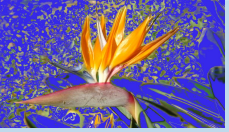
Vorlesung: Bio-Datenbanken

Kapitel 6: Schemaintegration

Dr. Dieter Sosna

24. Januar 2008



Allgemeines

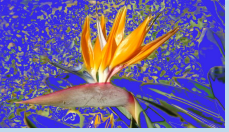
Metadatenbasiert

Instanzdatenbasiert

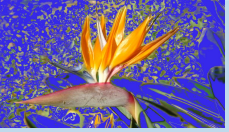
MOMA (reuse)

Qualität

Einige Folien, Graphiken wurden von Herrn A. Thor zur Verfügung gestellt. Danke.

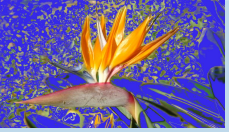


Allgemeines



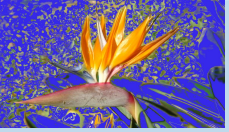
Zwischenstand und Ziel

- Bisher Integration von Instanzen: Gegeben eine Instanz in Quelle A, welche Einträge aus Quelle B gehören semantisch dazu?
  Lösung durch Vergleich charakterischer Werte (in der Art der Schlüsselkandidaten).
  Gleichheit - Ähnlichkeit.
  A (bzw. **B**). Komplexe Datentypen: Ähnlichkeit bei weitgehend übereinstimmendem Inhalt.
- Neuer Ansatz: In der Sprache der OO: Vergleich der Konzepte. Welche Konzepte in Datenquelle A entsprechen semantisch welchen Konzepten in Quelle B?
  Konzepte in rel. DB durch Tabellen, in ... beschrieben.
  Also: Welche Bestandteile / welche Daten aus DB 1 entspricht welchem Teil von BD 2?
- Vorgehensweisen: top down - bottom up als erster Ansatz;
  Kombinierte und andere Ansätze evt. besser.



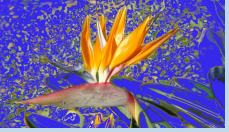
Literatur

Leser, Ulf und Naumann, Felix: Informationsintegration.
Architekturen und Methoden zur Integration verteilter und Heterogener Datenquellen.
dpunkt.verlag, Heidelberg, 2007.ISBN (978-) 3-89864-400-6.



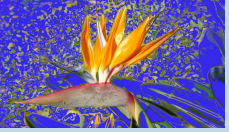
Begriff: Metadaten

- Daten: durch Zustände eines physikalischen Mediums dargestellte Information.
- Metainformationen: Beschreibung der Informationen, meist oder zum Teil in einer (externen) Metasprache.
- Bespiele an Hand des E/R-Modells:
  Attribut (kleinste strukturelle Einheit des Modells): Name, Wertevorrat, Kontext, Semantik.
  Dabei:
  Wertevorrat: Wertevorrat im eng. Sinn, Integritätsbed. auf Attributebene.
  Kontext: zu welcher größeren Struktur (E oder R) gehörig.
  Semantik: Erklärung der Bedeutung in der Miniwelt - meist in natürlicher Sprache.
  Entitätsmenge: ...



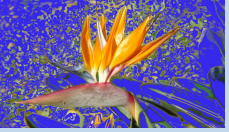
Metadaten (2)

- Metainformationen werden beim Übergang zur Implementierung im DBS *unvollständig* übernommen.
  Es fehlen:
  Erklärungen der Semantik, gleiche Semantik von Attributen kann in Ausnahmefällen z.B. durch Fremdschlüsseleigenschaft auch im DBS erkannt werden.
  Wertevorrat (natürlicher W.; nach Abb. auf vordefinierte Datentypen verloren)
- Potentielle Möglichkeiten zur Semantikbeschreibung:
  XML-Schema: Namensräume
  rtf ?



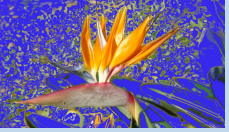
Schemaintegration - Definitionen

- Vorgelegt zwei Schemata **A,B**.
  Gesucht ein Schema **C** und Transformationen t_A, t_B (Schemakonstruktion) d_A, d_B (Datentransformationen) mit folgenden Eigenschaften:
  - **Vollständigkeit** (s. nächste Folie)
  - **Minimalität**
  - **Korrektheit**
- Definition: Schemaintegration bezeichnet den Prozeß des Findens von **C** und t_A, t_B, d_A, d_B .
  Wir zählen zur Schemaintegration auch die Anwendung der Transformationen d_A, d_B, mit denen die Daten aus **A,B** in **C** überführt werden.
- Komponenten der Schemaintegration:
  Schemamatching (Finden semantisch gleicher Konzepte in **A** und **B** bzw. in **A** und **C** bzw. in **B** und **C** und der Abbildungen t_A, t_B.
  Schemamapping (Finden der Abbildungen d_A, d_B.); Datentransformation.



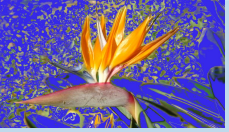
Begriff: Vollständigkeit, ...

- **Vollständigkeit:**
  Das Anwendungsgebiet von **C** umfasst die Anwendungsgebiete der Schemata **A** und **B**.
  Alle Beziehungen zwischen Konzepten in **A** bzw. **B**, können in **C** verlustfrei dargestellt werden. Alle Daten aus **A** bzw. **B** können in **C** dargestellt werden.
- **Minimalität:**
  Aus **C** kann kein Konzept entfernt werden ohne dass die Vollständigkeit verletzt wird. (Das bedeutet insbesondere, dass in **A** und **B** semantisch gleiche Konzepte in **C** nur einmal auftreten.)
- **Korrektheit:**
  Zu jedem Konzept der Schemata **A** und **B** existiert ein semantisch gleiches Konzept in **C**.
  Die Beziehungen zwischen Konzepten in **C**, eingeschränkt auf die Daten, die durch Tranformation aus **A** (bzw. **B**) entstanden sind, ist semantisch gleich der Beziehung zwischen den Konzepten in **A** (bzw. **B**).



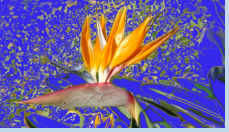
Bemerkungen

- Aus Vollständigkeit abgeleitet: **Mächtigkeit des Zielmodells**
  $\rightarrow$? operationale Vollständigkeit — XML-Schema.
- Widerspruch zwischen Automatisierung und Qualitätsanforderung.
  Ziel: Integration großer Schemata (10^3 Konzepte) erfordert berechnete Integration;
  Erkennung der Semantik aus syntaktischen und strukturellen Merkmalen nur partiell automatisierbar.
- deshalb Verständlichkeit der Ergebnisse:
  Mensch muß die Ergebnisse nachvollziehen, bewerten und korrigieren können.
  Dokumentation der Entsprechungen (welche, wie gewonnen, ...), der Transformationen, ...
  deshalb
  Integrationsalgorithmen müssen lernend sein bzw. korrigierbar sein,
  die Ergebnisse von Auswertungen in weitere Integrationsschritte einbeziehen.



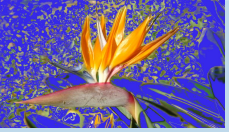
Bemerkungen (2)

- Integration ist gerichtet,
  d.h. die Abbildungen müssen nicht (per se) invertierbar sein,
  Beispiel: OUTER-JOIN-Invertierung:
  $A(a,b) = \{(-,1)\}, F(c,d) = \{(1,2),(4,5)\}$
  $(A\ OUTER-JOIN_{b=c}\ F)(a,b,d) = \{(-,1,2),(-,4,5)\}$
  Berechne aus $A\ OUTER-JOIN_{b=c}\ F$ wieder A !
  Konstruktion der Umkehrabbildung ist neue Aufgabe (Unterschied zu Schemaevolution)



Integrationsschritte

- **Integrationsvorbereitung:** Auswahl der Schemata, bzw. von Teilen davon, Festlegung von Reihenfolgen, anzuwendende Verfahren.
  IV wesentlich für Erfolg, da durch Arbeit des Menschen Semantik eingebracht wird.
- **Schemavergleich:** Ermittlung von Korrespondenzen: semantisch gleiche Elemente, Teilmengenbeziehungen,
  Erkennung von Heterogenitäten zwischen den Schemata: Namenskonflikte der Konzepte (Synonyme, Homonyme), Strukturelle Konflikte (Schlüsselalternativen, Normalformenunterschiede bei rel DB, Position in Ontologie (Buch (Autor (Name,Vorname), Titel, ...) vs. Autor(Name, Vorname, {Buch})).
- **Schemakonstruktion:** Ableitung des neuen Schema durch Vergleich der Korrespondenzen mit den alten Schemata.
  Konstruktion der Abbildungen von den Konzepten jedes alten Schema in des neue Schema.
  Konstruktion von Datenabbildungen (SQL-Befehle oder äquivalent).



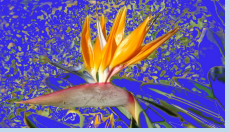
Korrespondenzbasierte Integration

Lit.: Conrad: Föd. DBS. Springer 1997.
Regeln zur Übernahme in das integrierte Schema

- Überhahme: Kategorie ohne Korrespondenz (mit Daten)
- Korespondierende Kategorien: Übernahme, Daten mit OUTER-JOIN.
- Gleiche (in beiden Ausgangsschemata) direkte Beziehungen übernehmen, Daten-JOIN.
- Beziehungen ohne Korrespondenz: Übernahme

Problem: Kategorien meist nicht identisch, sondern überlappend $\rightarrow$ Zersplitterung.

s

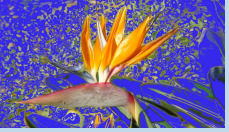


GIM-Modell

Schmitt, Ingo: Schemaintegration für d. Entwurf Föd.DB. Diss. 1998
GIM (Gener. Integrationsmodell) - Matrix und Schemaableitung

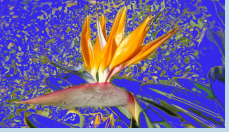
- Spalten: Minimale Zerlegung aller Objekte (aus Ausgangsschemata) in disjunkte Klassen: $\mathcal{A} \setminus \mathcal{B}, \mathcal{A} \cap \mathcal{B}, \mathcal{B} \setminus \mathcal{A}$
- Zeilen: Attribute - homogenisiert
- Felder: Wahrheitswerte: **w** = Attribut ist für Kategorie relevant.

Schemaableitung:
Umordnung der Matrix, so das große recheckige Bereichen mit **w**-Werten entstehen (dabei sind Überlappungen zugelassen.
Breite Rechtecke = Oberklassen, hohe Schmale = Unterklassen
Kritik:
1)Semantische Hauptarbeit: Homogenisierung der Attribute, außerhalb des Modells.
2)Klassenbildung formal, Semantik der Klassen ? K. modelliert evt. keine realen Objekte.



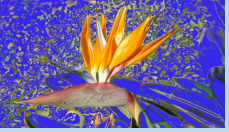
Datentransformation

- Zeitpunkt der DT: materialisierte Integration - sofort
  virtuelle Integration - Konstruktion eines Wrappers.
- Abbildungstypen im Mapping : 1:1, 1:N,N:1, N:M
- Leicht lösbar (auf Grund des Matching weitgehend automatisierbar):
  Wertkorrespondenzen vom Typ 1:1, N:1, 1:N bei einfachen Attributen (rel.DB, Simple-Type bei XML): 1:1-Transformationen (Umrechnungen, ...), funktionale Zusammenfassungen (Konkanation, ...), Extraktion.
- Schwierig (Automatisierung noch im Forschungbereich):
  M:N-Wertkorrespondenzen:
  Buch - Autor (vereinfacht)
  Buch({Autor(Name, Vorname)}, ISBN, Titel) vs.
  Person(Name, Vorname, geschrieb-Buch({(ISBN, Titel)})
  Korrespondenzen ?
- Schwierig: Korrespondenzen über mehrere Konzepte /Konzeptstufen
  Schemaheterogenität



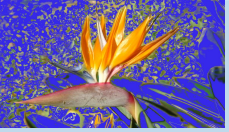
Matching komplexer Strukturen

- Ähnlichkeit der Probleme bei
  - Ontologiematching,
  - Schemamatching,
  - Matching komplexer Objekte
- In der Sprache der Objektorientierung:
  (Komplexe) Objektklassen und Beziehungen zwischen diesen.
- Unterschiede:
  Art der Objekte, Darstellung der Komplexität, Art (Typ) der Beziehungen.



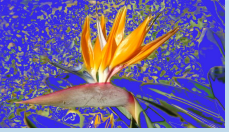
Ontologien

- το ον = das Seiende (philos.Begriff).
- ca. seit 1990 Informatik Beschreibung eines Anwendungsbereiches, der Begriffe und der Beziehungen untereinander.
  Eigenschaften:
  (1) Begriffe und Beziehungen eindeutig und unstrittig definiert
  (2) formal und genau: neues Wissen durch log. Schlüsse ableitbar.
- Top-Level-Ontologie: Fundamentale Beziehungen (nicht in dieser Vorlesung)
  Domänspezifische O. (Fachterminologie) - Metabeschreibung !.
  Überschneidungen der Gebiete → Anpassungen nötig. → Ontologiematching



Nutzen der Ontologien

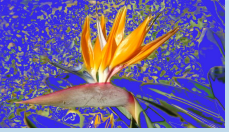
- Einheitliche Begriffswelt (kontrolliertes Vokabular):
  Übername des Vok. in Daten sichert Vergleichbarkeit von Daten,
  Hilfe bei Überwindung von Heterogenität (z.B. Synonyme
  kann bei Matchprozeduren helfen, die Semantik zu erhennen: Matching gegen Standard.
- Gene Ontologie: ca. 17000 Begriffe (Molekülchemie, (molekular-) biolog. Prozesse.
  Struktur: Konzepte, is-a- und part-of- Beziehung.
  Inhalte: von Experten erzeugt, Internationale Konsortium, sehr gut akzeptiert, da Nutzen offensichtlich - Quasistandard.
  Benutzung: tool-unterstützt.
- Praktisch wird Begriff der O. im stark erweiterten Sinn genutzt:
  Liste von Konzepten, Taxonomien, Tessauri, Polyhierarchien, Graphen



Ontologiebasierte Informationsintegration

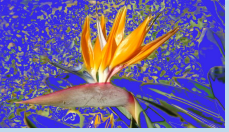
Lit: Leser, Naumann, a.a.O.

- 3 Schritte:
  Erstellung der globalen Ontologie
  Einordnung der Datenquellen
  Subsumption zur Anfragebearbeitung:
  Anfragen (z.B. nach Gleichheit , ...) als Konzepte formuliert. Alle Konzepte, die spezieller als das Anfragekonzept sind und eine Datenquelle repräsentieren, enthalten dann nur semantisch korrekze Objekte.
- Prakt. Anwendung /Realisierung bei Bio-DB: ?



Schemamatching

- Voraussetzung: Gegeben 2 Quellen mit zugehörigen Metadaten und Instanzdaten Ziel: Finden von semantisch gleichen Konzepten.
- im Bereich der Bio-DB als Besonderheit: vielfach Quellen auf Instanzniveau verbunden /vernetzt.
- 2 Ansätze:
  - Metadatenbasiert (Namen, Beschreibungen, Ontologie, Struktur (z.B. auch Fremdschlüsselbeziehungen))
  - Instanzbasiert
    Grundannahme: Zwei Konzepte sind ähnlich, wenn sie eine hinreichend große Anzahl gleicher oder zumindest (sehr) ähnlicher Elemente haben.

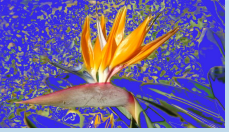


Mißerfolg erwartet

Gundproblem:
Aus formalen Merkmalen (Namensgleichheit, Strukturgleichheit, Häufigkeitsverteilung, ...) soll

auf semantische Ähnlichkeit

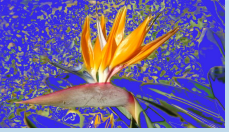
geschlossen werden.
Mit anderen Worten:
Schemamatching ist Forschungsgegenstand. Das Ziel der automatisierten Verfahren ist noch nicht erreicht.
Für jedes Verfahren lassen sich Negativbeispiele finden
$\rightarrow$ Kombination von Verfahren könnte Resultate verbessern. Problem: Wie kombinieren ?



Idee der bottom-up Verfahren

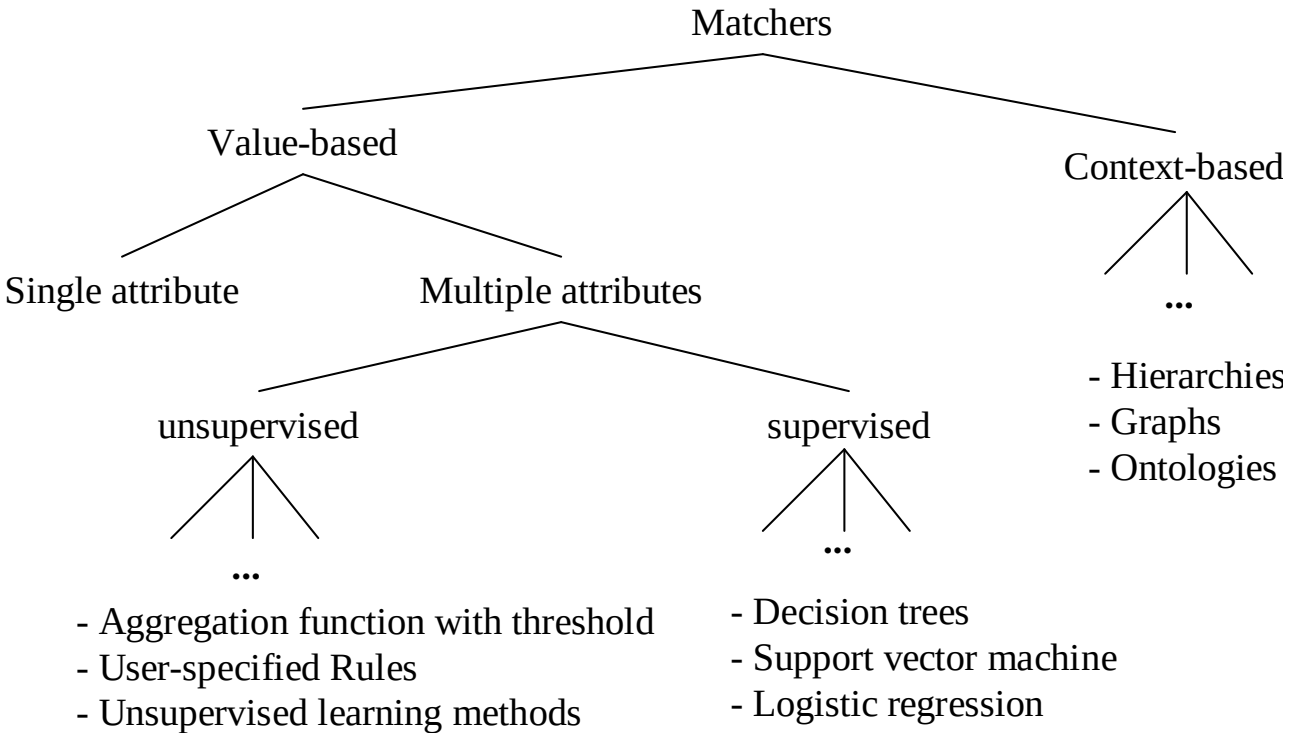
Vergleich zweier Objekte $a = (a_1, \ldots, a_m) \in A$ und $b = (b_1, \ldots, b_n) \in B$

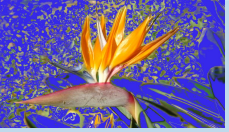
1. Bestimmung der Ähnlichkeitswerte
   - Ähnlichkeitsfunktionen zum Vergleich von Attributwerten
   - Verschiedene Funktionen, verschiedene Attributvergleiche möglich $\rightarrow$ mehrere Ähnlichkeitswerte
2. Anwendung der Matching-Regeln
   - Regel, die an Hand der Ähnlichkeitswerte bestimmt "Match" oder "kein Match"
   - Bsp: "Wenn Ähnlichkeit der Familienenamen 100% und Ähnlichkeit des Vornamens 80%, dann sind zwei Personen gleich."



Ansätze für Object-Matching

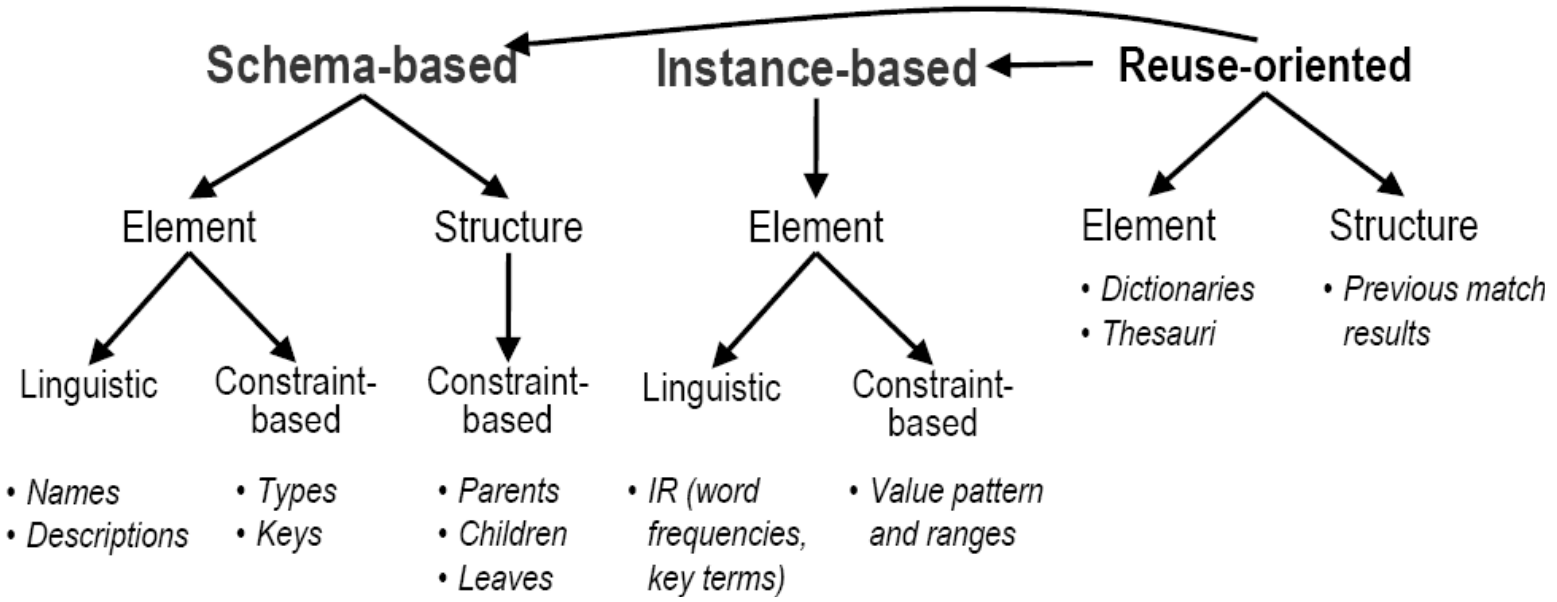
Viele verschiedene automatische Ansätze, die auch kombiniert werden können





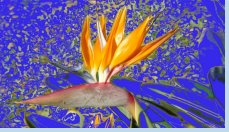
Ansätze für Schema-Matching

Viele verschiedene automatische Ansätze, die auch kombiniert werden können

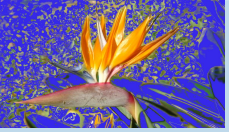


Publikationen:

- Rahm, E., P.A. Bernstein: A Survey of Approaches to Automatic Schema Matching. VLDB Journal 10 (4), 2001
- Do, H.-H., Rahm, E.: COMA - A System for Flexible Combination of Schema Matching Approaches. VLDB, 2002



Metadatenbasiert

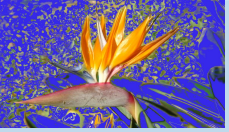


Metadaten: Indikatoren /Berechnungen für Gleichheit

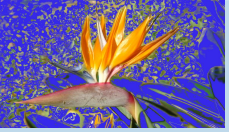
- Namensvergleiche:
  Gleichheit (! Homonyme, bei XML Lösung durch Namensräume)
  Gleichheit nach Normalisierung (Großschreibung, stemming, Übersetzung)
  Hyperonymie (hierarch. Beziehung is-a , Thessausus, Ontologie,Taxonomie)
  Ähnlichkeit
- Strukturvergleiche:
  Cupid (1): Schemata $\rightarrow$ Bäume . Konzepte ähnlich, wenn Eltern, Kinder, Brüder ähnlich sind; bei Blättern: Namensähnlichkeit.
  Similiarity-Flooding (2) : Schematapaar $\rightarrow$ Graphen. Startwert: Matrix der Ähnlichkeit. Iteration: Ähnlichkeit auf Nachbarn übertagen $\rightarrow$ Fixpunktproblem. Lsg. abh. von Anfangswerten! Unabh,. von Semantik.

(1) Madhavan, Bernstein, Rahm: Generic Schema matching with Cupid. Proc. VLDB, 2001.

(2) Melnik, Garcia-Moulina, Rahm: Similiarity Flooding: A Versatile Graph Matching Algorithm. Proc. Int. Conf. Data Eng. (ICDE), 2002.



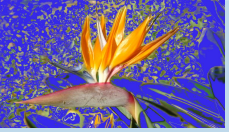
Instanzdatenbasiert



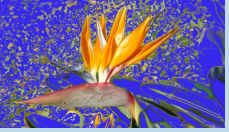
Instanzbasiertes Maching

setzt sets die Existenz von Instanzen in beiden Schemata voraus.

- Horizontale Matcher: Gleiche Konzepte in den Schemata durch Finden von Duplikaten erkannt.
  Vertikale Matcher: Extraktion von vorher definierten Merkmalen aus den Instanzen und Vergleich: z.B. statistische Merkmale (max, min, avg, var, covar, Clusterbildung, ...) aus den Werten der Attribute, aus Merkmalen der Attribute (Länge von Zeichenketten, ...)
- Erfahrungswert (Leser, Naumann, a.a.O) :
  Sind hinreichend viele (Statistik) Instanzen vorhanden (oder bei vert. Matchern theoret. Werte bekannt), so sind instanzbasierte Matcher (derzeit noch - D.S.) den metadatenbasierten überlegen.



MOMA (reuse)



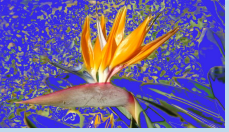
Mapping-Verarbeitung

Motivation

- Matching ist i.A. sehr aufwändig: Viele Ähnlichkeitsvergleiche, manuelle Überprüfung, ...
- Matching ist i.A. sehr schwierig: Welcher Match-Algorithmus? Welche Parameter? ...
- Match-Ergebnis ist "wertvoll" und sollte wiederverwendet werden

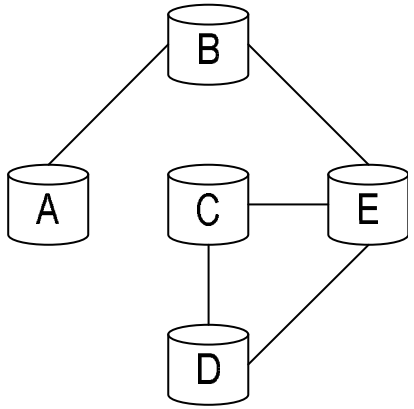
Ziele

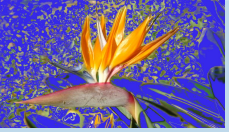
- Wiederverwendung von Match-Ergebnissen zur effizienten Berechnung neuer Match-Ergebnisse
- Kombination von Match-Ergebnissen zur Qualitätsverbesserung
- Bestimmung von Match-Ergebnissen, wenn kein geeignetes Ähnlichkeitsmaß zur Verfügung steht



Mapping-Verarbeitung: Beispiel

- Effiziente Berechnung: (A,E) mittels (A,B) und (B,E)
- Qualitätsverbesserung: Kombination von (D,E) direkt mit (D,C) + (C,E)
- Kein geeignetes Ähnlichkeitsmaß: (A,D) mittels (A,B) + (B,E) + (E,D)

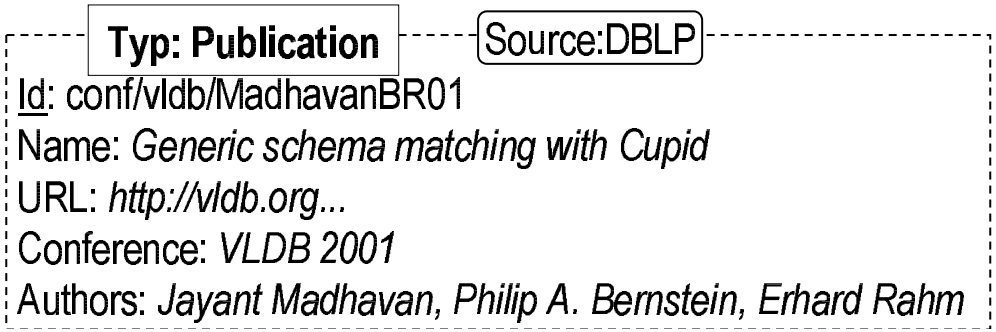


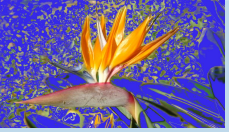


MOMA-Ansatz: Begriffe (1)

Definition: Datenquelle (Logische Datenquelle, LDS)

- Menge von Objektinstanzen
- Alle Objekte haben den gleichen semantischen Typ (z.B. Publikation)
- Jedes Objekt hat eine (innerhalb der Datenquelle) eindeutige Id und beliebige zusätzliche weitere Attribute
- Beispiel: Datenbanktabelle, Website, XML-Dokument, ...

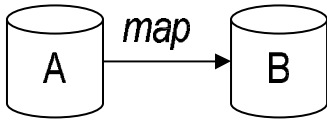




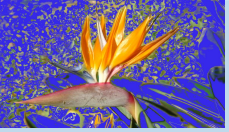
MOMA-Ansatz: Begriffe (2)

Definition: Same-Mapping

- $\{(a, b, s) | a \in A, b \in B, s \in [0, 1]\}$
- A und B sind Datenquellen, s ist Ähnlichkeitswert der Korrespondenz (a,b)
- Beispiel: Mapping-Tabelle, Web-Service, ...



A	B	s
a1	b1	1
a2	b2	0.9
a2	b3	0.3

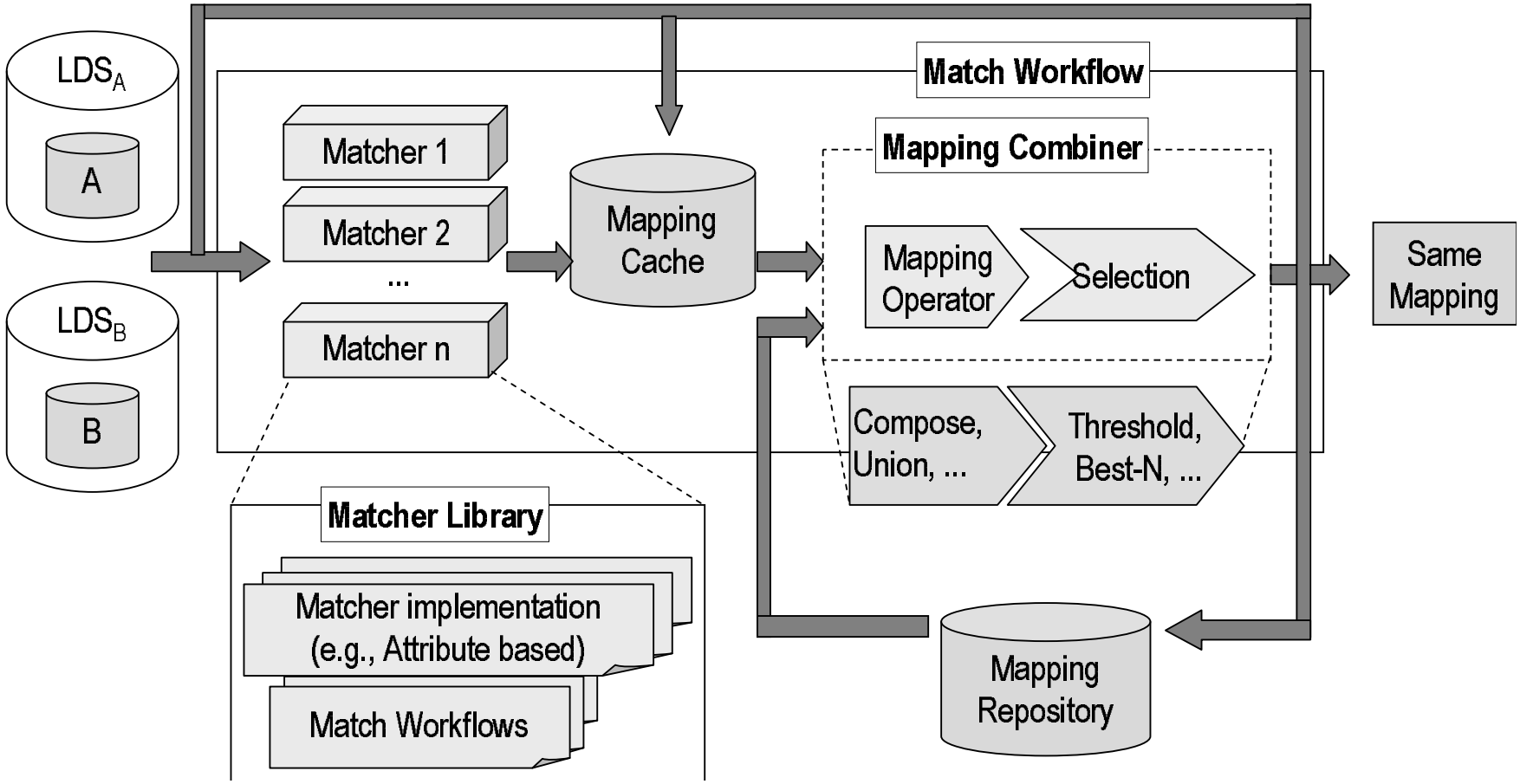


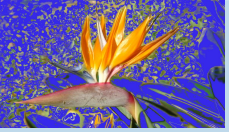
Mapping-Verarbeitung: MOMA-Ansatz

- Verarbeitung von Mappings und Objektinstanzen durch Operatoren
- Kombination der Operatorergebnisse durch Skriptsprache (iFuice*)
  - Prozedurale Programmiersprache mit Kontrollstrukturen (IF-THEN-ELSE, WHILE-DO)
  - Ergebnisse werden in Variablen gespeichert
  - Definition und Aufruf von Unterprozeduren
- MOMA = Mapping-based Object Matching
  - Definition und Ausführung von Match-Workflows
  - Eingabe: Objektinstanzen und Mappings, Ausgabe: Same-Mapping

* Rahm, E. et. al.: iFuice - Information Fusion utilizing Instance Correspondences and Mappings. WebDB, 2005

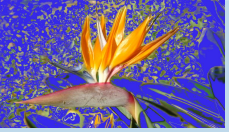
Thor, A., Rahm, E.: MOMA - A Mapping-based Object Matching System. CIDR, 2007





Operatoren: Übersicht (vereinfacht)

- Attributvergleich: $match(O_1, O_2, f) = map$
  - $\{(a, b, s) | a \in O_1, b \in O_2, s = f(a, b)\}$
  - f ist eine Match-Funktion, die für zwei Objekte den Ähnlichkeitswert ermittelt.
- Vereingung: $union(map_1, map_2) = map$
  - $\{(a, b, s) | (a, b, s_1) \in map_1 \vee (a, b, s_2) \in map_2\}$
- Durchschnitt: $intersect(map_1, map_2) = map$
  - $\{(a, b, s) | (a, b, s_1) \in map_1 \wedge (a, b, s_2) \in map_2\}$
- Komposition: $compose(map_1, map_2) = map$
  - $\{(a, b, s) | (a, x, s_1) \in map_1, (x, b, s_2) \in map_2\}$
- Weitere (Hilfs-)Operatoren
  - Selektion, z.B. alle Korrespondenzen deren Ähnlichkeitswert über einem Schwellwert liegen



Kombination: Vereinigung / Durchschnitt

- Ermittlung des kombinierten Ähnlichkeitswertes s durch Ähnlichkeitsfunktion $f(s_1, s_2)$
- Funktionen
  - Maximum (Max), Durchschnitt (Avg), Minimum (Min)
  - Ranked: $f(s_1, s_2) = s_1$, wenn $(a, b, s_1) \in map_1$, sonst s_2
- Umgang mit fehlenden Ähnlicheitswerten (relevant für Avg und Min)
  - Ignorieren oder "gleich Null setzen"

map1		
A	B	s
a1	b1	1
a2	b2	0.8

union (Max)		
A	B	s
a1	b1	1
a2	b2	0.8
a3	b3	0.6

union (Avg)		
A	B	s
a1	b1	0.8
a2	b2	0.8
a3	b3	0.6

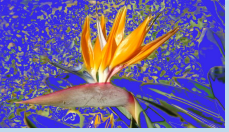
union (Min)		
A	B	s
a1	b1	0.6
a2	b2	0.8
a3	b3	0.6

map2		
A	B	s
a1	b1	0.6
a3	b3	0.6

union (Ranked)		
A	B	s
a1	b1	1
a2	b2	0.8
a3	b3	0.6

union (Avg-0)		
A	B	s
a1	b1	0.8
a2	b2	0.4
a3	b3	0.3

union (Min-0)		
A	B	s
a1	b1	0.6
a2	b2	0
a3	b3	0

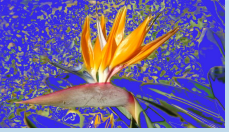


Kombination: Vereinigung / Durchschnitt (2)

- Evaluation für Publikationen von DBLP und ACM für drei attributbasierte Match-Verfahren

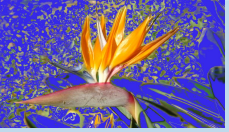
	Titel (Trigram)	Autoren (Trigram)	Jahr (Gleichheit)	Union-Avg (Filter:80%)
Precision	86,7%	38,0%	0,4%	97,3%
Recall	97,7%	87,9%	100,0%	93,9%
F-Measure	91,9%	53,1%	0,8%	95,5%

- Fazit
  - Kombination kann Match-Qualität steigern
  - Vereinigung verbessert Recall (evtl. auf Kosten der Precision)
  - Durchschnitt verbessert Precision (evtl. auf Kosten des Recalls)
  - Wahl der Ähnlichkeitsfunktion von Match-Problem abhängig



Komposition

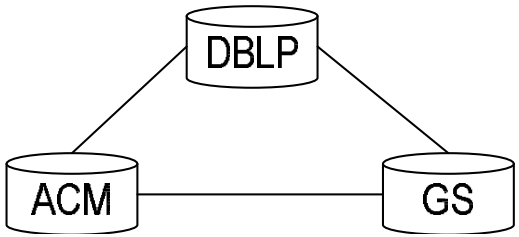
- $compose(map_1, map_2) = \{(a, b, s') | (a, x, s_1) \in map_1, (x, b, s_2) \in map_2\}$
- Ermittlung des kombinierten Ähnlichkeitswertes s durch zwei Ähnlichkeitsfunktionen, da Korrespondenz zwischen zwei Objekten bei Komposition durch mehrere Pfade erreicht werden kann
  - Horizontal: Bestimmung des Ähnlichkeitswerts eines Pfades
    - Min, Max, Avg, Left ($= s_1$), Right ($= s_2$)
  - Vertikal: Bestimmung des Ähnlichkeitswerts einer Korrespondenz aus den zugehörigen Pfad-Ähnlichkeitswerten
    - $Dice = 2 \cdot \frac{s(a,b)}{n(a)+n(b)}$
    - $DiceLeft = \frac{s(a,b)}{n(a)}$, $DiceRight = \frac{s(a,b)}{n(b)}$
    - $DiceMin = \frac{s(a,b)}{min(n(a)+n(b))}$
- Dabei sei
  - $s(a, b) =$ Summe der Ähnlichkeitswerte aller Pfade (a, b)
  - $n(a) =$ Anzahl der Korrespondenzen $(a, x) \in map_1$
  - $n(b) =$ Anzahl der Korrespondenzen $(x, b) \in map_2$

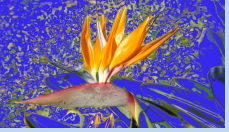


Komposition (2)

- Evaluation für Publikationen von DBLP, ACM und GS (F-Measure)

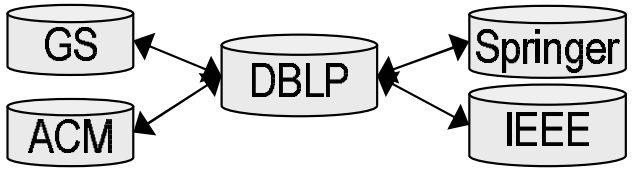
Mapping Compose via	DBLP - GS ACM	DBLP - ACM GS	GS - ACM DBLP
Direkt	81,3%	91,9%	35,3%
Compose	33,9%	63,7%	83,9%
Union	81,3%	91,6%	83,7%



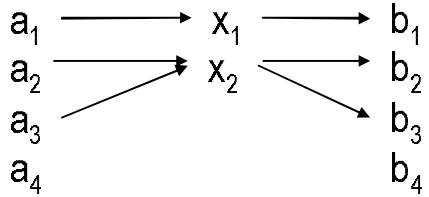


Komposition (3)

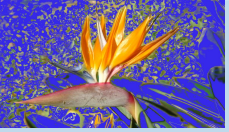
- Fazit
  - Komposition von Mappings ermöglicht effiziente Berechnung neuer Mappings
  - Besonders gut geeignet, falls Hub-Datenquelle vorhanden ist (Sternstruktur)
  - Fehlende Objekte in "mittlerer" Quelle führen zu fehlenden Korrespondenzen (Bsp: $a_4 - b_4$)
  - Komposition kann zu falschen Korrespondenzen führen (Bsp: $a_2 - b_3$)



Hub-Struktur

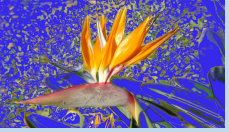


Problemfälle bei Komposition



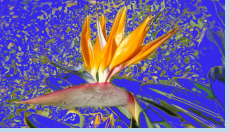
Neighborhood-Matcher: Motivation und Idee

- Motivation: Wertevergleich für heterogene Objekte schwierig
- Beispiel für gleiche Konferenzen
  - "Proceedings of the 27th International Conference on Very Large Databases" vs. "Proc. of VLDB 2001, Italy"
- Lösung 1: Match-Verfahren mittels Domänenwissen
  - Abkürzungen, z.B. VLDB = Very Large Databases
  - Zuordnungen, z.B. "VLDB 2001" = "27. VLDB"
  - ...
- Problem: Woher kommt Domänenwissen? Bei jeder Domäne anders!
- Lösung 2: Verwendung assoziierter Informationen
  - Beispiel: "Zwei Konferenzen sind gleich, wenn die Menge der zugehörigen Publikationen gleich sind."
  - Mögliche Abschwächungen: alle $\rightarrow$ viele, gleich $\rightarrow$ ähnlich



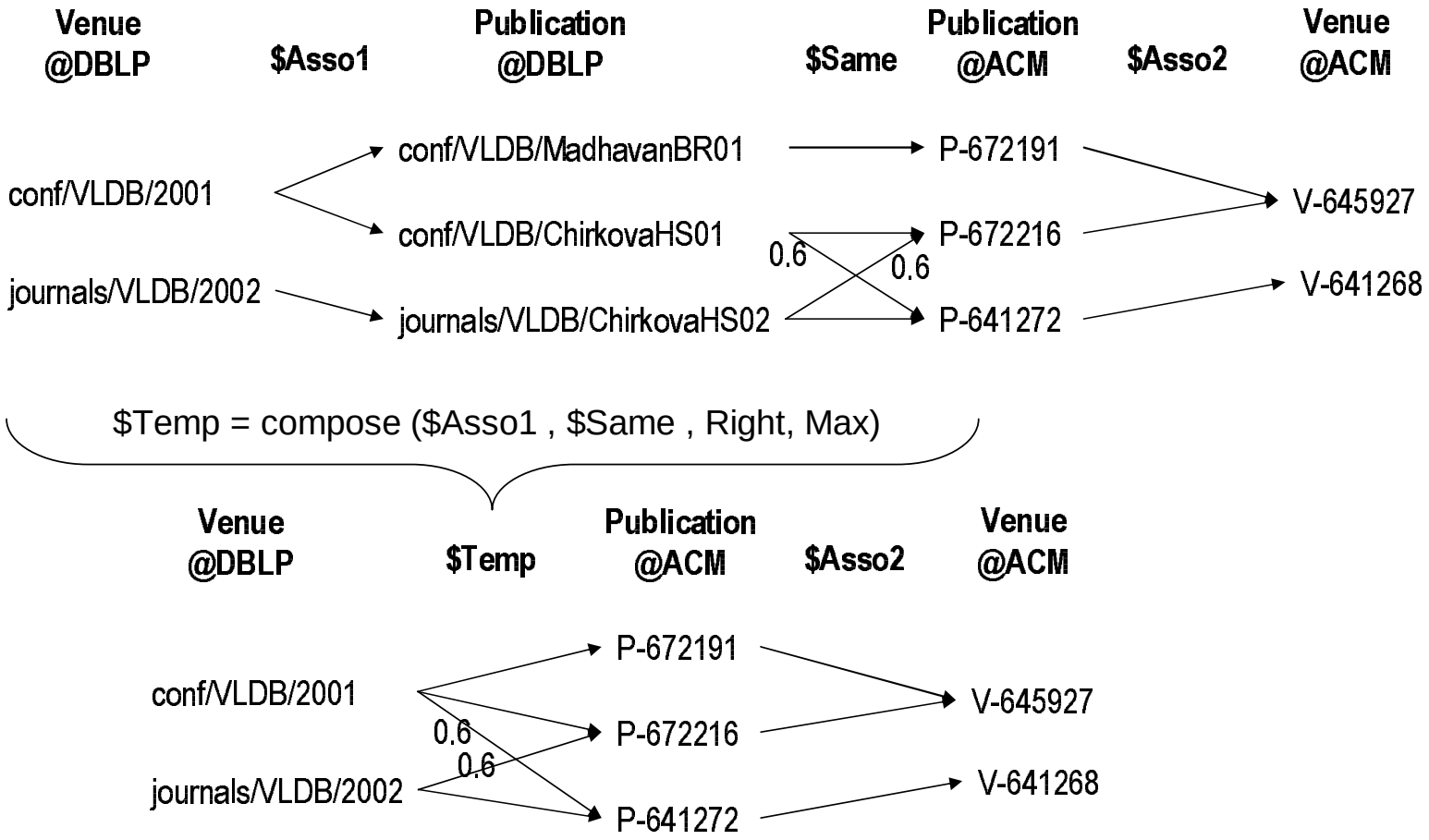
Neighborhood-Matcher: Match-Workflow

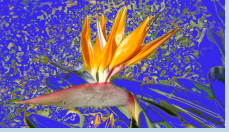
- Verwendung von Assoziations-Mappings
  - Syntax: Gleicher Struktur wie Same-Mappings; fester "Ähnlichkeitswert" = 1
  - Semantik: Korrespondenzen zwischen assoziierten Objekten, z.B. Publikationen - Venue
- Match-Workflow als Kompositon von drei Mappings
  - map1 und map3 sind Assoziations-Mappings; map2 ist ein Same-Mapping
- Idealfall (rechts) nicht immer erreicht, da
  - Assoziations-Mappings unvollständig, z.B. nicht alle Publikationen in jeder Datenquelle zu jedem Venue verfügbar
  - Same-Mapping fehlerhaft, z.B. als Ergebnis eines automatischen Match-Verfahrens



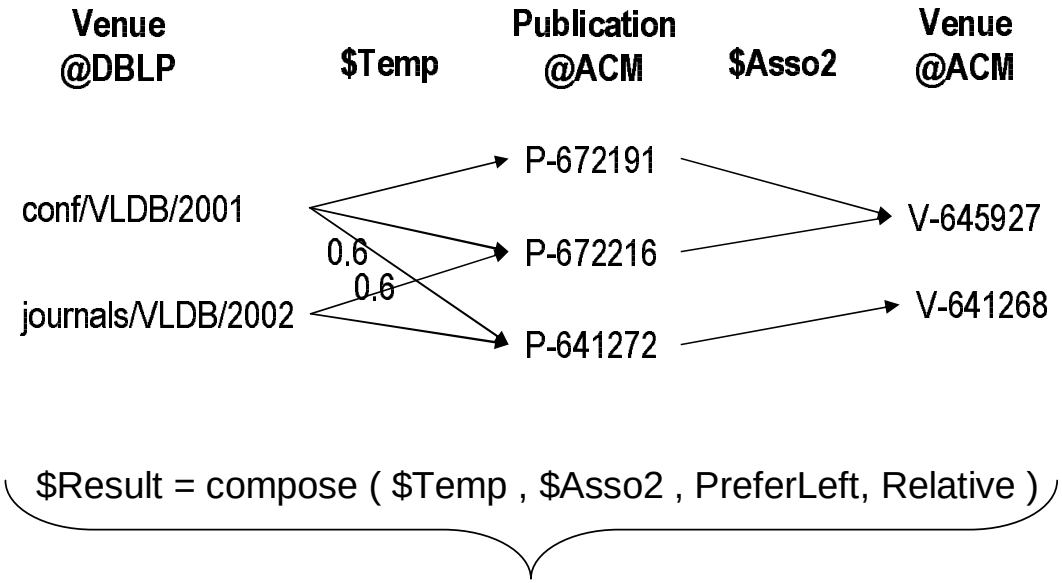
Neighborhood-Matcher: Beispiel (1)

- Ähnlichkeitswerte = 1 (solange nicht anders angegeben)

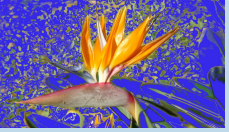




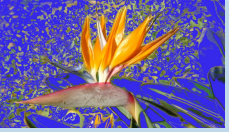
Neighborhood-Matcher: Beispiel (2)



Venue@DBLP	Publication@ACM	Ähnlichkeitswert s			
		DiceMin	DiceLeft	DiceRight	Dice
conf/VLDB/2001	V-645927	(1+1) / 2 = 1	(1+1) / 3 = 0.67	(1+1) / 2 = 1	2*(1+1) / (3+2) = 0.8
conf/VLDB/2001	V-641268	0.6 / 1 = 0.6	0.6 / 3= 0.2	0.6 / 1 = 0.6	2*0.6 / (3+1) = 0.3
journals/VLDB/2002	V-645927	0.6 / 2 = 0.3	0.6 / 2 = 0.3	0.6 / 2 = 0.3	2*0.6 / (2+2) = 0.3
journals/VLDB/2002	V-641268	1 / 1 = 1	1 / 2 = 0.5	1 / 1 = 1	2*1 / (2+1) = 0.67



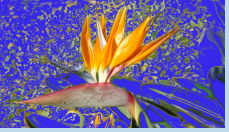
Qualität



Fehlerquellen

- Datenerfassung (Schreibfehler, Falschangaben (keine Wiedererkennung), Platzfüller bei Pflichtfeldern, wenn Wert fehlt, Meßfehler, ...)
- Alterung (Daten nicht aktuell und deshalb falsch)
- Tranformationsfehler, darunter auch falche Behandlung bei Integration !

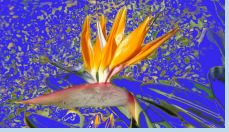
Folgen: Wert der Daten sinkt! Fehler führen zu Folgefehlern.
$\Rightarrow$ Fehlerbereinigung (Normalisierung, fehlende Werte korrekt Darstellen, ...
Diskussion: Fehler vs. Ausreißer $\rightarrow$ bisher keine LÖsung ohne den Experten



Qualitätskriterien - Auswahl

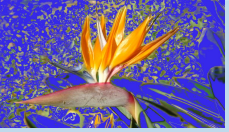
Was ist Qualität ? Meist nur *fitness for use*.

- Techn. Parameter (Verfügbarkeit, Zugriffszeit, Antwortzeit, Datenaktualität; Homogene Darstellung)
- Semantische P.: Metainformationen, Quellenangabe, Mehrwert durch Integration, Vollständigkeit, Fehlerfreiheit)
  Aber auch:
  Objektivität der Ziele, neutrale bzw. objektive Quellenwahl
  Reputation: bisher Quellen durch Experten bewertet, Wertung an Objekte (Attribute, Konzepte gebunden, Nutzerbefragungen.
- Bei Bio-DB haben sich einige (mit manuellem Einsatz) gepflegte Datenquellen etabliert. $\rightarrow$ Qualität durch Expertenwissen.
- Offene Fragen:
  Kann aus Qualität der Quellen und Ähnlichkeitsmaß auf Qualität des Ergebnisschema geschlossen werden?



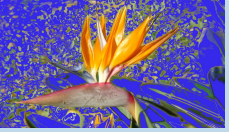
Qualitätsbewertung: Precision und Recall

- Annahme: Es gibt ein (z.B. manuell erstelltes) perfektes Match-Ergebnis (Mapping) map_{Perf}
- Verwendung der Qualitätsmaße aus dem Information Retrieval zur Bewertung von map_{Match}
  - $Precision = |map_{Match} \cap map_{Perf}|/|map_{Perf}|$
  - $Recall = |map_{Match} \cap map_{Perf}|/|map_{Match}|$
  - $F - Measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$
- Vorteile
  - Effektive Vergleichbareit verschiedener Match-Verfahren
  - Möglichkeit zur Optimierung von Match-Verfahren (z.B. Schwellwert bei Filterung)
- Nachteile
  - Vorhandensein des perfekten Mappings erforderlich



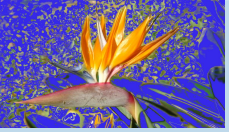
Qualitätsbewertung: Match Ratio und Match Coverage

- Perfektes Mapping nicht immer vorhanden
  - viele Datenquellen (P2P-Umfeld) + große Datenquellen → viele, große, manuell zu erstellende/verifizierende perfekte Mappings → großer Aufwand
  - perfektes Mapping nicht immer eindeutig
- Abschätzung von Precision und Recall durch Match Ratio und Match Coverage



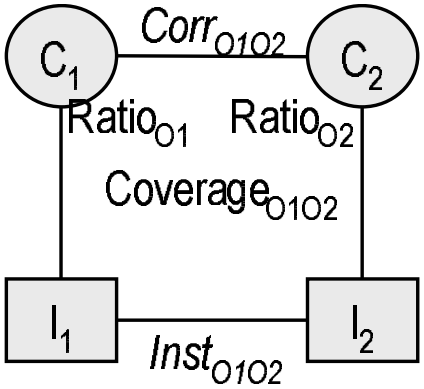
Qualitätsbewertung: Match Ratio und Match Coverage (2)

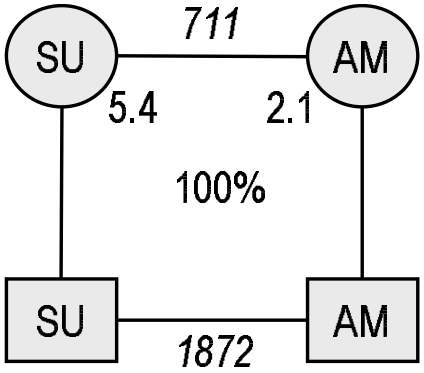
- Match Ratio ("Precision") $= \frac{|Corr_{O_1-O_2}|}{|C_{O_1}|}$ bzw. $= \frac{|Corr_{O_1-O_2}|}{|C_{O_2}|}$
  - Durchschnittliche Anzahl der Korrespondenzen (= Match-Partner) pro Konzept in O_i, dass einen mindestens einen Match-Partner hat
- Match Coverage ("Recall") $= \frac{|C_{O_1}|+|C_{O_2}|}{|C_{Base-O_1}|+|C_{Base-O_2}|}$
  - Anteil der Konzepte mit mind. einem Match-Partner im Vergleich zum Matching mit Baseline-Ähnlichkeit
- Dabei bedeuten
  - $|Corr_{O_1-O_2}|$ = Anzahl der Korrespondenzen des Mappings
  - $|C_{O_i}|$ = Anzahl der "gematchten" Konzepte (d.h. mind. ein Match-Partner) in O_i
  - $|C_{Base-O_i}|$ = Anzahl der "gematchten" Konzepte in O_i mit Baseline-Ähnlichkeit (d.h. Korrespondenz zwischen zwei Konzepten g.d.w. mind. eine gleiche Instanz zugeordnet)



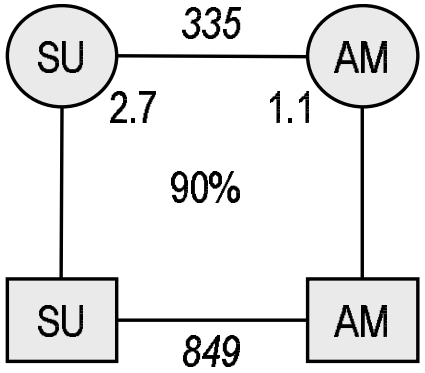
Qualitätsbewertung: Match Ratio und Match Coverage (Beispiel)

- Matching zwischen Produktkatalogen von Softunity und Amazon
  - Min-Ähnlichkeit sehr gut: ähnliche Match Coverage wie Baseline, geringere Match Ratio
  - Dice-Ähnlichkeit sehr restriktiv: Match Ratio ≈ 1, geringe Match Coverage

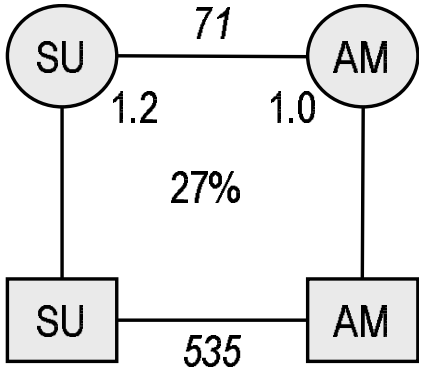




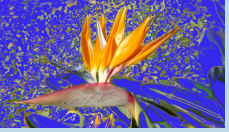
Baseline



Min (100%)



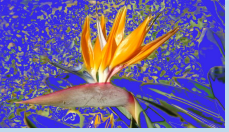
Dice (50%)



Qualität in P2P-Systemenn

- P2P-Systeme: Bewertung der Daten, des Servers, des Bewerters.
  Isolation von schlechten Peers, von schlechten Inhalten,
  Verfahren z.T. widerstandsfähig gegen Manipulation.
  Details: **Vorlesung** D. Sosna: P2P-Systeme und Datenbanken.
- Adaptionen:
  keine Angriffe, aber: verschiedene Lehrmeinungen, verschiedene Ontologien.
  Reputation durch Auswertung des Nutzungsverhaltens z.B. bei linkbasierten Systemen (indirekte Nutzung der Expertenkompenz)

... bisher nicht realisiert !



Zusammenfassung

- Schemaintegration: Ziel der Automatisierung nicht erreicht; Teilautomatisierung
- Ansätze: Metadatenbasiertes bzw instanzdatenbasiertes Mapping. Kombination verschiedener Ansätze erforderlich, Wiederverwendung von als gut erkannten Ergebnissen.
- Qualität: Begriffsbildungen vielfältig, z.T. weich, schwierige Einschätzung, subjektive Bewertung.