Publikationscrawler für Dokumentenserver

Die Erstellung eines Publikationscrawler mit Hilfe von Dapper und Apatar

Übersicht

- Die Grundidee
- Die Publikationsseiten
- Die Werkzeuge
- Der Aufbau des Publikationscrawler
- Die Arbeitsweise des Publikationscrawler
- Die Datenbank
- Die erfassten Publikationen
- Probleme bei der Datenextraktion
- Probleme bei der Datenintegration
- Derzeitige Vorgehensweise
- Lösungen und Verbesserungen
- Probleme und Fehler
- Ausblick
- Diskussion

Die Grundidee

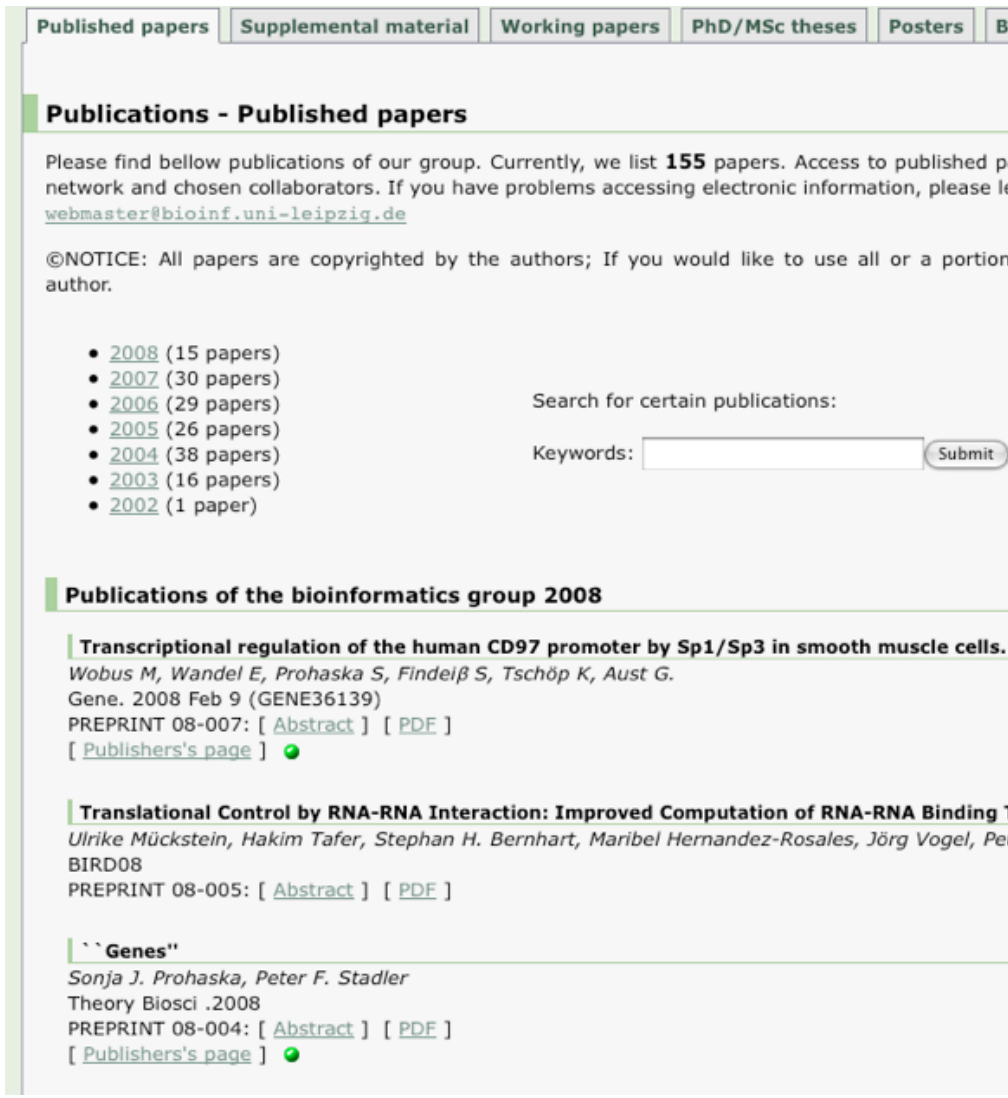
- Ursprüngliches Thema war eine Evaluierung von Dataflow Tools.
- Daraus entstand die Idee mit Hilfe von Extraktionstools und Dataflow Tools die Publikationen der einzelnen Abteilungen zu erfassen und zentral zu speichern.

Die Publikationsseiten

- Jede Abteilung benutzt ein eigenes System bzw. Format.
- Manche Webseiten sind sehr alt.
- Manche Abteilungen benutzen Masken und Eingabesysteme zum hinzufügen neuer Publikationen andere fügen diese manuell hinzu.

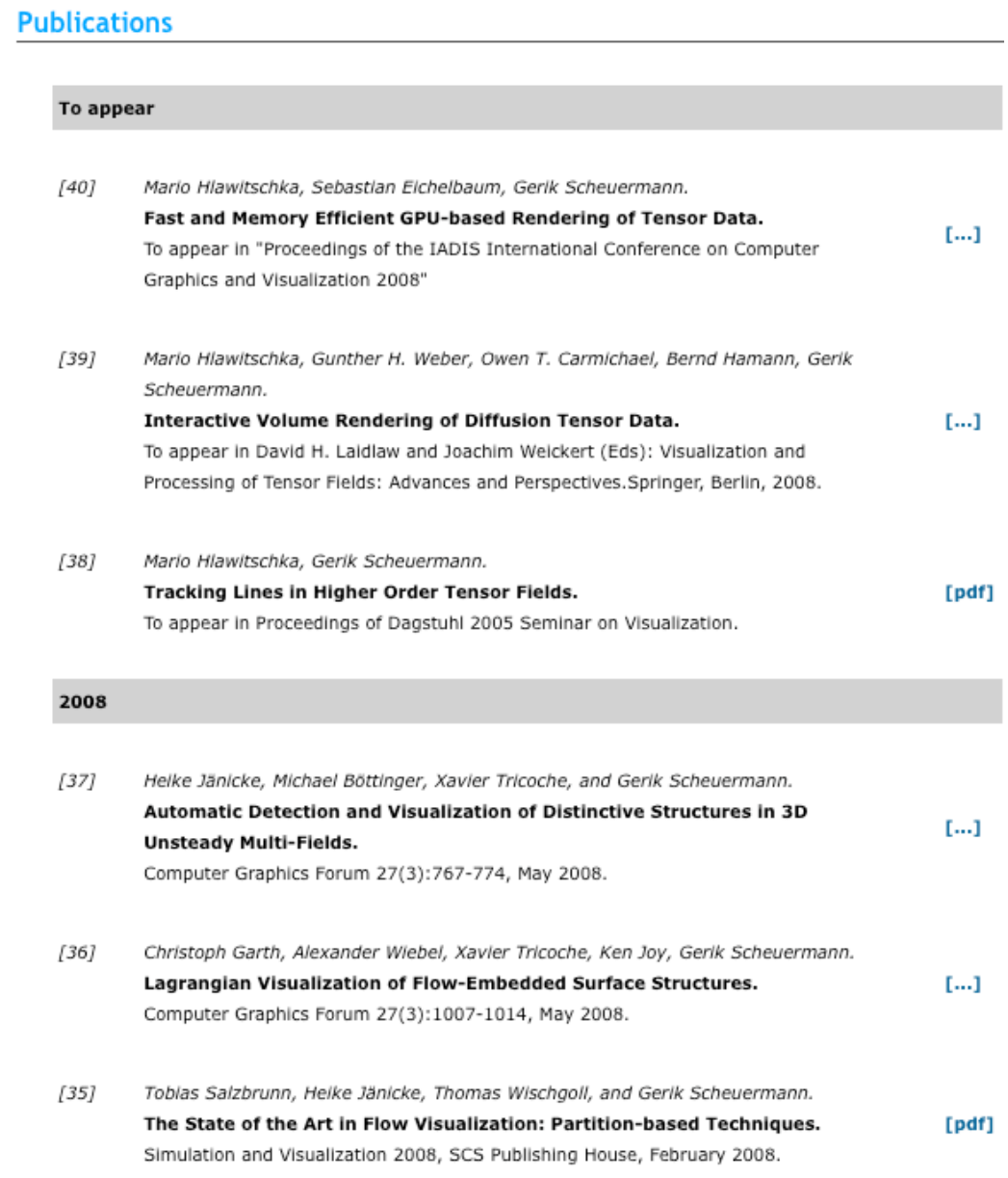
Die Publikationsseiten

- zwei Arten der Auflistung:
  - eine Seite für jedes Jahr
    - Datenbanken
    - Bioinformatik
    - Automatische Sprachverarbeitung
    - Betriebliche Informationssysteme



Die Publikationsseiten

- zwei Arten der Auflistung:
  - eine Seite für jedes Jahr
    - Datenbanken
    - Bioinformatik
    - Automatische Sprachverarbeitung
    - Betriebliche Informationssysteme
  - eine fortlaufende Liste
    - Angewandte Telematik / e-Business
    - Bild- und Signalverarbeitung
    - Computersysteme
    - Parallelverarbeitung und Komplexe Systeme
    - wird fortgestzt



Publications

To appear

[40] *Mario Hlawitschka, Sebastian Eichelbaum, Gerik Scheuermann.*
Fast and Memory Efficient GPU-based Rendering of Tensor Data. [...]
To appear in "Proceedings of the IADIS International Conference on Computer Graphics and Visualization 2008"

[39] *Mario Hlawitschka, Gunther H. Weber, Owen T. Carmichael, Bernd Hamann, Gerik Scheuermann.*
Interactive Volume Rendering of Diffusion Tensor Data. [...]
To appear in David H. Laidlaw and Joachim Weickert (Eds): Visualization and Processing of Tensor Fields: Advances and Perspectives.Springer, Berlin, 2008.

[38] *Mario Hlawitschka, Gerik Scheuermann.*
Tracking Lines in Higher Order Tensor Fields. [pdf]
To appear in Proceedings of Dagstuhl 2005 Seminar on Visualization.

2008

[37] *Heike Jänicke, Michael Böttinger, Xavier Tricoche, and Gerik Scheuermann.*
Automatic Detection and Visualization of Distinctive Structures in 3D Unsteady Multi-Fields. [...]
Computer Graphics Forum 27(3):767-774, May 2008.

[36] *Christoph Garth, Alexander Wiebel, Xavier Tricoche, Ken Joy, Gerik Scheuermann.*
Lagrangian Visualization of Flow-Embedded Surface Structures. [...]
Computer Graphics Forum 27(3):1007-1014, May 2008.

[35] *Tobias Salzbrunn, Heike Jänicke, Thomas Wischgoll, and Gerik Scheuermann.*
The State of the Art in Flow Visualization: Partition-based Techniques. [pdf]
Simulation and Visualization 2008, SCS Publishing House, February 2008.

Die Werkzeuge

- Dapper

  www.dapper.net

- Apatar

  www.apatar.com

Dapper

"Get any content from the web"

- Webapplikation
- erfordert keine Programmierkenntnisse
- große Auswahl an Ausgabeformaten (RSS, XML, HTML, CSV, JSON, ...)
- keinen direkten Einblick oder Manipulation der "dapp-logic"

Dapper Dapp Factory

1. Start
2. Collect Sample Pages
3. Select Content
4. Preview Feed
5. Save Feed

Next Step

Welcome to the Dapp Factory

Where is the content you want to use?

In a website | In an existing RSS Feed

Enter the URL of the website:

http://

Advanced

Choose a format (You can always change this later)

- Dapp XML
- RSS feed
- Google Gadget
- Netvibes Module
- Google Map
- iCalendar
- Image Loop

Not sur

If yo
exte
may

Dapper Dapp Factory

Welcome to the Dapp Factory

1 Start

2 Collect Sample Pages

3 Select Content

4 Preview Feed

5 Save Feed

Back | Next Step

Help

Use the virtual browser to add pages that look the same but contain different content.

For example, to build a Dapp for Flickr search results, enter http://www.flickr.com and then perform a few searches, adding each search results page to your basket. When you're done, click "Next Step".

Leave

http://dbs.uni-leipzig.de/de/publications

Page Added | Publikationen 2008 (7) | ... | Publikationen 2007 (19) |...

Abteilung Datenbanken Leipzig

am Institut für Informatik

Suchen

UNIVE

Startseite » Forschung » Publikationen

Inhalte

- Mitarbeiter
- Forschung
- Studium
- Service

Seiten mit Datumsangaben

- 24.6: News
- 25.6: Klausurergebnis DBS2 Zwischenklausur (26.05.08) und Praktische Klausur (09.06.08)

Neue Publikationen

- Analyzing the Evolution of Life Science Ontologies and Mappings
- Evaluating

Publikationen 2007 (19)

Drumm, C. ; Schmitt, M. ; Do, H.-H. ; Rahm, E.
QuickMig - Automatic Schema Matching for Data Migration Projects
Proc. ACM CIKM, Lisabon, Nov. 2007
2007-11

Rahm, Erhard
Model Management
Datenbank-Spektrum, Heft 23, 2007
2007-11

Do, H.-H. ; Rahm, E.
Matching Large Schemas: Approaches and Evaluation
Information Systems, Volume 32, Issue 6, September 2007, Pages 857-885
2007-09 [12 citations]

Hartung, M. ; Herre, H. ; Loebe, F. ; Rahm, E.
D-Grid Ontology - Semantic description of the D-Grid initiative
Poster for the D-Grid All Hands Meeting (AHM), Göttingen, September 2007
2007-09

Rahm, E. ; Thor, A. ; Aumueller, D.
Dynamic Fusion of Web Data
Proc. XSym07, Vienna, LNCS, Sep. 2007

This page has a login form. Click here to read about privacy.

Done

Dapper Dapp Factory

Click on the content you want. Click again to remove selected content

1 Start

2 Collect Sample Pages

3 Select Content

4 Preview Feed

5 Save Feed

Back | Next Step

Help

Click on the content you would like to include as a field. A field of content might be "Movie Title" or "Number of Results".

When you finish highlighting a field's content, save it by clicking "Save Field".

abcd

Select Inside

Abteilung Datenbanken Leipzig

am Institut für Informatik

Suchen

Startseite » Forschung » Publikationen

Inhalte

- Mitarbeiter
- Forschung
- Studium
- Service

Seiten mit Datumsangaben

- 24.6: News
- 25.6: Klausurergebnis DBS2 Zwischenklausur (26.05.08) und Praktische Klausur (09.06.08)

Neue Publikationen

- Analyzing the Evolution

Publikationen 2008 (7)

Hartung, M. ; Kirsten, T. ; Rahm, E.
Analyzing the Evolution of Life Science Ontologies and Mappings
Proc. of 5th Int. Workshop on Data Integration in the Life Sciences (DILS), Springer LNCS 510
2008-06

Massmann, S. ; Rahm, E.
Evaluating Instance-based Matching of Web Directories
11th International Workshop on the Web and Databases (WebDB 2008)
2008-06

Hartung, M.
Management von Ontologien in den Lebenswissenschaften
Tagungsband zum 20. GI-Workshop über Grundlagen von Datenbanken (20th GI-Workshop o
Databases), Apolda (Thüringen)
2008-05

Krefting, D. ; Bart, J. ; Beronov, K. ; Dzhimova, O. ; Falkner, J. ; Hartung, M. ; Hoheisel, A. ; K
Mohammed, Y. ; Peter, K. ; Rahm, E. ; Sax, U. ; Sommerfeld, D. ; Steinke, T. ; Tolxdorff, T. ; V
Weisbecker, A.

Preview selected content (7) Clear all

Clear Analyzing the Evolution of Life Science Ontologies and Mappings
Clear Evaluating Instance-based Matching of Web Directories
Clear Management von Ontologien in den Lebenswissenschaften
Clear MediGRID: Towards a user friendly secured grid infrastructure
Clear Ontologie-Matching von Produktkatalogen

Save Field

Content fields

authors Edit

Leave

Dapper Dapp Factory

Start

Collect Sample Pages

Select Content

Preview Feed

Save Feed

ck Next Step

elp

hen two or more fields are
lated to one another, they can
e grouped together.

or example, if you created
heater Name" and "Theater
ddress" fields, you can group
em together into a "Theater"
oup. The resulting Dapp will
ve one group for every movie
eater found.

ve

Your group has been saved

Extracted Content:

Publication

authors Hartung, M.;Kirsten, T.;Rahm, E.

title Analyzing the Evolution of Life Science Ontologies and Mappings

Publication

authors Massmann, S.;Rahm, E.

title Evaluating Instance-based Matching of Web Directories

Publication

authors Hartung, M.

title Management von Ontologien in den Lebenswissenschaften

Publication

authors Krefting, D.;Bart, J.;Beronov, K.;Dzhimova, O.;Falkner, J.;Hartung, M.;Hoheisel, A.;Knoch, T.A.;Lingner, T.;Mohammed, Y.;Peter, K.;Rahm, E.;Sax D.;Steinke, T.;Tolxdorff, T.;Vossberg, M.;Viezens, F.;Weisbecker, A.

title MediGRID: Towards a user friendly secured grid infrastructure

Publication

authors Massmann, S.

title Ontologie-Matching von Produktkatalogen

Publication

authors Rahm, E.

Content Fields Select all

- authors
- title

Save Group

Groups

Publication Edit

db_vorfuehr_dapp

Dapp creator options: Edit Dapp | Make public | Delete Dapp

Is this a Dapp for your site? Expose semantics to search engines

Data Mapping

Screenshot 0 seconds ago, http://dbs.uni-leipzig.de/publication/year/2008

Terms of Use

Content Owner
dbs.uni-leipzig.de

Direct Permission Required

Dapp Creator

see all Dapps by engsterhold

Use This Dapp

Choose format: XML Go

Create a Flash Widget

Create an Alert

Edit Dapp | Create from

Link Dapp's output to another Dapp

Set Input

URL: http://dbs.uni-leipzig.de/publication/year/2008

Update Input

Dapp Preview

Publication

authors Hartung, M. ; Kirsten, T. ; Rahm, E.

title Analyzing the Evolution of Life Science Ontologies and Mappings

Publication

authors Massmann, S. ; Rahm, E.

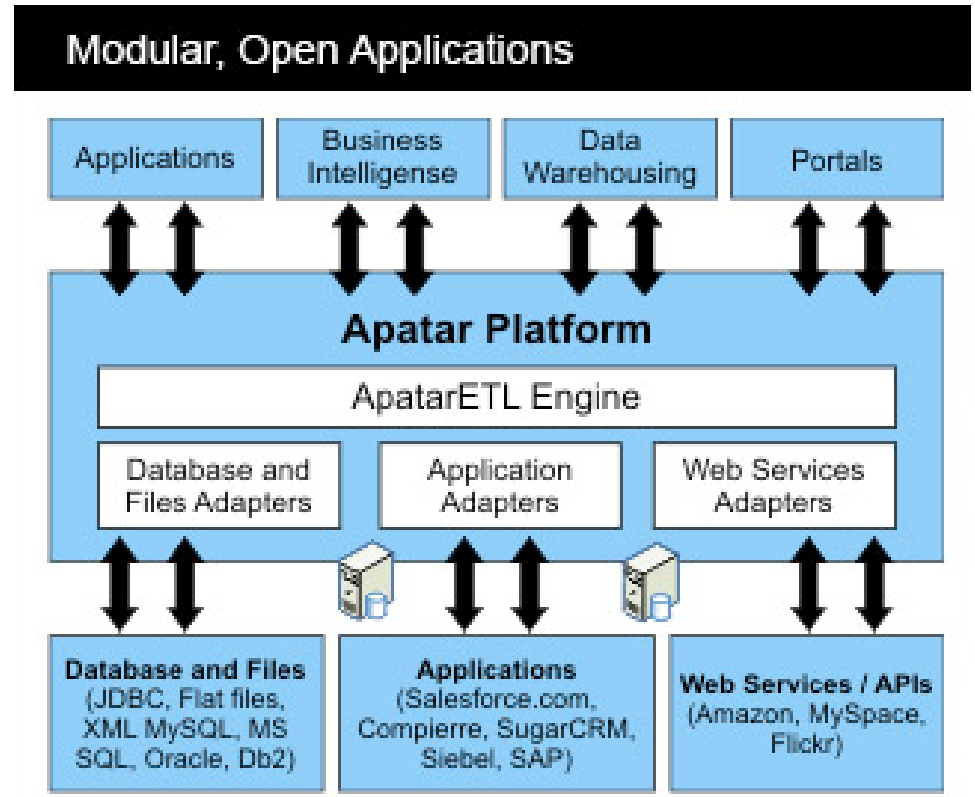
title Evaluating Instance-based Matching of Web Directories

<elements xsi:noNamespaceSchemaLocation="http://www.dapper.net/websiteServices/generate-dapp-xsd.php?
dappName=db_vorfuehr_dapp">
<dapper>
<dappName>db_vorfuehr_dapp</dappName>
<dappTitle>db_vorfuehr_dapp</dappTitle>
<urls>
<url>http://dbs.uni-leipzig.de/publication/year/2008</url>
<url>http://dbs.uni-leipzig.de/publication/year/2007</url>
</urls>
<applyToUrl>http://dbs.uni-leipzig.de/publication/year/2008</applyToUrl>
<executionTime>0.028</executionTime>
<ranAt>2008-06-24 09:10:31</ranAt>
<encoding>utf-8</encoding>
<ranEventChain>false</ranEventChain>
<inputVars/>
</dapper>
<Publication groupName="Publication" type="group">
<authors dataType="RawString" fieldName="authors" originalElement="span" type="field">Hartung, M. ; Kirsten, T. ; Rahm, E.</authors>
<title dataType="RawString" fieldName="title" href="http://dbs.uni-leipzig.de/de/publication/title/
analyzing_the_evolution_of_life_science_ontologies_and_mappings" originalElement="a" type="field">
Analyzing the Evolution of Life Science Ontologies and Mappings
</title>
</Publication>
<Publication groupName="Publication" type="group">
<authors dataType="RawString" fieldName="authors" originalElement="span" type="field">Massmann, S. ; Rahm, E.</authors>
<title dataType="RawString" fieldName="title" href="http://dbs.uni-leipzig.de/de/publication/title/
evaluating_instance_based_matching_of_web_directories" originalElement="a" type="field">
Evaluating Instance-based Matching of Web Directories
</title>
</Publication>
<Publication groupName="Publication" type="group">
<authors dataType="RawString" fieldName="authors" originalElement="span" type="field">Hartung, M.</authors>
<title dataType="RawString" fieldName="title" href="http://dbs.uni-leipzig.de/de/publication/title/
management_von_ontologien_in_den_lebenswissenschaften" originalElement="a" type="field">
Management von Ontologien in den Lebenswissenschaften
</title>
</Publication>
</elements>

Apatar

"Integrate your information between on-premise or on-demand data sources and applications"

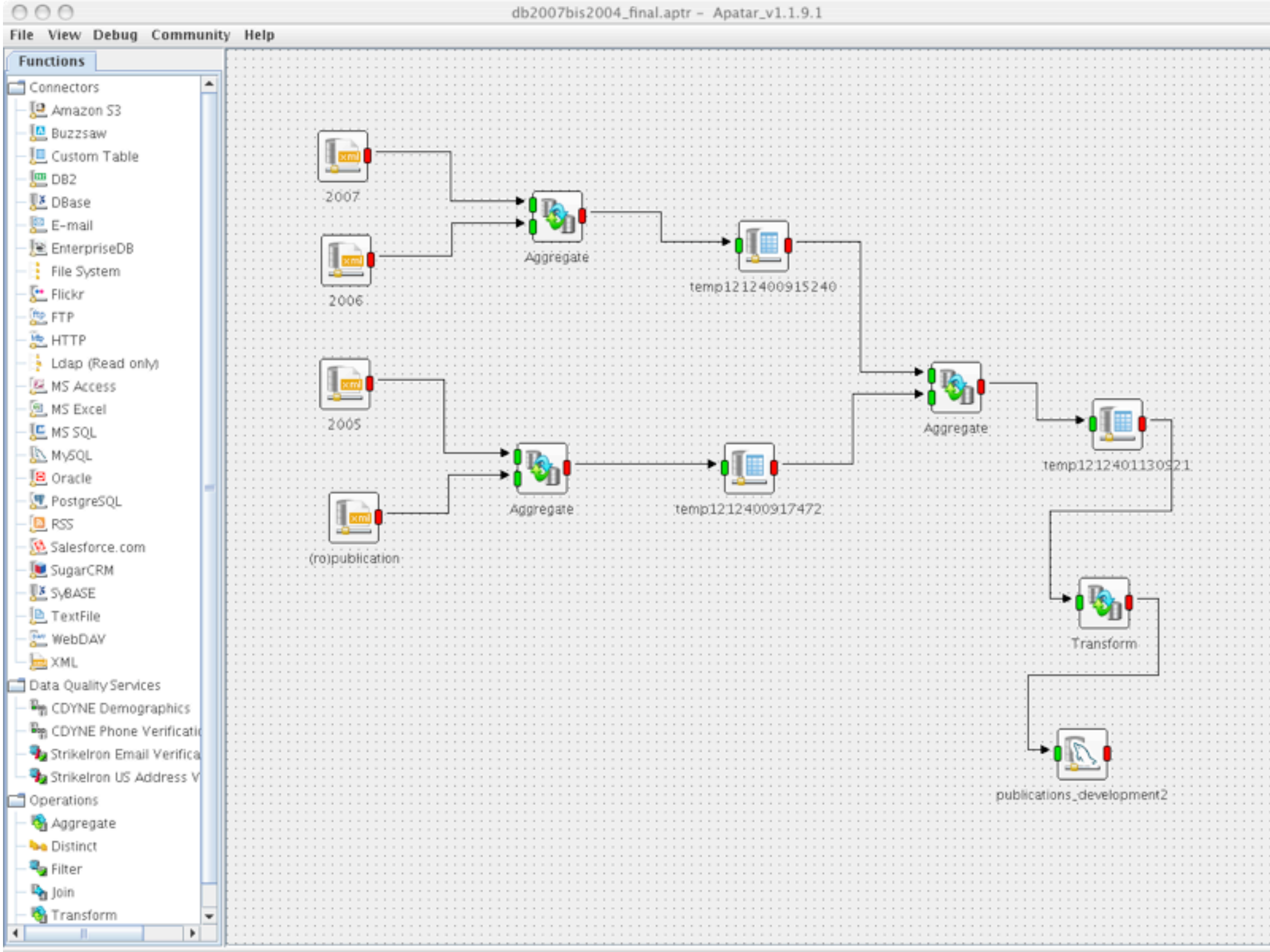
- open source
- Plattformunabhängig, da in Java geschrieben.
- modular und damit leicht erweiterbar.
- wenig Programmierkenntnisse nötig. Das meiste funktioniert per Drag & Drop über die GUI.



Apatar

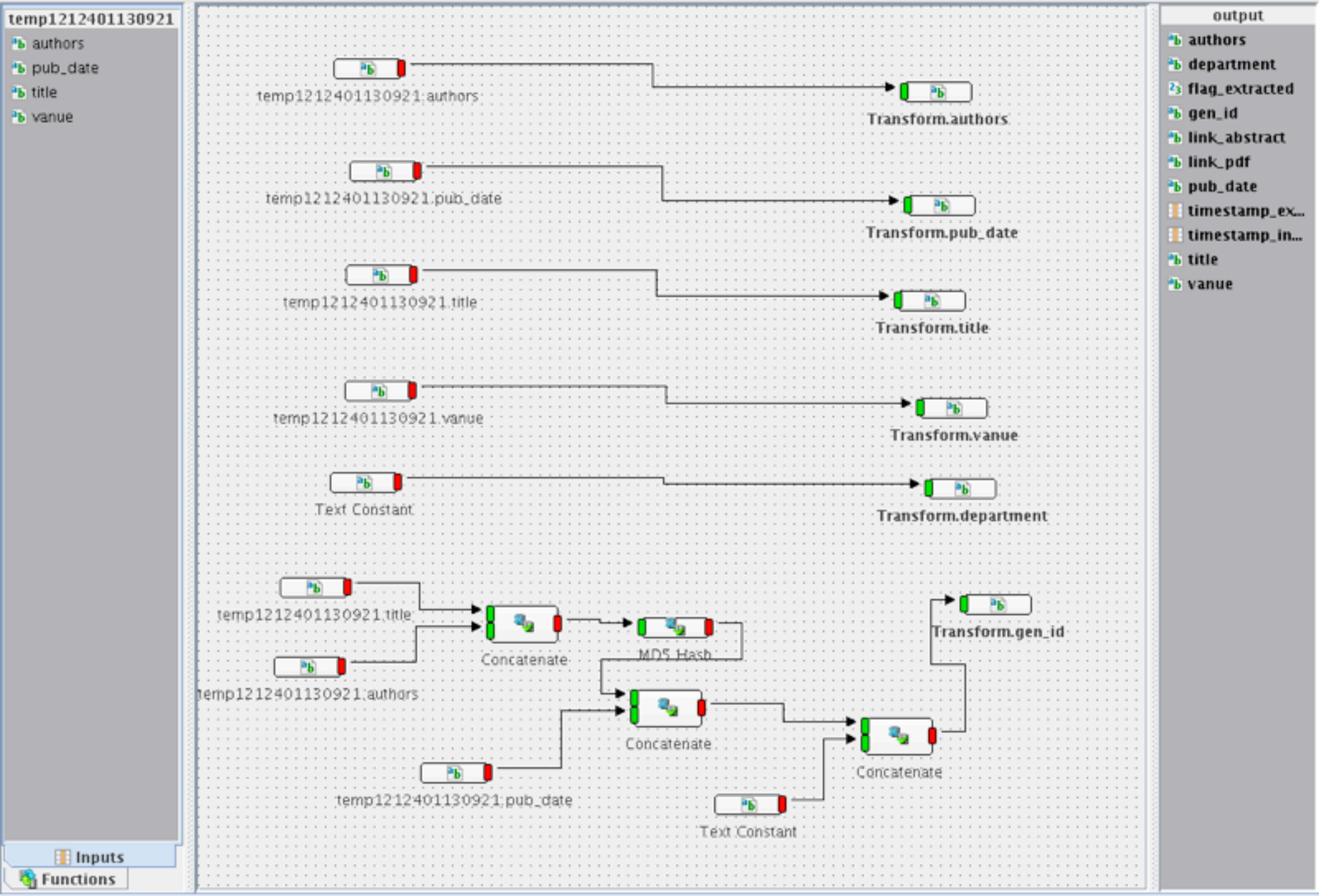
- Ist eigentlich eine Business Appliaktion.
- In diese Richtung läuft auch hauptsächlich die Entwicklung und der Support.
- Besitzt aber viele Funktionen zur Manipulation von Strings.

Featured Apatar Users


db2007bis2004_final.aptr - Apatar_v1.1.9.1
File View Debug Community Help
Functions
Connectors
Amazon S3
Buzzsaw
Custom Table
DB2
DBase
E-mail
EnterpriseDB
File System
Flickr
FTP
HTTP
Ldap (Read only)
MS Access
MS Excel
MS SQL
MySQL
Oracle
PostgreSQL
RSS
Salesforce.com
SugarCRM
SyBASE
TextFile
WebDAV
XML
Data Quality Services
CDYNE Demographics
CDYNE Phone Verificatic
StrikeIron Email Verifica
StrikeIron US Address V
Operations
Aggregate
Distinct
Filter
Join
Transform
2007
2006
2005
(ro)publication
Aggregate
Aggregate
temp1212400915240
temp1212400917472
Aggregate
temp1212401130921
Transform
publications_development2

Transformation

Edit Output



View operation guide

Ok

Cancel

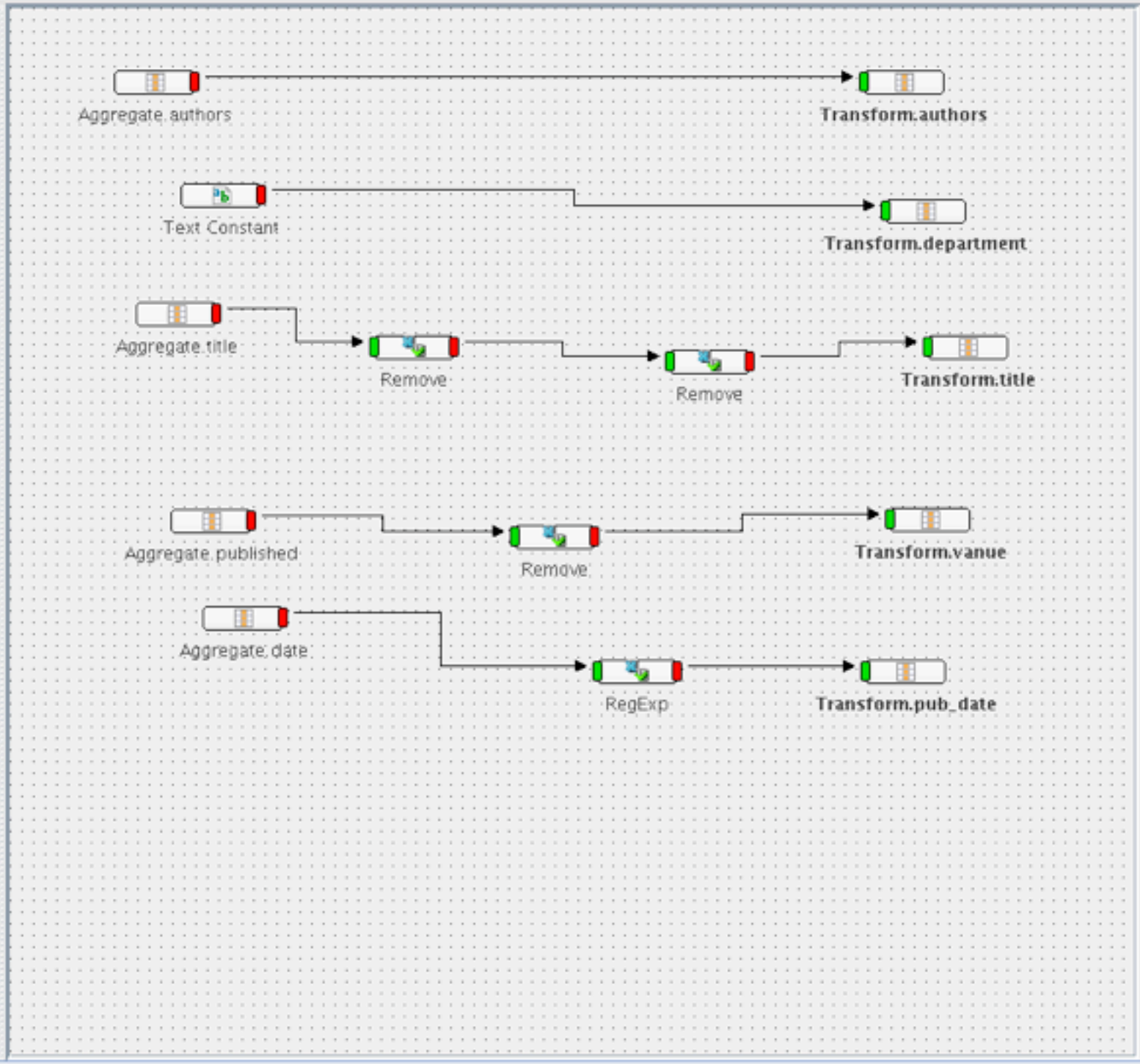
Transformation

Edit Output

Aggregate
- authors
- block
- date
- published
- title

Inputs

Functions

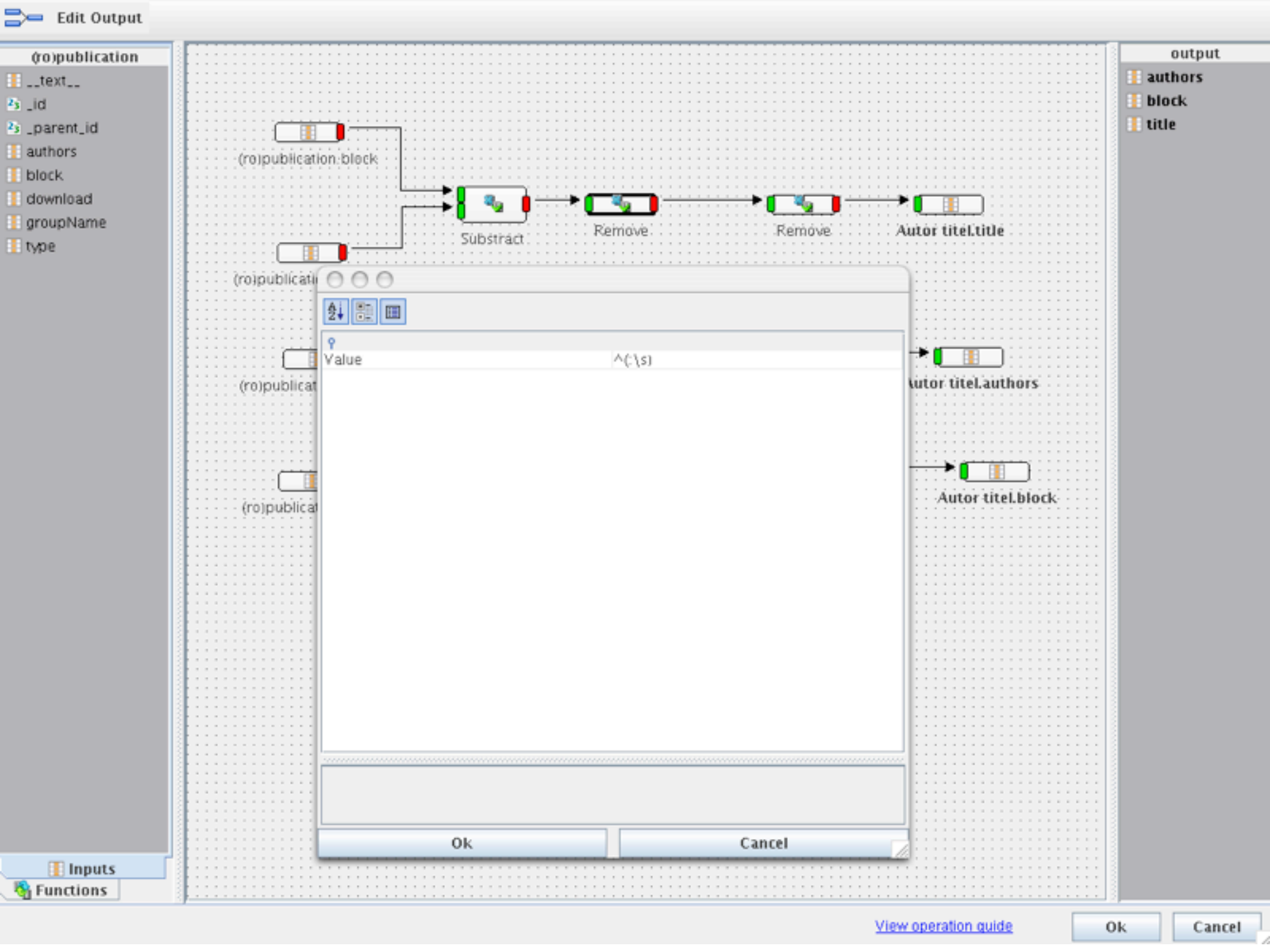


output
- authors
- department
- link_abstract
- link_pdf
- pub_date
- title
- vanue

View operation guide

Ok

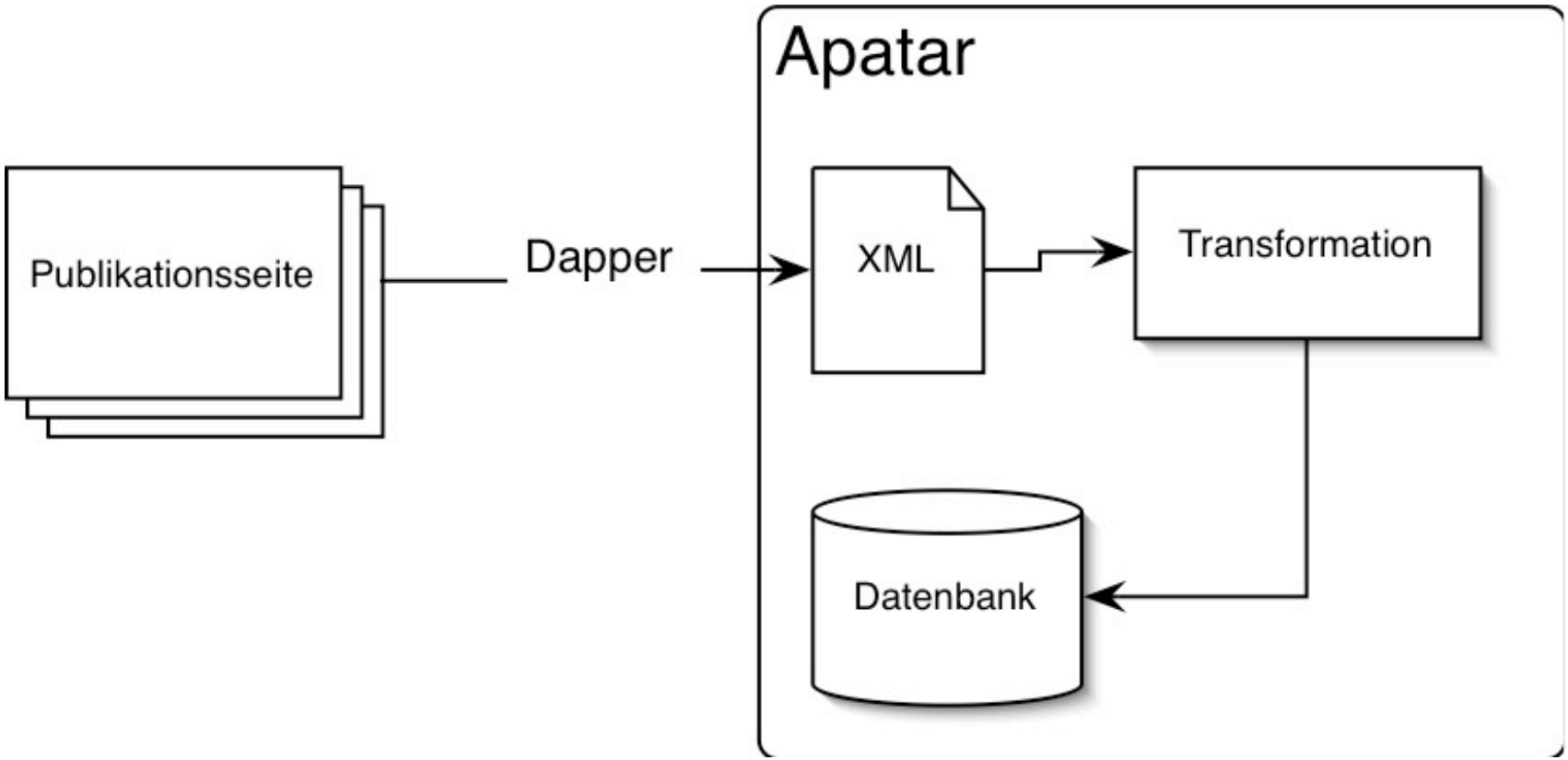
Cancel


Edit Output
(ro)publication
__text__
_id
_parent_id
authors
block
download
groupName
type
(ro)publication.block
Substract
Remove
Remove
Autor titel.title
Value
^(:\s)
Autor titel.authors
Autor titel.block
Ok
Cancel
Inputs
Functions
output
authors
block
title
View operation guide
Ok
Cancel

Der Aufbau des Publikationscrawler

1. Für jede Publikationsseite wird ein Dapp erstellt.
2. Dieses Dapp wird in Apatar eingelesen über den XML Connector.
3. Nach den benötigten Transformationen werden die Daten über den einen Connector in die Datenbank geschrieben. Zur Zeit eine mySQL Datenbank.

Der Aufbau des Publikationscrawler



Arbeitsweise des Publikationscrawler

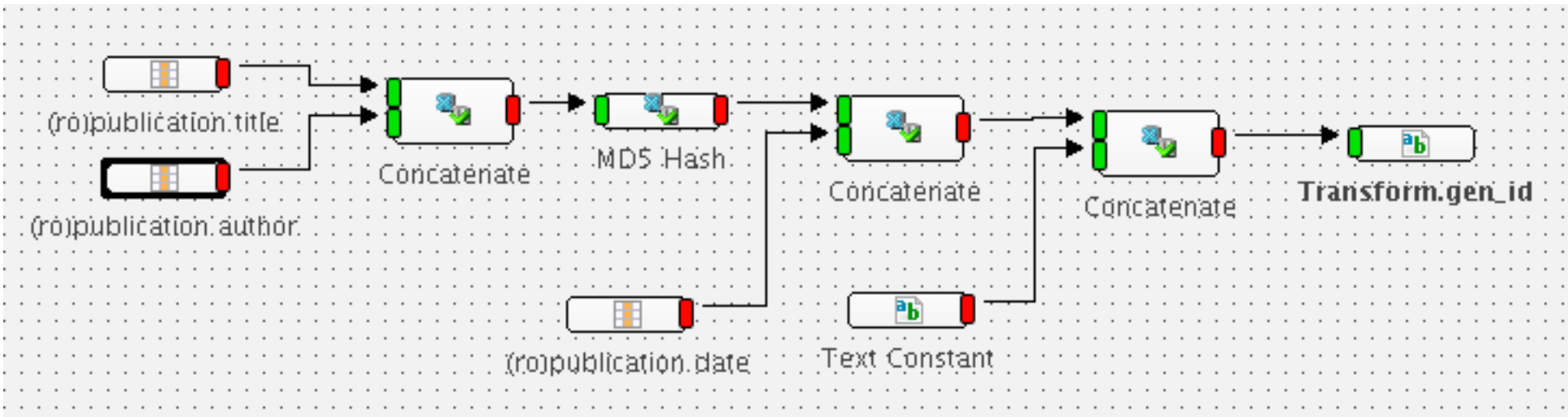
- Die Publikationsseiten vergangener Jahre müssen nur einmal sauber erfasst werden.
- Die aktuellen Seiten sollen einmal die Woche abgefragt werden.
- Apatar besitzt eine Schedulerfunktion, womit sich die entsprechenden Datamaps aufrufen lassen.

Die Datenbank

```
CREATE TABLE `publications_development2` (
    `gen_id` CHAR(100) DEFAULT '' NOT NULL,
    `title` TEXT(65535) NOT NULL,
    `authors` TEXT(65535) NOT NULL,
    `vanue` TEXT(65535),
    `pub_date` TEXT(65535),
    `link_pdf` TEXT(65535),
    `link_abstract` TEXT(65535),
    `department` TEXT(65535),
    `timestamp_insert` TIMESTAMP DEFAULT 'CURRENT_TIMESTAMP' NOT NULL,
    `timestamp_extract` TIMESTAMP DEFAULT '0000-00-00 00:00:00' NOT NULL,
    `flag_extracted` TINYINT(3) DEFAULT '0',
     PRIMARY KEY (`gen_id`)
);
```

Der Primärschlüssel

Eine Publikation soll durch den Primärschlüssel eindeutig identifizierbar sein. Titel und Autoren bestimmen meistens ein Werk eindeutig, aber die Datensätze können zu lang für den Primärschlüssel sein. Also werden Titel und Autoren mit dem MD5-Algorithmus verschlüsselt und um mögliche Kollisionen minimieren werden noch das Jahr und die Abteilung hinzugefügt.



Back Next

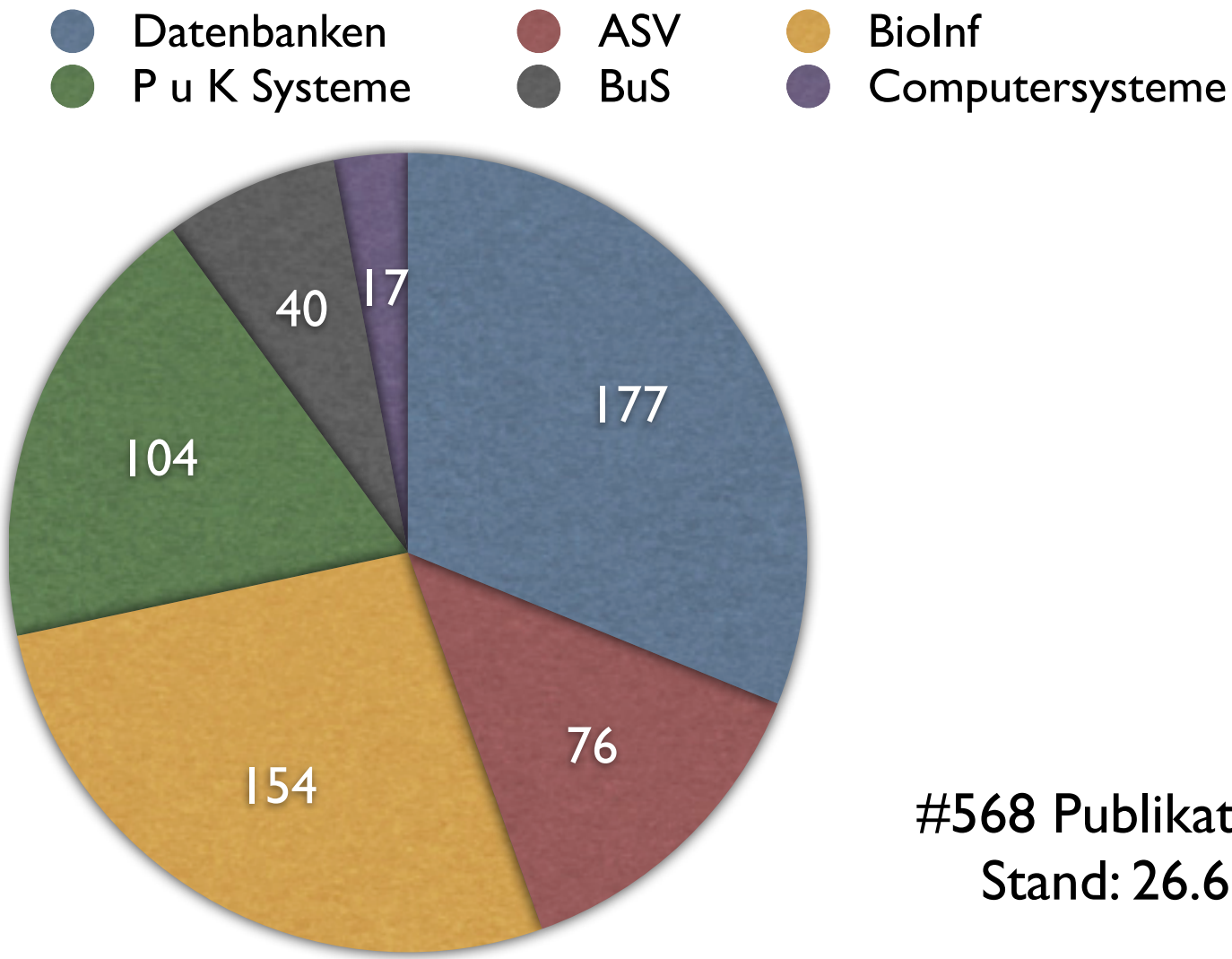
```
SELECT * FROM publications_development2 p
```

Execute Stop

Query

#gen_id	title	authors	vanue	pub
9efee1b2607b368d998093cc647d61f32000	Data Warehouse Scenarios for Model	Bernstein, P.A. ; Rahm, E.	Proc. 19th Int. Conf. on Entity-Relationship	200
79386039acbf0b5b3f8c08ab5d7f84e62000D	On Metadata Interoperability in Data	Do, H. H. ; Rahm, E.	Technical Report 01-2000. Dept. of	200
816cdf9be1976a2c17640e66d3b3d7822000	Evaluierung von Data Warehouse-	Do, H. H. ; Stöhr, T. ; Rahm, E. ; Müller, R. ;	Proc. Data Warehousing (DW) 2000,	200
4c27fdfede51d263c2d1f7832a3578162000D	Dealing with Logical Failures for	Müller, R. ; Rahm, E.	In Etzion, O.; Scheuermann, P. (Eds.):	200
20f54f61b240b11a8a436d42917087a02000	Data Cleaning: Problems and Current	Rahm, E. ; Do, H.H.	IEEE Techn. Bulletin on Data Engineering, Dec.	200
fd06ab2ee7364d28e668d86f3881c132000D	Annotationen in Dokumenten einer verteilten	Sosna, D.	Ifl-Report, 2000	200
989fc141b0c6f948ede2aa9d7e6559802000D	Multi-Dimensional Database Allocation for	Stöhr, T. ; Märtens, H. ; Rahm, E.	Proc. 26th Intl. Conference on Very Large	200
d007c8508c1fbf29c45a3939b85fd80f2000D	OLAP-Auswertung von Web-Zugriffen	Stöhr, T. ; Rahm, E. ; Quitzsch, S.	Proc. GI-Workshop Internet-Datenbanken,	200
81f0d32499c093d2ea1413db68571b7e2007	QuickMig - Automatic Schema Matching for	Drumm, C. ; Schmitt, M. ; Do, H.-H. ; Rahm,	Proc. ACM CIKM, Lisabon, Nov. 2007	200
6c375806496d32f63ed696eaf3e1393b2007	Model Management	Rahm, Erhard	Datenbank-Spektrum, Heft 23, 2007	200
7afccf421835d66078f3f01b1a9f7d0a2007-0	Matching Large Schemas: Approaches and	Do, H.-H. ; Rahm, E.	Information Systems, Volume 32, Issue 6,	200
132d6083b1c2fd36c2b6235bdd274f0f2007-	D-Grid Ontology - Semantic description of	Hartung, M. ; Herre, H. ; Loebe, F. ; Rahm, E.	Poster for the D-Grid All Hands Meeting	200
86943e12bc6bc016812624c38753b4c32007	Dynamic Fusion of Web Data	Rahm, E. ; Thor, A. ; Aumueller, D.	Proc. XSym07, Vienna, LNCS, Sep. 2007	200
1e28eb7429790f67fc514d5dba918dc12007-	Data Integration Support for Mashups	Thor, Andreas ; Aumueller, David ; Rahm,	Sixth International Workshop on Information	200
16fbf1d6e74b8876ac2625c3154ca7a02007-	Caravela: Semantic Content Management with	Aumueller, David ; Rahm, Erhard	E. Franconi, M. Kifer, and W. May (Eds.): ESWC	200
32d0d6ea680e42c6176fa906243cc1332007	Parameterized XPath Views	Böhme, Timo ; Rahm, Erhard	Proc. BNCOD, LNCS	200
c553151381653ca7f6f846110b0e5b572007	Instance-based matching of large life science	Kirsten, T. ; Thor, A. ; Rahm, E.	Proc. DILS 2007, LNCS	200
8d9f365a99a7d10b5128494f0ed384f42007-	A Grid Middleware for Ontology Access	Hartung, M. ; Rahm, E.	1st German e-Science Conference, 2007.	200
979a8f80008633fbb50dcad8afdd38362007-	Auf dem Weg zur individualisierten Medizin -	Sax, U. ; Weisbecker, A. ; Falkner, J. ; Viezens,	Telemed 2007 - Electronic Health Record und	200
e7d9bde43fd90a95e0bafc8fcb912922007-0	Instance Matching with COMA++	Engmann, D. ; Massmann, S.	BTW 2007 Workshop: Model Management	200
9d2928d04def5a12b3edd55b0a60bcc72007	Automatisierte Umsetzung von komplexen	Hartung, M.	BTW Workshop "Model Management und	200
bb750c4e088320492620bf88fcf3fc0f2007-0	Datenbanksysteme in Business, Technologie	Jarke, M. ; Seidl, T. ; Quix, C. ; Kensche, D. ;	ISBN 3-86130-929-7, Verlagshaus Mainz,	200
241f7eebbccf872df343b761c2b7e0c52007-	Instance-based matching of hierarchical	Thor, A. ; Kirsten, T. ; Rahm, E.	Proc. of 12. GI-Fachtagung für	200
c9f933ac9afd78e78e9b5e5932ee92f82007-	FUNC: a package for detecting significant	Prufer, Kay ; Muetzel, Bjoern ; Do, Hong-Hai ;	BMC Bioinformatics 2007, 8:41	200
dc3a6f5f2284dbdf7a7c9178333f82e2007-0	Analyse von Zitierungshäufigkeiten für die	Köpcke, H. ; Rahm, E.	Datenbank-Spektrum, 7. Jahrgang, Heft 20	200
f7487a3ab6e5ecf01d1e6eef8ddc599c2007-0	The GeWare data warehouse platform for the	Rahm, Erhard ; Kirsten, Toralf ; Lange, Jörg	Journal of Integrative Bioinformatics, 4(1):47,	200
b0950e5d5e63352fcdcdcd81413c2e4c2007-	MOMA - A Mapping-based Object Matching	Thor, A. ; Rahm, E.	Proc. of the 3rd Biennial Conference on	200
944441cfb444dee9bc8588367c79b952006-	COMA++: Results for the Ontology	Massmann, S. ; Engmann, D. ; Rahm, E.	International Workshop on Ontology	200
875a71d709fb4214e32137058a7c30202006	BioFuice: Mapping-based data integration in	Kirsten, Toralf ; Rahm, Erhard	Proc. of 3rd Int. Workshop on Data	200
d321e5319a50db05d97b9bc0adba301f2006	An integrated platform for analyzing	Kirsten, Toralf ; Lange, Jörg ; Rahm, Erhard	EDBT-Workshop Information Integration in	200
d8593055ebac609304229bb77bc6b7062006	LOTS - Online-Training an der Universität	Böhme, T. ; Rahm, E. ; Sosna, D.	Workshop on e-Learning 2006, HTWK Leipzig	200
46990765959ebef97e353228ea2410d2006D	Schema Matching and Mapping-based Data	Do, Hai Hong	Verlag Dr. Müller (VDM), ISBN	200
556ec8ddcc2c725b1bcc23a3115f282006Dat	An Online Bibliography on Schema Evolution	Rahm, Erhard ; Bernstein, Philip A.	Sigmod Record, Dec. 2006	200
55fb0e6c4d6c997f63db5d7c6c318f462005-	BioFuice: A decentralized Approach to	Kirsten, T. ; Rahm, E.	Proc 4th Research Festival for Life Sciences,	200
a79e301c19ca00cb14abb77c3a43dc572005-	An integrated platform for analyzing clinical	Lange, J. ; Kirsten, T. ; Rahm, E. ; Berger, H. ;	Proc 4th Research Festival for Life Sciences,	200
f2cae88e99546b426f3fda9f8e5f32b72005-1	Citation analysis of database publications	Rahm, E. ; Thor, A.	ACM Sigmod Record	200
85de9b1044bcd27e0ec36ec83efb30192005-	Adaptive Website Recommendations with	Thor, A. ; Golovin, N. ; Rahm, E.	VLDB Journal 14(4)	200
ce547a22df45c737127b25cc8a235cd22005-	Towards a Semantic Wiki Experience -	Aumueller, D. ; Auer, S.	1st Workshop on The Semantic Desktop, Next	200
10581640bda59adf44b26d82c61a87b02005	Automatic Optimization of Web	Golovin, N. ; Rahm, E.	5th Int. Conf. on Web Engineering (ICWE)	200
9c0c3c3d64a72bb6a6afbbc11eaf6e6f2005-0	Hybrid Integration of Molecular-biological	Kirsten, T. ; Körner, C. ; Do, H.H. ; Rahm, E.	Proc. 2nd International Workshop on Data	200

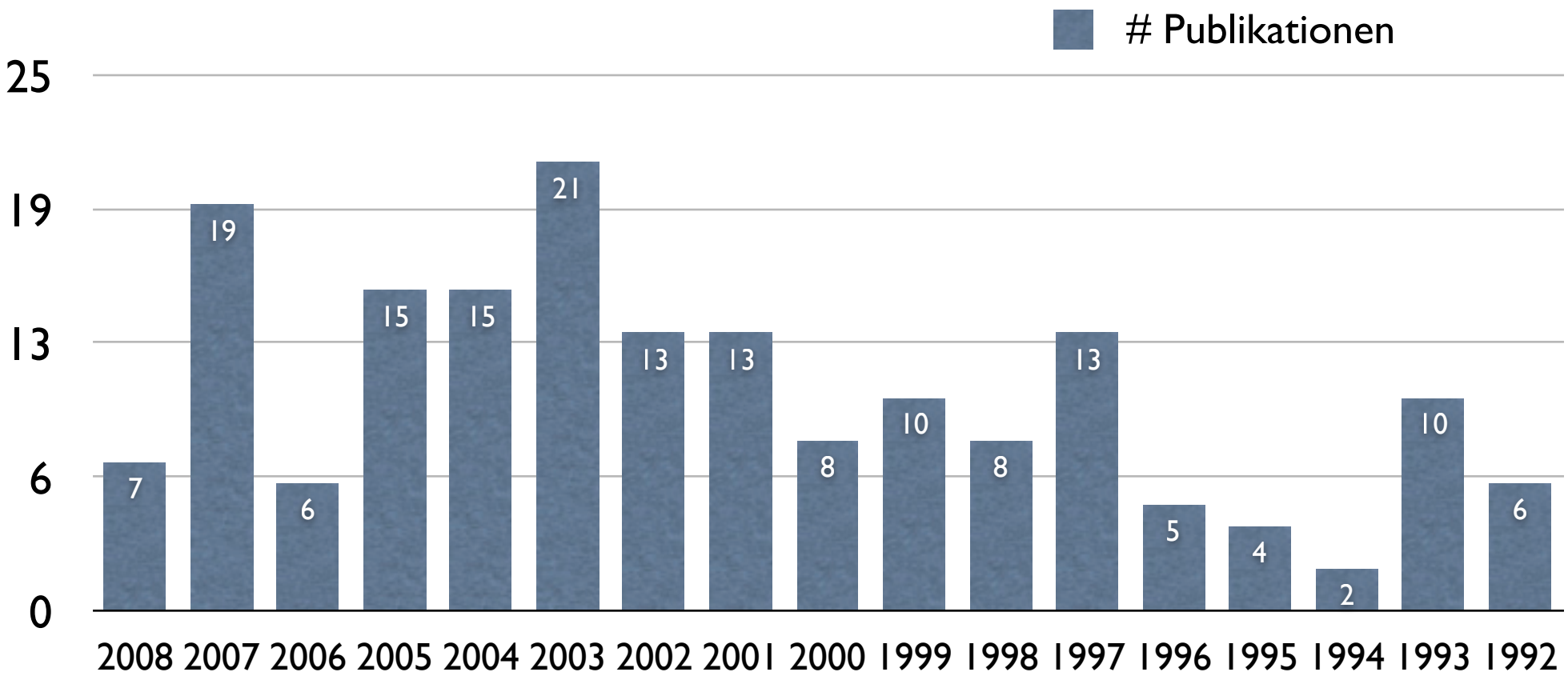
Die momentan erfassten Publikationen



#568 Publikationen
Stand: 26.6.08

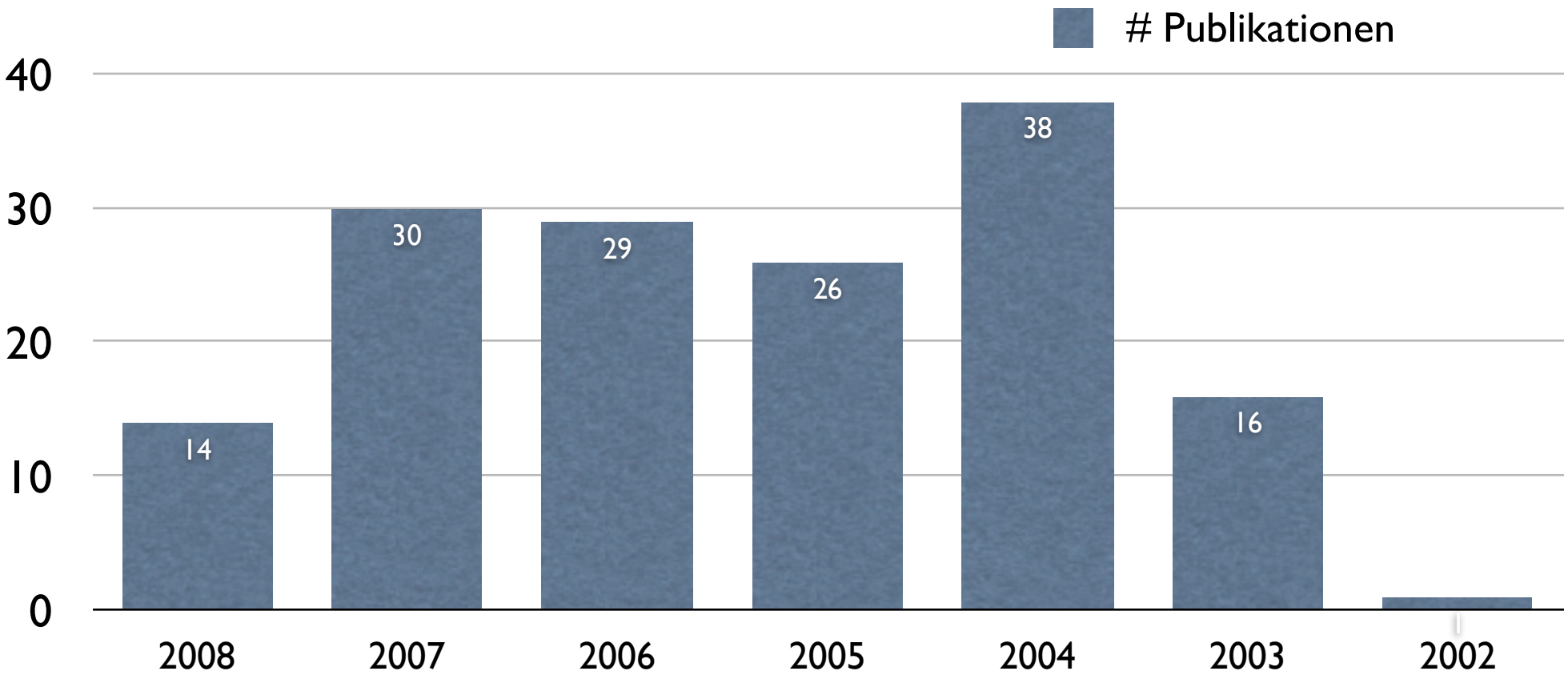
Datenbanken

- Das Dapp erkennt Autoren, (verlinkten) Titel, Venue, Erscheinungsjahr.
- In Apatar ist nur ein einfaches Mapping auf die Datenbank nötig.



Bioinformatik

- Das Dapp erkennt Autoren, Titel, Venue, Link auf ein Abstrakt, Link auf ein PDF.
- In Apatar müssen noch die Jahre dem Mapping hinzugefügt werden.



Automatische Sprachverarbeitung

- Das Dapp erkennt Autoren, Titel, Venue, Link auf ein PDF.
- In Apatar müssen noch Titel und Venue nachbearbeitet werden. Das Erscheinungsdatum muss noch aus dem Venue Block extrahiert werden.

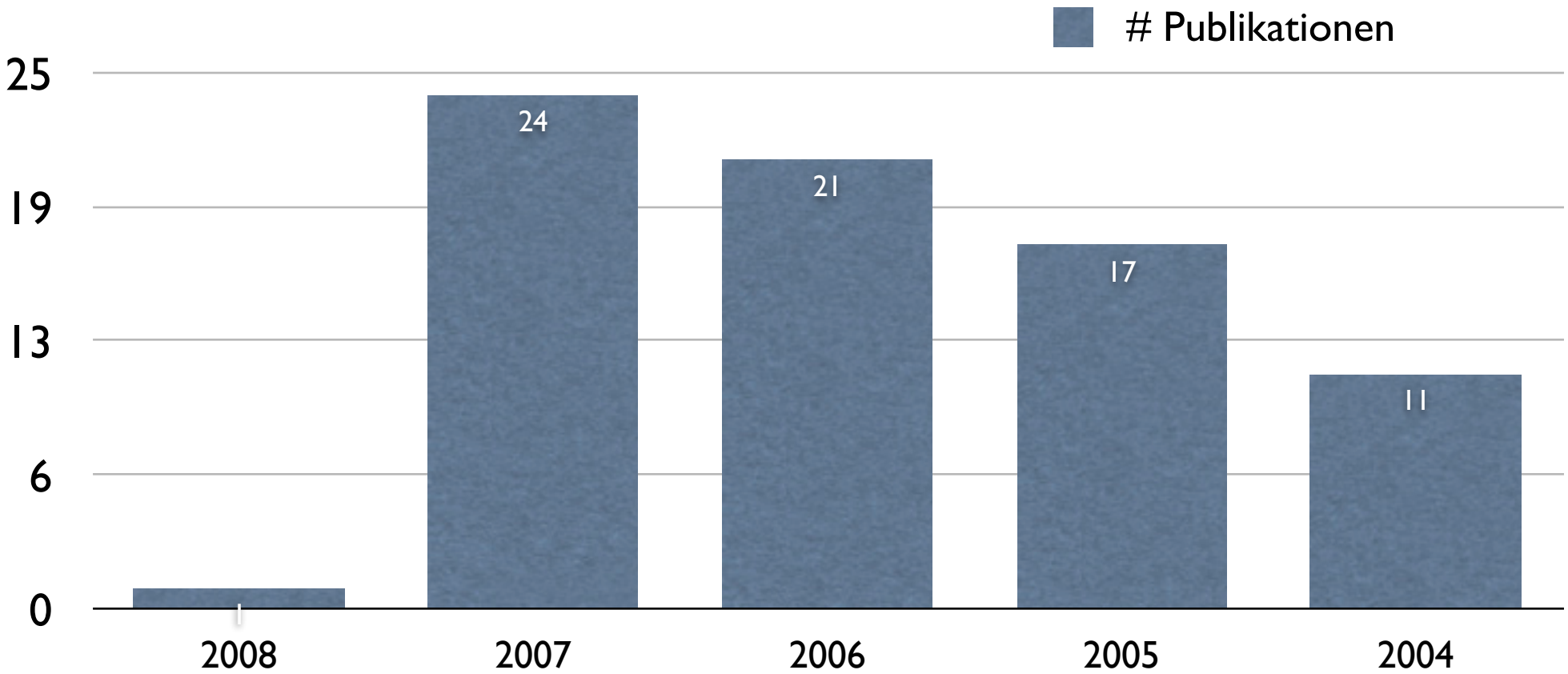
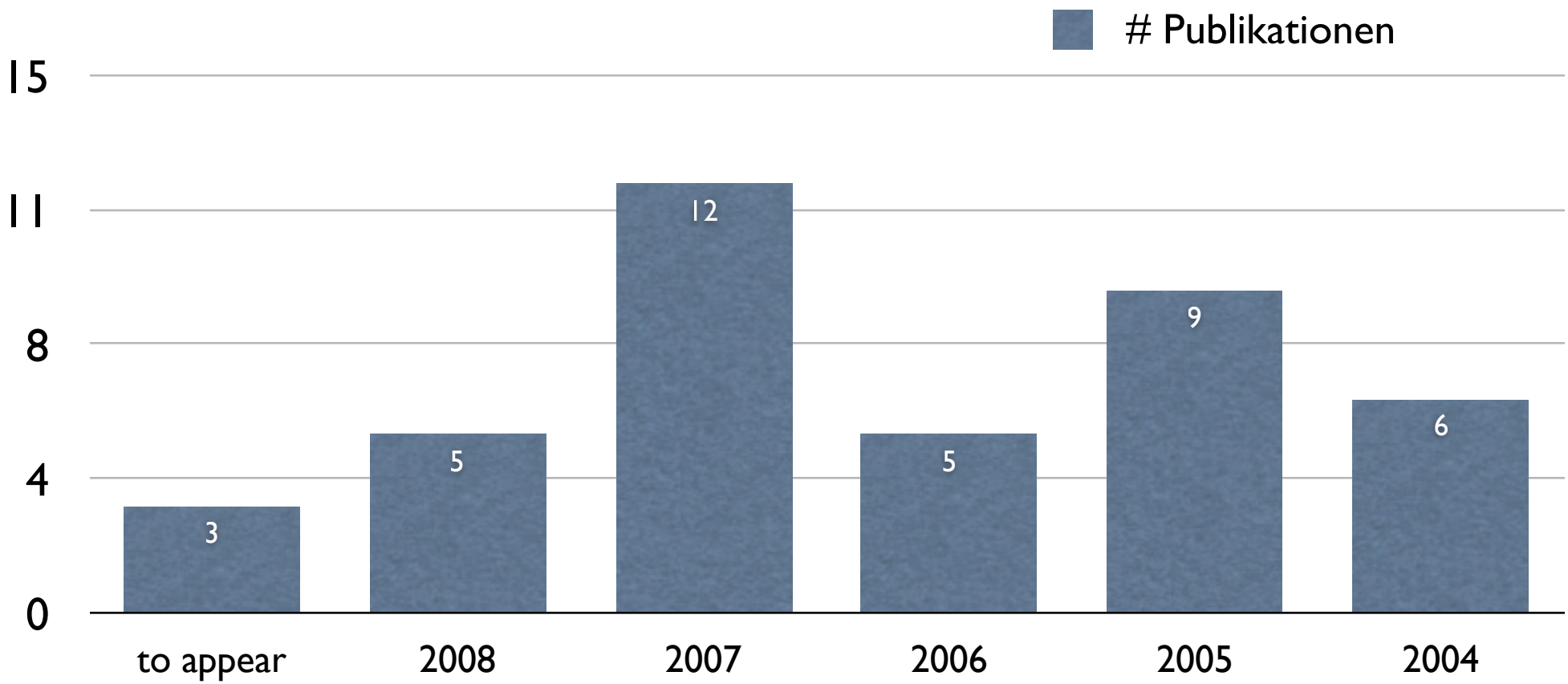


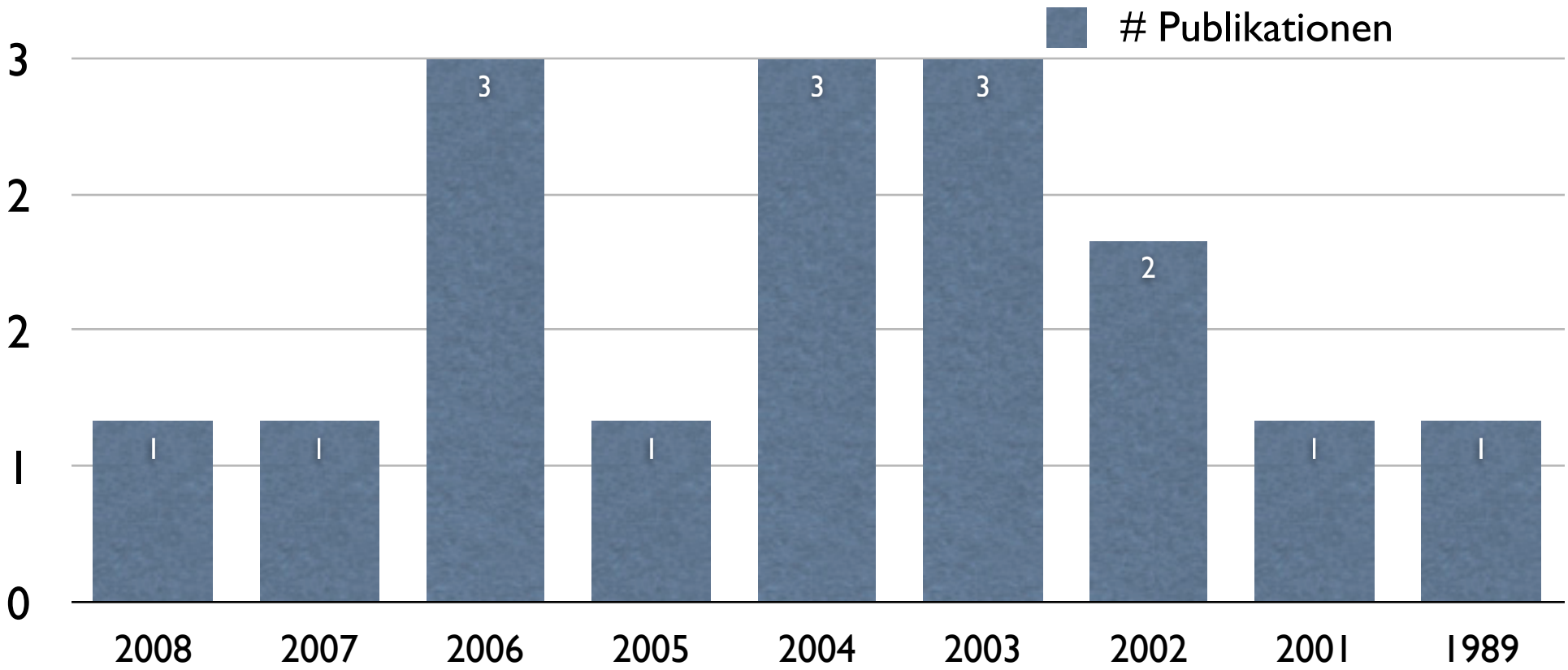
Bild- und Signalverarbeitung

- Das Dapp erkennt Autoren, Titel, Venue, Link auf ein PDF.
- In Apatar kann dank des Join-Operators jeder Publikation dem zugehörigen Jahr zugeordnet werden.



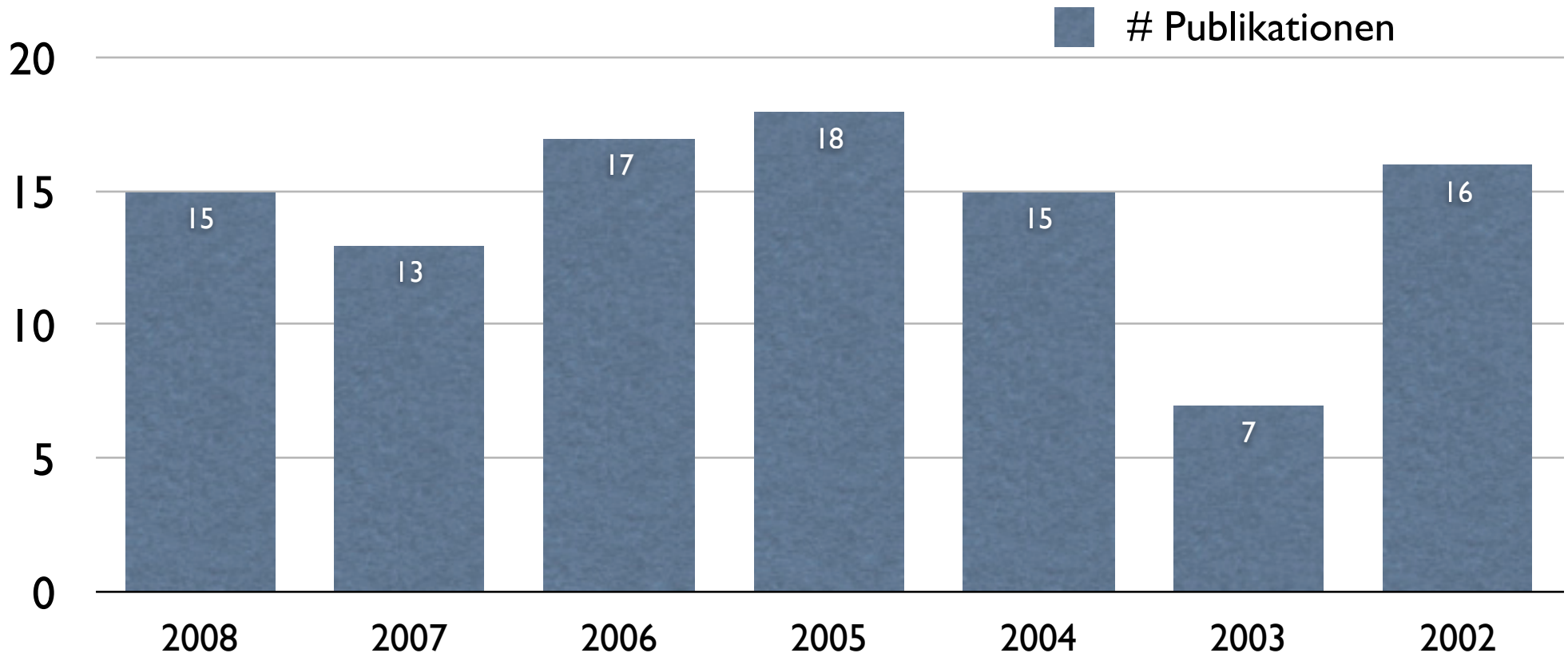
Computersysteme

- Das Dapp erkennt nur den Autor, den gesamten Block und Downloadlink.
- In Apatar müssen noch Titel, Venue und das Erscheinungsjahr mit Hilfe regulärer Ausdrücke und String Manipulation erkannt werden.



Parallelverarbeitung und Komplexe Systeme

- Das Dapp erkennt nur den Autor, den Titel, den gesamten Block und Downloadlink.
- In Apatar müssen noch Venue und das Erscheinungsjahr mit Hilfe regulärer Ausdrücke und String Manipulation erkannt werden.



Probleme bei der Datenextraktion

- Damit der Publikationscrawler überhaupt funktioniert, ist es wichtig, dass Dapper die benannten Felder (authors, title, venue usw.) richtig und vollständig erkennt. Auch für zukünftige Publikationen.
- Je weniger Elemente Dapper sauber erkennt umso mehr muss in Apatar nachgearbeitet werden.

Probleme bei der Datenextraktion

Jede Abteilung verwendet eine andere Struktur zur Auflistung der Publikationen. Je weniger die einzelnen für den Crawler interessanten Elemente durch Html-Elemente eingeschlossen werden, um so schlechter kann Dapper die Elemente erkennen. Um so mehr müssen diese Elemente mit Hilfe von Regulären Ausdrücken nachbearbeitet werden.

Beispiele

- Datenbanken

  Sämtliche für den Crawler relevante Felder sind mit einem umgeben. So kann Dapper alle Felder erkennen.

- Computersysteme

  Eine Publikation findet sich innerhalb eines <p>-Blocks. Html-Elemente können nicht für die Auftrennung von Autor, Titel usw. herangezogen werden. Nur eine Publikation als ganzes kann sicher von Dapper erkannt werden. Zur Auftrennung werden Textstrukturen benutzt. Die Autoren enden mit einem Doppelpunkt, der Titel mit einem Punkt. Es gibt keine Garantie, dass das immer zutrifft, bzw. auch in Zukunft so sein wird. Eine vollständige und fehlerfreie Erfassung der Publikationen wird wohl nicht möglich sein.

Die Venue Erkennung bei Computersysteme

Computersysteme Dapp Version 2 (details)

▼ publication

block Wilhelm G. Spruth: The Design of a Microprocessor. Springer-Verlag, Berlin 1989, ISBN 3-540-51395-7. Einzelheiten finden Sie hier

authors Wilhelm G. Spruth

▼ publication

block Udo Kebschull, Paul Herrmann, Wilhelm G: Spruth: Einführung in z/OS und OS/390. Oldenbourg-Verlag, 2002, ISBN 3-486-27214-4. Einzelheiten finden Sie hier

authors Udo Kebschull, Paul Herrmann, Wilhelm G

▼ publication

block U. Kebschull, W. G. Spruth: Kommerzielle Großrechner als Ausbildungsaufgabe an Universitäten und Fachhochschulen. Informatik Spektrum, Band 24, Heft 3, 2001, S. 140-144.

authors U. Kebschull, W. G. Spruth

▼ publication

block Wilhelm G. Spruth, Erhard Rahm: Sysplex-Cluster Technologien für Hochleistungs-Datenbanken. Datenbank-Spektrum, Heft 3, 2002, S. 16-26. download

authors Wilhelm G. Spruth, Erhard Rahm

download download

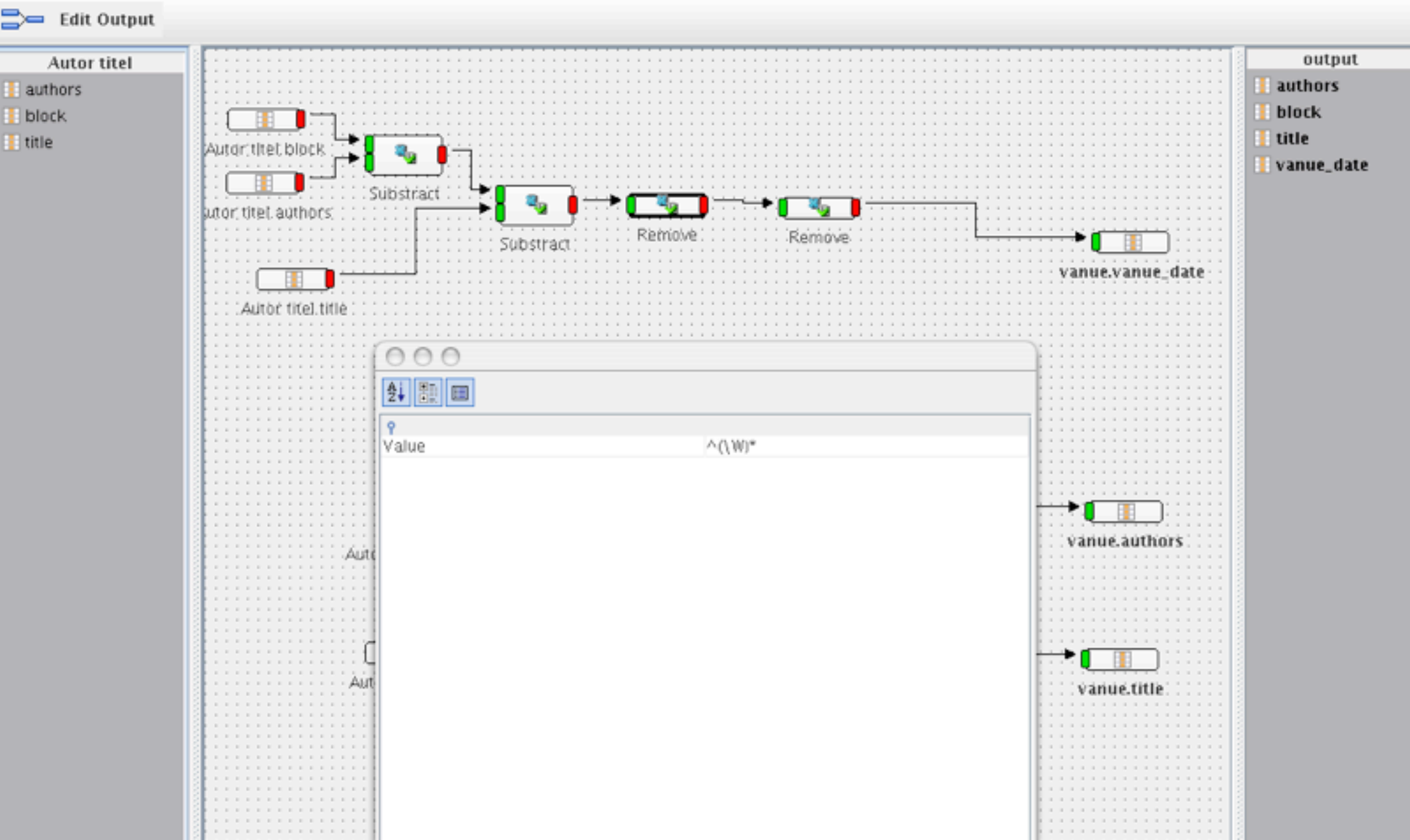
▼ publication

block Joachim Franz, Wilhelm G. Spruth: Reengineering von Kernanwendungssystemen auf Großrechnern. Informatik Spektrum, Band 26, Nr. 2, April 2003, S. 83-93. download Die Originalpublikation ist unter www.springerlink.com verfügbar.

authors Joachim Franz, Wilhelm G. Spruth

download download

Die Venue Erkennung bei Computersysteme



Probleme bei der Datenextraktion

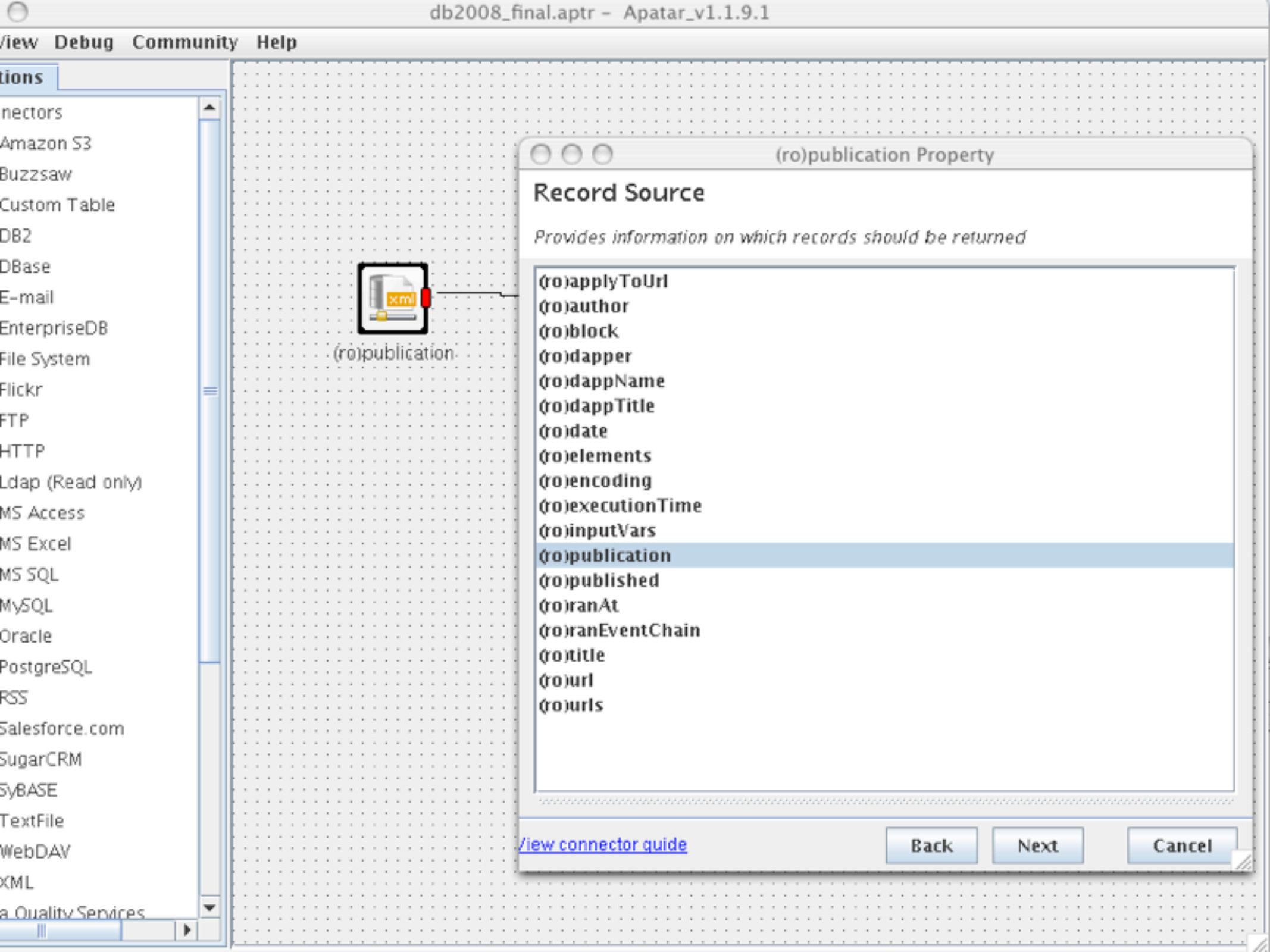
- Ein Dapp für mehrer Seiten gleicher Art (Jahresseiten) erkennt für eine Seite nicht alle Elemente.
- Das Dapp für alte Jahresseiten kann nachbearbeitet werden, so dass alle Elemente erkannt werden.
- Wichtig ist das Dapper alle Elemente des aktuellen Jahres und der kommenden Jahre erkennt.

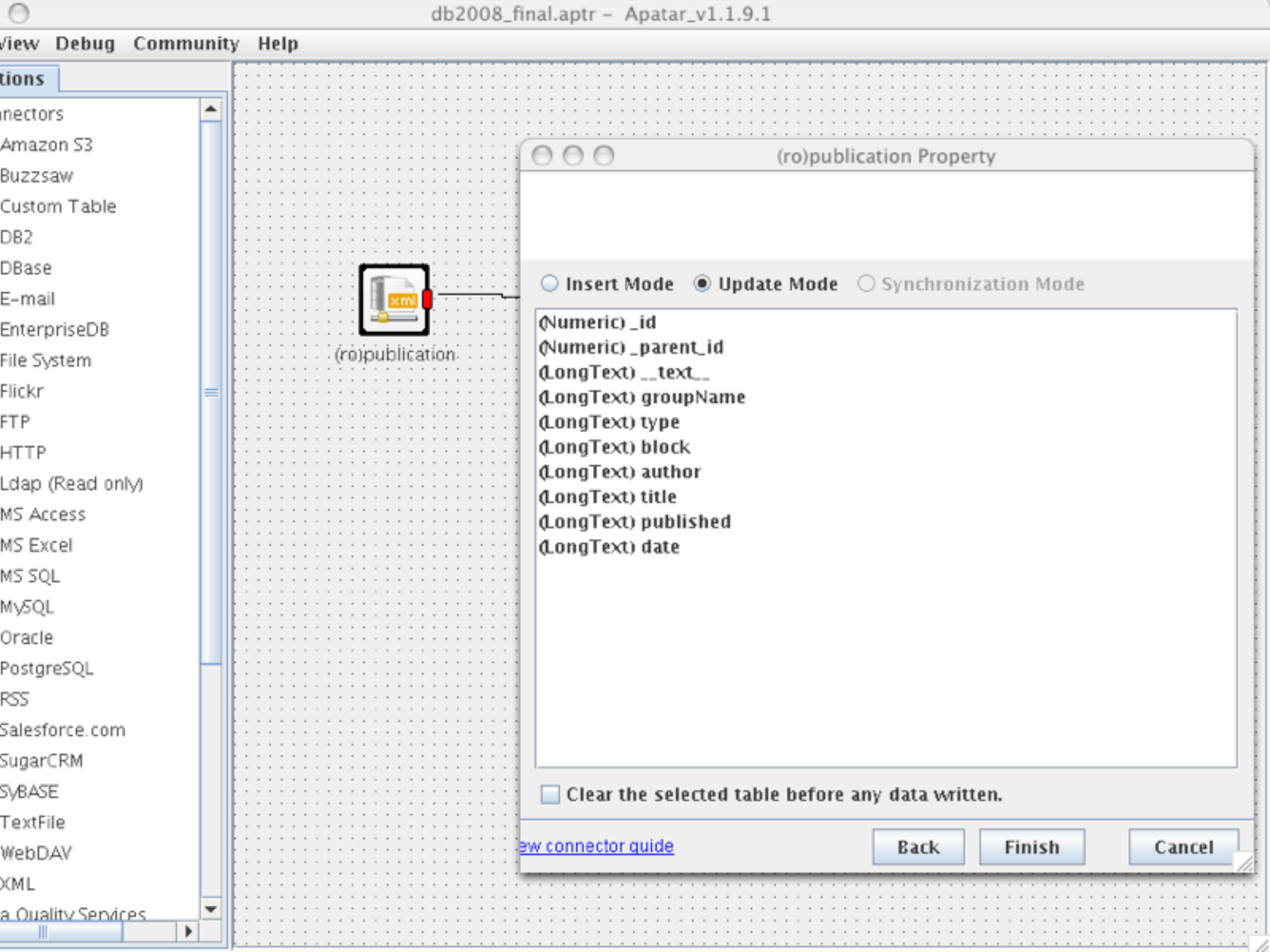
Probleme bei der Datenextraktion

- Dapper hat sich weiterentwickelt. Inzwischen ist der Beta-Tag verschwunden.
- Die “Logik” funktioniert schneller und zuverlässiger.
- Zur Zeit größtes Manko ist, dass sich keine Subgruppen erstellen lassen.

Probleme bei der Datenintegration

- Apatar arbeitet Tabellenbasiert.
- Beim Einlesen muss ein Wurzelelement ausgewählt werden.
- Die Attribute des Wurzelelements und die Kindelemente bilden die Spalten der Tabelle.
- Die Attribute der Kindelemente werden nicht eingelesen.
- Bei mehreren gleichartigen Kindelementen wird nur das letzte eingelesen.


db2008_final.aptr - Apatar_v1.1.9.1
View Debug Community Help
tions
nectors
Amazon S3
Buzzsaw
Custom Table
DB2
DBase
E-mail
EnterpriseDB
File System
Flickr
FTP
HTTP
Ldap (Read only)
MS Access
MS Excel
MS SQL
MySQL
Oracle
PostgreSQL
RSS
Salesforce.com
SugarCRM
SyBASE
TextFile
WebDAV
XML
a Quality Services
(ro)publication
(ro)publication Property
Record Source
Provides information on which records should be returned
(o)applyToUrl
(o)author
(o)block
(o)dapper
(o)dappName
(o)dappTitle
(o)date
(o)elements
(o)encoding
(o)executionTime
(o)inputVars
(o)publication
(o)published
(o)ranAt
(o)ranEventChain
(o)title
(o)url
(o)urls
View connector guide
Back
Next
Cancel


db2008_final.aptr - Apatar_v1.1.9.1
View Debug Community Help
tions
nectors
Amazon S3
Buzzsaw
Custom Table
DB2
DBase
E-mail
EnterpriseDB
File System
Flickr
FTP
HTTP
Ldap (Read only)
MS Access
MS Excel
MS SQL
MySQL
Oracle
PostgreSQL
RSS
Salesforce.com
SugarCRM
SyBASE
TextFile
WebDAV
XML
a Quality Services
xml
(ro)publication
(ro)publication Property
Insert Mode
Update Mode
Synchronization Mode
(Numeric) _id
(Numeric) _parent_id
(LongText) __text__
(LongText) groupName
(LongText) type
(LongText) block
(LongText) author
(LongText) title
(LongText) published
(LongText) date
Clear the selected table before any data written.
ew connector guide
Back
Finish
Cancel

No.	_id	_parent_id	__text__	groupName	type	block	author	title	published	date
	20	1	...	publication	group	Hartung, ...	Hartung, ...	Analyzing ...	Proc. of 5t...	2008-06
	26	1	...	publication	group	Massman...	Massman...	Evaluating...	11th Inter...	2008-06
	32	1	...	publication	group	Hartung, ...	Hartung, M.	Managem...	Tagungsb...	2008-05
	38	1	...	publication	group	Krefting, ...	Krefting, ...	MediGRID...	Future Ge...	2008-05
	44	1	...	publication	group	Massman...	Massman...	Ontologie...	Datenban...	2008-02
	50	1	...	publication	group	Rahm, E. ...	Rahm, E.	Logging a...	Definition...	2008-01
	56	1	...	publication	group	Rahm, E. ...	Rahm, E.	Comparin...	Proc. Aca...	2008

Probleme bei der Datenintegration

- Dapper behandelt eine URL zu einem Downloadlink als Attribut. Diese Information geht leider beim einlesen in Apatar verloren.
- Abhilfe schafft der Join-Operator mit einem Join über die _id der Publikationselemente und der _parent_id der Downloadelemente.
- Leider ist der Join-Operator in Apatar nur ein Equi-Join. Somit gehen alle Datensätze ohne Downloadlink verloren.

Derzeitige Vorgehensweise

1. Alle Datensätze werden ohne URL Link in die Datenbank geladen.
2. In einer zweiten Datamap werden die Publikationen, die einen URL besitzen mit Hilfe der Join-Operation herausgefiltert.
3. An Hand des erzeugten Schlüssels werden die Publikationen in der Datenbank mit den URLs aktualisiert.

- Aufwendig, Unübersichtlich und Fehleranfällig für kleine Änderungen.

Mögliche Lösungen und Verbesserungen

- Da Apatar open source ist bietet es sich an Lösungen für dieses Problem zu programmieren und in Apatar zu integrieren.
- Eine Möglichkeit ist einen outer-join Operator zu programmieren.
- Eine andere Möglichkeit ist den XML-Connector zu verändern, dass er die Attribute der Kindelemente erkennt (Das plane ich im Moment).

Weiter Probleme und Fehler in Apatar

- Apatar ist noch eine Beta-Applikation.
- Zur Zeit funktioniert die Anbindung an die DB2 Datenbank noch nicht.
- Probleme mit manchen Charset Encodings. Deswegen konnte das Dapp für die Abteilung Angewandte Telematik / e-Business noch nicht geladen werden.
- Die Verwendung Regulärer Ausdrücke ist nicht einheitlich.

Ausblick

- Überarbeiten der Datenbank.
- Anpassen von Apatar, so dass Links in einem Schritt hinzugefügt werden können.
- Lösen des Encoding-Problems.
- Erstellen von RSS-Feeds.
- Hinzufügen der noch fehlenden Abteilungen.
- Dapper beobachten, ob es weiterhin sauber Publikationen erkennen kann.
- Schnittstelle / Framework einrichten, um die Publikationen in den Dokumentenserver zu laden.

Diskussion

selbstgeschriebenes Programm - vorgefertigte Tools

- Selbstgeschriebenes Programm
  - \+ gezielt an den Anwendungsfall angepasst.
  - \+ Bessere Kontrolle über die Funktionen und Einblick in die Funktionsweise.
  - \- Ein guter Wrapper ist wahrscheinlich sehr aufwendig zu programmieren.
- vorgefertigte Tools
  - \+ bereits vorgefertigte Framework.
  - \+ die Idee ist aus einer Bewertung solcher Tools entstanden.
  - \+ ausreichend modifizierbar, dank open source.
  - \+ die Tools sind da, um benutzt zu werden.
  - \- die Geschäftsmodelle können sich ändern.
  - \- keine Garantie das die Service immer zu Verfügung stehen.
  - \- keine 100% Anpassung an den Anwendungsfall. Können fehlerhaft sein.

Fragen, Anmerkungen, Kommentare

Ende.