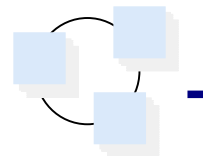


Diplomarbeitsthema:

Integration lokaler Daten in iFuice

Bearbeiter: Sarah Gebhardt

Betreuer: Andreas Thor



Warum eine Integration lokaler Daten?

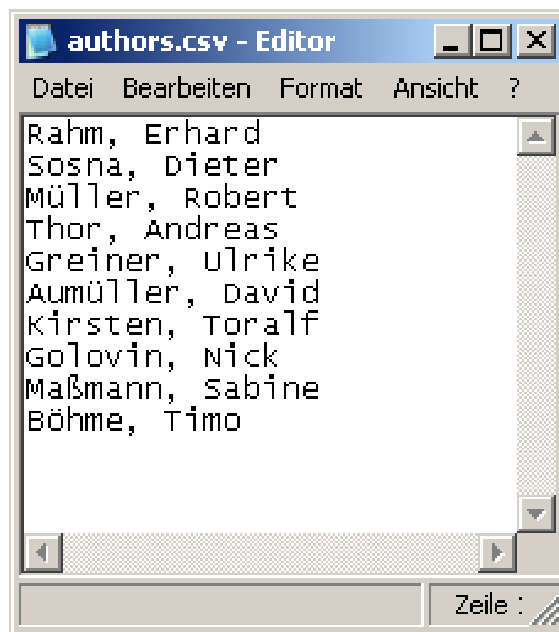
andere Listen im Web,
aber nicht für mich relevant

Welche Informationen sind
überhaupt für mich
interessant

Viele Infos im Web, aber
zu mühsam einzeln
rauszusuchen

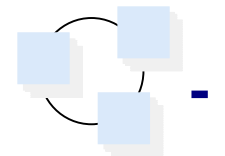
schon gesuchte Infos
nicht mehr aktuell

Private Daten gemeinsam
mit Webdaten darstellen
und verknüpfen



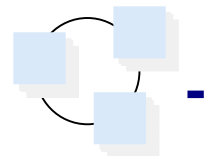
Konkret:

- Welche Autoren interessieren mich überhaupt?
- Von welchen Autoren besitze ich Publikationen
- Wie kann ich stets die aktuellen Publikationslisten der für mich interessanten Autoren abfragen?
- Wie bekomme ich zusätzliche Informationen zu Autoren (Homepage, E-mail)
- ...



Überblick

- 1. Motivation
- 2. iFuice
 - Kurzeinführung
- 3. Architektur
 - Lokale Datenquellen
 - Wrapper
 - Extractor
 - Manipulator
- 4. Beispiel
- 5. Zusammenfassung / Ausblick



iFuice

Information Fusion utilizing Instance Correspondences and Peer Mappings

□ Integrationsplattform

- Datenquellen durch Mappings verbunden
- Mappings durch Mediator gesteuert
- High-Level-Operatoren

□ Siehe Paper und Vortrag von Herrn Andreas Thor

The screenshot shows the iFuice application window. The top menu bar includes 'File' and 'Tools'. Below it are buttons for 'Connect to...', 'Open script', 'Save script', and 'Reference'. The main area is divided into a script editor on the left and a results pane on the right.

Script Editor:

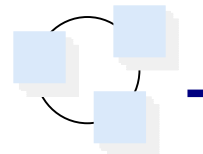
```
Enter script here:  
#sigmod95 := queryInstances {Conference@DBLP, series=  
#sigmod95pubs := map (#sigmod95, DBLP.ConfPub)  
#sigmod95auth := traverse (#sigmod95, DBLP.ConfPub, DI  
#sigmod95temp := getInstances(domain(#sigmod95pubs))  
#sigmod95temp1 := getInstances(range(#sigmod95pubs))
```

Results Pane:

Variable: sigmod95pubs (MappingResult) Set as Tree Root

Source Object ID	Target Object ID
conf/sigmod/95	conf/sigmod/bm095
conf/sigmod/95	conf/sigmod/GatzuGD95
conf/sigmod/95	conf/sigmod/Elison95
conf/sigmod/95	conf/sigmod/Inelink95
conf/sigmod/95	conf/sigmod/Bamp95
conf/sigmod/95	conf/sigmod/Chaudhuri95
conf/sigmod/95	conf/sigmod/Jonker95
conf/sigmod/95	conf/sigmod/WoelkIJOTU95
conf/sigmod/95	conf/sigmod/UH95
conf/sigmod/95	conf/sigmod/UM95
conf/sigmod/95	conf/sigmod/UM95

Source Attribute	Source Value	Target Attribute	Target Value
series	SIGMOD	title	Carrot and InfoSeuth: Da...
title	Proceedings of the 1995 A...	xml_key	conf/sigmod/WoelkIJOTU95
xml_key	conf/sigmod/95	year	1995
year	1995		



Datenstruktur in iFuice

Physische Datenquellen (Physical Data Sources - PDS)

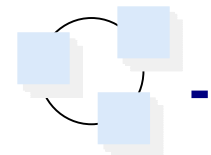
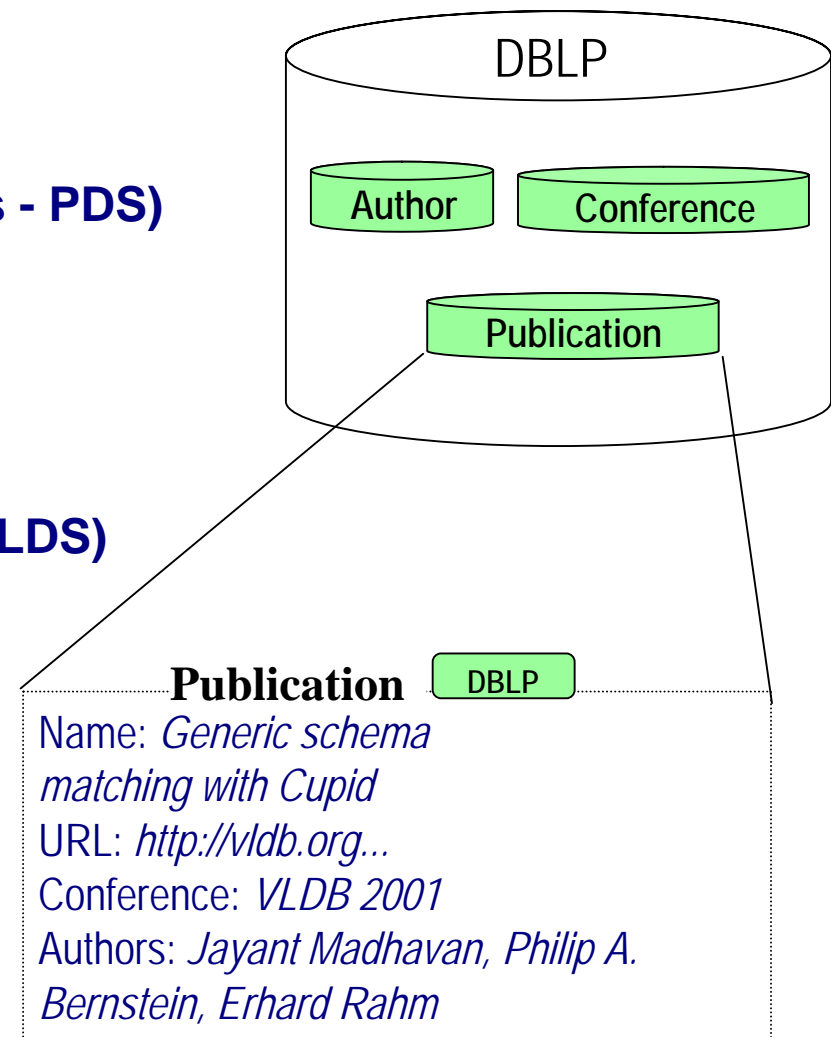
- Webdaten (DBLP), lokale Daten (Dateien)
- Aufgeteilt in logische Datenquellen

Logische Datenquellen (Logical Data Sources - LDS)

- definiert durch eine PDS und einen Objekttyp
- enthalten Objektinstanzen

Objektinstanzen (Object Instances)

- beziehen sich auf ein Objekt der realen Welt
- besitzen eine Menge von Attributen
- ein Attribut entspricht einer eindeutigen ID



Mappings

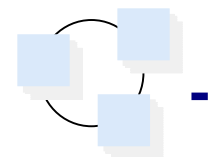
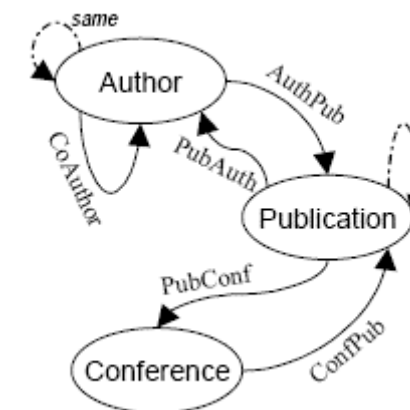
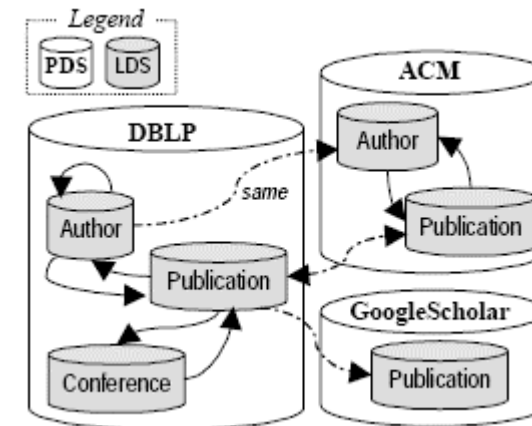
Gerichtete Beziehung zwischen LDS

Metadaten: Bedeutung des Mappings

- **Semantische Mapping Typen**
 - z.B. “publications of author” (PubAuth)
- **Same Mappings vs. Association Mappings**
 - **same** = “Gleichheitsbeziehung” zwischen PDS
 - e.g. LocalAuthors (id) → DBLP authors (id)

Instanzen: Instanzkorrespondenzen

- Materialisiert: Mappingtabellen (z.B. XML-Datei)
- On-the-fly: Ergebnis einer Berechnung (z.B. durch einen Web-Service)



Typische lokale Datenquellen

Unstrukturierte Daten

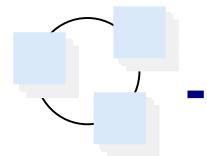
- PDF, Word, PowerPoint
 - Informationen: Publikationen, Berichte, Präsentationen, Skripte

Semistrukturierte Daten

- E-mails
 - Informationen: Korrespondenzen, Informationen (z.B. Name <-> E-mail Adresse)

Strukturierte Daten

- Excel-Tabellen, CSV, Adressbücher
 - Informationen: Listen von Objekten mit ihren Informationen, Verbindungen zwischen Objekten in z.B. Kreuztabellen (Objektkorrespondenzen)



Anforderungen

Finden / Extraktion

- lokale Datenquellen suchen / durchsuchen und Daten extrahieren

Manipulation

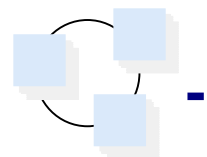
- extrahierte und importierte Daten bearbeiten

Export / Import

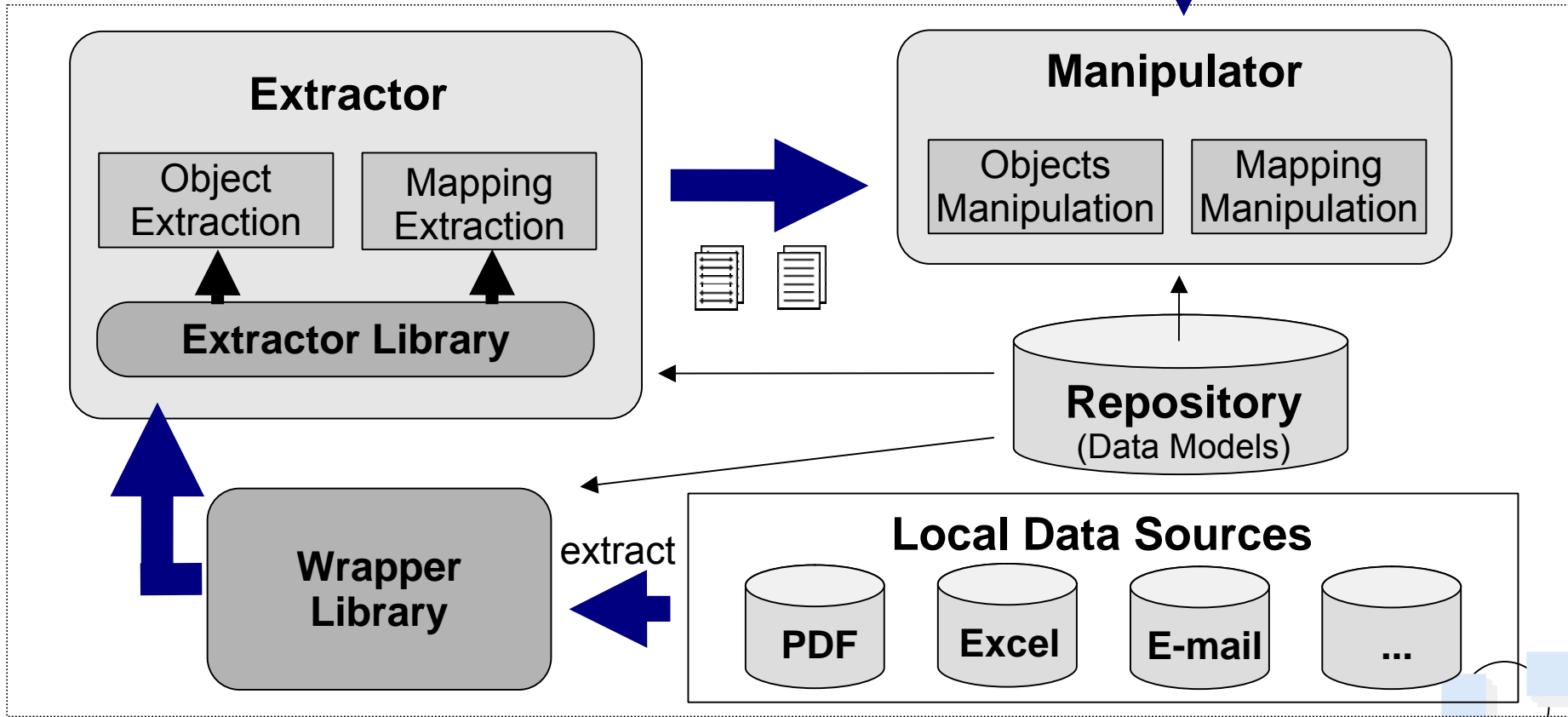
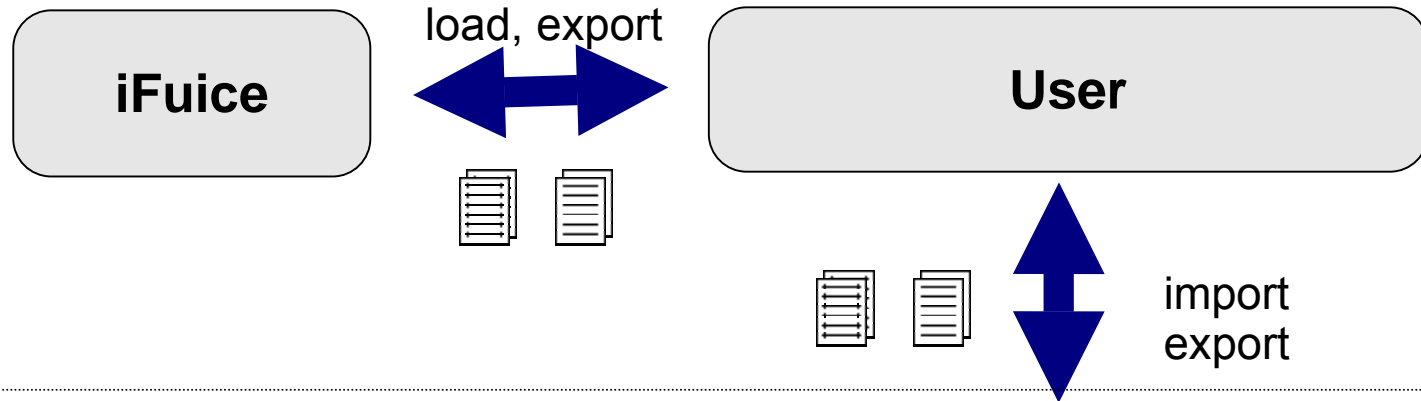
- Objektinstanzen u. Mappings, Format kompatibel mit iFuice (XML)

Aufbau

- modular, erweiterbar



- MappingResults (XML-Format)
- Objectinstances (XML-Format)



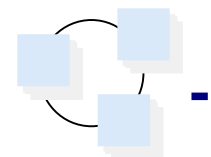
Wrapper Library

Wrapper Module zur Aufbereitung für Extractor

- lokale Daten auslesen / umwandeln
- jeweils einheitliches Zielformat
 - z.B. CSV, Excel, Calc (OO) -> einheitliches Tabellenformat
- ein Extractor Modul für mehrere Formate

Probleme

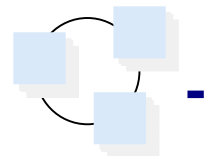
- wie kommt man an Daten
 - Verzeichnis lesen, Datei lesen, Desktop Search (Google Desktop Search)
- Format der Daten
 - offengelegt (PDF), nicht offengelegt (Excel, Word), Bibliotheken
- Struktur



Wrapper Library (Fortsetz.)

PDF Wrapper Modul

- Finden von PDF-Dateien über GDS / Angabe von Verzeichnis
- Grundlage PostScript, originalgetreue Darstellung, Druckausgabe, plattformunabhängig
- Probleme:
 - Struktur, Passwortschutz, gescannte PDFs
- existierende Bibliotheken:
 - frei: pdftohtml, **Pdfbox**, xpdf, itext, ..., kommerziell: Adobe SDK, JPedal, ...
- Erweiterung Text-Extrahierungs-Klasse von Pdfbox
 - fontsize, font-family, text-align, position
- Ergebnis für Extractor
 - Menge von Line-Objekten, die TaggedText-Objekte enthalten

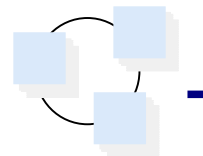


Wrapper Library (Fortsetz.)

Excel Wrapper Modul

- einzelne Datei einlesen
- Probleme:
 - kein öffentliches Format -> viele nicht unterstützte Objekte
- existierende Bibliotheken:
 - frei: POI, **JExcel**, Open Office, ..., kommerziell: Microsoft Excel, ...
- Ergebnis für Extractor
 - Tabellen-Objekt mit Row- und Cell-Objekten

- **Vorteil Wrapper Library:**
solange Schnittstellenformat zu einem Extractor Modul erzeugt wird, ist Implementierung des Wrapper Moduls egal



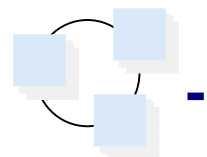
Extractor Library

Extractor Module zum Extrahieren von Daten

- extrahieren von Objektinstanzen und Mappings aus Wrapper - Extractor Schnittstellenformat
- Benutzerinteraktion
- Umwandlung in Manipulator Schnittstellenformat
- Implementierung egal, wenn entsprechende Schnittstellenformate eingehalten

Probleme

- Datenspezifisch



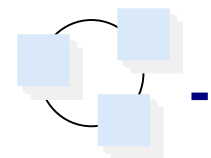
Extractor Library (Fortsetz.)

PDF Extractor Modul

- Extrahieren von Titeln und Autoren
 - aufgrund Position, Größe, Abstand, Ausrichtung, Stil
- Probleme:
 - Reihenfolge der Elemente, Kodierung, Vielfältigkeit
- Ergebnisse:
 - Qualität abhängig von Domäne, Kodierung

Table Extractor Modul

- Einfache Tabelle -> Objektinstanzen, Kreuztabelle -> Objektinstanzen + Mappings
- Benutzerinteraktion
 - Bedingungen für Mappings (z.B. regulärer oder arithmetischer Ausdruck)



Manipulator

Aufbereitung Objektinstanzen...

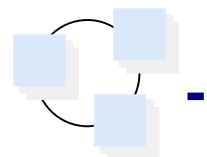
- **Objekte**
 - sortieren, filtern, löschen, einfügen
- **Attribute**
 - löschen, hinzufügen, editieren

und Mappings

- **Mappings löschen, hinzufügen sortieren**
- **Gruppieren nach PDS**
- **LDS bearbeiten, Attribute editieren**

Import / Export zu iFuice

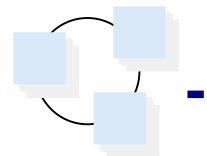
- **Datenformat: XML-Datei**
-



Schnittstelle zu iFuice

XML-Austauschformat für Objectinstances

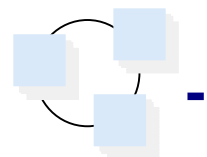
```
<OBJECTINSTANCES objecttype="Authors" datasource="FavAuthors" idattribute="id">
  <OBJECT>
    <DISPLAYVALUE>Erhard Rahm</DISPLAYVALUE>
    <ATTRIBUTE>
      <NAME>id</NAME>
      <VALUE>Id34</VALUE>
    </ATTRIBUTE>
    <ATTRIBUTE>
      <NAME>Last Name</NAME>
      <VALUE>Rahm</VALUE>
    </ATTRIBUTE>
    <ATTRIBUTE>
      <NAME>First Name</NAME>
      <VALUE>Erhard</VALUE>
    </ATTRIBUTE>
  </OBJECT>
</OBJECTINSTANCES>
```



Schnittstelle zu iFuice (Fortsetz.)

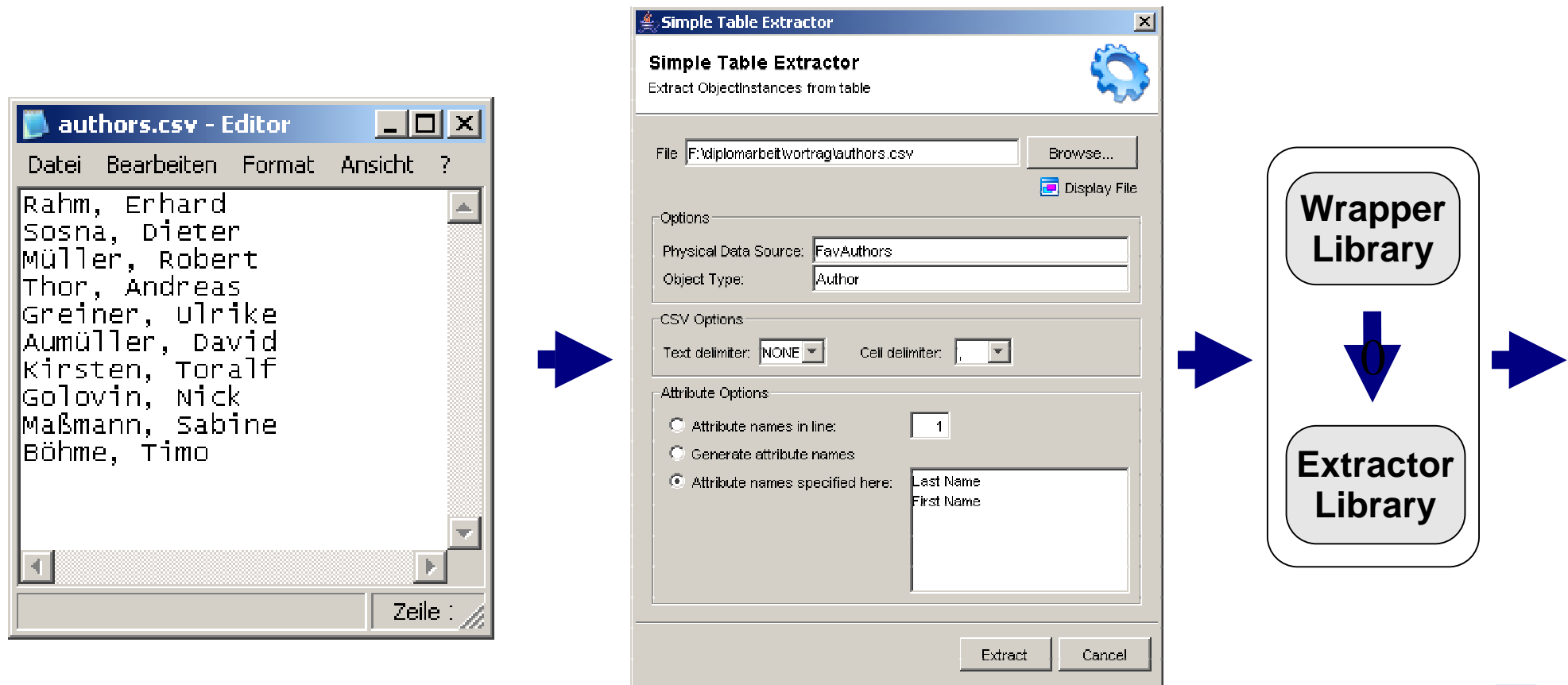
XML-Austauschformat für MappingResults

```
<MAPRESULT inputdatasource="FavAuthors" inputobjecttype="Authors" inputidattribute="id"
outputdatasource="DBLP" outputobjecttype="Authors" outputidattribute="key">
  <MAP confidence="0.907071">
    <OBJECT datasource="FavAuthors">
      <DISPLAYVALUE>Erhard Rahm</DISPLAYVALUE>
      ...
    </OBJECT>
    <OBJECT datasource="DBLP">
      <DISPLAYVALUE>Rahm:Erhard</DISPLAYVALUE>
      ...
    </OBJECT>
  </MAP>
</MAPRESULT>
```



Was haben für mich interessante Autoren publiziert?

Lokale Daten extrahieren



Was haben für mich interessante Autoren publiziert? (Fortsetz.)

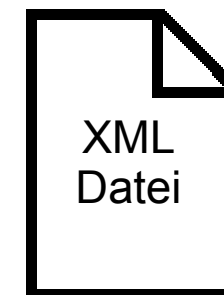
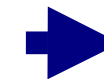
Lokale Daten aufbereiten

Id	Display Value	Last Name	First Name
Id6	David Aumüller	Aumüller	David
Id10	Timo Böhme	Böhme	Timo
Id8	Nick Golovin	Golovin	Nick
Id5	Ulrike Greiner	Greiner	Ulrike
Id7	Toralf Kirsten	Kirsten	Toralf
Id9	Sabine Maßmann	Maßmann	Sabine
Id3	Robert Müller	Müller	Robert
Id1	Erhard Rahm	Rahm	Erhard
Id2	Dieter Sosna	Sosna	Dieter
Id4	Andreas Thor	Thor	Andreas

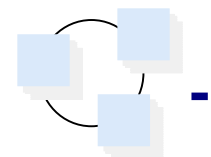
Attribute Name	Attribute Value
Id	Id10
Display Value	Timo Böhme
Last Name	Böhme
First Name	Timo

10 Objects, Selected: Object 2

export



ObjectInstances
FavAuthors



Was haben für mich interessante Autoren publiziert? (Fortsetz.)

Mapping EINMAL erzeugen: Lokale Autoren -> DBLP

- `$FavAuthors := loadInstances(Author@FavAuthors)`
- `$FavAuthorsDBLP := map($FavAuthors, ViaGoogle2DBLP)`



Was haben für mich interessante Autoren publiziert? (Fortsetz.)

Mappings EINMAL aufbereiten

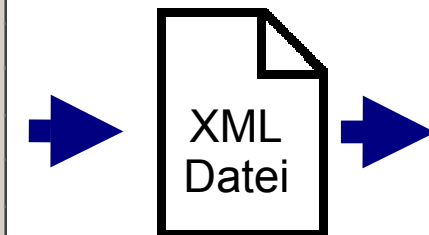
-> erzeugte Datei nutzbar für alle weiteren Anfragen an iFuice

```
$FavAuthorsDBLPCleaned := loadMapResult (FavAuthors2DBLP@Local)
$FavPubs := map (range ($FavAuthorsDBLPCleaned), DBLP.AuthPub)
$Result := compose ($FavAuthorsDBLPCleaned, $FavPubs)
```

Source Attribute	Source Value	Target Attribute	Target Value
id	id3	id	Müller:Robert
Display Value	Robert Müller	Display Value	Müller:Robert
Last Name	Müller	DBLP Author Website	www.informatik.uni-trier.de/~infa/DBLP/author/Mueller:Robert
First Name	Robert		

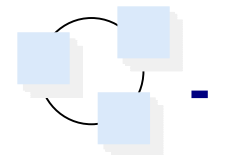
Autoren -> Publikationen

Autoren -> Publikationen-> Citations



MappingResult
FavAuthorsDBLPCleaned

Autoren -> ACM



Wer publizierte als Mitglied des Programmkomitees? Mapping aus lokalen Daten extrahieren

authorsconf.xls - OpenOffice.org 1.1.2

Datei Bearbeiten Ansicht Einfügen Format Extras Daten Fenster Hilfe

	A	B	C	D
1		SIGMOD	BTW	VLDB
2	Rahm	x	x	
3	Sosna	x	x	x
4	Müller		x	
5	Thor			x
6	Aumüller	x		x
7	Greiner		x	x
8	Golovin	x		x
9	Kirsten	x	x	
10	Maßmann		x	x
11	Böhme		x	

Tabelle1 Tabelle2 Tabelle3



Simple Table Extractor

Cross Table Extractor
Extract Mappings from cross table

File: F:\diplomarbeit\vortrag\authorsconf.xls Browse...

Display File

CSV Options
Text delimiter: NONE Cell delimiter: ;

Horizontal Objects
Physical Data Source: AuthorsConfLocal
Object Type: Conference
Attribute Name: Name Configure Objects...

Vertical Objects
Physical Data Source: AuthorsConfLocal
Object Type: Author
Attribute Name: Last Name Configure Objects...

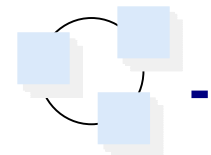
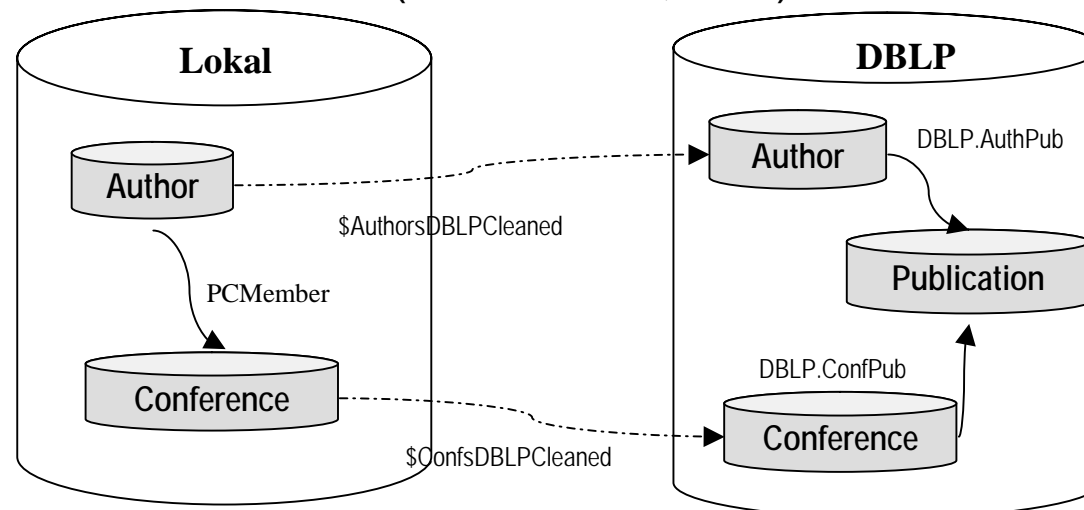
Mappings
Mapping name: PCMember
Condition: Arithmetic Expression:
 Regular Expression: x
Direction: Horizontal -> Vertical
 Vertical -> Horizontal

Create Mapping Finish Cancel

Wer publizierte als Mitglied des Programmkomitees? (Fortsetz.)

□ Daten verknüpfen

- **gewünschtes Ergebnis:**
Autoren, die im Programmkomitee waren und gleichzeitig publizierten
- $\$M1 := \text{map}(\text{range} (\$AuthorsDBLPCleaned), \text{DBLP.AuthPub})$
- $\$M2 := \text{map}(\text{range} (\$ConfsDBLPCleaned), \text{DBLP.ConfPub})$
- $\$M3 := \text{compose}(\$AuthorsDBLPCleaned, \$M1, \text{inverse}(\$M2), \text{inverse}(\$ConfsDBLPCleaned))$
- Autoren -> Konferenzen in denen sie publiziert haben mit Hilfe FavAuthorsDBLPCleaned
- $\$AuthorIsPCMember := \text{intersect} (\$PCMember, \$M3)$



Zusammenfassung u. Ausblick

Integration lokaler Daten in iFuice

- Daten finden, extrahieren, aufbereiten
- importieren / exportieren <-> iFuice
- basierend auf Wrapper und Extractor Modulen
- Probleme vom Anfang gelöst
 - relevante Daten finden (z.B. Autoren, von denen man Publikationen besitzt)
 - nach einmaliger Extraktion und Aufbereitung immer aktuelle Informationen
 - Darstellung lokaler Daten zusammen mit externen Daten ...

Ausblick

- neue Möglichkeiten der Extraktion lokaler Daten
- Verknüpfung des Personal Information Management mit Webdaten

