Smartes Harvesten von Literaturdaten

Das DFG-Projekt Smart Harvesting II

Philipp Schaer, Technische Hochschule Köln (University of Applied Sciences), Cologne, Germany

Version: 2018-05-23

Technology
Arts Sciences
TH Köln

Das Projektteam

GESIS

Brigitte Mathiak

Nadine Dulisch

dblp

Michael Ley

Christopher Michels

TH Köln

Philipp Schaer

Mandy Neumann

GESIS

- Leibniz-Institut für Sozialwissenschaften
- Größte Infrastruktureinrichtung für die Sozialwissenschaften in Europa
- Zusammenschluss des
  - Informationszentrums Sozialwissenschaften,
  - des Zentralarchivs für empirische Sozialforschung und
  - des Zentrums für Umfragen, Methoden und Analysen.
- **Usecase für Harvesting**
  - 45.600 Volltexte in **SSOAR**
  - Akquise neuer Volltexte (viele davon von kleineren Verlagen, aber auch Self-Archiving und Kooperationen mit großen Verlagen)

dblp – computer science bibliography

- „Die Personennormdatei für die Informatik“
- Offene Daten für Recherche und Forschung
- Flache (nicht inhaltliche) bibliografische Erschließung und Nachweis qualitativ hochwertiger Metadaten
  - > 4,1 Mio. Publikationen,
  - > 2 Mio. Autoren,
  - > 5.000 Konferenzbände und
  - > 1.500 Journale
- DOIs, ORCIDs, Google Scholar Profile, etc.

dblp computer science bibliography

SCHLOSS DAGSTUHL Leibniz-Zentrum für Informatik

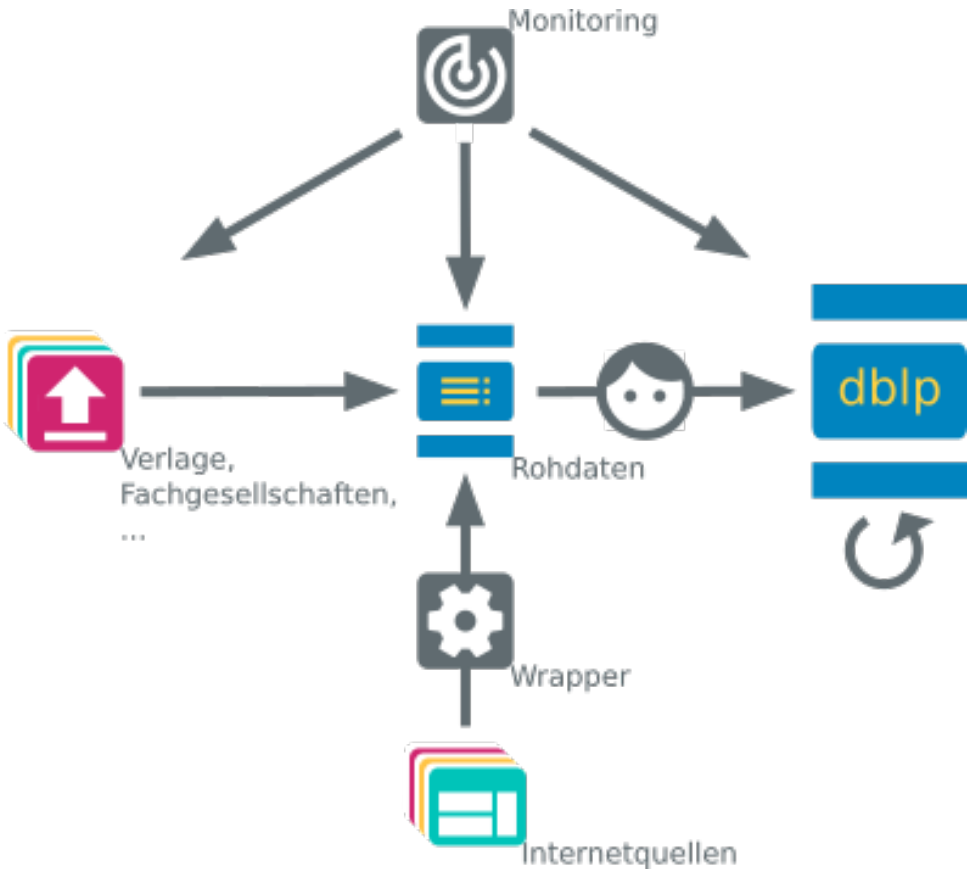
Universität Trier

Technische Hochschule Köln

- Größte Hochschule für ang. Wissenschaften mit über 26.000 Studierenden
- Institut für Informationswissenschaft
  - 3 BA-Studiengänge: **Data and Information Science**, Bibliothek und digitale Kommunikation, Online-Redaktion
  - 2 MA-Studiengänge: Library and Information Science, Markt und Medienforschung
- Professur für Information Retrieval seit 07/2016
  - Forschung: Web Information Extraction, Retrieval Evaluation, Living Labs, digitale Bibliotheken, Bias in Web Search Engines
  - Projektförderungen u.a. durch

DFG Google | DIGITAL NEWS INITIATIVE Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

Datenfluss in dblp



Was ist Web Harvesting?

“Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites.”

WIKIPEDIA
The Free Encyclopedia

Beispiel:


```
@article{bathverantwortung,
title={Verantwortung und
Un/Verf{\"u}gbarkeit},
author={Bath, Corinna and Mei{\ss}ner,
Hanna and Trinkaus, Stephan and
V{\"o}lker, Susanne} }
```

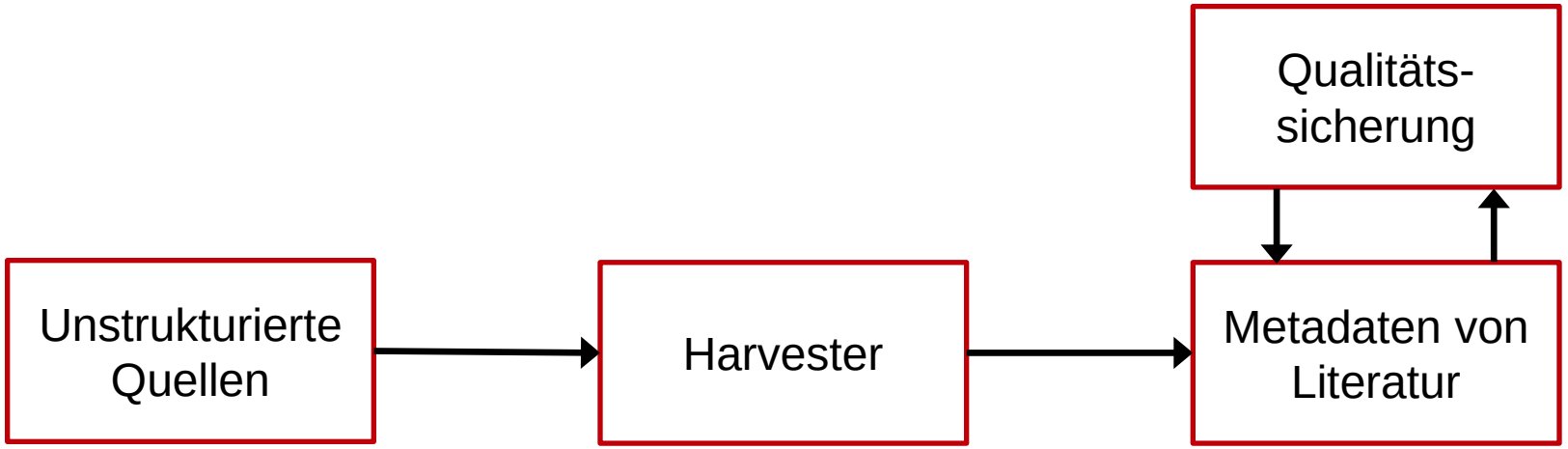
Grundproblem

Wir sind an Quellen interessiert, die

- **nicht durch Schnittstellen**, wie z.B. OAI-PMH, abbildbar sind und
- daran die **dazugehörigen Harvesting-Prozesse** zu verbessern.

Beispiel:

- Ein kleiner Verlag, ein Open Access-Journal oder eine Konferenz möchte die Metadaten teilen, aber verfügt nicht über das Resourcen oder das Know-How strukturierte Daten zu liefern.



dblp vs. GESIS

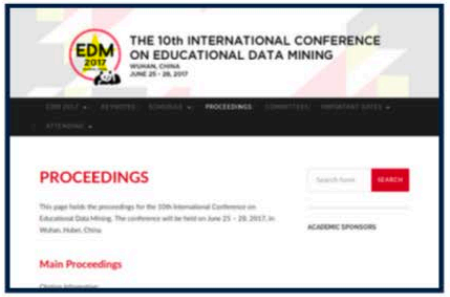
Harvesting für **dblp**

- 130 Wrapper decken etwa 90% der wichtigsten Verlage für dblp ab
- Wrapper basierten auf Java-Code und regulären Ausdrücken
- Große Probleme bei der Erweiterung und der Wartung

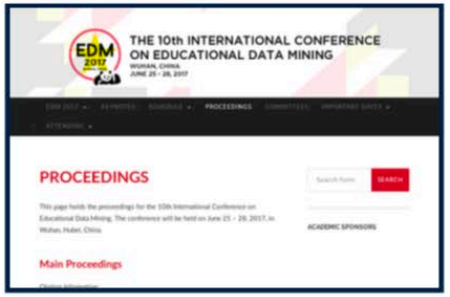
Harvesting für **GESIS**

- Anpassung auf große Verlage für GESIS ist trivial (Springer, etc.)
- Die Anzahl kleiner Verlage ist in den Sozialwissenschaften signifikant höher als in der Informatik
- 34,4% der relevanten Publikationen (Artikel) verteilen sich auf mehr als 1.000 Zeitschriften (2001–2005)

Konferenzen – ein besonderes Problem



Konferenzen – ein besonderes Problem



Wie macht man das nun „Smart“?

„Smarte“ Wrapper (dblp, TH)

- Schwerpunkt der ersten Projektphase, technisch, Java-basiert
- Seit Smart Harvesting II basierend auf **OXPath**
- **Interactive Wrapper**: Bezieht den Faktor Mensch mit ein

Datenqualität (GESIS)

- Autorendisambiguierung
- Linked Open Data Infrastruktur
- Plausibilitätsprüfung
- Entity Recognition

Monitoring (dblp, TH)

- Wie verwaltet man viele 100 Quellen?
- **Scheduling** von Harvesting-Vorgängen

Wie macht man das nun „Smart“?

„Smarte“ Wrapper (dblp, TH)

- Schwerpunkt der ersten Projektphase, technisch, Java-basiert
- Seit Smart Harvesting II basierend auf **OXPath**
- **Interactive Wrapper**: Bezieht den Faktor Mensch mit ein

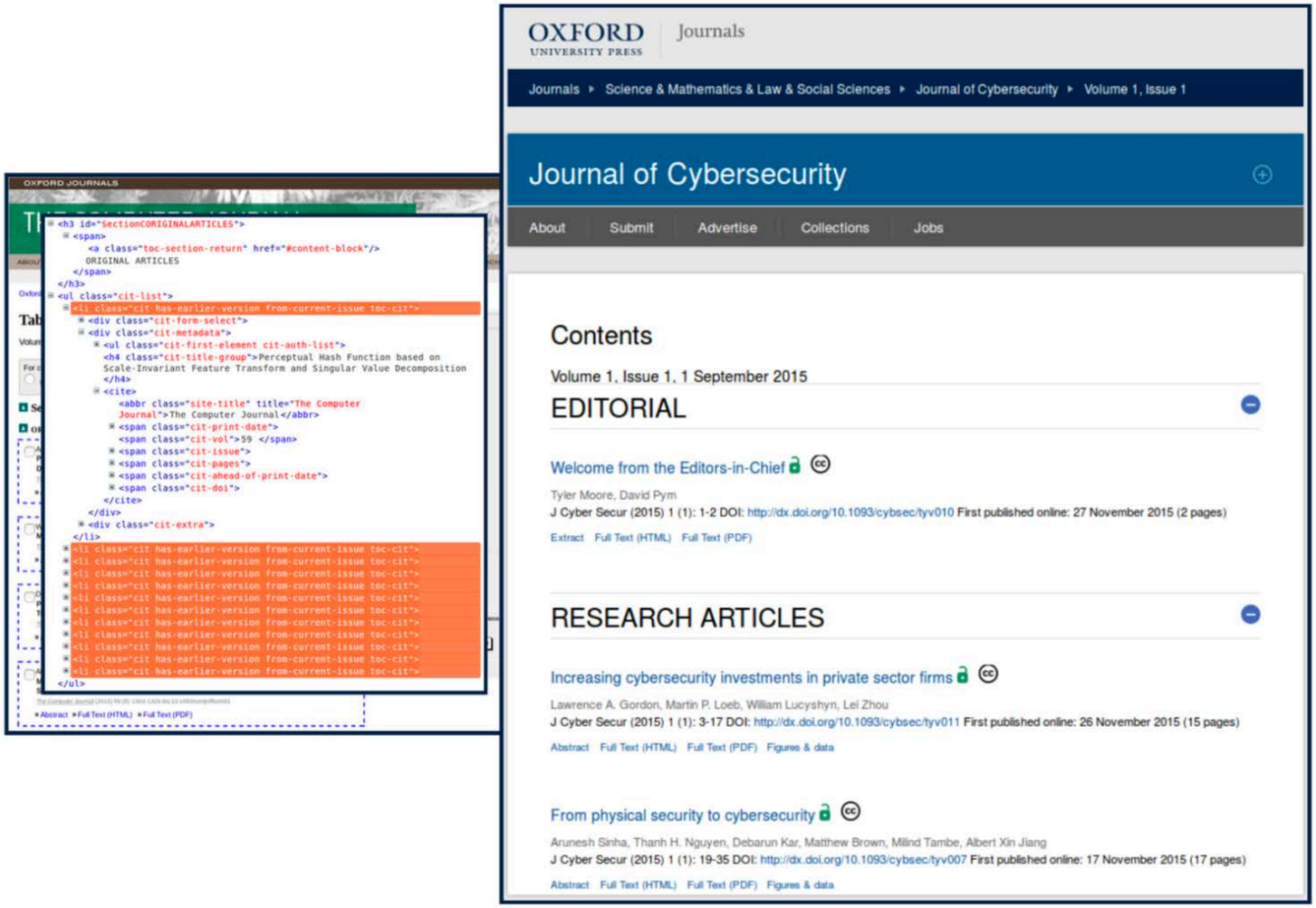
Datenqualität (GESIS)

- Autorendisambiguierung
- Linked Open Data Infrastruktur
- Plausibilitätsprüfung
- Entity Recognition

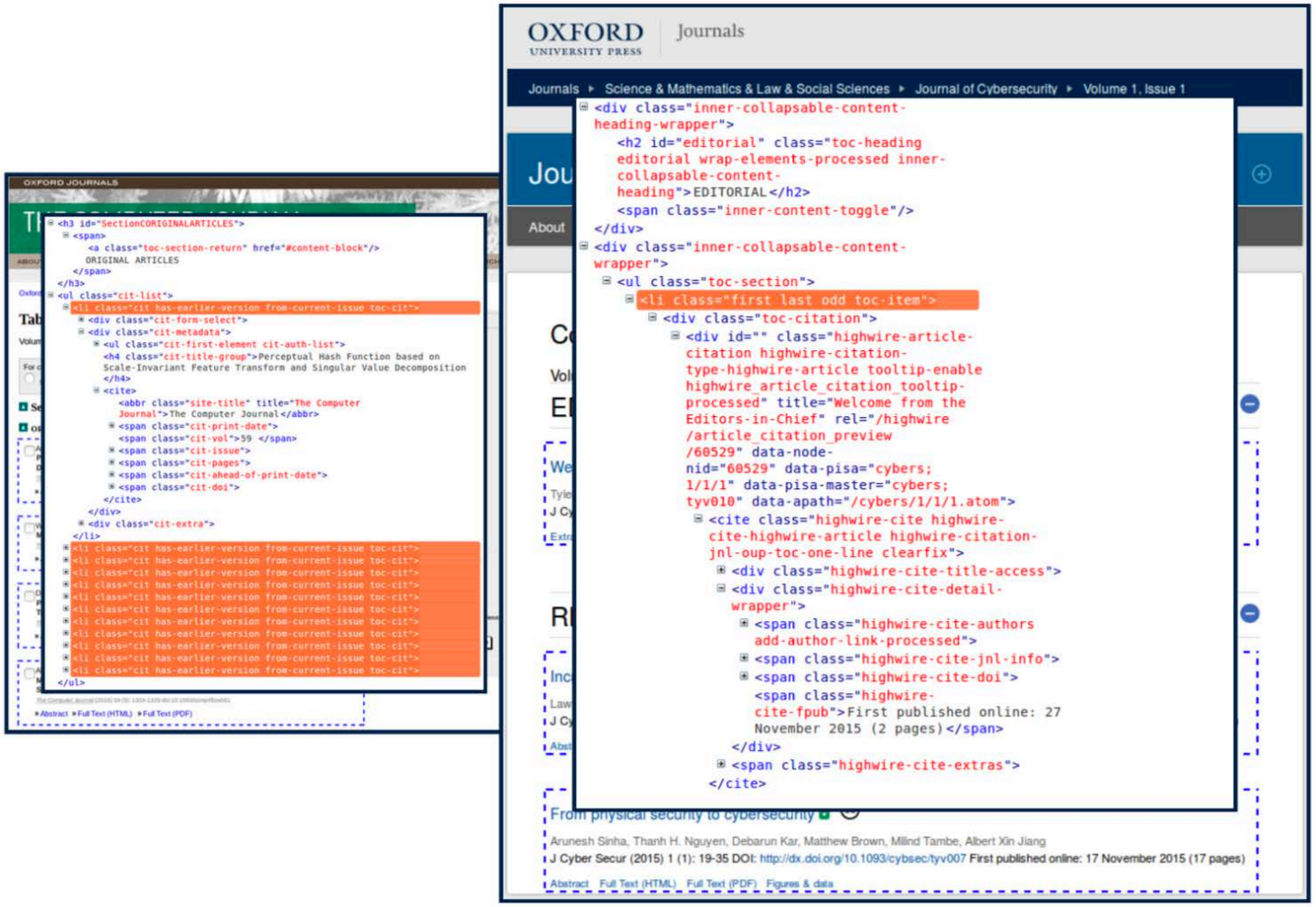
Monitoring (dblp, TH)

- Wie verwaltet man viele 100 Quellen?
- **Scheduling** von Harvesting-Vorgängen

Grundlage für OXPath - XPath

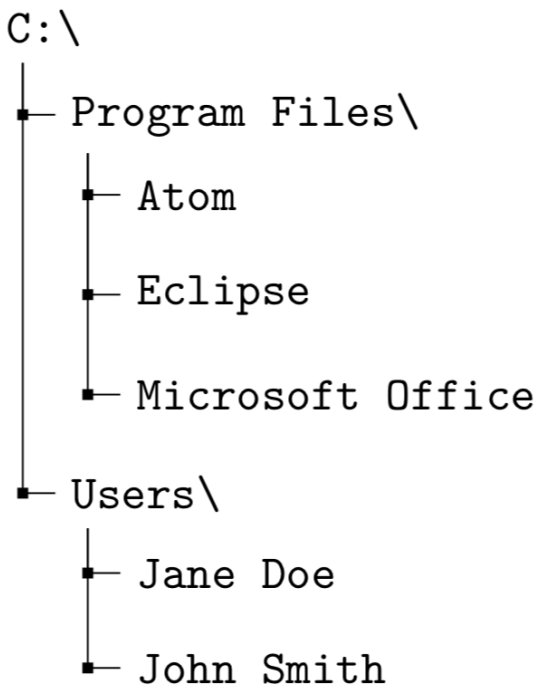


Grundlage für OXPath - XPath



XPath

- Abfragesprache für XML
- XML-Dokument als Baum von Knoten
- XPath-Ausdrücke als Lokalisierungspfade



Dateipfad-Beispiele

```
1 C:\Program Files\Microsoft Office
2 C:\Users\Jane Doe
```

XPath in a Nutshell

XML-Datei

```xml
<?xml version="1.0" encoding="UTF-8"?>
<results>
  <record class="current">
    <volume>30</volume>
    <issue>11</issue>
    <year>2016</year>
    <url>http://.../tadr20/30/11</url>
  </record>
  <record>
    <volume>30</volume>
    <issue>10</issue>
    <year>2016</year>
    <url>http://.../tadr20/30/10</url>
  </record>
  <record>
    <volume>30</volume>
    <issue>9</issue>
    <year>2016</year>
    <url>http://.../tadr20/30/9</url>
  </record>
</results>
```

XPath Ausdruck

```
/results/record[@class="current"]
```

Ergebnismenge

```xml
(
  <record class="current">
    <volume>30</volume>
    <issue>11</issue>
    <year>2016</year>
    <url>[...]</url>
  </record>
)
```

Was fügt OXPath hinzu?

Aktionen:

- Ausfüllen von Formularfeldern
- Klicks auf Links, Buttons etc.

Extraktion:

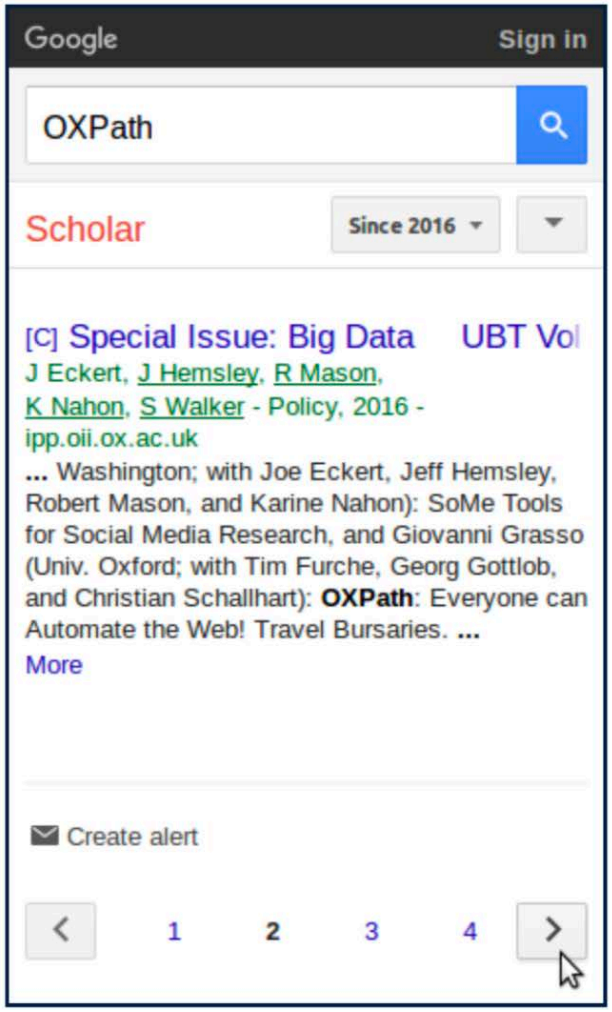
- Extraktionsmarker an ausgewählten Knoten
- Funktionen zur Manipulation der zu extrahierenden Daten

Iteration:

- Schleifen, z.B. für Paginierung

XPath	OXPath
Statisches Web	Dynamisches Web
Pures HTML	AJAX
Kompletter Inhalt	Content on demand

OXPath-Beispiel



OXPath-Ausdruck

```
1 doc('https://scholar.google.com')
2   //input[@id='gs_hdr_tsi']/{"OXPath"}
3   /../following-sibling::button/{click/}
4     //*[@id='gs_res_ab_yy-b']/{click}
5       //following::*[@role='menuitemradio'][contains(.,
            '2016')]/{click/}
6     /(//*[@id='gs_nm']/button[2][not(@disabled)]/{click/})*
7       //div[@class='gs_ri']//h3/a:<title=string(.)>
```

XML-Ausgabe

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3   <title>Tim Furche, Georg Gottlob, [...]</title>
4   <title>Special Issue: Big Data [...]</title>
5   <!--[...]-->
6 </results>
```

Beispiele für die praktische Nutzung

Umstellung der dblp auf OXPath

- Michels, Fayzrakhmanov, Ley, Sallinger und Schenkel (2017): OXPath-Based Data Acquisition for dblp. JCDL. ACM.

OXPath für Bibliothekare

code{4}lib JOURNAL

- Neumann, Steinberg und Schaer (2017): Web-Scraping for Non-Programmers: Introducing OXPath for Digital Library Metadata Harvesting. In: Code4Lib Journal (38).

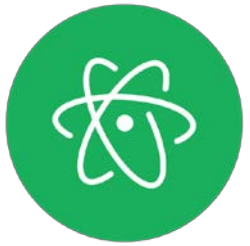
Erweiterungen von Retrieval-Testkollektionen

- Schaer und Neumann (2017): Enriching Existing Test Collections with OXPath. CLEF. Springer.

Toolbox rund um OXPath

Im Rahmen des Projektes wurde eine Reihe von Tools entwickelt um die Arbeit mit OXPath zu vereinfachen.

Atom-Modul

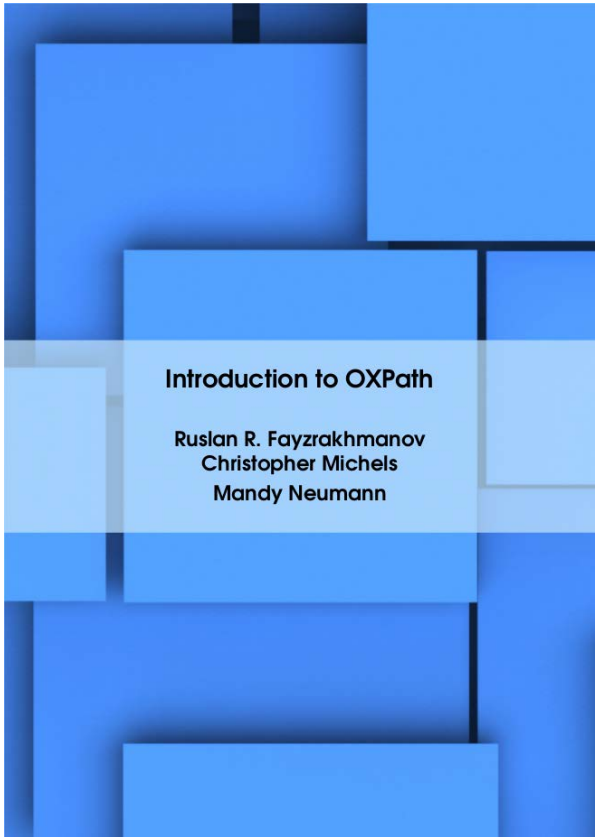


- Syntax-Hervorhebung für Schlüsselwörter
- Für verbesserte Fehlererkennung und Lesbarkeit
- Soll Einstiegshürden mindern

Docker-Container

- Ursprüngliches OXPath nur unter Ubuntu
- Durch Docker auch unter Windows/Mac
- Alle Abhängigkeiten in Container erfüllt

OXPath – The Missing Manual



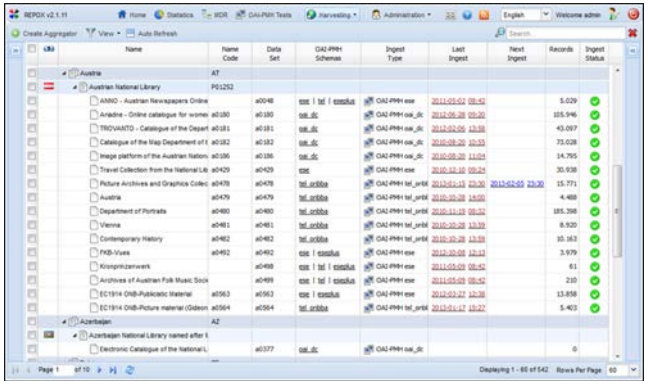
- Unterstützt durch Teile des ursprünglichen Entwicklungsteams von OXPath aus Oxford
- Enthält:
  - eine Zusammenfassung zu XPath
  - Einrichtungs- und Nutzungsanweisungen für OXPath
  - Liste aller verfügbaren Action-Schlüsselwörter
  - Liste aller Funktionen für Extraktion und DOM-Navigation
  - **Einstiegsbeispiele aus der bibliographischen Domäne**

http://www.oxpath.org/papers/2017-IntroductionToOxpath-ed1.pdf

Monitoring

Harvesting „en gros“ denken!

- Im Zweifelsfalle werden viele 100 Quellen und dazugehörige Wrapper verwendet.
- Im OAI-Umfeld gibt es Tools wie z.B. REPOX.

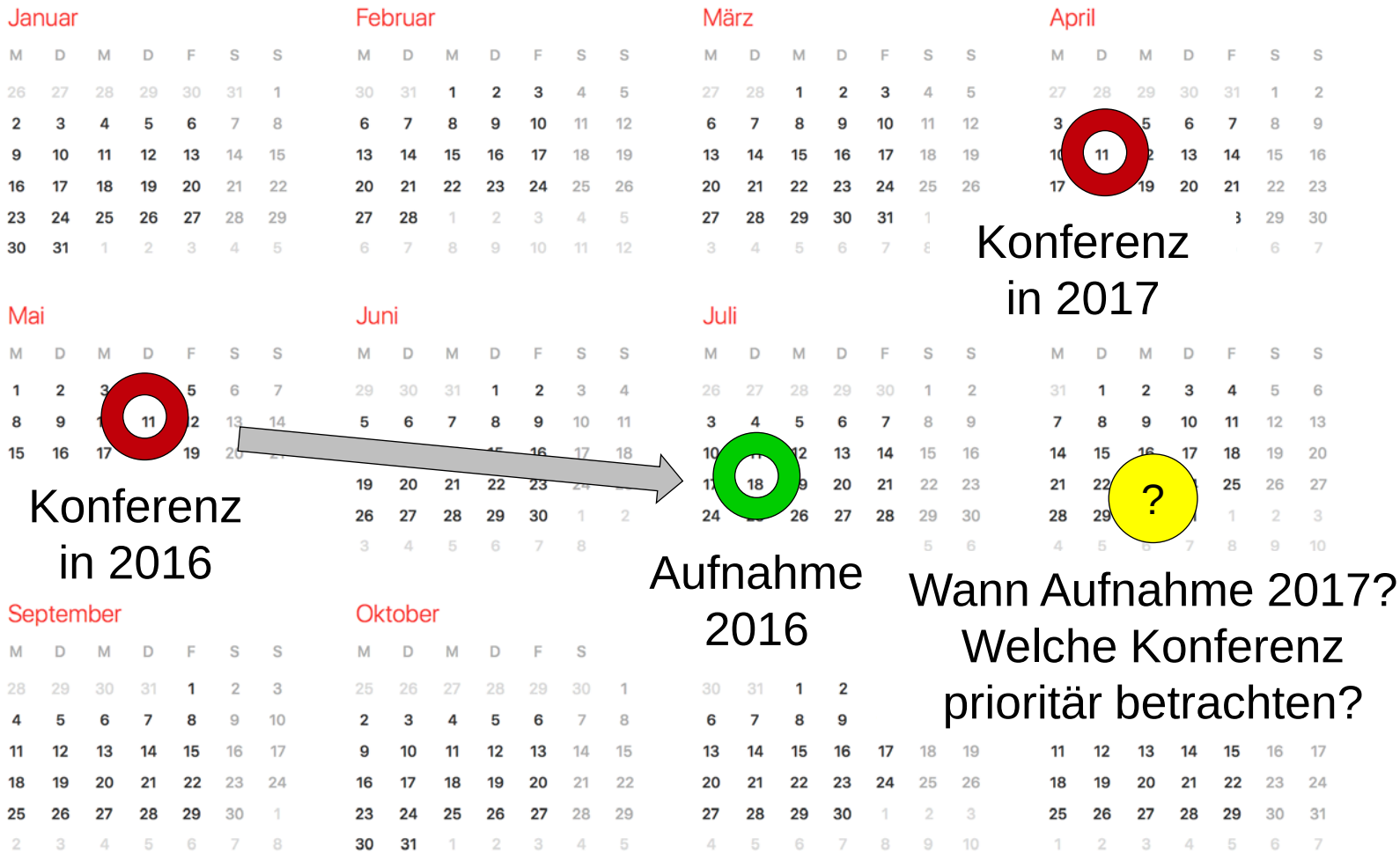


Allerdings:

- Jede Nacht jede Quelle anfragen ist sinnlos, da viele Quellen (z.B. Konferenzen) nur jährlich erscheinen.
- Auch die Wartung der Wrapper sollte nur stattfinden, wenn nötig.

Gebraucht werden daher **„smarte“ Monitoring-Ansätze**.

Szenario: Aufnahme von Konferenzen



Ranking von Harvesting-Kandidaten

Experiment in dblp: Wie können wir alle Konferenzen so ranken, dass die **für Neuaufnahme dringlichsten** ganz oben stehen?

Datensets

- Historische dblp Daten
  - Datum der Aufnahme einer Konferenz über Jahre hinweg
  - Ort einer Konferenz
  - Ko-Autorenschaften
- Microsoft Academic Graph
  - Zitationsraten
- CORE Konferenz-Ratings

Merkmale für das Ranking

Faktoren zur Bestimmung der „Dringlichkeit“

- $\Delta(c)$ Erwartetes nächstes Auftreten
- w_{delay} Maß für “Überfälligkeit“
- w_r Rating der Konferenz
- w_{i} Internationalität der Konferenzen
- w_{d} Wahrscheinlichkeit der Diskontinuität
- w_{c} Zitationshäufigkeit
- w_{prm} Autorenprominenz basierend auf Ko-Autorenschaften

c	$\Delta(c)$	w_{delay}	w_r	w_i	w_d	w_{cit}	w_{prm}
jcdl	3	4	1.88	1.192	0.250	1.029	1.312
tpdl	0	4	1.63	1.577	0.250	1.024	1.352
icadl	0	4	1.75	1.385	0.250	1.009	1.347
dl	146	1	1.00	1.039	0.004	1.091	1.445

Evaluation mittels Pseudorelevanz

Evaluationsjahr

- Sliding window über alle Monate von 2016

Vergleichsdaten

- “Goldstandard” definiert mittels intervallbasierter Pseudorelevanz
- Berechnung von NDCG für jeden Monat
- Vergleich des Einflusses der verschiedenen Merkmale
- Vergleich der **Baseline** (Ranking nur nach Delay)
  + **einem weiteren Gewichtungsfaktor**

Ranking von Harvesting-Kandidaten

system	ndcg-10	ndcg-20	ndcg-100	ndcg-200
baseline	0.530	0.545	0.505	0.439
conf. rating	0.739**	0.716**	0.645***	0.597***
internationality	0.616	0.632	0.608***	0.575***
discontinued	0.713**	0.686***	0.643***	0.594***
citations	0.588	0.575	0.554***	0.548***
prominence	0.681**	0.662**	0.608***	0.577***

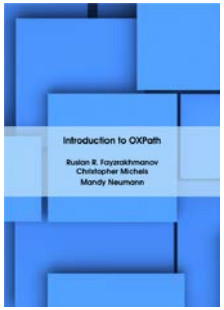
Ergebnisse in Neumann et al. (2018) - JCDL 2018

- Preprint: https://arxiv.org/abs/1804.06169

Vielen Dank! Gibt es Fragen?

Save the Date

- OXPath-Hands-on-Lab und Vortrag
- **13.6. und 15.6.2018 @ Bibliothekartag 2018**



OXPath-Tutorial

- http://www.oxpath.org/papers/2017-IntroductionToOxpath-ed1.pdf

Try it out

- OXPath als **Docker-Container** https://github.com/irgroup/oxpath_docker
- Syntax-Modul für **Atom** https://atom.io/packages/language-oxpath