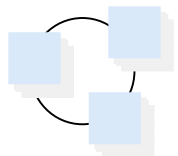
Diplomarbeitsthema:

Integration lokaler Daten in iFuice

Bearbeiter: Sarah Gebhardt

Betreuer: Andreas Thor



Warum eine Integration lokaler Daten?

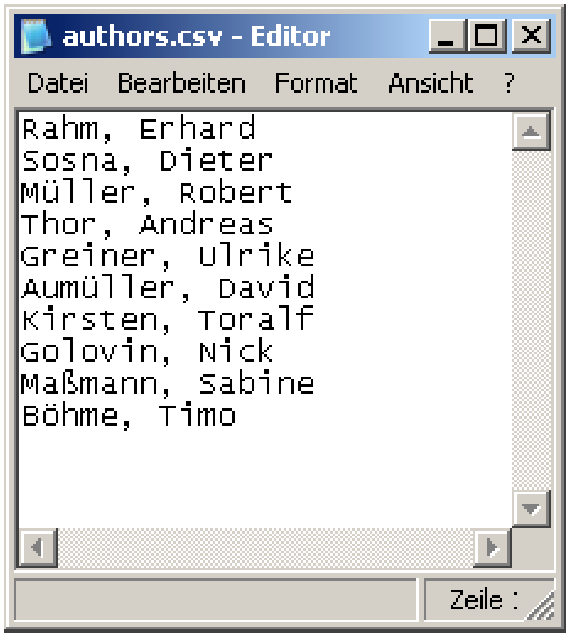
Viele Infos im Web, aber zu mühsam einzeln rauszusuchen

andere Listen im Web, aber nicht für mich relevant

Welche Informationen sind überhaupt für mich interessant

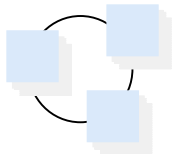
schon gesuchte Infos nicht mehr aktuell

Private Daten gemeinsam mit Webdaten darstellen und verknüpfen



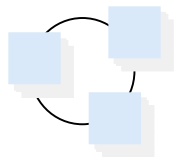
Konkret:

- 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

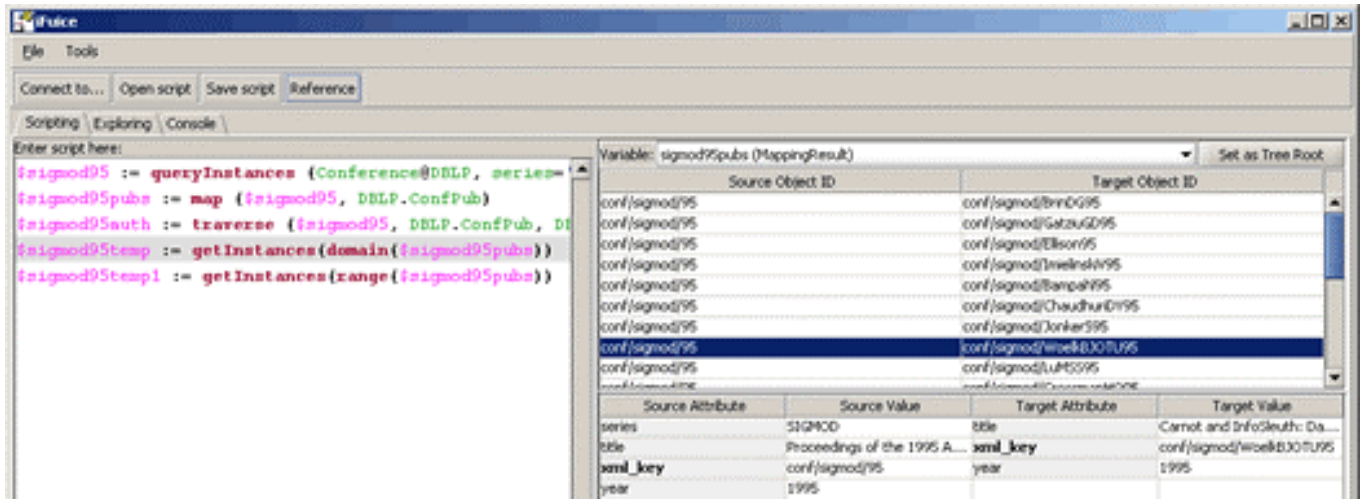
- 1. Motivation
- **2. iFuice**
  - **Kurzeinführung**
- **3. Architektur**
  - **Lokale Datenquellen**
  - **Wrapper**
  - **Extractor**
  - **Manipulator**
- **4. Beispiel**
- **5. Zusammenfassung / Ausblick**



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**



Datenstruktur in iFuice

Physische Datenquellen (Physical Data Sources - PDS)

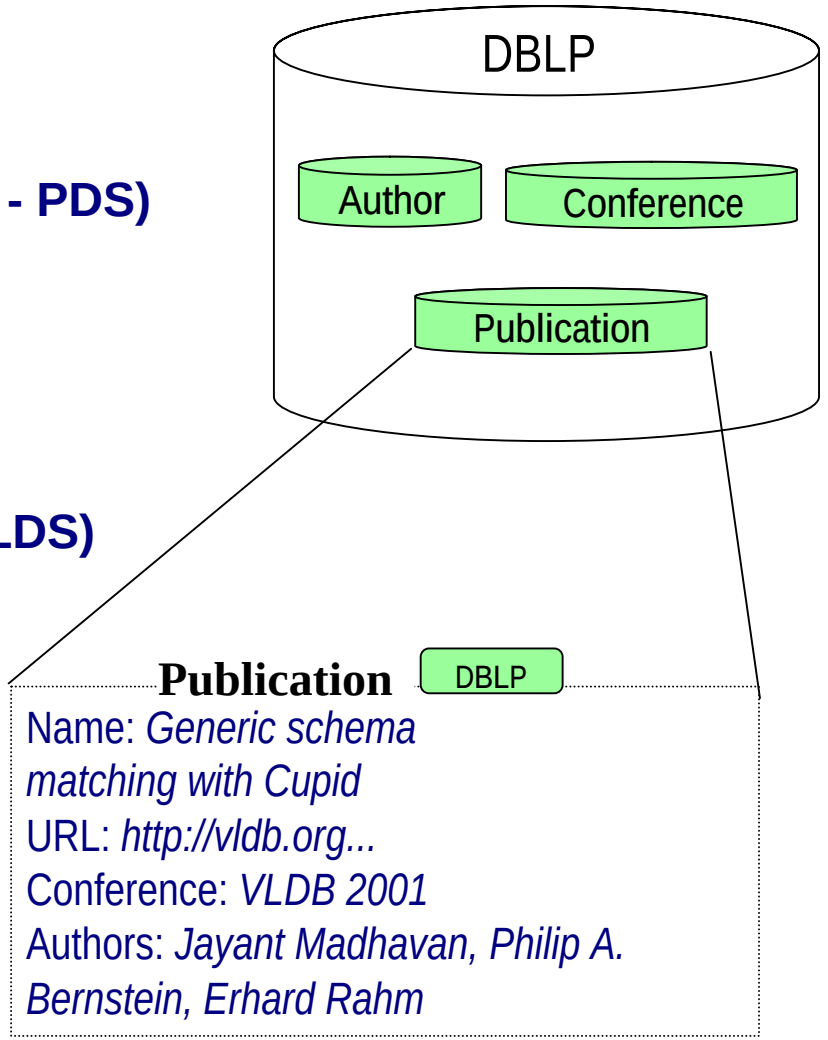
- Webdaten (DBLP), lokale Daten (Dateien)
- Aufgeteilt in logische Datenquellen

Logische Datenquellen (Logical Data Sources - LDS)

- definiert durch eine PDS und einen Objekttyp
- enthalten Objektinstanzen

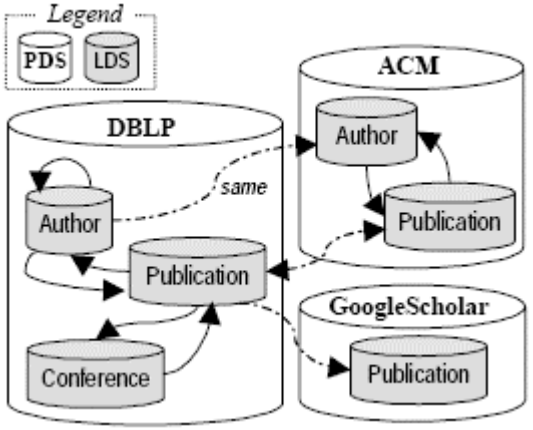
Objektinstanzen (Object Instances)

- 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)**

Instanzdaten: Instanzkorrespondenzen

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

Typische lokale Datenquellen

Unstrukturierte Daten

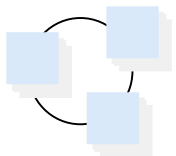
- PDF, Word, PowerPoint
  - Informationen: Publikationen, Berichte, Präsentationen, Skripte

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

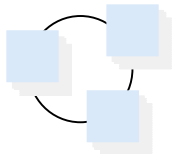
- extrahierte und importierte Daten bearbeiten

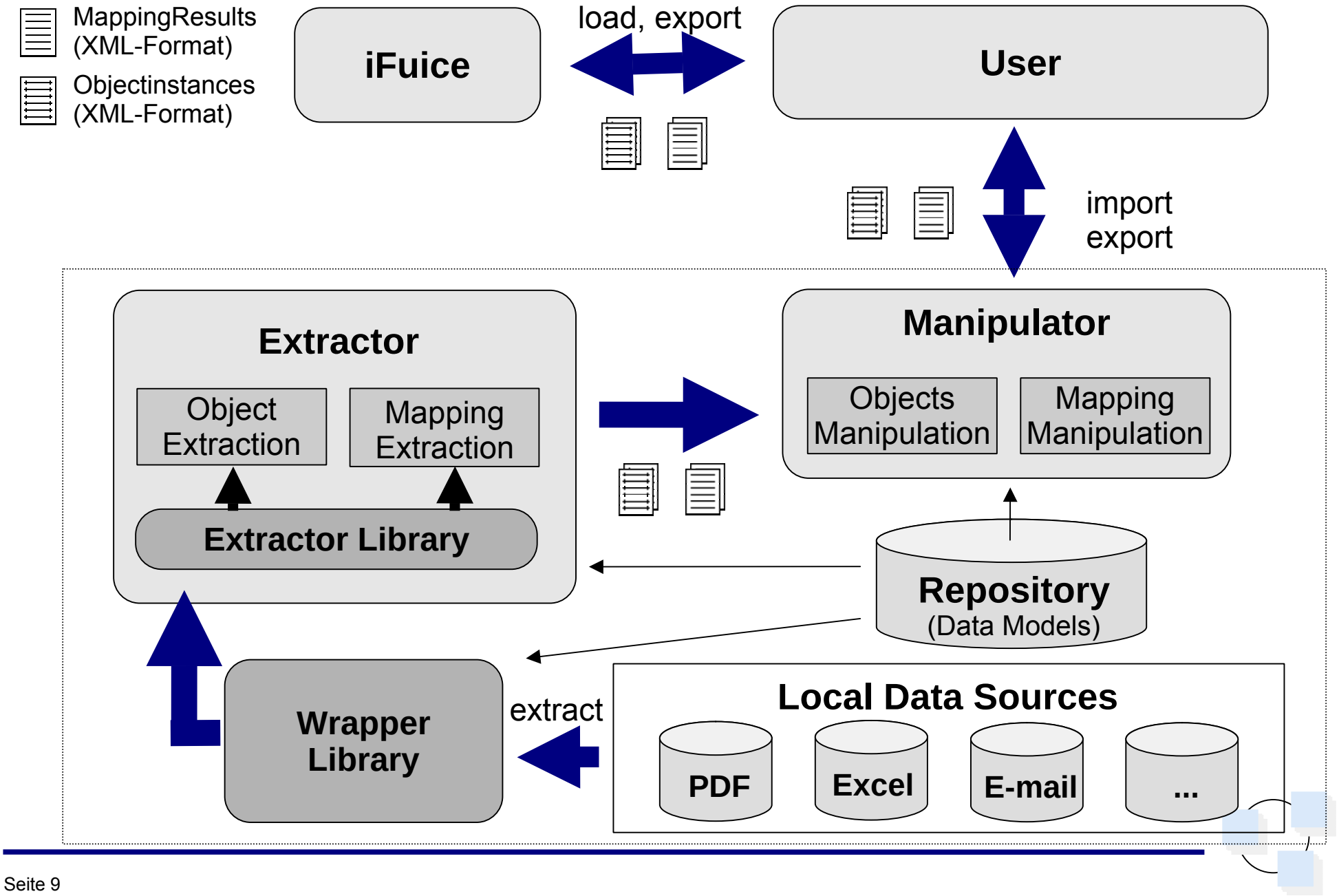
Export / Import

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

Aufbau

- modular, erweiterbar




MappingResults
(XML-Format)
Objectinstances
(XML-Format)
iFuice
load, export
User
import
export
Extractor
Object
Extraction
Mapping
Extraction
Extractor Library
Manipulator
Objects
Manipulation
Mapping
Manipulation
Repository
(Data Models)
Wrapper
Library
extract
Local Data Sources
PDF
Excel
E-mail
...

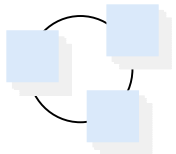
Wrapper Library

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

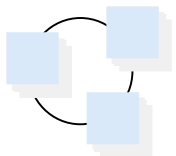
- wie kommt man an Daten
  - Verzeichnis lesen, Datei lesen, Desktop Search (Google Desktop Search)
- Format der Daten
  - offengelegt (PDF), nicht offengelegt (Excel, Word), Bibliotheken
- Struktur



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

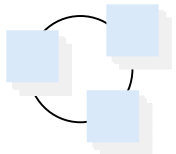


Wrapper Library (Fortsetz.)

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



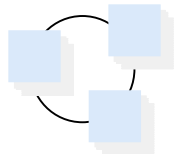
Extractor Library

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



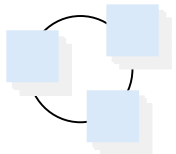
Extractor Library (Fortsetz.)

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...

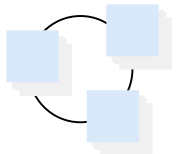
- **Objekte**
  - sortieren, filtern, löschen, einfügen
- **Attribute**
  - **löschen, hinzufügen, editieren**

und Mappings

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

Import / Export zu iFuice

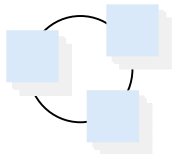
- **Datenformat: XML-Datei**



Schnittstelle zu iFuice

XML-Austauschformat für Objectinstances

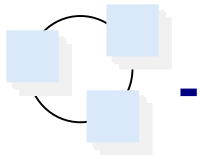
```
<OBJECTINSTANCES objecttype="Authors" datasource="FavAuthors" idattribute="id">
   <OBJECT>
      <DISPLAYVALUE>Erhard Rahm</DISPLAYVALUE>
      <ATTRIBUTE>
         <NAME>id</NAME>
         <VALUE>Id34</VALUE>
      </ATTRIBUTE>
      <ATTRIBUTE>
         <NAME>Last Name</NAME>
         <VALUE>Rahm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



Schnittstelle zu iFuice (Fortsetz.)

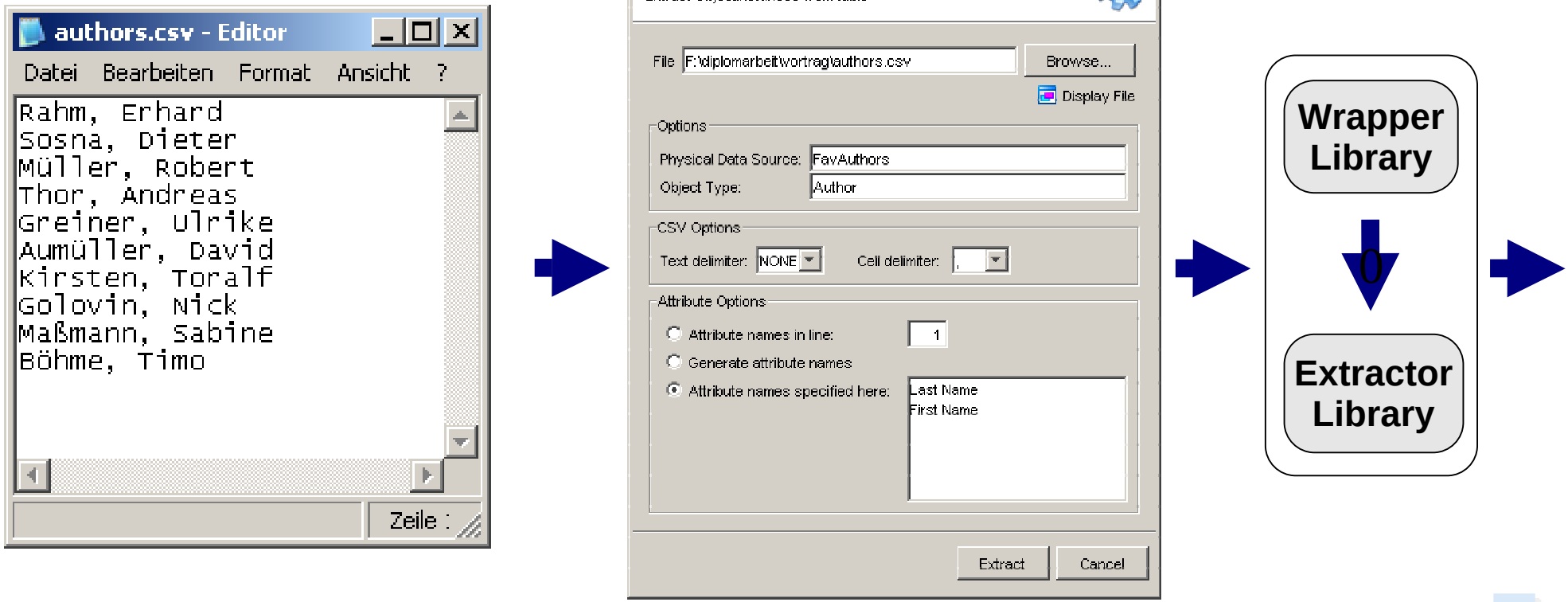
XML-Austauschformat für MappingResults

```
<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>
```



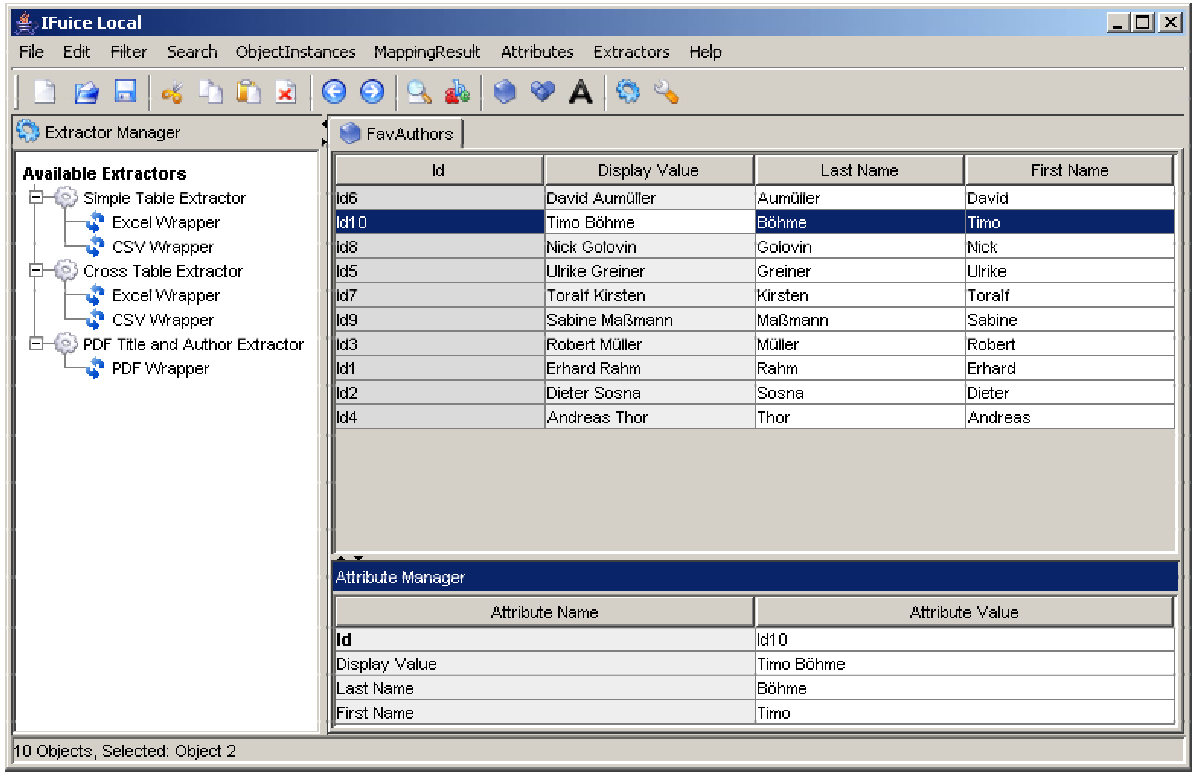
Was haben für mich interressante Autoren publiziert?

Lokale Daten extrahieren



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

Lokale Daten aufbereiten



export

XML Datei

ObjectInstances
FavAuthors

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

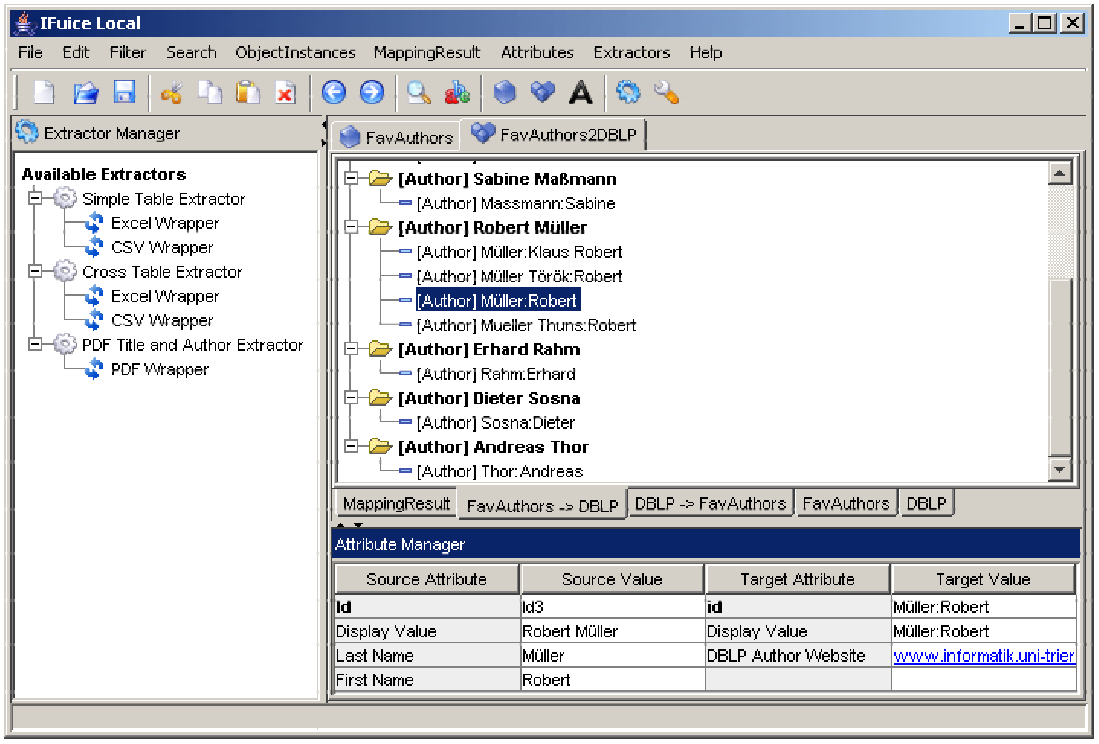
Mapping EINMAL erzeugen: Lokale Autoren -> DBLP

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

Was haben für mich interressante 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)
```



XML Datei

MappingResult
FavAuthorsDBLPCleaned

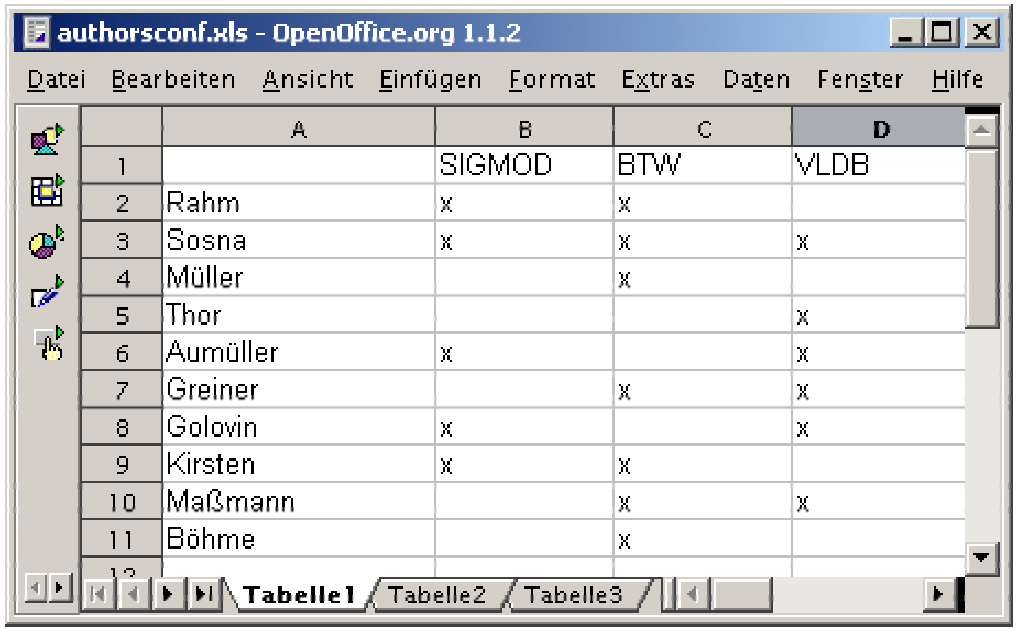
Autoren -> Publikationen

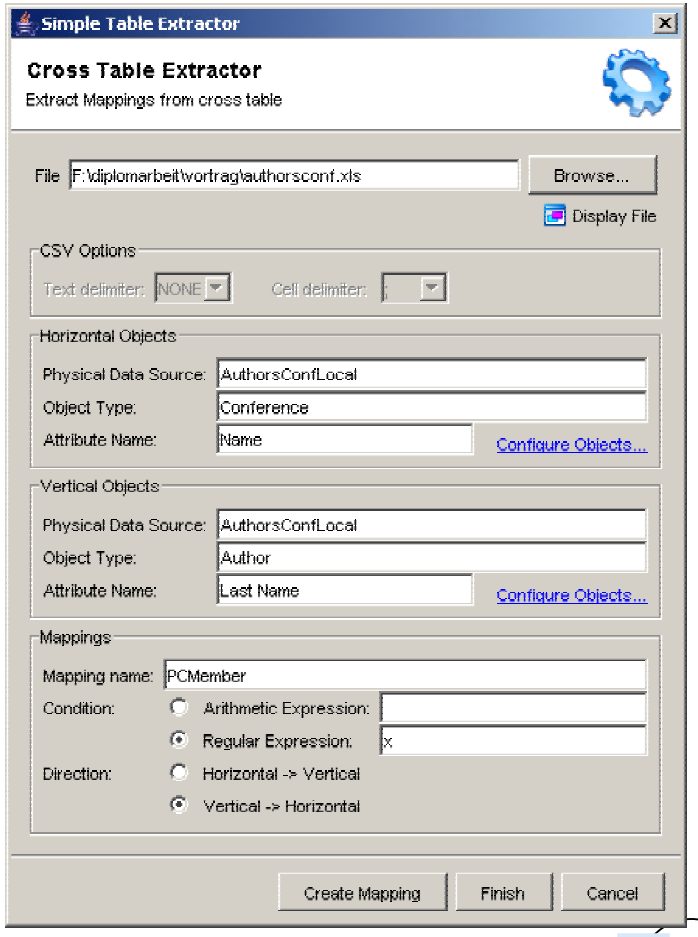
Autoren -> Publikationen-> Citations

Autoren -> ACM

Wer publizierte als Mitglied des Programmkomitees?

Mapping aus lokalen Daten extrahieren

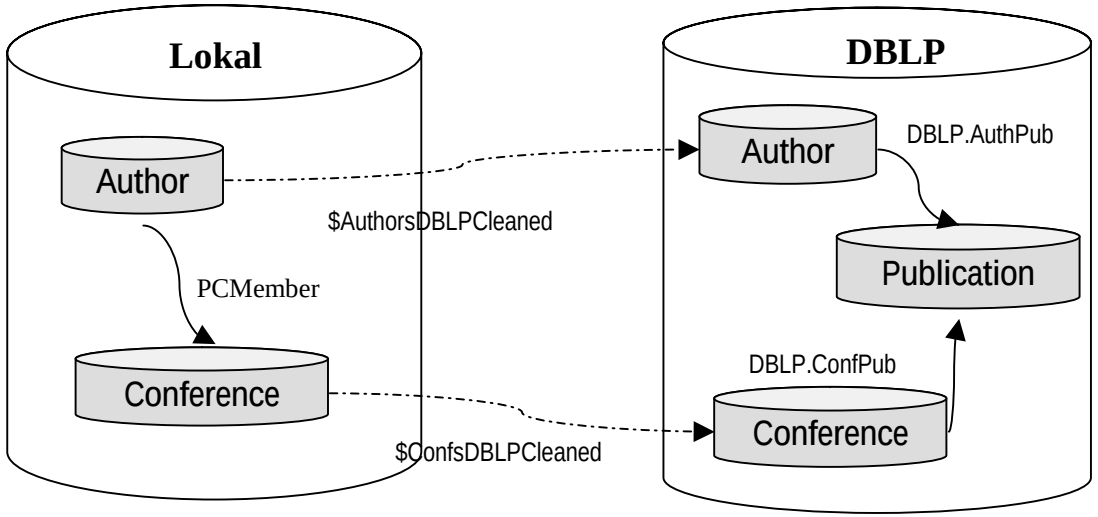




Wer publizierte als Mitglied des Programmkomitees? (Fortsetz.)

Daten verknüpfen

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