Towards identifying thinktanks

Chris Mitchell
www.epicicons.com

← Ideenhochburg

Zingst 2008, David Aumüller

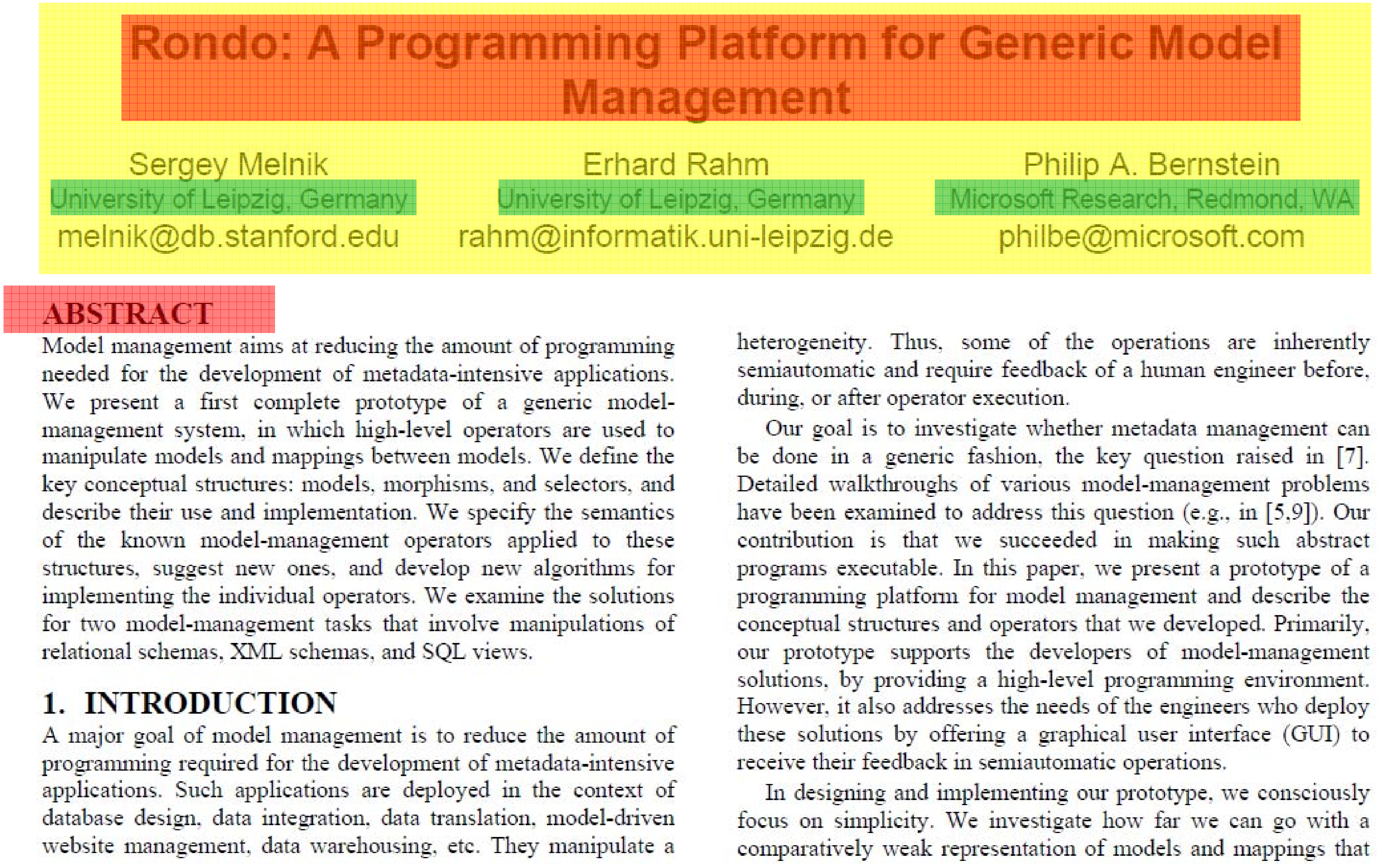
Woher stammen die Papiere …

… zu einer bestimmten Tagung?

… zu einem bestimmten Sachgebiet?

Überblick

- Papers → Zuordnung zu Orten / Forschungseinrichtungen
- Quellen (Daten/Services)
  - Papers: PDF → Volltext (text/plain)
  - Webservice: Google Scholar, …
  - Datenbanken: Länder, Städte
- Werkzeuge
  - Framework im Aufbau (Prototyp in Perl)



Rondo: A Programming Platform for Generic Model Management

Sergey Melnik
University of Leipzig, Germany
melnik@db.stanford.edu

Erhard Rahm
University of Leipzig, Germany
rahm@informatik.uni-leipzig.de

Philip A. Bernstein
Microsoft Research, Redmond, WA
philbe@microsoft.com

ABSTRACT

Model management aims at reducing the amount of programming needed for the development of metadata-intensive applications. We present a first complete prototype of a generic model-management system, in which high-level operators are used to manipulate models and mappings between models. We define the key conceptual structures: models, morphisms, and selectors, and describe their use and implementation. We specify the semantics of the known model-management operators applied to these structures, suggest new ones, and develop new algorithms for implementing the individual operators. We examine the solutions for two model-management tasks that involve manipulations of relational schemas, XML schemas, and SQL views.

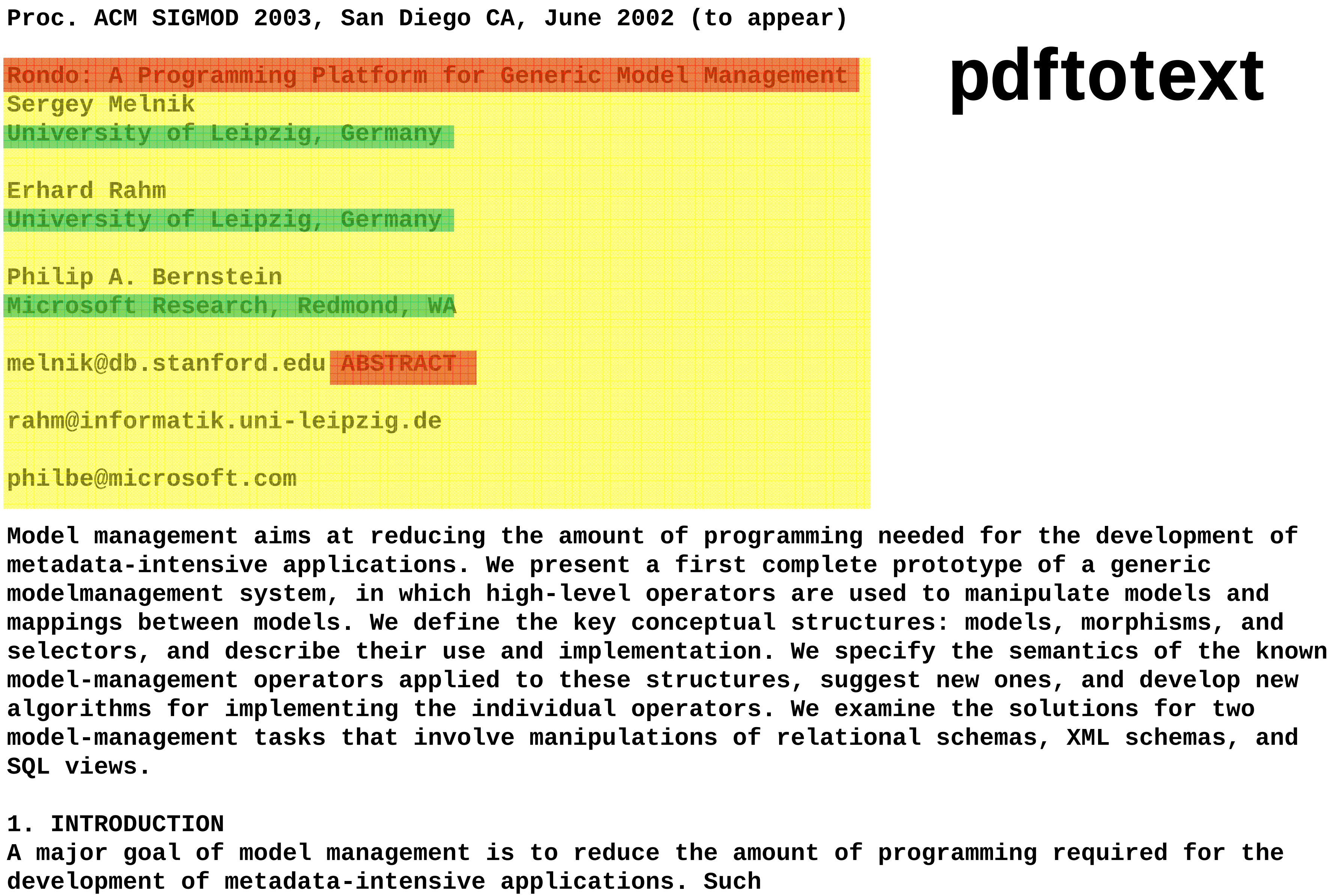
1. INTRODUCTION

A major goal of model management is to reduce the amount of programming required for the development of metadata-intensive applications. Such applications are deployed in the context of database design, data integration, data translation, model-driven website management, data warehousing, etc. They manipulate a heterogeneity. Thus, some of the operations are inherently semiautomatic and require feedback of a human engineer before, during, or after operator execution.

Our goal is to investigate whether metadata management can be done in a generic fashion, the key question raised in [7]. Detailed walkthroughs of various model-management problems have been examined to address this question (e.g., in [5,9]). Our contribution is that we succeeded in making such abstract programs executable. In this paper, we present a prototype of a programming platform for model management and describe the conceptual structures and operators that we developed. Primarily, our prototype supports the developers of model-management solutions, by providing a high-level programming environment. However, it also addresses the needs of the engineers who deploy these solutions by offering a graphical user interface (GUI) to receive their feedback in semiautomatic operations.

In designing and implementing our prototype, we consciously focus on simplicity. We investigate how far we can go with a comparatively weak representation of models and mappings that

pdftotext

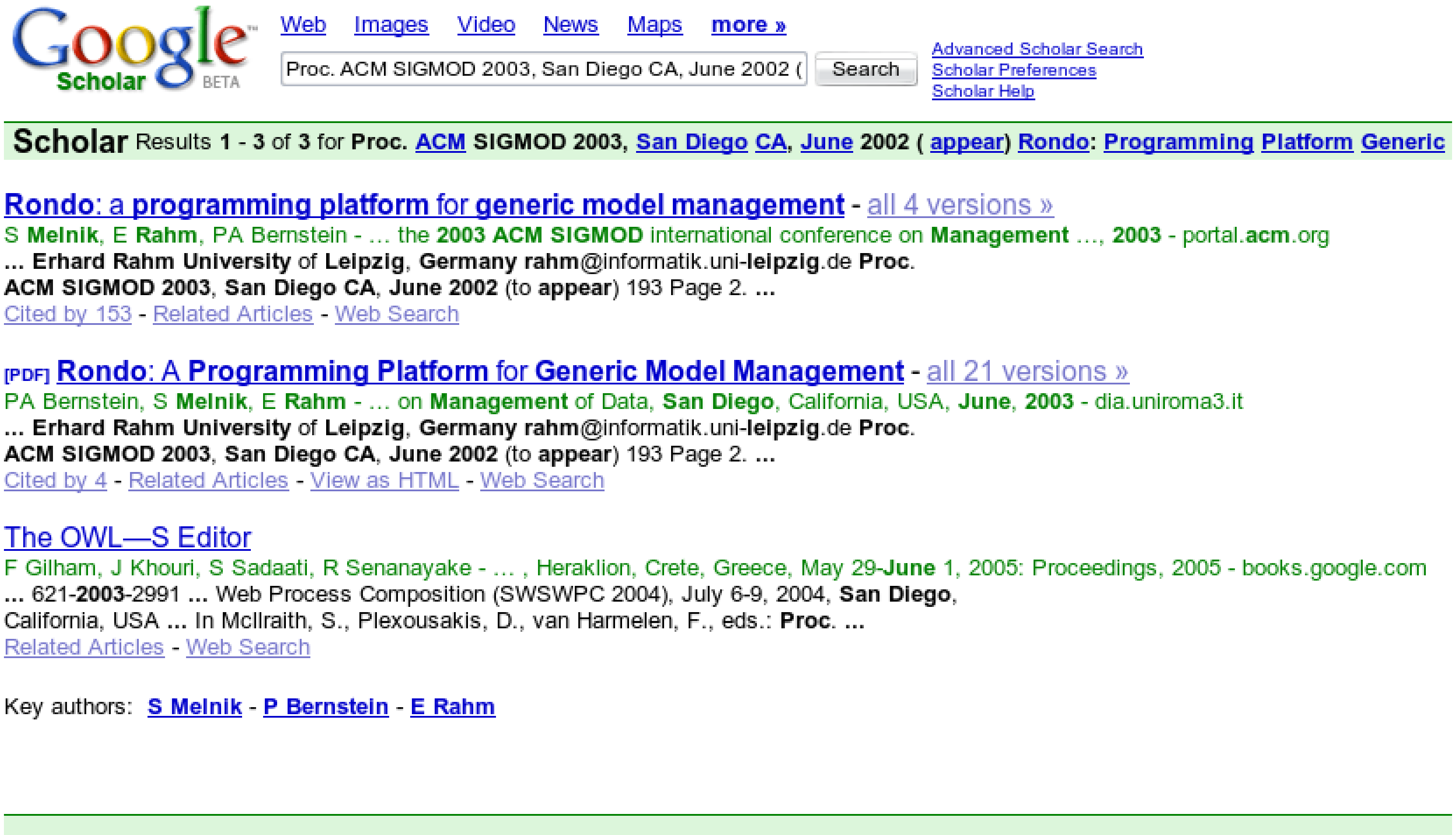


```
Proc. ACM SIGMOD 2003, San Diego CA, June 2002 (to appear)

Rondo: A Programming Platform for Generic Model Management
Sergey Melnik
University of Leipzig, Germany

Erhard Rahm
University of Leipzig, Germany

Philip A. Bernstein
Microsoft Research, Redmond, WA

melnik@db.stanford.edu ABSTRACT

rahm@informatik.uni-leipzig.de

philbe@microsoft.com
Model management aims at reducing the amount of programming needed for the development of
metadata-intensive applications. We present a first complete prototype of a generic
modelmanagement system, in which high-level operators are used to manipulate models and
mappings between models. We define the key conceptual structures: models, morphisms, and
selectors, and describe their use and implementation. We specify the semantics of the known
model-management operators applied to these structures, suggest new ones, and develop new
algorithms for implementing the individual operators. We examine the solutions for two
model-management tasks that involve manipulations of relational schemas, XML schemas, and
SQL views.

1. INTRODUCTION
A major goal of model management is to reduce the amount of programming required for the
development of metadata-intensive applications. Such
```

Typische Metadaten: Web-Service

- Zuordnung PDF und Scholar-Einträge: Anfrage über Volltext (plain text)
  - Dokumentanfang als Query
    - Reihe von Einträgen als Resultset
  - Matchen des Ergebnisses (Scholar Results) mit Volltext
    - Test: Scholar-Titel auf erster Seite des Volltexts?

Volltext-Query: Scholar Results



Proc. ACM SIGMOD 2003, San Diego CA, June 2002 (

Scholar Results **1** - **3** of **3** for **Proc. ACM SIGMOD 2003, San Diego CA, June 2002 (appear) Rondo: Programming Platform Generic**

Rondo: a programming platform for generic model management - all 4 versions »
S **Melnik**, E **Rahm**, PA Bernstein - … the **2003 ACM SIGMOD** international conference on **Management** …, **2003** - portal.**acm**.org
… **Erhard Rahm University** of **Leipzig**, **Germany rahm**@informatik.uni-**leipzig**.de **Proc**.
ACM SIGMOD 2003, **San Diego CA**, **June 2002** (to **appear**) 193 Page 2. …
Cited by 153 - Related Articles - Web Search

[PDF] Rondo: A Programming Platform for Generic Model Management - all 21 versions »
PA Bernstein, S **Melnik**, E **Rahm** - … on **Management** of Data, **San Diego**, California, USA, **June**, **2003** - dia.uniroma3.it
… **Erhard Rahm University** of **Leipzig**, **Germany rahm**@informatik.uni-**leipzig**.de **Proc**.
ACM SIGMOD 2003, **San Diego CA**, **June 2002** (to **appear**) 193 Page 2. …
Cited by 4 - Related Articles - View as HTML - Web Search

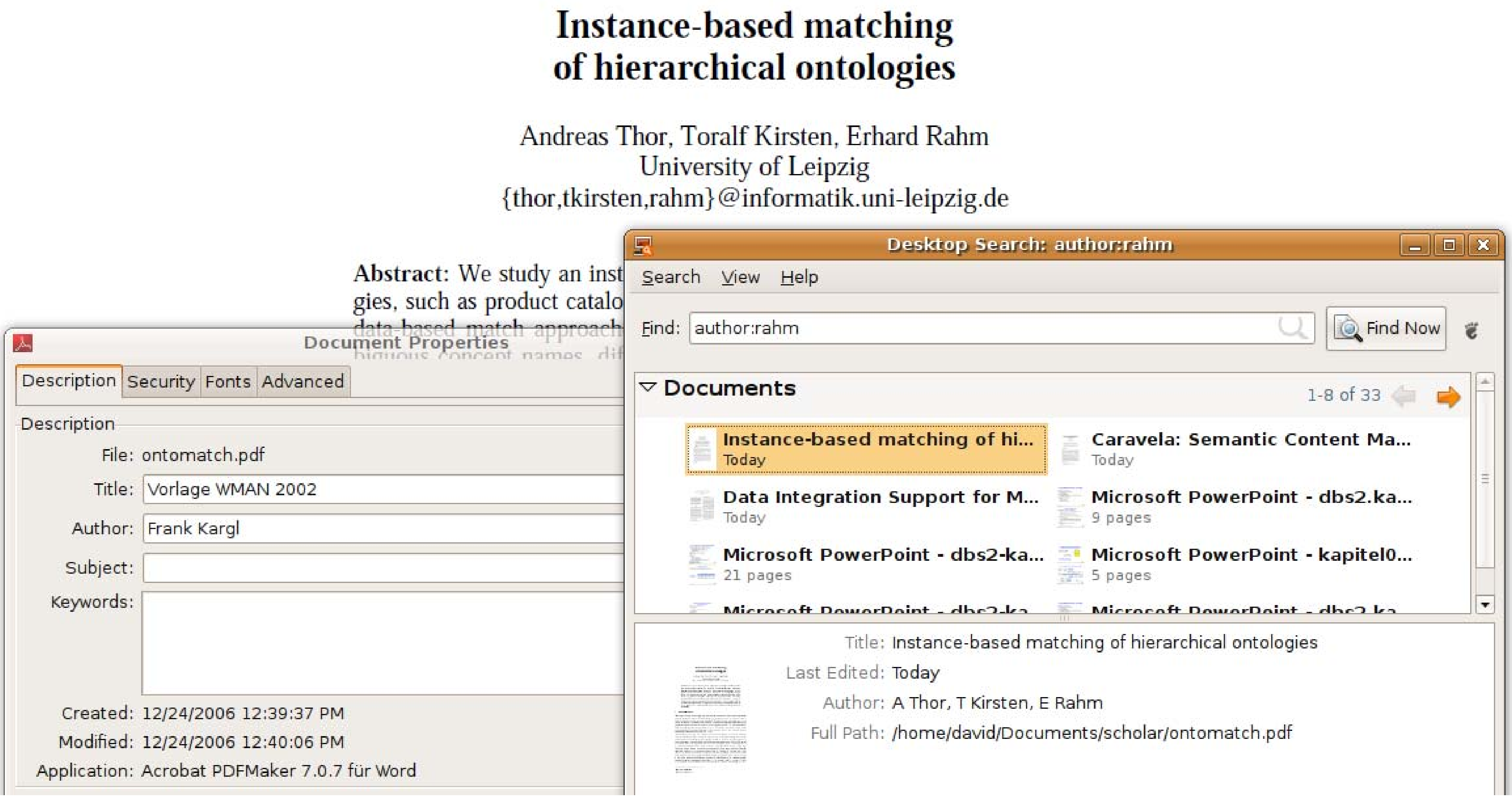
The OWL—S Editor
F Gilham, J Khouri, S Sadaati, R Senanayake - … , Heraklion, Crete, Greece, May 29-**June** 1, 2005: Proceedings, 2005 - books.google.com
… 621-**2003**-2991 … Web Process Composition (SWSWPC 2004), July 6-9, 2004, **San Diego**,
California, USA … In McIlraith, S., Plexousakis, D., van Harmelen, F., eds.: **Proc**. …
Related Articles - Web Search

Key authors: **S Melnik** - **P Bernstein** - **E Rahm**

Attribute

• Titel	Scholar
• Autorennamen	Scholar
• Jahr	Scholar
• Erscheinungsort	Scholar
• Emailadressen	?
• Affiliation/Zugehörigkeit	**?**
– Einrichtung, **Ort**, Adresse	
– **pro Paper** / pro Autor	

[Exkurs: Anwendung in Desktop-Suche]



Instance-based matching of hierarchical ontologies

Andreas Thor, Toralf Kirsten, Erhard Rahm
University of Leipzig
{thor,tkirsten,rahm}@informatik.uni-leipzig.de

Abstract: We study an inst
gies, such as product catalo
data-based match approach
biguous concept names, dif

Document Properties

Description

File: ontomatch.pdf

Title: Vorlage WMAN 2002

Author: Frank Kargl

Subject:

Keywords:

Created: 12/24/2006 12:39:37 PM

Modified: 12/24/2006 12:40:06 PM

Application: Acrobat PDFMaker 7.0.7 für Word

Desktop Search: author:rahm

Find: author:rahm

Documents 1-8 of 33

Instance-based matching of hi... Today

Caravela: Semantic Content Ma... Today

Data Integration Support for M... Today

Microsoft PowerPoint - dbs2.ka... 9 pages

Microsoft PowerPoint - dbs2-ka... 21 pages

Microsoft PowerPoint - kapitel0... 5 pages

Microsoft PowerPoint - dbs2-ka...

Microsoft PowerPoint - dbs2 ka...

Title: Instance-based matching of hierarchical ontologies

Last Edited: Today

Author: A Thor, T Kirsten, E Rahm

Full Path: /home/david/Documents/scholar/ontomatch.pdf

Abgleich Volltext – Ortsdatenbank

  - Ortsdatenbank mit 3 Mio. Tupel/Orte
- Welche Orte werden (wo) genannt?
  - Wort für Wort-Vergleich? +schnell, –prob:
      - SELECT country, city WHERE „Los“ = [city]
      - SELECT country, city WHERE „Angeles“ = [city]
  - Enthaltenstest über SQL-Join/Like-Anfrage:
    - WHERE volltext LIKE % city %
      - WHERE „foo Los Angeles bar“
        LIKE „% Los Angeles %“

Suchraum „Volltext“ groß

- Suche zeitaufwändig
- Ergebnisse unbefriedigend
  - Proc. ACM SIG… Rondo: A progr. platform for gen. model mgmt. …

```
country            city      region
Burkina Faso       Rondo     43
Haiti              Rondo     06
Haiti              Rondo     07
Tanzania           Rondo     07
United States      Rondo     AR
United States      Rondo     MI
United States      Rondo     MO
United States      Rondo     VA
```

- Kontext: Zwischen Titel und Abstract stehen die Orte / Autor-Affiliations

Suchraum verringern

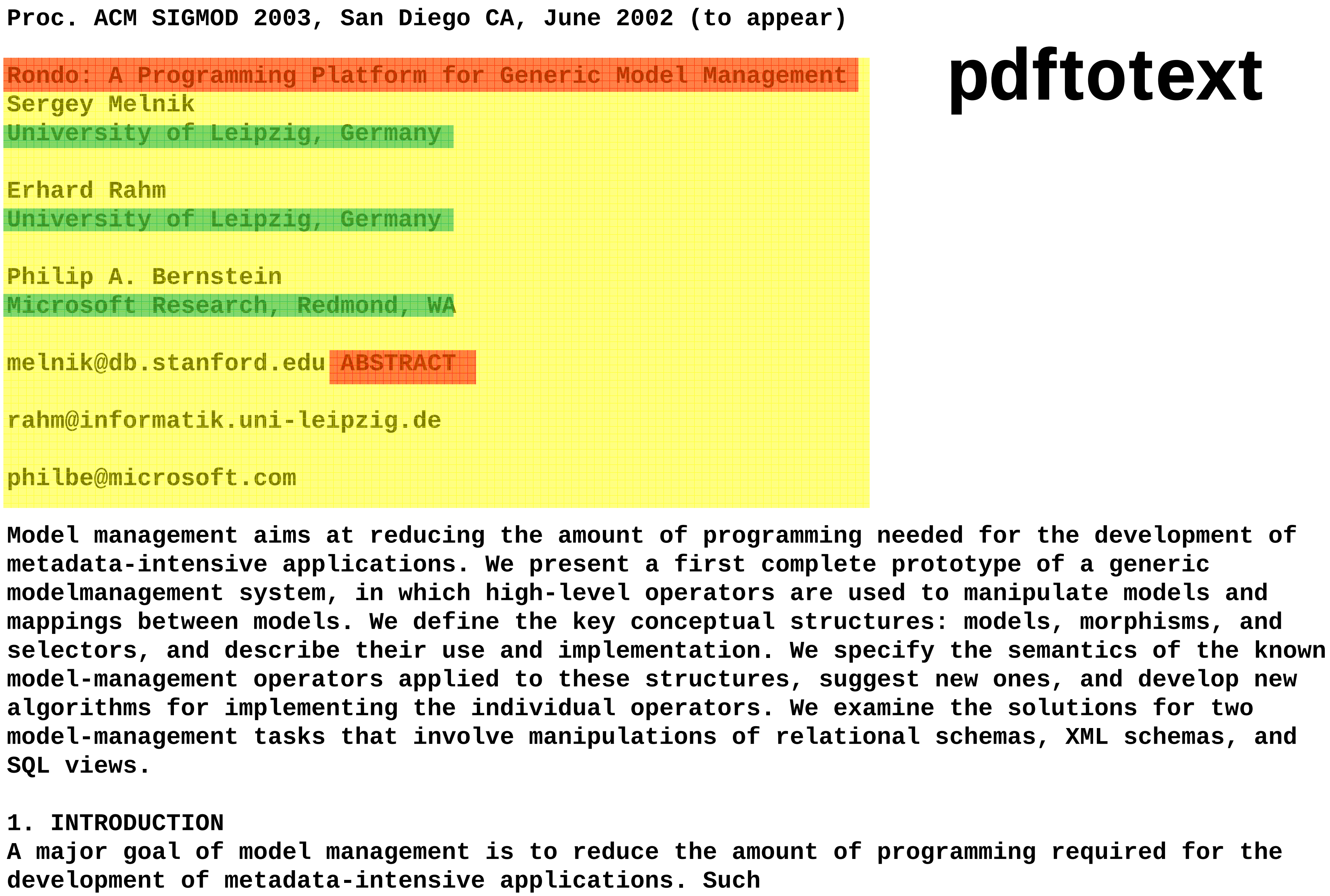
- **Volltext** → Kopf „AuthorBlob“
  - Titel, Autoren, Affil., Emails, …
  - Text mit potentieller Ortsangabe syntaktisch kürzen, weniger Worte →
  - Anzahl Orts-Kandidaten geringer

- **Ortsdatenbank** mit ca. 3 Mio Tupel
  - Country, City, Region, Population, lat/long
  - Einschränken nach Land, Population

Bekannte Metadaten ausblenden

- Scholar Autorname z.B.: „PA Bernstein“
  - Regular expression erstellen:
    - P[^ -]*?[-]A[^ -]*? Bernstein
  - Erkennt Varianten
    - P. A. Bernstein
    - Phil A. Bernstein
    - Philip A. Bernstein
    - Philip Alan Bernstein
  - (So Extraktion des vollen Namen möglich)
- Emailadressen ausblenden (analog)

pdftotext



```
Proc. ACM SIGMOD 2003, San Diego CA, June 2002 (to appear)
Rondo: A Programming Platform for Generic Model Management
Sergey Melnik
University of Leipzig, Germany

Erhard Rahm
University of Leipzig, Germany

Philip A. Bernstein
Microsoft Research, Redmond, WA

melnik@db.stanford.edu ABSTRACT

rahm@informatik.uni-leipzig.de

philbe@microsoft.com
Model management aims at reducing the amount of programming needed for the development of
metadata-intensive applications. We present a first complete prototype of a generic
modelmanagement system, in which high-level operators are used to manipulate models and
mappings between models. We define the key conceptual structures: models, morphisms, and
selectors, and describe their use and implementation. We specify the semantics of the known
model-management operators applied to these structures, suggest new ones, and develop new
algorithms for implementing the individual operators. We examine the solutions for two
model-management tasks that involve manipulations of relational schemas, XML schemas, and
SQL views.

1. INTRODUCTION
A major goal of model management is to reduce the amount of programming required for the
development of metadata-intensive applications. Such
```

Was bleibt vom Volltext?

University of Leipzig, Germany

University of Leipzig, Germany

Microsoft Research, Redmond, WA

Treffer in Ortsdatenbank

CC	Country	city	region	pop	lat	long
de	Germany	Leipzig	13	492637	51.3	12.3333
ua	Ukraine	Leipzig	17	0	46.3022	29.0197
au	Australia	Redmond	08	0	-34.8667	117.683
us	United States	Redmond	CO	0	40.4789	-105.04
us	United States	Redmond	OR	18807	44.2728	-121.173
us	United States	Redmond	PA	0	41.0833	-80.0353
us	United States	Redmond	UT	0	39.0061	-111.861
us	United States	Redmond	WA	47264	47.6742	-122.12
us	United States	Redmond	WV	0	38.8042	-82.1342
gh	Ghana	Wa	11	50268	10.05	-2.48333
tr	Turkey	Of	61	31968	40.95	40.2667
us	United States	University	IL	0	37.9364	-88.6094
us	United States	University	NC	0	36.0372	-79.0358
us	United States	University	WA	0	47.6667	-122.309
au	Australia	Research	07	0	-37.7	145.183

Weitere Einschränkungen nötig

- Stoppworte ausblenden, z.B. *of, and,* …
- Population > 0

„Mehrdeutigkeiten“

- Gleicher Ortsname, mehrere Länder (z.B. 313 mal „San Jose“)
  - Erst Land herausfinden
    - Abgleich mit Ortsdatenbank (LIKE % Landname %)
    - Emailadresse Domain
  - Community Feedback
    - Manuell
    - Web-Service, z.B. Google/Wikipedia Anfrage um bekanntestes Land einer Stadt mit gleichem Namen zu identifizieren

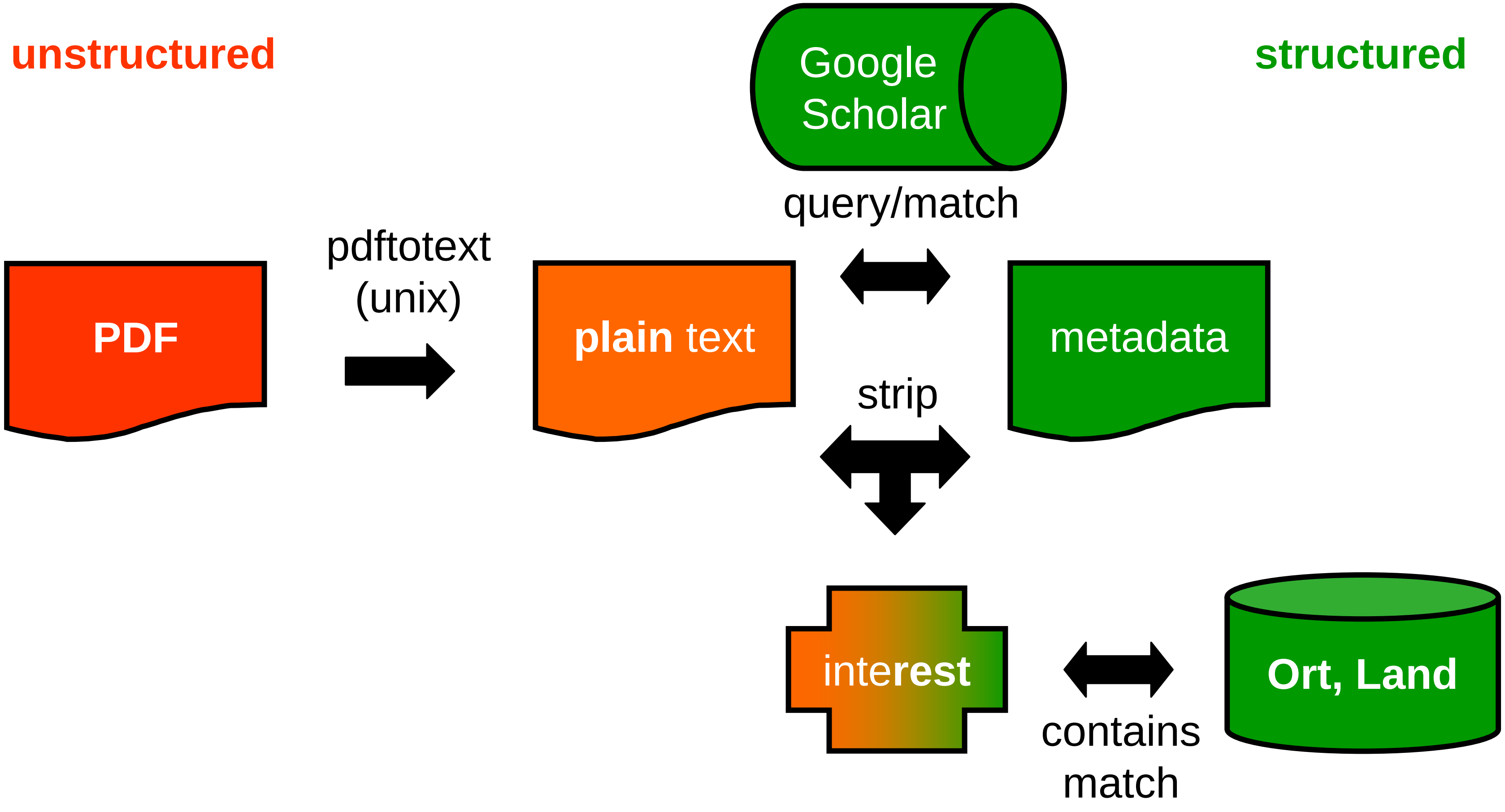
Enthaltenstest möglichst spezifisch

- Musterreihenfolge
  - LIKE % city, state, USA %
  - LIKE % city, state
  - LIKE % city, country %
  - LIKE % city %
- Iterativ:
  - Auf Muster prüfen
  - Treffer in authorblob entfernen/ausblenden

Orte/Datensätze ausschließen

- Anzahl zu vergleichender Orte verringern
  - Mehrdeutigkeiten
    Anzahl der„false positives“ verringern
- Emailadressen weisen auf Land hin
  - .de → Deutschland
  - .edu → U.S.A.? – Ausnahmen!
  - .com/.org/… → …

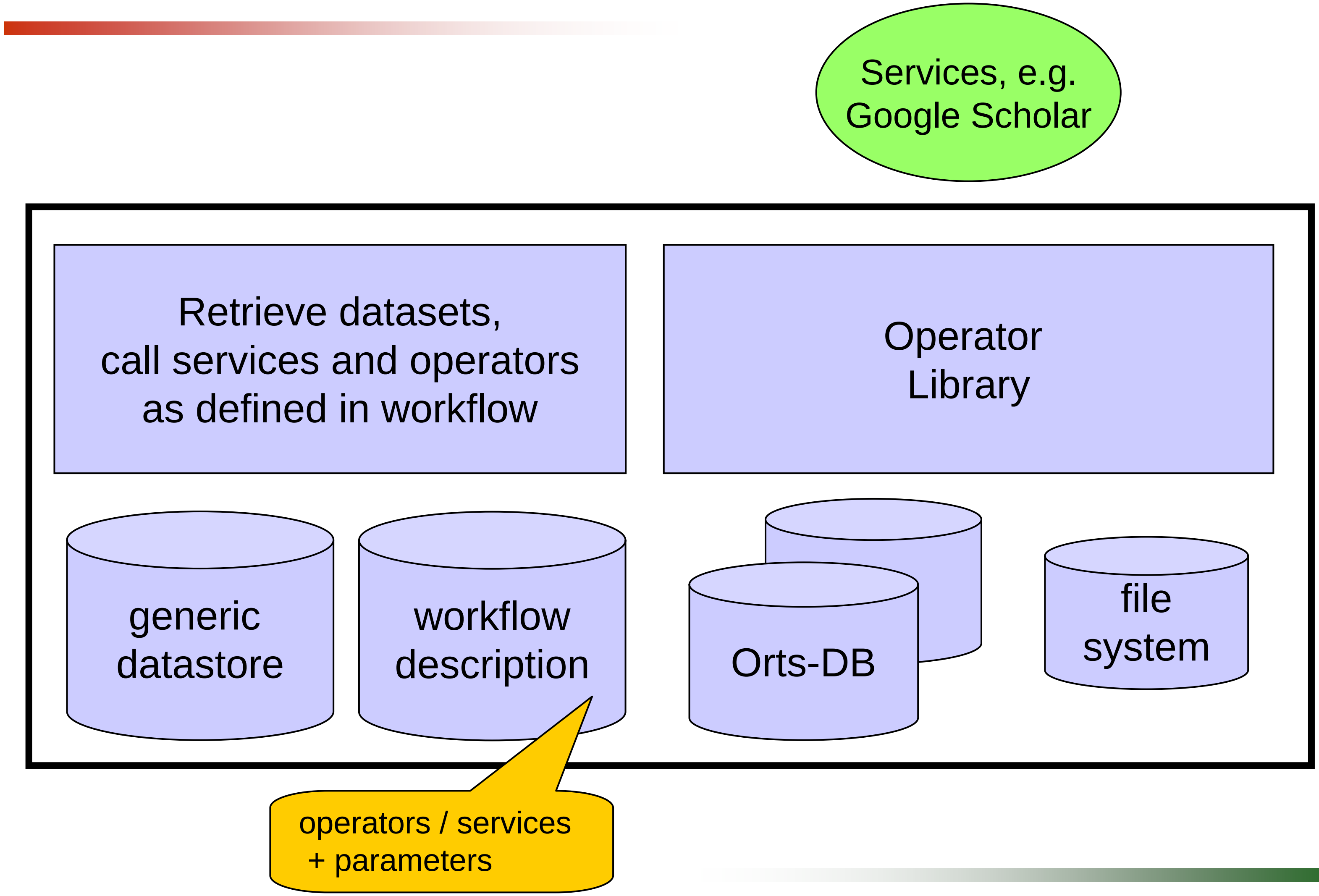
Workflow Beispiel



Sukzessive Datenextraktion

- schrittweise Extraktion anhand schon extrahierter Einheiten
  - Syntaktische und semantische Regeln
  - Operatoren und (Web-) Service Anfragen
- Operatoren
  - substring, eg. [title] (…) “Abstract“
  - strip, eg. [authorblob] \ emails, authors
  - contains, eg. [authorblob] contains {city}
  - match / rank?

towards a framework



Perl Prototype

- Workflow
  - In/out über DB statt pipes (?)
    - Primary key / ID reicht als Übergabeparameter
    - Primary key ist (local) URL to PDF
- Datenbank
  - Zwei Tabellen: simple + multiple values
  - Somit Ad-Hoc Ergänzungen möglich
    - Attribut-Wert-Tripel / Quadrupel:
    - Filename | Attribute | Delta | Value

Ortsdatenbank ungenügend zur Erkennung der „Ideenhochburgen“

- Viele Papiere: lediglich Angaben über Institutsnamen, keine Ortsangaben
  - → Institutsdatenbank für Enthaltenstest / Matching erstellen/heranziehen

- Vorgestellte Extraktion hilfreich zur Erstellung dieser Institutsdatenbank?

Zusammenfassung

- PDF Volltext als plain text
- Mapping
  - lokale PDFs zu Google Scholar
  - Metadaten: Titel, Autoren, Venue, Jahr
- Extraktion
  - Potentielle Affiliations (Einrichtung/Ort)
- Zuordnung Papier – Orte der Autoren

Ausblick - Ansatz

- Evaluation
- Ansatz verallgemeinern
  - Operatoren, Framework
  - Übergangswahrscheinlichkeiten
    - p(Autor->Affil) vs. p(Autor->Autor)
  - Automatische Regelerzeugung?
- Re-Use von erkannten Zuordnungen
- Anwendung auf andere Gebiete?

Ausblick – Biblio-Domäne

- Mehr Services nutzen, z.B. GMaps
- Web Ad-Hoc Analyse (Mashup)
- Neue Dimensionen in Caravela
  - Ort: Kontinent > Land > Stadt
  - Einrichtungsart: Uni > Inst > Dep | Firma
  - Interaktive Landkarte
- Analyse der „Woher stammen Papiere zu [Tagung|Thema|...]“-Fragen
- Pro-Autor Zuordnung statt pro-Papier