



# Webscraping für die Metadatengewinnung

Das DFG-Projekt Smart Harvesting II

# Das Projektteam

**GESIS**



Brigitte Mathiak

**dblp**



Michael Ley

**TH Köln**



Philipp Schaer



Nadine Dulisch



Christopher Michels



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.
- **Use Case für Harvesting**
  - 46.000 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,
  - > 5.400 Konferenzbände und
  - > 1.500 Journale
- DOIs, ORCIDs, Google Scholar Profile, etc.



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 (P. Schaer):
  - Forschung: Web Information Extraction, Retrieval Evaluation, Living Labs, digitale Bibliotheken, Bias in Web Search Engines
  - Projektförderungen u.a. durch



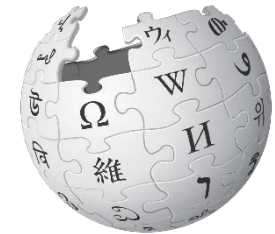
Ministerium für  
Kultur und Wissenschaft  
des Landes Nordrhein-Westfalen



# Was ist Web Harvesting?

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

→ **automatisch** und **gezielt**



**WIKIPEDIA**  
The Free Encyclopedia

## Neuerscheinungen



Corinna Bath, Hanna Meißner, Stephan Trinkaus, Susanne Völker (Hrsg.)

**Verantwortung und Un/Verfügbarkeit**

2017 - 259 Seiten - 30,00 €  
ISBN: 978-3-89691-248-0

zum Inhalt

```
@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} }
```



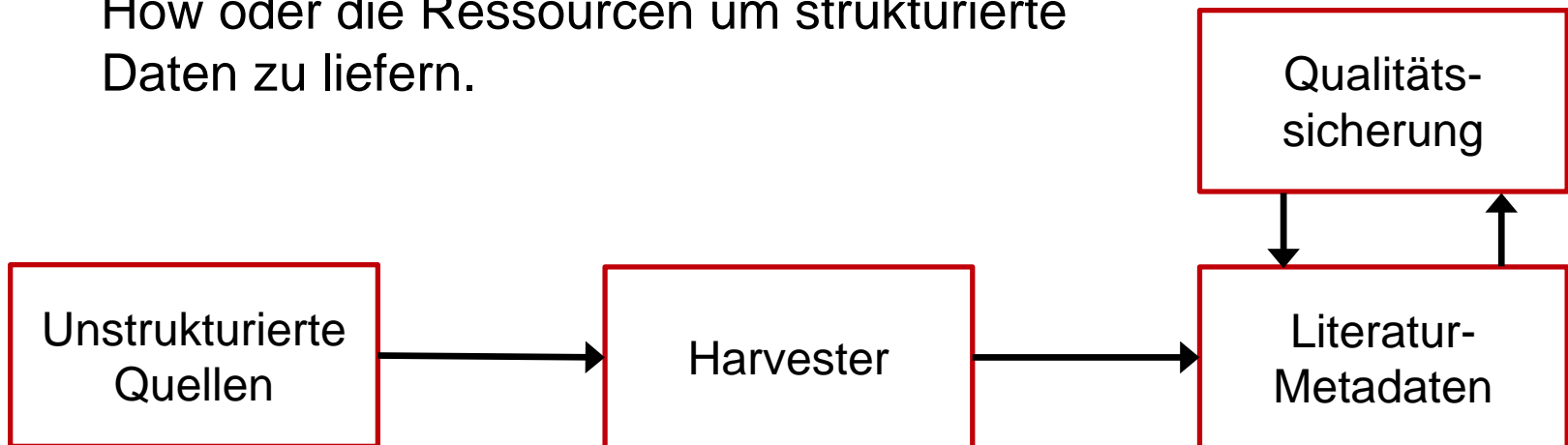
# Grundproblem

Wir sind an Quellen interessiert, die

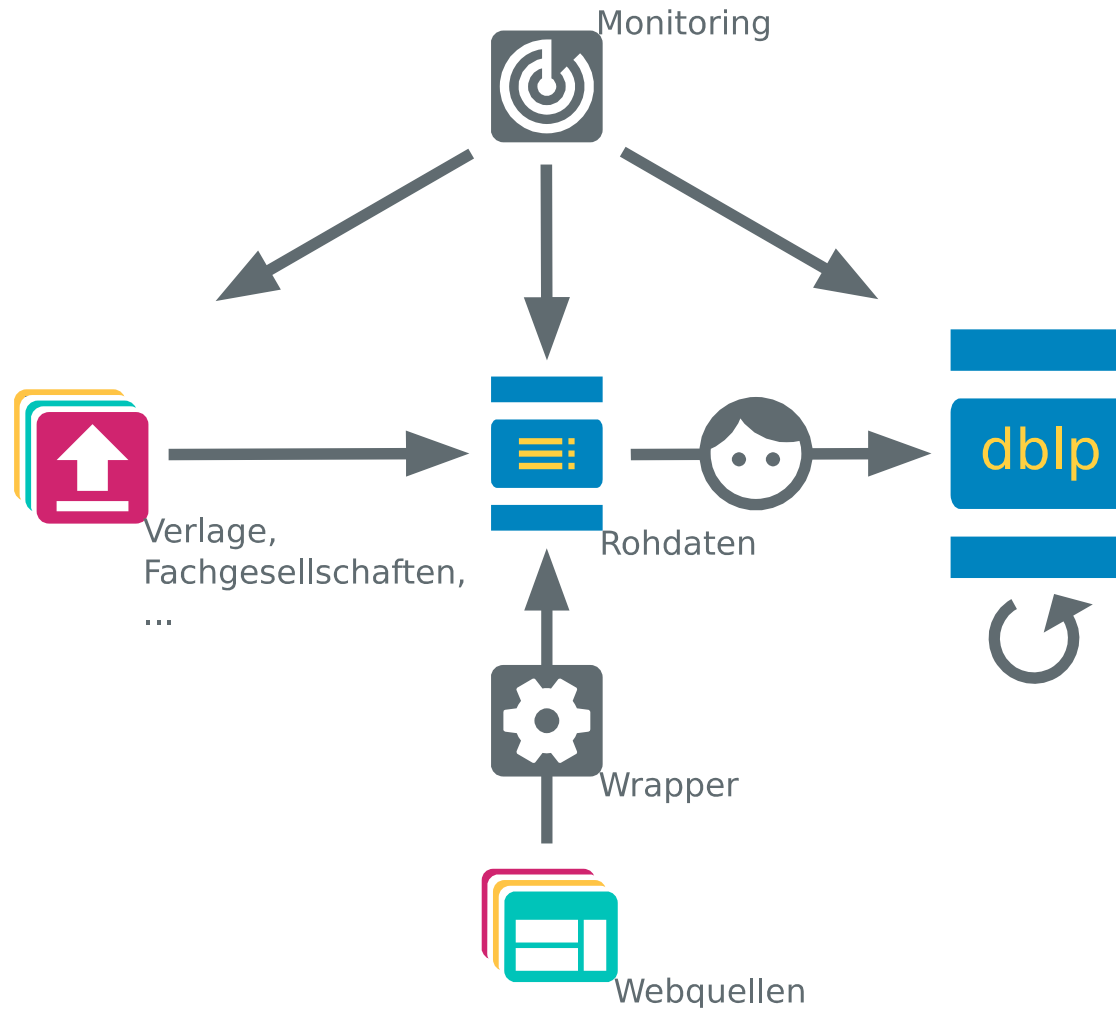
- **nicht durch Schnittstellen**, wie z.B. OAI-PMH, abbildbar sind, sowie daran
- die **dazugehörigen Harvesting-Prozesse** zu verbessern.

Beispiel:

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



# Datenfluss in dblp





# 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)

# Wie macht man das nun „Smart“?

„Smarte“ Wrapper (dblp, TH):

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

Datenqualität (GESIS):

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

Monitoring (dblp, TH):

- Wie verwaltet man viele 1000 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 **XPath**
- **Interactive Wrapper**: Bezieht den Faktor Mensch mit ein

Datenqualität (GESIS):

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

Monitoring (dblp, TH):

- Wie verwaltet man viele 1000 Quellen?
- **Scheduling** von Harvesting-Vorgängen

# Grundlage für XPath - XPath

OXFORD JOURNALS

## THE COMPUTER JOURNAL

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Science & Mathematics > Computer Journal > Volume 59 Issue 9

### Table of Contents

Volume 59 Issue 9 September 2016

For checked items  
☐ view abstracts ☐ download to citation manager

#### Section C

##### ORIGINAL ARTICLES

☐ Arambam Neelima and Kh Manglem Singh  
**Perceptual Hash Function based on Scale-Invariant Feature Transform and Singular Value Decomposition**  
*The Computer Journal* (2016) 59 (9): 1275-1281 doi:10.1093/comjnl/bw079  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

☐ Wei Ni  
**Minimized Error Propagation Location Method Based on Error Estimation**  
*The Computer Journal* (2016) 59 (9): 1282-1288 doi:10.1093/comjnl/bw081  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

☐ D. Thenmozhi and Chandrabose Aravindan  
**Paraphrase Identification by Using Clause-Based Similarity Features and Machine Translation Metrics**  
*The Computer Journal* (2016) 59 (9): 1289-1302 doi:10.1093/comjnl/bw083  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

☐ Alok Kumar Singh Kushwaha and Rajeev Srivastava  
**Maritime Object Segmentation Using Dynamic Background Modeling and Shadow Suppression**  
*The Computer Journal* (2016) 59 (9): 1303-1329 doi:10.1093/comjnl/bw091  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

« Previous | Next »

**This Issue**  
 September 2016 59 (9)



» Index By Author  
 » Front Matter (PDF)  
 » Table of Contents (PDF)  
 » Back Matter (PDF)

» Section C

» ORIGINAL ARTICLES

Find articles in this issue containing these words:

[Advance Access](#)

# Grundlage für XPath - XPath

OXFORD JOURNALS

## THE COMPUTER JOURNAL

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Oxford Journals > Science & Mathematics > Computer Journal > Volume 59 Issue 9

### Table of Contents

Volume 59 Issue 9 September 2016

For checked items  
☐ view abstracts ☐ download to citation manager

#### Section C

##### ORIGINAL ARTICLES

☐ Arambam Neelima and Kh Manglem Singh  
**Perceptual Hash Function based on Scale-Invariant Feature Transform and Singular Value Decomposition**  
*The Computer Journal* (2016) 59 (9): 1275-1281 doi:10.1093/comjnl/bxw079  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

☐ Wei Ni  
**Minimized Error Propagation Location Method Based on Error Estimation**  
*The Computer Journal* (2016) 59 (9): 1282-1288 doi:10.1093/comjnl/bxw081  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

☐ D. Thenmozhi and Chandrabose Aravindan  
**Paraphrase Identification by Using Clause-Based Similarity Features and Machine Translation Metrics**  
*The Computer Journal* (2016) 59 (9): 1289-1302 doi:10.1093/comjnl/bxw083  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

☐ Alok Kumar Singh Kushwaha and Rajeev Srivastava  
**Maritime Object Segmentation Using Dynamic Background Modeling and Shadow Suppression**  
*The Computer Journal* (2016) 59 (9): 1303-1329 doi:10.1093/comjnl/bxw091  
[» Abstract](#) [» Full Text \(HTML\)](#) [» Full Text \(PDF\)](#)

« Previous | Next »

**This Issue**  
September 2016 59 (9)



» Index By Author  
 » Front Matter (PDF)  
 » Table of Contents (PDF)  
 » Back Matter (PDF)

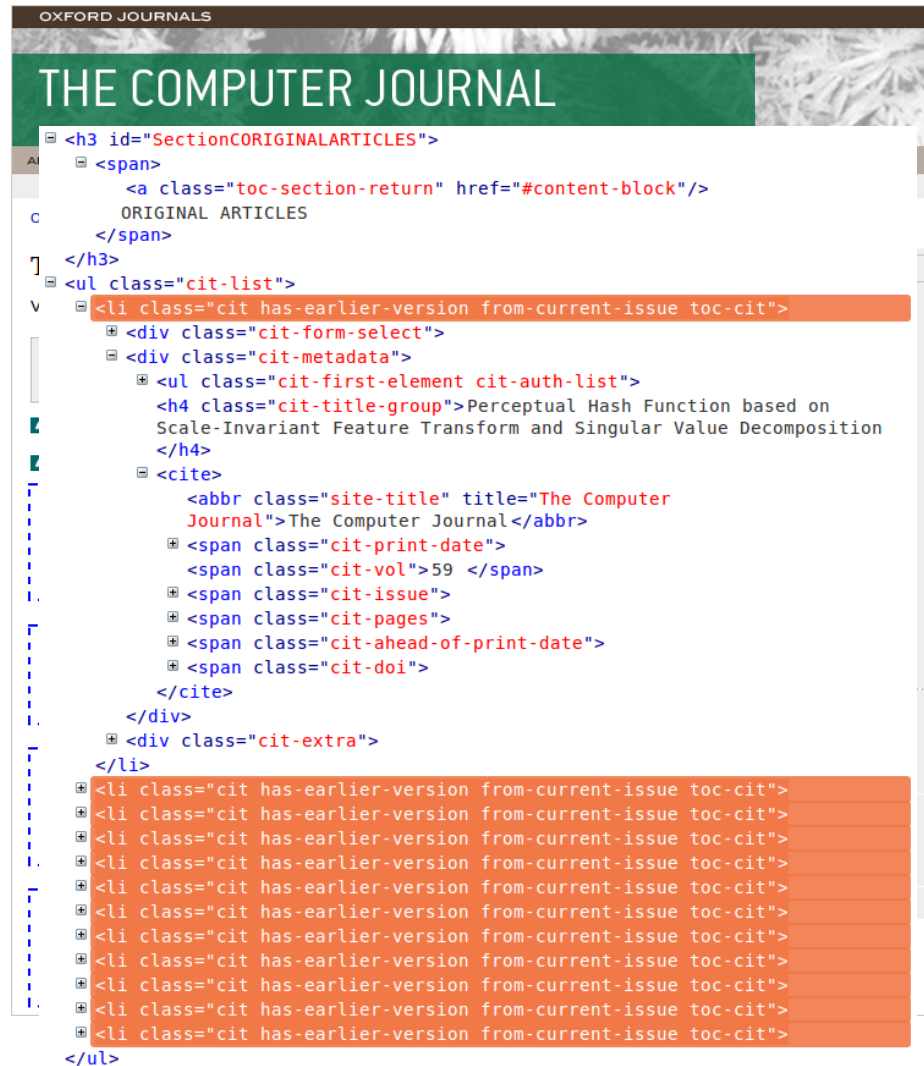
» Section C

» ORIGINAL ARTICLES

Find articles in this issue containing these words:

Advance Access

# Grundlage für XPath - XPath



```

OXFORD JOURNALS

THE COMPUTER JOURNAL

<h3 id="SectionCORIGINALARTICLES">
  <span>
    <a class="toc-section-return" href="#content-block"/>
      ORIGINAL ARTICLES
    </span>
  </h3>
  <ul class="cit-list">
    <li class="cit has-earlier-version from-current-issue toc-cit">
      <div class="cit-form-select">
        <div class="cit-metadata">
          <ul class="cit-first-element cit-auth-list">
            <h4 class="cit-title-group">Perceptual Hash Function based on
              Scale-Invariant Feature Transform and Singular Value Decomposition
            </h4>
            <cite>
              <abbr class="site-title" title="The Computer
                Journal">The Computer Journal</abbr>
              <span class="cit-print-date">
                <span class="cit-vol">59 </span>
                <span class="cit-issue">
                  <span class="cit-pages">
                    <span class="cit-ahead-of-print-date">
                      <span class="cit-doi">

```

# XPath

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

```
C:\  
├─ Program Files\  
│   ├── Atom  
│   ├── Eclipse  
│   └─ Microsoft Office  
└─ Users\  
    ├── Jane Doe  
    └─ John Smith
```

## Dateipfad-Beispiele

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

# XPath in a Nutshell

## XML-Datei

```

1 <?xml version="1.0" encoding="UTF-8"?>
2 <results>
3   <record class="current">
4     <volume>30</volume>
5     <issue>11</issue>
6     <year>2016</year>
7     <url>http://.../tadr20/30/11</url>
8   </record>
9   <record>
10    <volume>30</volume>
11    <issue>10</issue>
12    <year>2016</year>
13    <url>http://.../tadr20/30/10</url>
14  </record>
15  <record>
16    <volume>30</volume>
17    <issue>9</issue>
18    <year>2016</year>
19    <url>http://.../tadr20/30/9</url>
20  </record>
21 </results>

```

## XPath Ausdruck

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

## Ergebnismenge

```

1 (
2   <record class="current">
3     <volume>30</volume>
4     <issue>11</issue>
5     <year>2016</year>
6     <url>[...]</url>
7   </record>
8 )

```



# 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

# XPath-Beispiel

The screenshot shows a Google Scholar search result for the query "XPath". The search bar at the top contains "XPath" and a magnifying glass icon. Below the search bar, the "Scholar" logo is visible, along with a filter "Since 2016" and a dropdown arrow. The search results list a paper titled "[C] Special Issue: Big Data UBT Vol" by J Eckert, J Hemsley, R Mason, K Nahon, and S Walker. The abstract mentions "SoMe Tools for Social Media Research" and "XPath: Everyone can Automate the Web!". At the bottom, there is a "Create alert" button and a pagination bar with numbers 1, 2, 3, 4 and navigation arrows.

## XPath-Ausdruck

```

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

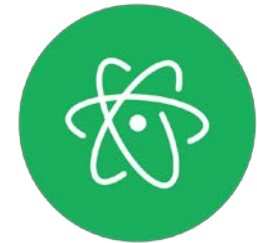
```

# Toolbox rund um XPath

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

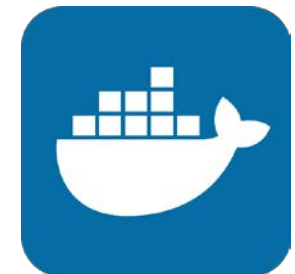
## Atom-Modul

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

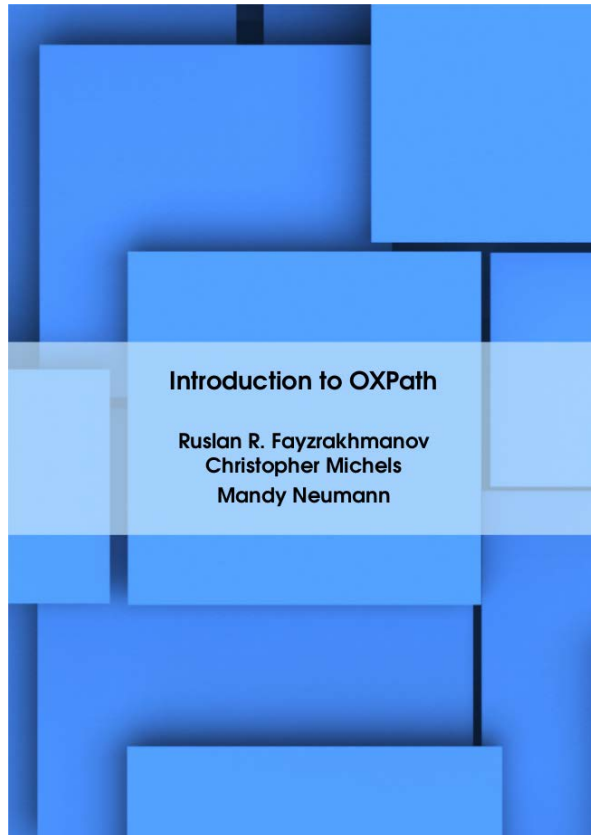


## Docker-Container

- Ursprüngliches XPath nur unter Linux
- Durch Docker auch unter Windows/Mac
- Alle Abhängigkeiten in Container erfüllt



# XPath – The Missing Manual



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

<http://www.xpath.org/papers/2017-IntroductionToXPath-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.

Name	Name Code	Data Set	OAI-PMH Schema	Ingest Type	Last Ingest	Next Ingest	Records	Ingest Status
Austrian National Library	AT	PS1252						
ANNO - Austrian Newspapers Online	a0046	ess   set   essda	OAI-PMH-ess		2012-03-01 08:51		5,029	✓
Arxiphe - Online catalogue for women	a0180	ess_da	OAI-PMH-ess_da		2012-03-28 09:20		85,946	✓
Technische - Catalogue of the Depot	a0181	ess_da	OAI-PMH-ess_da		2012-03-28 12:18		43,587	✓
Catalogue of the Top Department of L	a0182	ess_da	OAI-PMH-ess_da		2012-03-28 12:23		73,028	✓
Image platform of the Austrian Nation	a0186	ess_da	OAI-PMH-ess_da		2012-03-28 12:24		34,765	✓
Travel Collection from the National Lib	a0429	ess	OAI-PMH-ess		2012-03-28 09:24		30,928	✓
Picture Archives and Graphics Collec	a0478	set_daiba	OAI-PMH-set_daiba		2012-03-15 13:30	2012-04-05 13:30	15,771	✓
Austria	a0479	set_daiba	OAI-PMH-set_daiba		2012-03-28 14:50		4,488	✓
Department of Portraits	a0480	set_daiba	OAI-PMH-set_daiba		2012-03-28 09:24		85,268	✓
Vienna	a0481	set_daiba	OAI-PMH-set_daiba		2012-03-28 12:23		8,920	✓
Contemporary History	a0482	set_daiba	OAI-PMH-set_daiba		2012-03-28 12:23		89,343	✓
PDB-Vienn	a0482	ess   setda	OAI-PMH-ess		2012-03-28 12:23		5,979	✓
Kronprinzwerk	a0489	ess   set   setda	OAI-PMH-ess		2012-03-28 09:24		61	✓
Archives of Austrian Folk Music Soc	a0489	ess   set   setda	OAI-PMH-ess		2012-03-28 09:24		210	✓
EC1914 ONB-Publicist Material	a0563	ess   setda	OAI-PMH-ess		2012-03-27 12:38		13,858	✓
EC1914 ONB-Picture material (Gibson	a0564	set_daiba	OAI-PMH-set_daiba		2012-03-17 12:27		5,403	✓
Austrian National Library named after L	a0377	ess_da	OAI-PMH-ess_da				0	✓

## Allerdings:

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

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

# Szenario: Aufnahme von Konferenzen

Januar

M	D	M	D	F	S	S
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Februar

M	D	M	D	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	1	2	3	4	5
6	7	8	9	10	11	12

März

M	D	M	D	F	S	S
27	28	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

April

M	D	M	D	F	S	S
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7

Mai

M	D	M	D	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21

Konferenz  
in 2015

Juni

M	D	M	D	F	S	S
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

Juli

M	D	M	D	F	S	S
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

August

M	D	M	D	F	S	S
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

September

M	D	M	D	F	S	S
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

Oktober

M	D	M	D	F	S	S
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

November

M	D	M	D	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3
4	5	6	7	8	9	10

Dezember

M	D	M	D	F	S	S
27	28	29	30	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31
1	2	3	4	5	6	7

# Szenario: Aufnahme von Konferenzen

Januar

M	D	M	D	F	S	S
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Februar

M	D	M	D	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	1	2	3	4	5
6	7	8	9	10	11	12

März

M	D	M	D	F	S	S
27	28	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

April

M	D	M	D	F	S	S
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7

Mai

M	D	M	D	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21

Konferenz  
in 2015

Juni

M	D	M	D	F	S	S
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

Juli

M	D	M	D	F	S	S
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Aufnahme  
2015

August

M	D	M	D	F	S	S
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

September

M	D	M	D	F	S	S
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

Oktober

M	D	M	D	F	S	S
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

M	D	M	D	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	1	2	3
4	5	6	7	8	9	10

Dezember

M	D	M	D	F	S	S
27	28	29	30	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31
1	2	3	4	5	6	7

# Szenario: Aufnahme von Konferenzen

Januar

M	D	M	D	F	S	S
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Februar

M	D	M	D	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	1	2	3	4	5
6	7	8	9	10	11	12

März

M	D	M	D	F	S	S
27	28	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

April

M	D	M	D	F	S	S
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Konferenz  
in 2016

Mai

M	D	M	D	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21

Konferenz  
in 2015

Juni

M	D	M	D	F	S	S
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

Juli

M	D	M	D	F	S	S
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Aufnahme  
2015

M	D	M	D	F	S	S
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

September

M	D	M	D	F	S	S
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

Oktober

M	D	M	D	F	S	S
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Dezember

M	D	M	D	F	S	S
27	28	29	30	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31
1	2	3	4	5	6	7



# Szenario: Aufnahme von Konferenzen

Januar

M	D	M	D	F	S	S
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

Februar

M	D	M	D	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	1	2	3	4	5
6	7	8	9	10	11	12

März

M	D	M	D	F	S	S
27	28	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2
3	4	5	6	7	8	9

April

M	D	M	D	F	S	S
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Konferenz  
in 2016

Mai

M	D	M	D	F	S	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21

Konferenz  
in 2015

Juni

M	D	M	D	F	S	S
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

Juli

M	D	M	D	F	S	S
26	27	28	29	30	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

Aufnahme  
2015

M	D	M	D	F	S	S
31	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3
4	5	6	7	8	9	10

Wann Aufnahme 2016?  
Welche Konferenz  
prioritär betrachten?

September

M	D	M	D	F	S	S
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

Oktober

M	D	M	D	F	S
25	26	27	28	29	30
1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	1	2	3	4	5

30	31	1	2
3	4	5	6
7	8	9	10
11	12	13	14
15	16	17	18
19	20	21	22
23	24	25	26
27	28	29	30
31	1	2	3
4	5	6	7
8	9	10	11

# Ranking von Harvesting-Kandidaten

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

## Datasets:

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

# Historische Daten zu Konferenzreihen

stream ▾	I	M	O	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002
<a href="#">conf/3dgis</a>			08 131													2.6				
<a href="#">conf/3dic</a>	1	11	8			9.9	8.2	11.2	9.3		8.2	9.6	10.4							
<a href="#">conf/3dica</a>	1	02	17			8.0	8.6	8.7	8.5	8.4	6.4	6.6		5.2		5.0		8.1		5.5
<a href="#">conf/3dim</a>	1	10			17.3	22.7	16.5	24.6	20.0	18.9	15.3				11.7		14.0		14.5	
<a href="#">conf/3dor</a>	1	05	2		18.1	29.0	21.2	22.8	25.5	21.9	20.9	35.9	18.8	13.4						
<a href="#">conf/3dph</a>			12 91										14.0							
<a href="#">conf/3dpvt</a>	2	06	49							18.9?	15.3?					16.2		15.2		15.3
<a href="#">conf/3dtv</a>	1	07			15.1	14.0	15.1	18.0	8.5	17.9										
<a href="#">conf/3dui</a>	1	03	4		12.9	14.7	15.1	16.5	16.1	14.5	10.3	11.4	14.4	9.9	12.0	13.3				
<a href="#">conf/3pgcic</a>	1	11	8			10.3	17.5	9.3	18.2	11.5	16.4	15.2								
<a href="#">conf/5gu</a>																				
<a href="#">conf/a2cwic</a>			09 82									1.7								
<a href="#">conf/a4cloud</a>			06 37					4.5												
<a href="#">conf/aaa-idea</a>	3	06	109													10.2	7.1			
<a href="#">conf/aaai</a>	1	02		22.3	22.6	25.5	22.6	25.6	20.0	22.9	19.0	21.5		20.1	20.0	17.6	17.0	15.0		12.9
<a href="#">conf/aaaifs</a>	1	11	56							5.8	7.6	17.1	5.0	3.5						
<a href="#">conf/aaaiss</a>	1	03	52						5.5	4.3	3.5	6.3	12.0	10.5	6.5	6.4	7.7			
<a href="#">conf/aaate</a>	2	09		4.4		4.0														
<a href="#">conf/aacc</a>			10 153															2.8		
<a href="#">conf/aadebug</a>	2	09	130													7.2		3.4?		
<a href="#">conf/aadios</a>			06 61							2.8										
<a href="#">conf/aaecc</a>	2	02	89										7.7		8.0	9.9		7.2		
<a href="#">conf/aafd</a>	2		49							2.2		4.6		3.5		6.3				
<a href="#">conf/aaim</a>	1	06	13			19.2		13.9	16.3	17.4	13.8	13.5	17.4	25.0	13.7	13.0	9.6			
<a href="#">conf/aaip</a>			09 94										4.5							
<a href="#">conf/aamas</a>			181																	5.7
<a href="#">conf/ab</a>	1	07	108											9.5	11.1					
<a href="#">conf/abials</a>	5	06	61											6.7				6.3?		

**I** = interval, **M** = month, **O** = overdue

# Merkmale für das Ranking

Faktoren zur Bestimmung der „Dringlichkeit“:

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

$c$	$\Delta(c)$	$w_{\text{delay}}$	$w_r$	$w_i$	$w_d$	$w_{\text{cit}}$	$w_{\text{prm}}$
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

# Ranking von Harvesting-Kandidaten

Vergleich der Baseline (Ranking nur nach Delay) mit jeweils Delay + einem weiteren Gewichtungsfaktor. Pseudo-Relevanz basierend auf tatsächlichen Aufnahmedaten aus 2016.

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

- <https://doi.org/10.1145/3197026.3197069>



# Zukünftige Projektarbeiten

„Smarte“ Wrapper (dblp, TH):

- **Interactive Wrapper:** Entwicklung eines GUI-basierten Tools zum Web-Harvesting von Metadaten, abgestimmt auf Bedarfe der Ziel-Nutzergruppe

Monitoring (dblp, TH):

- **Scheduling** von Harvesting-Vorgängen: Weiterführung der Forschung
- Überführung in eine Anwendung



# Vielen Dank! Gibt es Fragen?



## Einstieg zu XPath:

- Hands-on-Lab digital „Smart Harvesting mit XPath“
- **15.6.2018 @ Bibliothekartag 2018**



## XPath-Tutorial

- <http://www.xpath.org/papers/2017-IntroductionToXPath-ed1.pdf>

## Try it out

- XPath als **Docker-Container**  
[https://github.com/irgroup/xpath\\_docker](https://github.com/irgroup/xpath_docker)
- Syntax-Modul für **Atom**  
<https://atom.io/packages/language-xpath>

