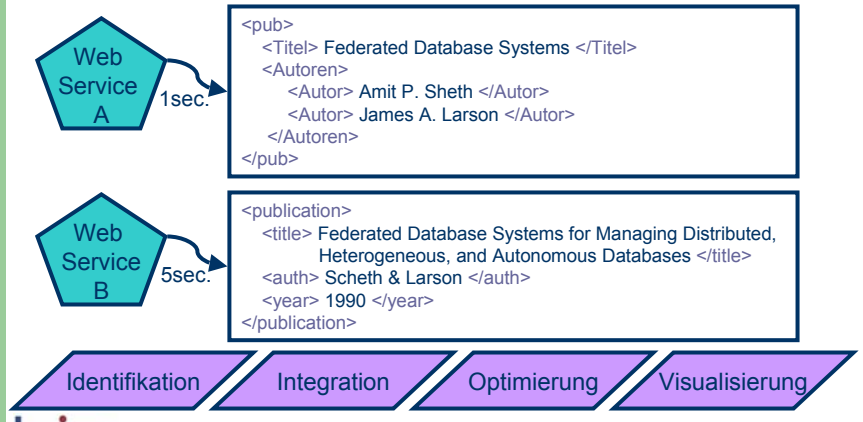


# Lösung Unterschiedlicher Semantik






# Beispiel der Informationsintegration



# Objektidentifikation

Identifikation | Integration | Optimierung | Visualisierung

- Erkennung mehrere (ähnlicher) Darstellungen des gleichen Objekts
- Auch
  - Duplicate Detection
  - Data Cleansing
  - Record Linkage
- Domänenspezifische Algorithmen
  - Adressdaten
  - Mikrobiologische Daten
- Manchmal gibt es IDs
  - Bücher: ISBN
  - Personen: SSN / Personalausweisnummer
  - Webseiten: URL



# Objektidentifikation

Identifikation | Integration | Optimierung | Visualisierung

- Das klassische Problem
  - Finde Duplikate innerhalb einer Tabelle.
  - Sehr große Datenmenge
    - kein quadratischer Algorithmus
    - kein Hauptspeicher-Algorithmus
- Forschung
  - Merge/Purge Technik (Hernandez & Stolfo 1998)
    - Sorted-Neighborhood Methode
  - Record Linkage (Fellegi & Sunter, 1969)
  - u.a.
- Industrie
  - Trillium, Vality, ETI, et al.
  - Algorithmen sind gut gehütete Geheimnisse

# Sorted Neighborhood



- Idee (Hernandez & Stolfo 1998)
  - Daten geschickt partitionieren.
  - Nur innerhalb dieser Partitionen Duplikate suchen.
- Algorithmus
  1. Create Key:
     - Schlüssel mittels relevanter Feldern erzeugen.
  2. Sort:
     - Daten nach dem Schlüssel sortieren.
  3. Merge:
     - Fenster (der Größe *w*) über sortierte Tupel schieben.
     - Nur Tupel innerhalb des Fensters miteinander vergleichen.



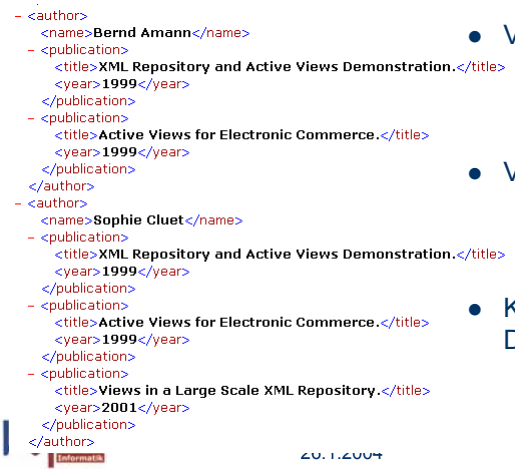
# Objektidentifikation in XML Daten



- Data Warehouse Duplicates
  - Ausnutzung hierarchischer Daten
  - Nur Star-Schema
  - (Ananthakrishna, Chauduri & Ganti 2002)
- XML Duplikate
  - Kinder verschiedener Typen (Snowflake-Schema)
  - Was ist ein Duplikat?
  - Semistrukturierte Daten
  - Schemalose Daten

# Objektidentifikation in XML Daten





```
- <author>
    <name>Bernd Amann</name>
  - <publication>
      <title>XML Repository and Active Views Demonstration.</title>
      <year>1999</year>
    </publication>
  - <publication>
      <title>Active Views for Electronic Commerce.</title>
      <year>1999</year>
    </publication>
  </author>
- <author>
    <name>Sophie Cluet</name>
  - <publication>
      <title>XML Repository and Active Views Demonstration.</title>
      <year>1999</year>
    </publication>
  - <publication>
      <title>Active Views for Electronic Commerce.</title>
      <year>1999</year>
    </publication>
  - <publication>
      <title>Views in a Large Scale XML Repository.</title>
      <year>2001</year>
    </publication>
  </author>
```

- Vergleiche <author>
  - Mit Subelementen (<publication>)?
  - Wie tief?
- Vergleiche <publication>
  - Mit parallelen Elementen (<year>)?
  - Schema, oder Daten?
- Kurz: Was ist ein Duplikat?



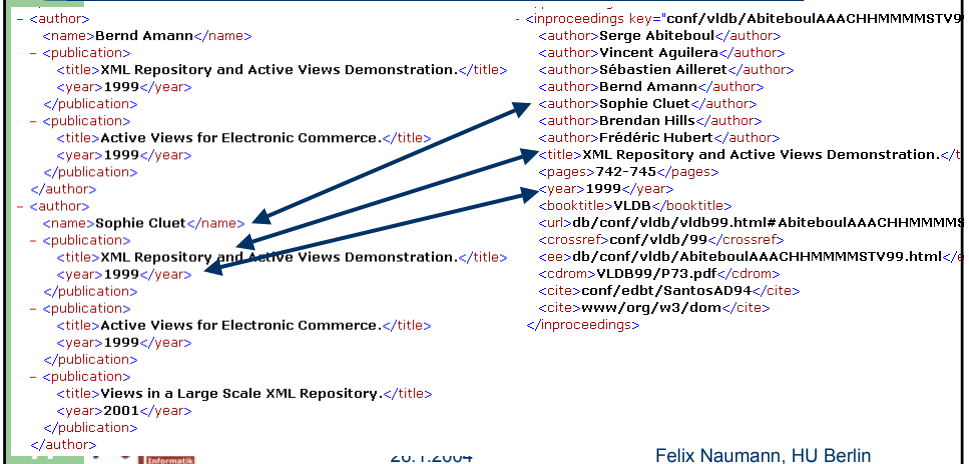
# Objektidentifikation in XML Daten



- Entwicklung eines Vergleichsmaßes für XML Objekte
  - Basierend auf IDFs
  - Symmetrisch
  - Einbeziehung beliebig tiefer Nachkommen
  - Effiziente Speicherung
  - Effizienter Vergleich
  - Prä-Selektion durch „korrekten“ Filter
- Future Work
  - Berücksichtigung von Schemainformationen
  - Annahme bisher: XML Daten gleichen Schemas
  - Jetzt: XML Daten heterogener Schemata

# Objektidentifikation in XML Daten

Identifikation | Integration | Optimierung | Visualisierung






# Duplikat-gesteuertes Schema Mapping

Identifikation | Integration | Optimierung | Visualisierung

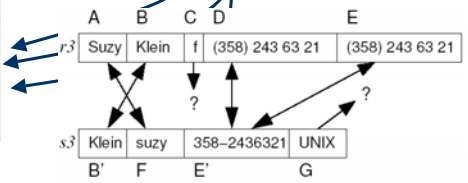
- Umgekehrte Idee
- Schema Matching
  - Klassisches Problem der Informationsintegration:
  - Finde Korrespondenzen zwischen zwei (heterogenen) Schemata
- Herkömmliche Lösungen
  - Auswertung von Attributnamen, Attributwerten und Struktur (Rahm & Bernstein 2001)
- Nun
  - Finde Duplikate (trotz mangelnder Schemata)
  - Korrespondenzen zwischen gleichen Attributwerten
  - In Kooperation mit Alexander Bilke

# Duplikat-gesteuertes Schema Mapping



| R | A | B | C | D | E |
|---|---|---|---|---|---|
| $r_1$ | John | Doe | m | (408) 757 33 39 | (408) 757 33 38 |
| $r_2$ | Joe | Smith | m | (249) 361 56 16 | (249) 234 23 66 |
| $r_3$ | Suzy | Klein | f | (358) 243 63 21 | (358) 243 63 21 |
| $r_4$ | Sam | Adams | m | (541) 812 71 00 | (541) 812 11 64 |
| $r_5$ | Mark | Spitz | m | (901) 831 93 11 | (901) 861 23 82 |
| $r_6$ | Jim | Beam | - | (782) 123 89 57 | (781) 188 37 44 |
| $r_7$ | Kate | Moss | f | (124) 965 45 65 | - |
| $r_8$ | Sam | Wong | f | (124) 495 56 70 | (999) 999 99 99 |
| $r_9$ | John | Dean | m | (369) 366 36 24 | (367) 366 36 25 |

| S | B' | F | E' | G |
|---|---|---|---|---|
| $s_1$ | Doe | jdoe | 408-9182043 | XP |
| $s_2$ | Deen | jdean | 369-3663625 | XP |
| $s_3$ | Klein | suzy | 358-2436321 | UNIX |
| $s_4$ | Adams | adams | 541-8121164 | W2000 |
| $s_5$ | Wong | howard | 923-6363443 | Linux |
| $s_6$ | Kurz | itsme | - | UNIX |






# Objektidentifikation – Zusammenfassung

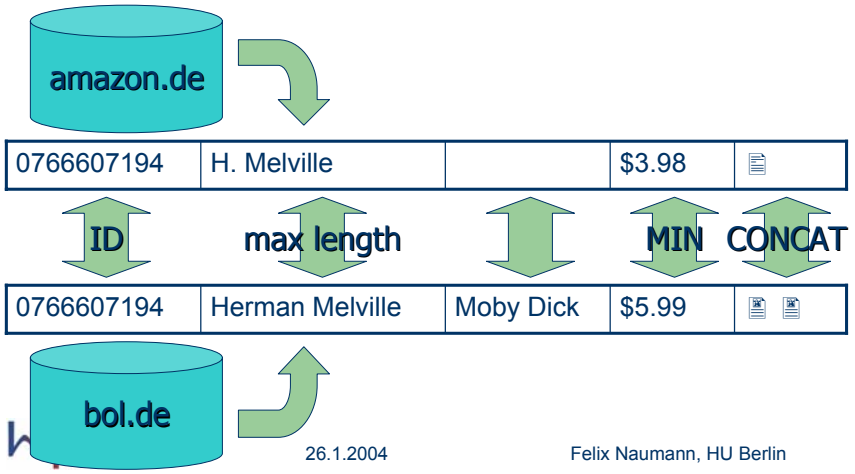


- Relationale Objekt Identifikation
  - Viele Algorithmen
  - No magic
- Identifikation in XML Daten
  - Wahl der Granularität
  - Nutzung struktureller Information
  - Hilfestellung für Schema Matching
- Agenda
  - Erweiterung existierender Algorithmen
  - Entwicklung neuer Ähnlichkeitsmaße und Algorithmen
  - Effiziente Ausführung
    - Datenstruktur
    - Algorithmen

**Jetzt wissen wir, WAS integriert werden soll. Bloß WIE?**

# Objektintegration



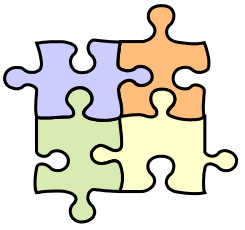





# Relationale Objektintegration



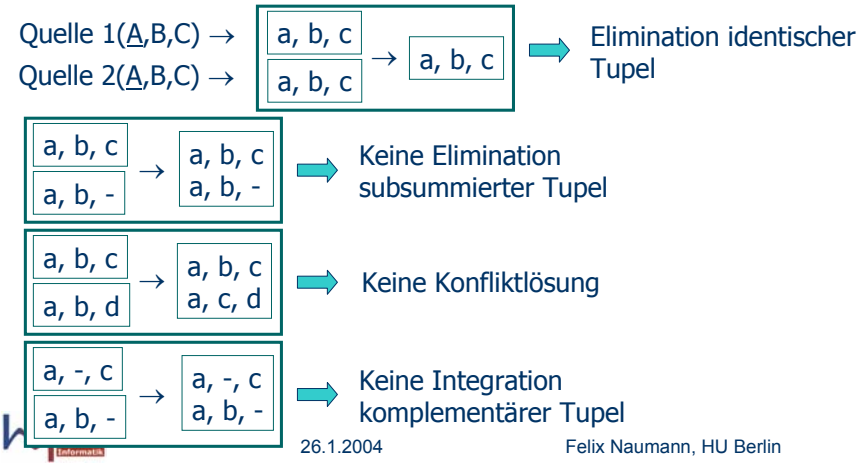
- Union ∪
  - Duplikat-Eliminierung
- Outer union ⟇ (Codd 1979)
  - Union bei heterogenen Schemata
- Minimum union ⊕ (Ullmann 1989, Galindo-Legaria 1994)
  - Eliminierung subsummierter Tupel
- Merge union
  - Duplikatintegration
  - Konfliktlösung



Weitere Operatoren
-Generalized union
-... ??? JENS ???

# Wie Union integriert

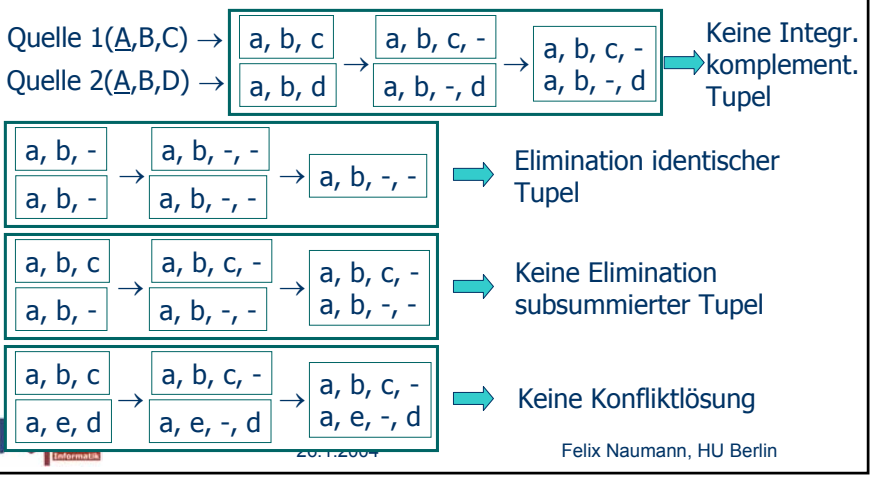






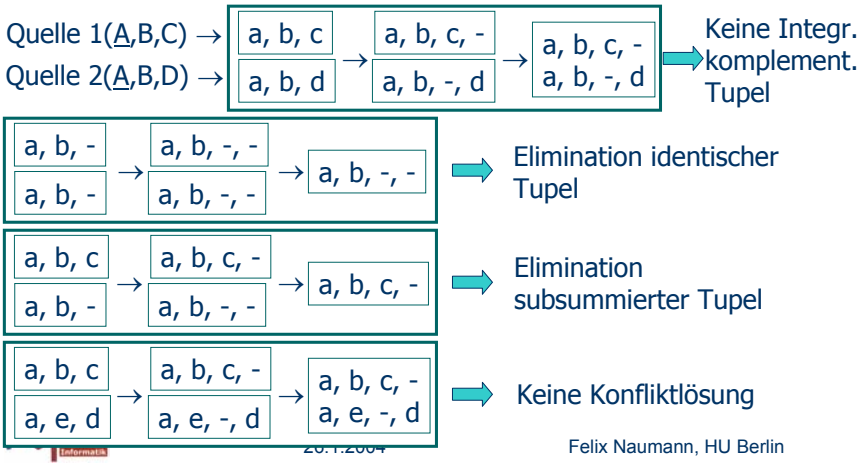

# Wie Outer Union integriert





# Wie Minimum Union integriert

Identifikation | Integration | Optimierung | Visualisierung



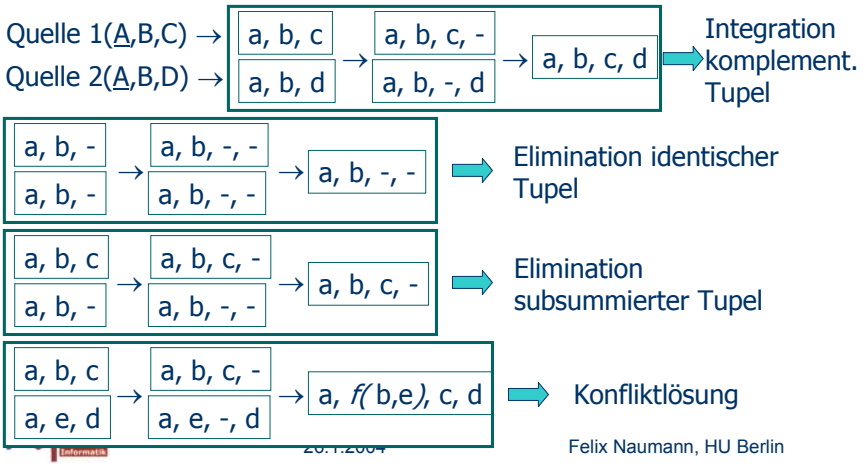







# Wie Merge Union integrieren sollte

Identifikation | Integration | Optimierung | Visualisierung



# Konfliktlösung



$$f(x,y) := \begin{cases} \perp & if\ x = \perp\ and\ y = \perp \\ x & if\ y = \perp\ and\ x \neq \perp \\ y & if\ x = \perp\ and\ y \neq \perp \\ g(x,y) & else \end{cases}$$

null = unbekannt

interne Konfliktlösungsfunktion



# Konfliktlösungsfunktionen



- Numerisch:
  SUM, AVG, MAX, MIN, …
- Nicht-numerisch:
  MAXLENGTH, CONCAT, AnnCONCAT,…
- Spezialisiert:
  RANDOM, COUNT, CHOOSE, FAVOR, MaxIQ,…
- Domänen-spezifisch

…

# Implementierung von Merge Union



- Merging = Gruppierung & Aggregation
  - User-defined Grouping (zur Objektidentifikation)
    - Meist nicht möglich
    - Gruppen typischerweise sehr klein
  - User-defined Aggregation (zur Konfliktlösung)
    - Meist nicht möglich
    - Input nicht nur Werte des aggregierten Attributs
- Relationales Modell
  - Relationale Algebra / SQL
  - Programm
  - Erweiterung einer föderierten Datenbank
- XML Datenmodell
  - XML Query Algebra / XQuery
  - Programm
- Effizienz
- User Spezifikation / Interaktion



# Integration - Zusammenfassung



- "Wahre" Integration auf Datenwert-Niveau
- Vereinigen ist nicht Integrieren
  - Redundanzen
  - Widersprüche
  - Komplemente
- Agenda
  - Merge-Operatoren für relationale und XML Daten
  - Konfliktlösung
  - Implementierung

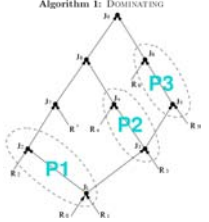
**Jetzt wissen wir, WIE integriert werden soll.**
**Aber WIE GUT schaffen wir das?**

# Optimierung - Ziele



1. Integrationsalgorithmen
   - Stand-alone Algorithmen zur schnellen Ausführung von Integrationsoperationen
2. Anfrageoptimierung
   - Strategien zur optimalen Ausführung komplexer Anfragen mit Integration
   - Redundanz ⇒ größerer Suchraum
   - *Merge union* Operator mit unklaren Kosten
3. Optimierungsziel
   - Paradigmenwechsel von *Zeit* zu *Inhalt*

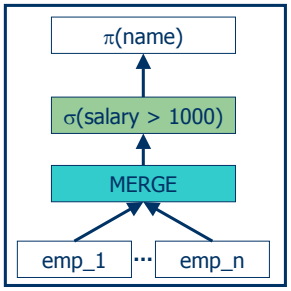

**Input:** Query $Q$; sources $S = \{s_1, s_2, \ldots, s_n\}$; costs $\{c_1, c_2, \ldots, c_n\}$; limit $L$
**Output:** Result for $Q$
1: $Answer \leftarrow \{\}$;
2: $R \leftarrow S$; {remaining sources}
3: $U \leftarrow 0$; {used cost}
4: **while** $U < (L - C)$ and $R$ is not empty **do**
5: $greedy \leftarrow greedySequenceCoverage(R)$;
6: $single \leftarrow singleLargestCoverage(R)$;
7: **if** $greedy > single$ **then**
8: $Next \leftarrow maxGreedySequence(R)$;
9: **else**
10: $Next \leftarrow maxSource(R)$;
11: $R \leftarrow R - \{Next\}$;
12: Execute $Q$ at $Next$;
13: **if** $Next$ is available **then**
14: Collect result into $Answer$;
15: $U \leftarrow U + c_{Next}$;
16: Return $Answer$;
**Algorithm 1:** DOMINATING

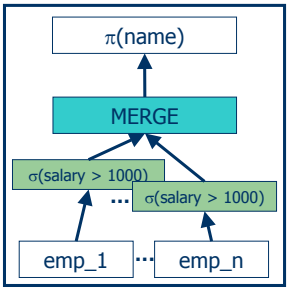





# Anfrageoptimierung mit Merging



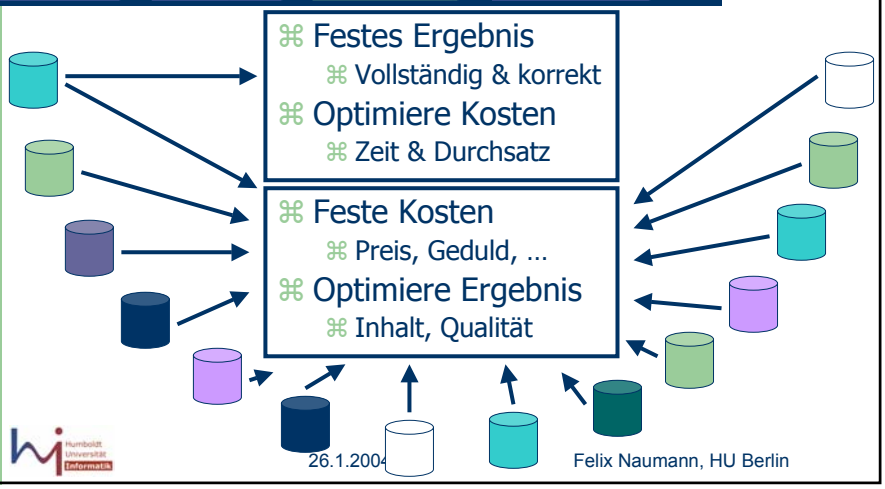



- Korrektheit?
- Vollständigkeit?
- Effizienz?

# Optimierung – Ein Paradigmenwechsel

Identifikation | Integration | Optimierung | Visualisierung






# Optimierung - Zusammenfassung

Identifikation | Integration | Optimierung | Visualisierung

- Behandlung Integrierender Operatoren
  - Stand-alone
  - In Anfragepläne
- Paradigmenwechsel der Optimierung
  - Nutzeranforderungen
- Agenda
  - Multi-dimensionale Optimierung
  - SQL, XQuery
  - Optimierung in DBMS

**Jetzt wissen wir, WIE wir schnell und gut integrieren.**
**Aber FÜR WEN tun wir das?**

# Visualisierung – Ziele



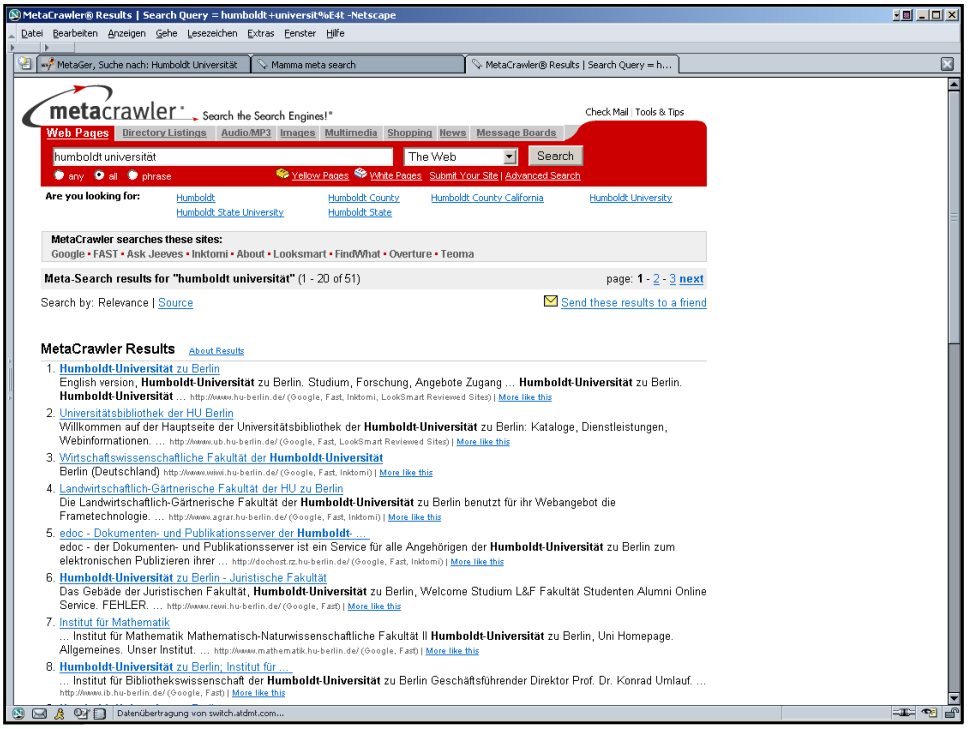
- Darstellung integrierter Ergebnisse
  - Viele Quellen
  - Viele Operationen / Transformationen
- Nutzerfreundliche Darstellung
- Drill-Down
  - Metadaten
- Kein kleinster gemeinsamer Nenner!

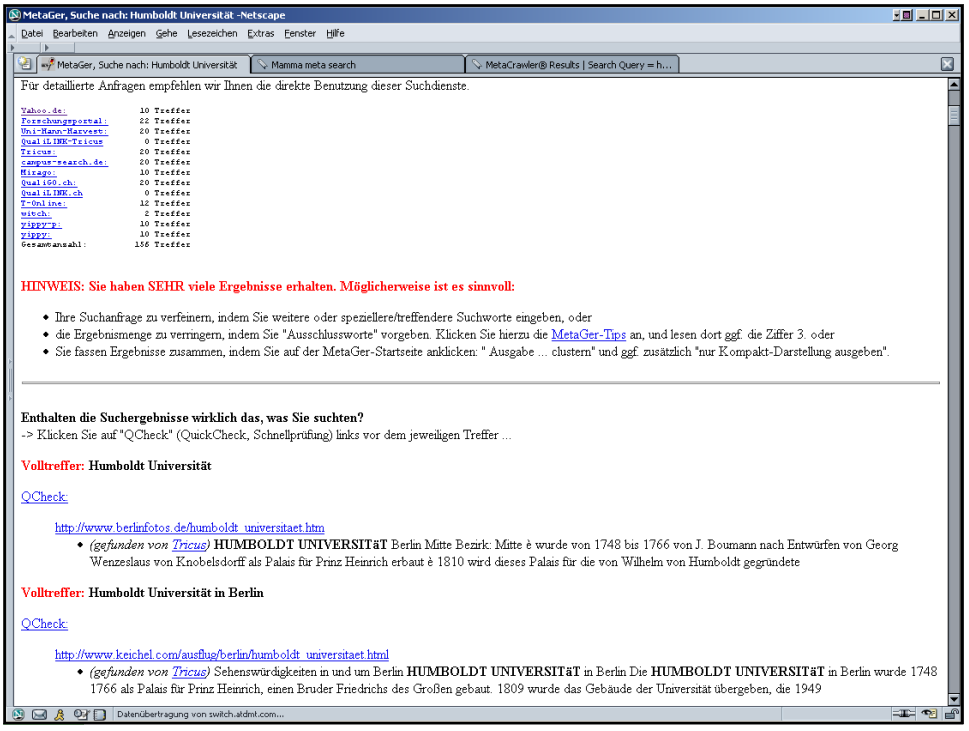


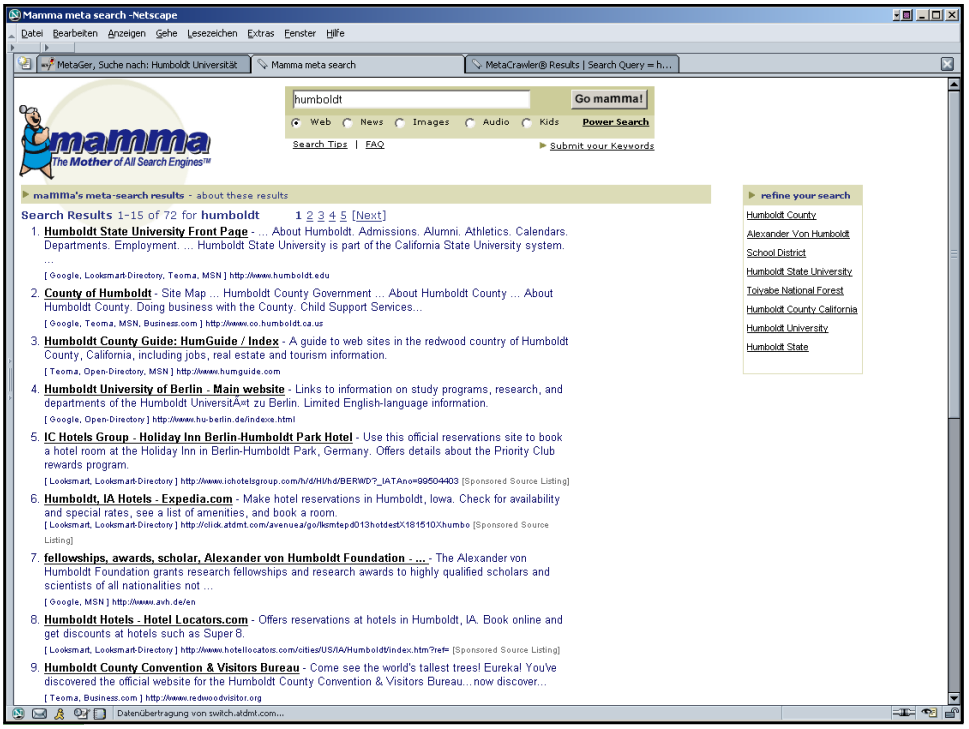
# Visualisierung – Metasuchmaschinen



- MetaGer
- Mamma
- MetaCrawler
  - Google: Title, Summary, Descr., Category, URL, Size
  - Fast: Title, Summary, Descr., URL, Size
  - Inktomi: Title, Summary, URL


MetaCrawler® Results | Search Query = humboldt+universit%E4t -Netscape
Datei Bearbeiten Anzeigen Gehe Lesezeichen Extras Fenster Hilfe
MetaGer, Suche nach: Humboldt Universität
Mamma meta search
MetaCrawler® Results | Search Query = h...
metacrawler
Search the Search Engines!®
Check Mail | Tools & Tips
Web Pages
Directory Listings
Audio/MP3
Images
Multimedia
Shopping
News
Message Boards
humboldt universität
The Web
Search
any
all
phrase
Yellow Pages
White Pages
Submit Your Site | Advanced Search
Are you looking for:
Humboldt
Humboldt County
Humboldt County California
Humboldt University
Humboldt State University
Humboldt State
MetaCrawler searches these sites:
Google • FAST • Ask Jeeves • Inktomi • About • Looksmart • FindWhat • Overture • Teoma
Meta-Search results for "humboldt universität" (1 - 20 of 51)
page: 1 - 2 - 3 next
Search by: Relevance | Source
Send these results to a friend
MetaCrawler Results
About Results
1. Humboldt-Universität zu Berlin
English version, Humboldt-Universität zu Berlin. Studium, Forschung, Angebote Zugang ... Humboldt-Universität zu Berlin. Humboldt-Universität ... http://www.hu-berlin.de/ (Google, Fast, Inktomi, LookSmart Reviewed Sites) | More like this
2. Universitätsbibliothek der HU Berlin
Willkommen auf der Hauptseite der Universitätsbibliothek der Humboldt-Universität zu Berlin: Kataloge, Dienstleistungen, Webinformationen. ... http://www.ub.hu-berlin.de/ (Google, Fast, LookSmart Reviewed Sites) | More like this
3. Wirtschaftswissenschaftliche Fakultät der Humboldt-Universität
Berlin (Deutschland) http://www.wiwi.hu-berlin.de/ (Google, Fast, Inktomi) | More like this
4. Landwirtschaftlich-Gärtnerische Fakultät der HU zu Berlin
Die Landwirtschaftlich-Gärtnerische Fakultät der Humboldt-Universität zu Berlin benutzt für ihr Webangebot die Frametechnologie. ... http://www.agrar.hu-berlin.de/ (Google, Fast, Inktomi) | More like this
5. edoc - Dokumenten- und Publikationsserver der Humboldt- ...
edoc - der Dokumenten- und Publikationsserver ist ein Service für alle Angehörigen der Humboldt-Universität zu Berlin zum elektronischen Publizieren ihrer ... http://dochost.rz.hu-berlin.de/ (Google, Fast, Inktomi) | More like this
6. Humboldt-Universität zu Berlin - Juristische Fakultät
Das Gebäde der Juristischen Fakultät, Humboldt-Universität zu Berlin, Welcome Studium L&F Fakultät Studenten Alumni Online Service. FEHLER. ... http://www.rewi.hu-berlin.de/ (Google, Fast) | More like this
7. Institut für Mathematik
... Institut für Mathematik Mathematisch-Naturwissenschaftliche Fakultät II Humboldt-Universität zu Berlin, Uni Homepage. Allgemeines. Unser Institut. ... http://www.mathematik.hu-berlin.de/ (Google, Fast) | More like this
8. Humboldt-Universität zu Berlin, Institut für ...
... Institut für Bibliothekswissenschaft der Humboldt-Universität zu Berlin Geschäftsführender Direktor Prof. Dr. Konrad Umlauf. ... http://www.ib.hu-berlin.de/ (Google, Fast) | More like this
Datenübertragung von switch.atdmt.com...


MetaGer, Suche nach: Humboldt Universität -Netscape
Datei Bearbeiten Anzeigen Gehe Lesezeichen Extras Fenster Hilfe
MetaGer, Suche nach: Humboldt Universität
Mamma meta search
MetaCrawler® Results | Search Query = h...
Für detaillierte Anfragen empfehlen wir Ihnen die direkte Benutzung dieser Suchdienste.
Yahoo.de: 10 Treffer
Forschungsportal: 22 Treffer
Uni-Mann-Harvest: 20 Treffer
QualiLINK-Tricus 0 Treffer
Tricus: 20 Treffer
campus-research.de: 20 Treffer
Mirago: 10 Treffer
Quali60.ch: 20 Treffer
QualiLINK.ch 0 Treffer
T-Online: 12 Treffer
witch: 2 Treffer
zippy-p: 10 Treffer
zippy: 10 Treffer
Gesamtanzahl: 156 Treffer
HINWEIS: Sie haben SEHR viele Ergebnisse erhalten. Möglicherweise ist es sinnvoll:
Ihre Suchanfrage zu verfeinern, indem Sie weitere oder speziellere/treffendere Suchworte eingeben, oder
die Ergebnismenge zu verringern, indem Sie "Ausschlussworte" vorgeben. Klicken Sie hierzu die MetaGer-Tips an, und lesen dort ggf. die Ziffer 3. oder
Sie fassen Ergebnisse zusammen, indem Sie auf der MetaGer-Startseite anklicken: " Ausgabe ... clustern" und ggf. zusätzlich "nur Kompakt-Darstellung ausgeben".
Enthalten die Suchergebnisse wirklich das, was Sie suchten?
-> Klicken Sie auf "QCheck" (QuickCheck, Schnellprüfung) links vor dem jeweiligen Treffer ...
Volltreffer: Humboldt Universität
QCheck:
http://www.berlinfotos.de/humboldt_universitaet.htm
(gefunden von Tricus) HUMBOLDT UNIVERSITäT Berlin Mitte Bezirk: Mitte è wurde von 1748 bis 1766 von J. Boumann nach Entwürfen von Georg Wenzeslaus von Knobelsdorff als Palais für Prinz Heinrich erbaut è 1810 wird dieses Palais für die von Wilhelm von Humboldt gegründete
Volltreffer: Humboldt Universität in Berlin
QCheck:
http://www.keichel.com/ausflug/berlin/humboldt_universitaet.html
(gefunden von Tricus) Sehenswürdigkeiten in und um Berlin HUMBOLDT UNIVERSITäT in Berlin Die HUMBOLDT UNIVERSITäT in Berlin wurde 1748 1766 als Palais für Prinz Heinrich, einen Bruder Friedrichs des Großen gebaut. 1809 wurde das Gebäude der Universität übergeben, die 1949
Datenübertragung von switch.atdmt.com...




# HiQIQ Meta Search Engine

Identifikation | Integration | Optimierung | Visualisierung

http://www.icdt.org/ (14 [ Altavista ] 3 [ Northern Light ] 1 [ Fast ] 1 [ Hotbot ] 3 [ Google ] )

*five search engines found this page*

ICDT Home Page [ Altavista ]
ICDT Home Page [ Northern Light ]
ICDT HomePage [ Fast ]
ICDT HomePage [ Hotbot ]
ICDT HomePage [ Google ]

*the descriptions differ*

description Download your FREE evaluation copy at www.auscomp.com... [ Altavista ]
Islamic Centre for Development of Trade. Islamic Centre for Development of Trade web site was built to facilitate your business and trade with OIC. Trade... [ Northern Light ]
BusinessOpportunities|EconomicOperators|TradeGuides|Events&Fairs|Indicators|Statistics|Publications
Islamic Centre for Development of Trade has built this web site to facilitate your business and trade with OIC member states. For Members and NFP [ Fast ]
BusinessOpportunities|EconomicOperators|TradeGuides|Events&Fairs|Indicators|Statistics|Publications
Islamic Centre for Development of Trade has built this web site to facilitate your business and trade
IC member states. For Members and NFP ... [ Hotbot ]
*data conflict*
sOpportunities|EconomicOperators|TradeGuides|Events&Fairs|Indicators|Statistics|Publications
Centre for Development of Trade has built this web site to ... [ Google ]

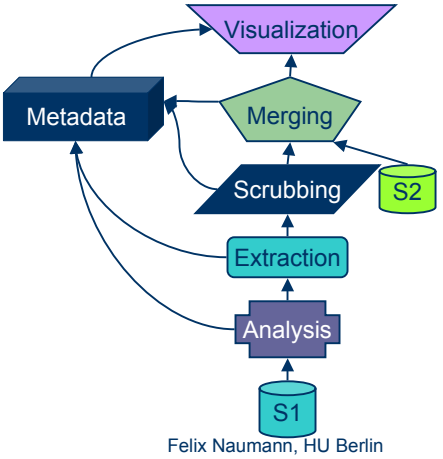
date 07/19/1999 [ Altavista ] 07/19/1999 [ Northern Light ] 1/18/2000 [ Hotbot ] size 8K [ Altavista ]
8k [ Google ] language English [ Altavista ] category Non-profit site [ Northern Light ] rating 94% [ Northern Light ]

*different attributes from different engines*

# Metadaten der Integration



- Herkunft
  - Data lineage
    - z.B. (Cui & Widom 2001)
  - Durch Operatoren hindurch
- Transformationen
  - Aggregation / Integration
- Konflikte
- Datenqualität
- Ranking





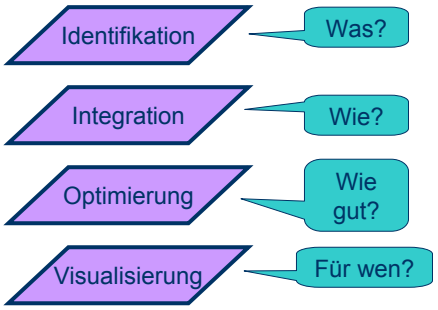

# Visualisierung – Zusammenfassung



- Zeige mehr, nicht weniger
- Metasuchmaschinen
- Metadaten
  - Data Lineage
  - Data Quality
  - Transformationen
- Ziele
  - Spezifikation notwendiger Metadaten
  - Automatisierte Erfassung
  - Metadaten-Verwaltung
  - Drill-Down Visualisierung

**Jetzt kennen wir alle notwendigen Komponenten zur effektiven Informationsintegration.**

# Informationsintegration - Zusammenfassung




www.informatik.hu-berlin.de/mac
naumann@informatik.hu-berlin.de



# Literatur


- Fellegi & Sunter 1969
  - *A theory of record linkage*. Journal of the American Statistical Association, 64: 1183-1210.
- Codd 1979
  - *Extending the Relational Database Model to Capture more Meaning*. ACM Transactions on Database Systems (TODS), 4(4): 397-434.
- Galindo-Legaria 1994
  - *Outerjoins as Disjunctions*. ACM SIGMOD Conference, 348-358.
- Hernandez & Stolfo 1998
  - *Real-world data is dirty: Data cleansing and the merge/purge problem*. Data Mining and Knowledge Discovery, 2(1): 9-37.
- Rahm & Bernstein 2001
  - *A survey of approaches to automatic schema matching*. VLDB Journal 10(4), 334-350.
- Cui & Widom 2001
  - *Lineage tracing for general data warehouse transformations*. VLDB Jorunal 12(1): 41-58.
- Ananthakrishna, Chauduri & Ganti 2002
  - *Eliminating Fuzzy Duplicates in Data Warehouses*. VLDB Conference: 586-597.