



**Hasso  
Plattner  
Institut**

IT Systems Engineering | Universität Potsdam

Seminar - Einführung

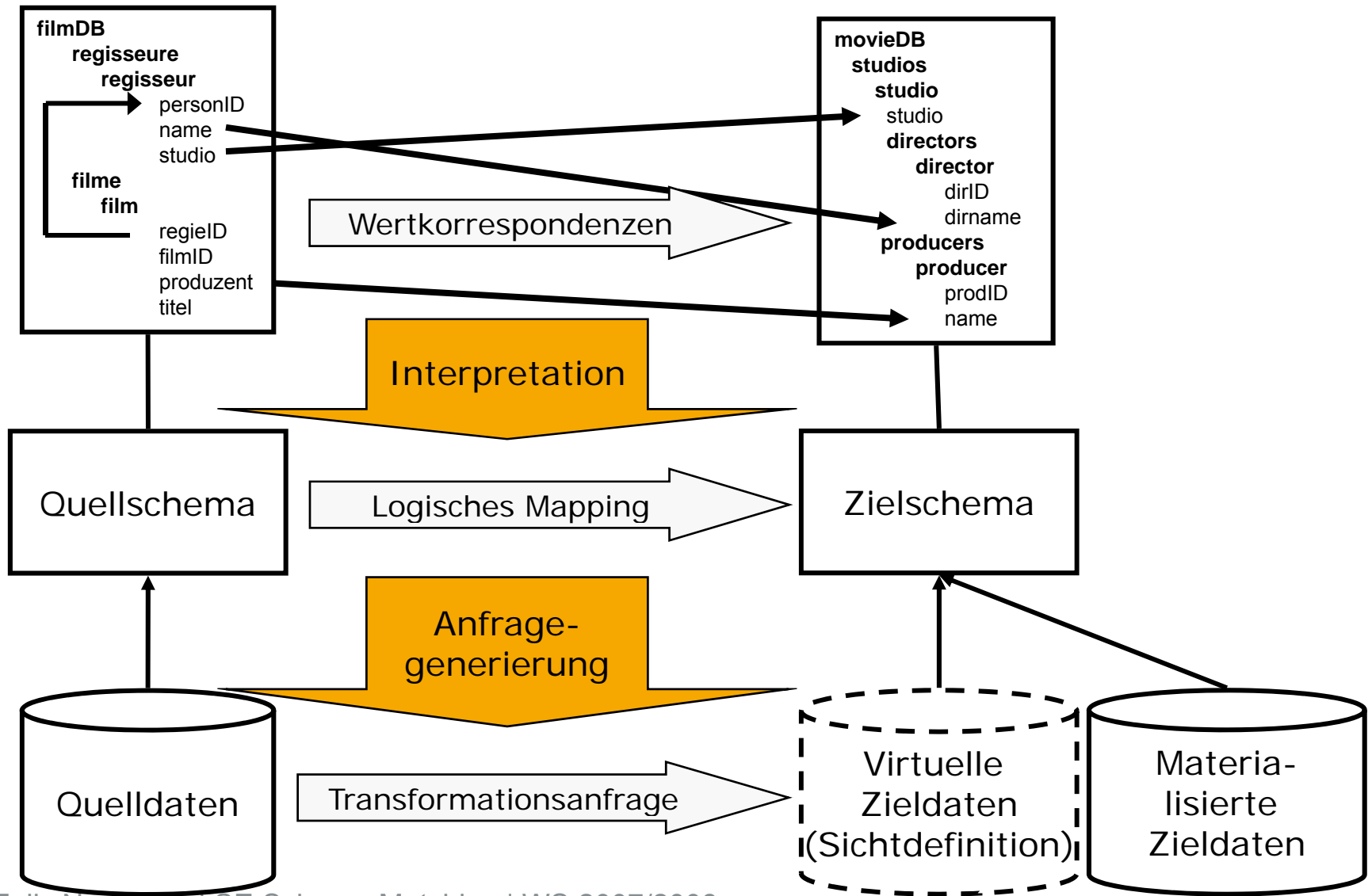
## Schema Matching

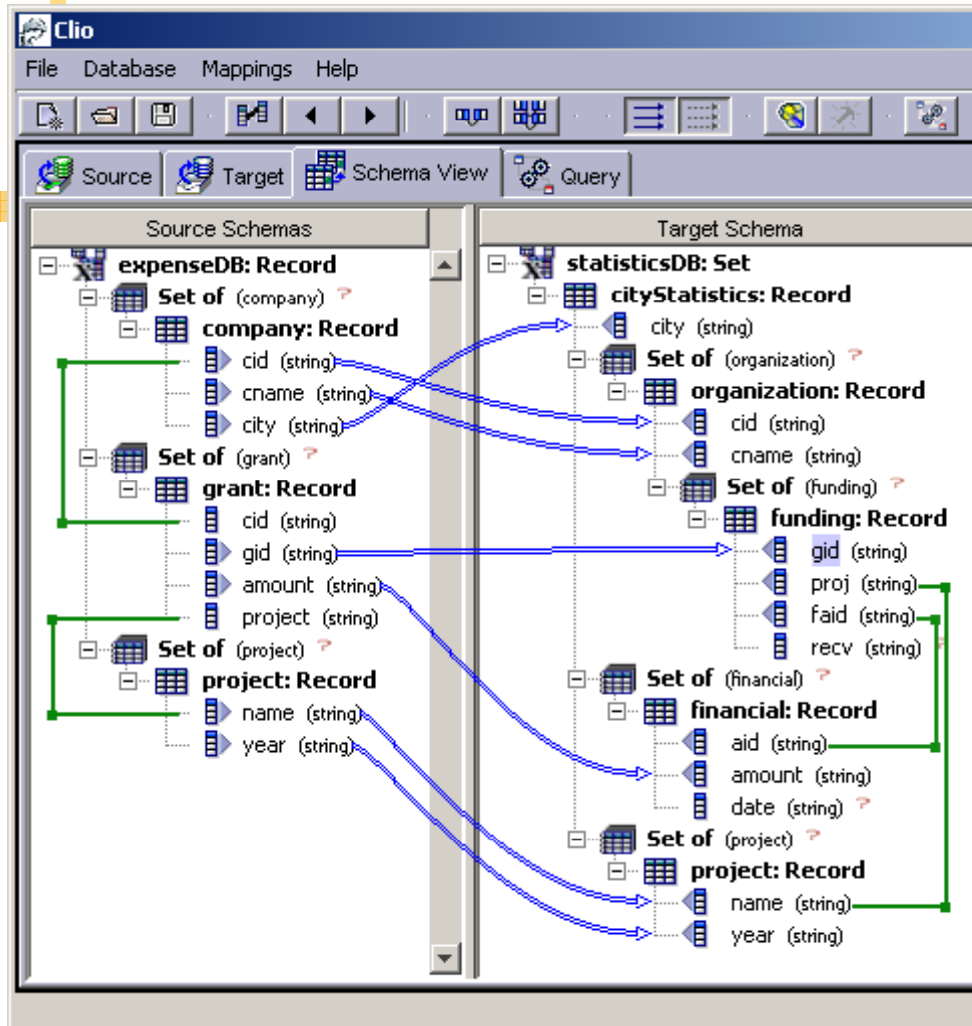
24.10.2007

Felix Naumann

# Schema Mapping im Kontext

2





Clío

File Database Mappings Help

Source Target Schema View Query

XQuery XSL SQL Query

```

RETURN
  $x0L1/project/text() = $x1L1/name/text() AND
  $x2L1/cid/text() = $x0L1/cid/text() AND
  $x2/city/text() = $x2L1/city/text()
RETURN
<organization>
  <cid> $x0L1/cid/text() </cid>,
  <cname> $x2L1/cname/text() </cname>,
  distinct (
  FOR
    $x0L2 IN $doc/expenseDB/grant ,
    $x1L2 IN $doc/expenseDB/project ,
    $x2L2 IN $doc/expenseDB/company
  WHERE
    $x0L2/project/text() = $x1L2/name/text() AND
    $x2L2/cid/text() = $x0L2/cid/text() AND
    $x2L1/cname/text() = $x2L2/cname/text() AND
    $x2L1/city/text() = $x2L2/city/text() AND
    $x0L1/cid/text() = $x0L2/cid/text()
  RETURN
    <funding>
      <gid> $x0L2/gid/text() </gid>,
      <proj> $x0L2/project/text() </proj>,
      <faid> "SK267(", $x0L2/project/text(), " ,
    </funding> )
  </organization> ,
  distinct (
  FOR
    $x0L1 IN $doc/expenseDB/grant ,
    $x1L1 IN $doc/expenseDB/project ,
    $x2L1 IN $doc/expenseDB/company
  WHERE

```

Preview Execute Query Copy to Clipboard

No File

# Schema Matching – Motivation

4

- Große Schemas
  - > 100 Tabellen, viele Attribute
  - Bildschirm nicht lang genug
- Unübersichtliche Schemas
  - Tiefe Schachtelungen
  - Fremdschlüssel
  - Bildschirm nicht breit genug
  - XML Schema
- Fremde Schemas
  - Unbekannte Synonyme
- Irreführende Schemas
  - Unbekannte Homonyme
- Fremdsprachliche Schemas
- Kryptische Schemas
  - |Attributnamen| ≤ 8 Zeichen
  - |Tabellennamen| ≤ 8 Zeichen

The screenshot shows the Clío application window with a menu bar (File, Schemas, Mappings, Help) and a toolbar. The main area is split into 'Source schemas' and 'Target Schema'. The 'Source schemas' pane displays a hierarchical tree of XML Schema elements. The tree includes elements like `author: Record`, `pubMed: Record`, `Set [0,*] ?`, `ev: Record`, `citedInBook: Record`, `Record`, `editors: Record`, `editor: Record`, `author: Record`, `Set [1,*]`, `Set [0,*] ?`, `ev: Record`, and `observations: Record`. Each element is annotated with various attributes such as `@ name (string)`, `@ value (unsignedInt) ?`, `@ status (NMTOKEN) ?`, `@ medlineID (unsignedInt) ?`, `@ type (string) ?`, `@ href (anyURI)`, `@ role (anyURI) ?`, `@ title (string) ?`, `@ label (string) ?`, `@ value() (string) ?`, `@ xlink:type (string) ?`, `@ role (anyURI) ?`, `@ title (string) ?`, `@ volume (unsignedInt) ?`, `@ year (gYear)`, `@ first (unsignedInt)`, `@ last (unsignedInt)`, `@ ISBN (string) ?`, `@ publisher (string)`, `@ city (string)`, `@ country (string) ?`, `@ title (string)`, `@ bookTitle (string)`, `@ name (string)`, `@ href (anyURI)`, `@ role (anyURI) ?`, `@ title (string) ?`, `@ label (string) ?`, `@ value() (string) ?`, `@ type (string) ?`, `@ href (anyURI)`, `@ role (anyURI) ?`, `@ title (string) ?`, `@ label (string) ?`, `@ value() (string) ?`, `@ type (string) ?`, `@ href (anyURI)`, `@ role (anyURI) ?`, `@ title (string) ?`, `@ label (string) ?`, `@ value() (string) ?`, and `@ xlink:type (string) ?`.

Two yellow callout boxes are overlaid on the image:

- The first callout box, pointing to a vertical scrollbar on the right side of the source schema tree, contains the text: "Man beachte die Scrollbar!".
- The second callout box, pointing to a deeply nested `Set [1,*]` element, contains the text: "Man beachte die Schachtelungstiefe!".

er  
ut

# Schema Matching – Motivation

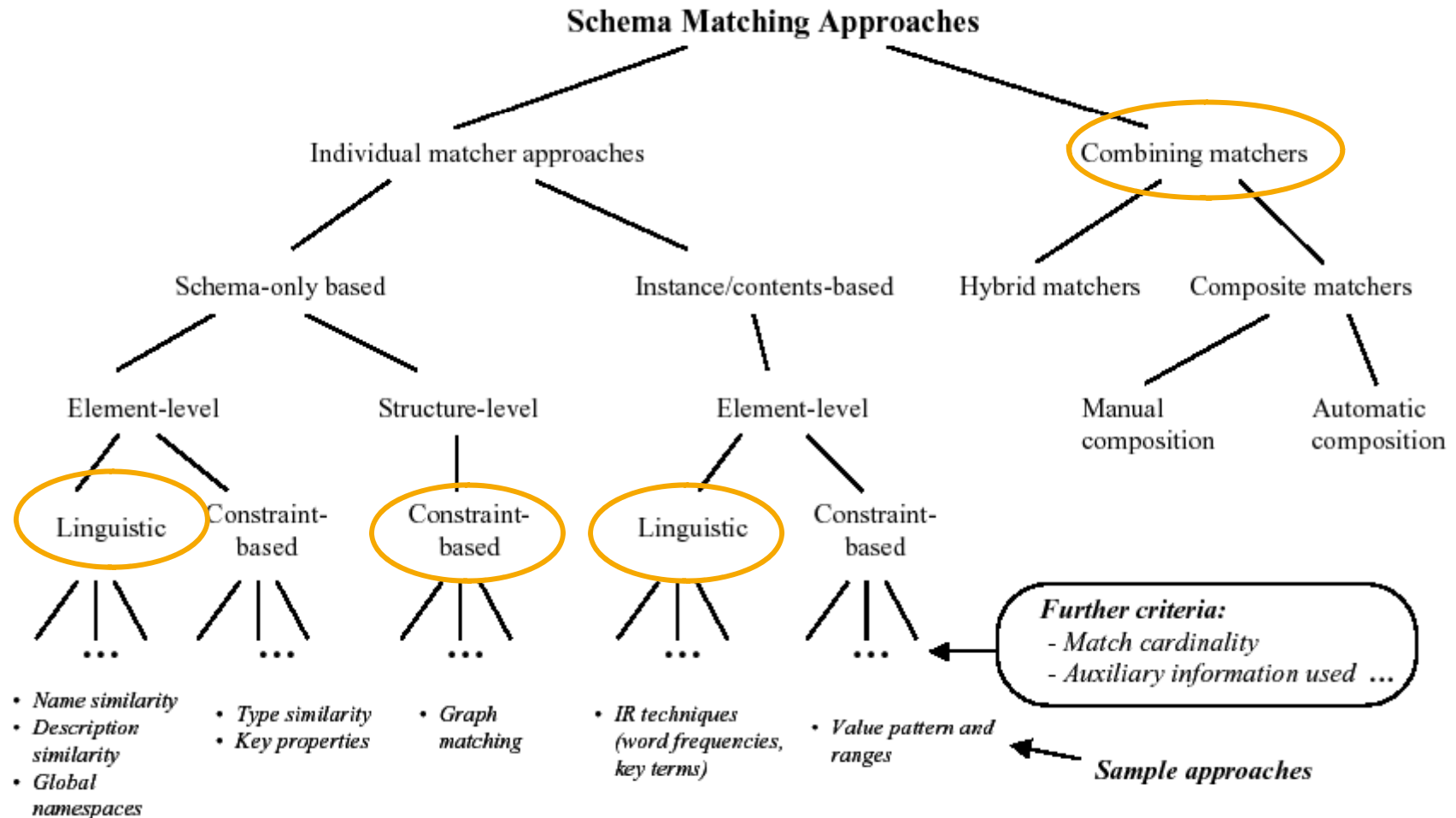
6

## Die Folgen

- Falsche Korrespondenzen (false positives)
- Fehlende Korrespondenzen (false negatives)
- Frustration
  - User verlieren sich im Schema
  - User verstehen Semantik der Schemas nicht

# Schema Matching Klassifikation nach [RB01]

7



# Schema Matching Klassifikation

8

Schema Matching basierend auf

- Namen der Schemaelemente (*label-based*)
- Darunterliegende Daten (*instance-based*)
- Struktur des Schemas (*structure-based*)
- Mischformen



## Schema Matching – Label-based

9

Gegeben zwei Schemata mit Attributmengen A und B

Kernidee:

- Bilde Kreuzprodukt aller Attribute aus A und B.
- Für jedes Paar vergleiche Ähnlichkeit bezgl. Attributnamen (Label).
  - Z.B. Edit-distance
- Ähnlichste Paare sind Matches

Probleme:

- Effizienz
- Auswahl der besten Matches (globales Matching)
  - Iterativ?
  - Stable Marriage?
- Synonyme und Homonyme werden nicht erkannt

## Schema Matching – Label-based

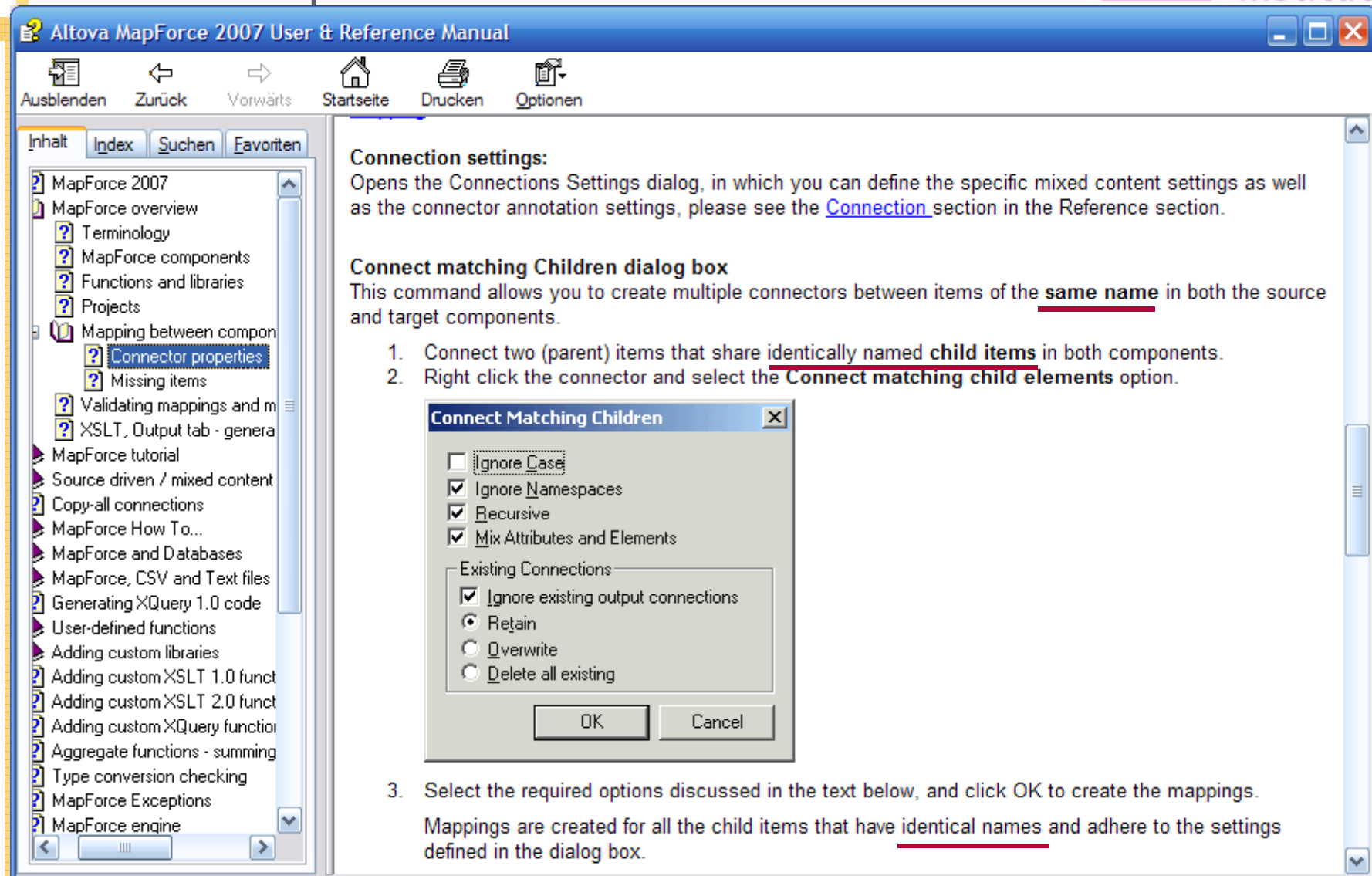
10

Stand der Technik in kommerziellen Produkten

- Label-based
- Namensgleichheit
- Kein globales Matching
- Keine Ähnlichkeitsmaße
- Kein Instanz-basiertes Matching

# Altova MapForce 2007

11



**Altova MapForce 2007 User & Reference Manual**

**Connection settings:**  
 Opens the Connections Settings dialog, in which you can define the specific mixed content settings as well as the connector annotation settings, please see the [Connection](#) section in the Reference section.

**Connect matching Children dialog box**  
 This command allows you to create multiple connectors between items of the same name in both the source and target components.

1. Connect two (parent) items that share identically named child items in both components.
2. Right click the connector and select the **Connect matching child elements** option.

**Connect Matching Children**

Ignore Case

Ignore Namespaces

Recursive

Mix Attributes and Elements

Existing Connections:

Ignore existing output connections

Retain

Overwrite

Delete all existing

OK Cancel

3. Select the required options discussed in the text below, and click OK to create the mappings.  
 Mappings are created for all the child items that have identical names and adhere to the settings defined in the dialog box.

# Schema Matching – Instance-based

12

Gegeben zwei Schemata mit Attributmengen A und B, jeweils mit darunterliegenden Daten.

Kernidee

- Für jedes Attribute extrahiere interessante Eigenschaften der Daten
  - Buchstabenverteilung, Länge, etc.
- Bilde Kreuzprodukt aller Attribute aus A und B.
- Für jedes Paar vergleiche Ähnlichkeit bzgl. der Eigenschaften

Probleme

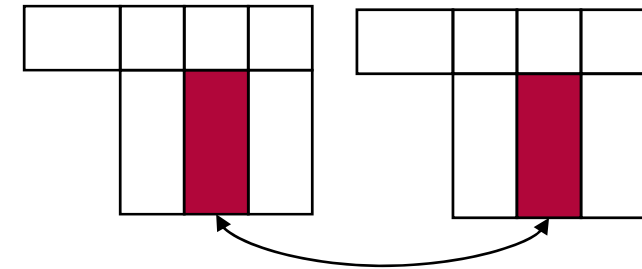
- Auswahl der Eigenschaften
- Datenmenge: Sampling
- Vergleichsmethode, z.B. Naive Bayes
- Gewichtung (Maschinelles Lernen)

# Instanz-basiertes Schema Matching

13

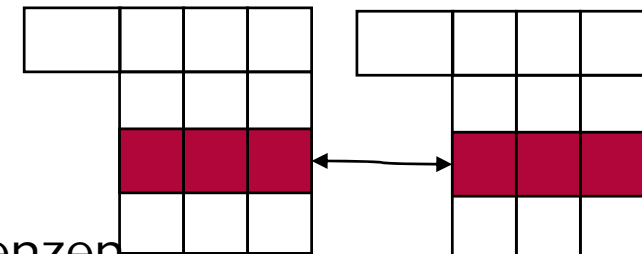
Herkömmliche Lösung: Vertikal

- Vergleich von Spalten
- = Attribut-Klassifikation
- [ICDE'02] u.v.a.m.



Neue Lösung: Horizontal

- Vergleich von Zeilen
- = Duplikaterkennung
  - trotz fehlender Attribut-korrespondenzen
- [ICDE'05]



## Schema Matching – Structure-based

14

- Gegeben zwei Schemata mit Elementmengen A und B.
- Kernidee
  - Nutze (komplexe) Struktur des Schemas aus.
  - Hierarchieebene
  - Elementtyp (Attribut, Relation, ...)
  - Nachbarschaftsbeziehungen

## Schema Matching – Mischformen

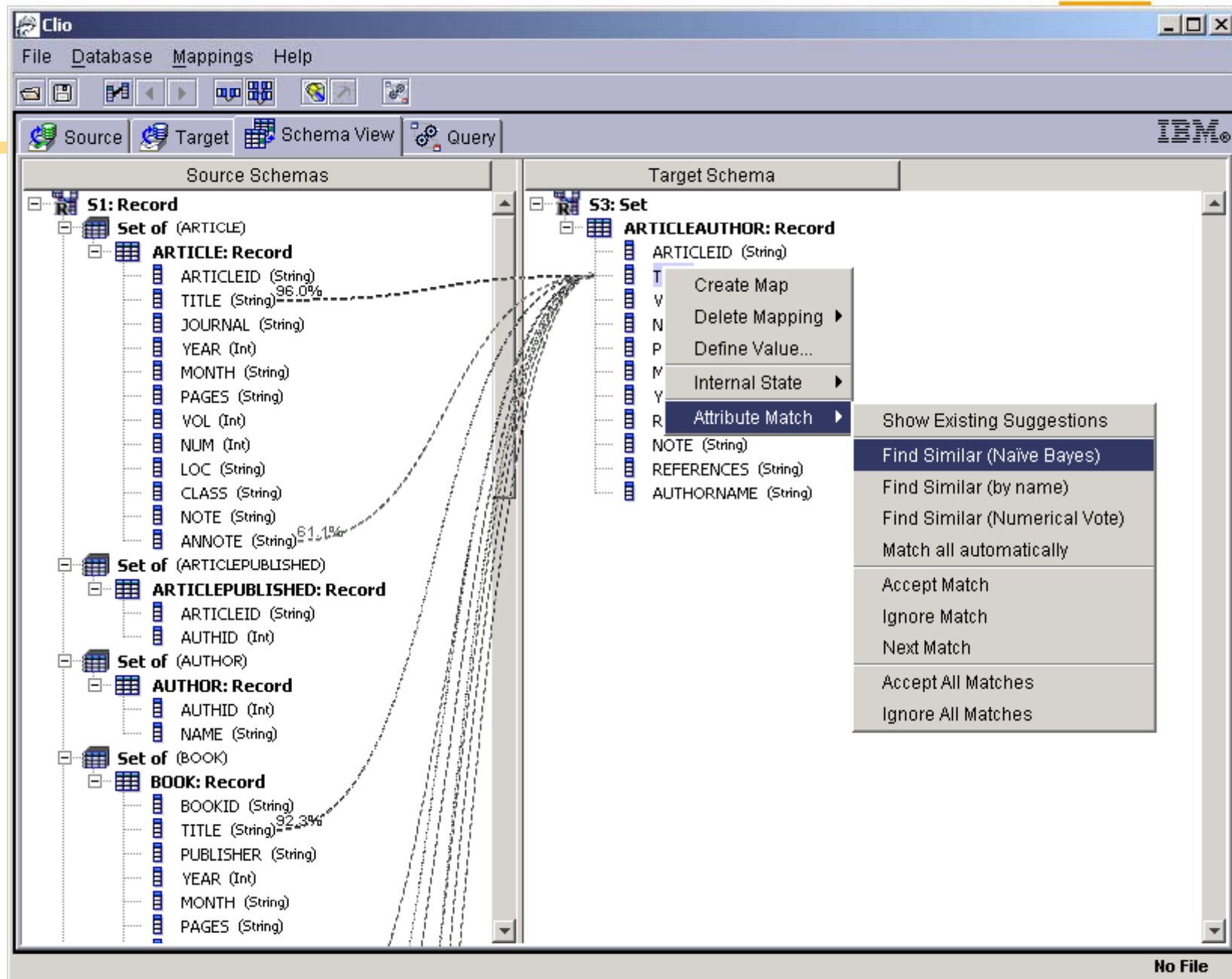
15

### Hybrid

- Gleichzeitige Anwendung mehrerer Techniken
- Bsp: Instance-based + Datentypvergleich

### Composite

- Repertoire bekannter Techniken (inkl. hybrider Techniken)
- Kombination dieser unabhängigen Verfahren
- Bsp: Durch Gewichtung
- Bsp: Durch automatisches Lernen
  - Des besten Verfahrens
  - Einer guten Gewichtung





# Schema Matching – Erweiterungen

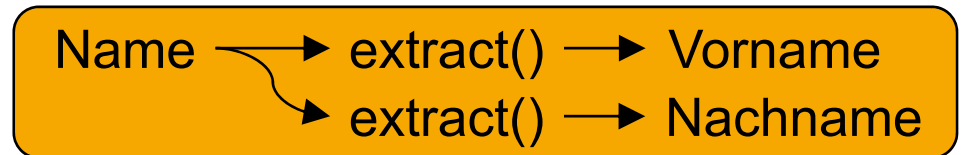
17

- n:1 und 1:n Matches
  - Viele Kombinationsmöglichkeiten
  - Viele Funktionen denkbar
    - Mathematische Operatoren, Konkatination, etc.
  - Parsingregeln
- n:m Matching?
- Matching in komplexen Schemata
  - Ziel: Finde Mapping, nicht Korrespondenzen

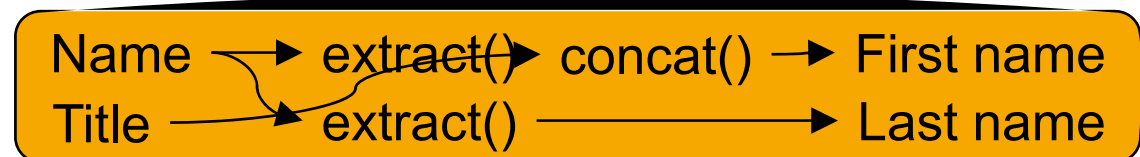
## n:1 Matching



## 1:n Matching



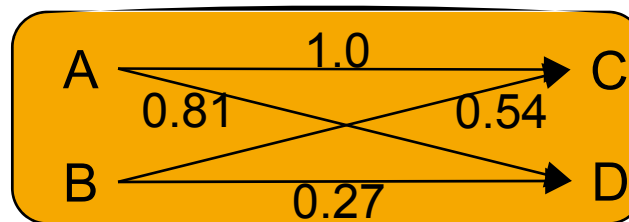
## m:n matching



## Schema Matching – Erweiterungen

18

- Global matching
  - Matche nicht nur einzelne Attribute (oder Attributmengen)
  - Sondern komplette Tabellen oder komplette Schemata
  - Stable Marriage Problem
  - Maximum Weighted Matching



# Schema Matching Kritik

19

- Few (public) use cases
- No good benchmarks
  - Too many and too toy-ish
  - Minimum requirement: 2 interesting schemata + mapping
- Few available prototypes
- No commercial implementation
  - To speak of
- Real schemata are too large
  - Toy examples work fine
  - Screen not wide or long enough
  
- The YAM Effect
  - YAM – Yet Another Matcher

# Phil Bernstein and Sergey Melnik – SIGMOD 2007 keynote

20

- Past goal: Improved precision and recall
  - Big productivity gains are unlikely
- Better goals
  - Return top-k, not best overall match
  - Avoid the tedium. Manage work.
    - Help with scrolling
  - HCI – handle large schemas
  - User studies – what would improve productivity