



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

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.
- **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, ORCID, Google Scholar Profile, etc.



SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

 **Universität Trier**

Technische Hochschule Köln

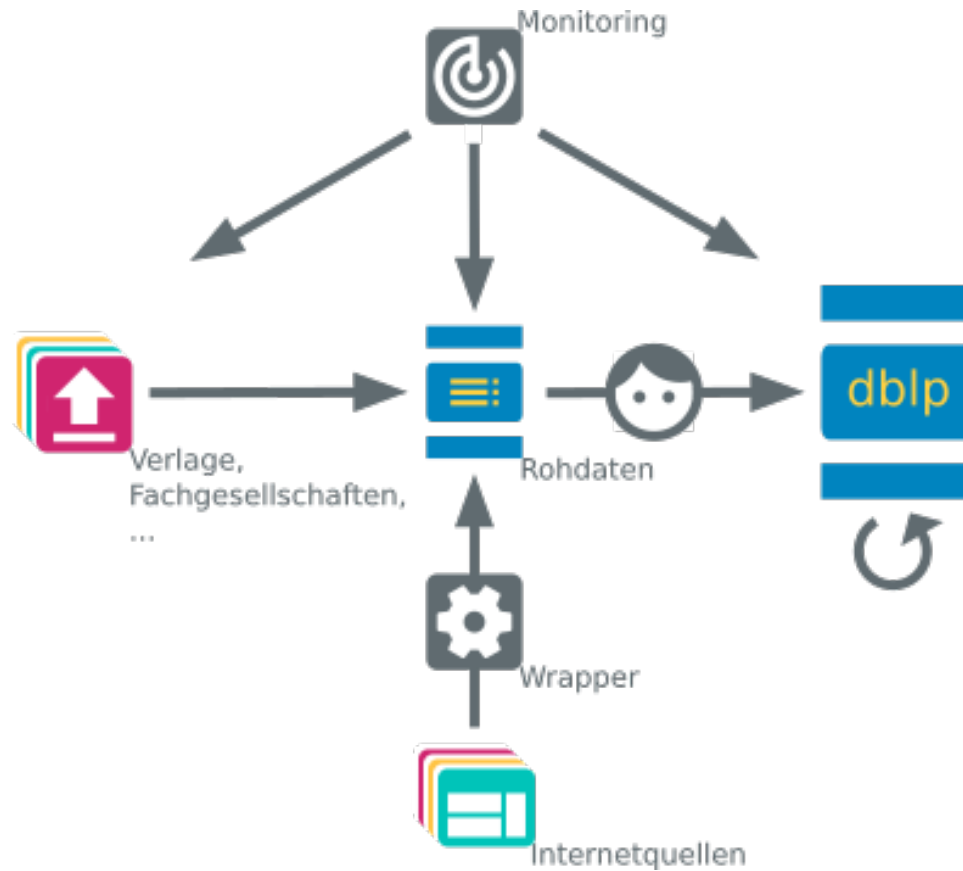
- Größte Hochschule für ang. Wissenschaften mit über 26.000 Studierenden
- Institut für Informationswissenschaft
 - 3 BA-Studiengänge: **Data and Information Science**, Bibliothek und digitale Kommunikation, Online-Redaktion
 - 2 MA-Studiengänge: Library and Information Science, Markt und Medienforschung
- Professur für Information Retrieval seit 07/2016
 - 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



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:

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} }
```



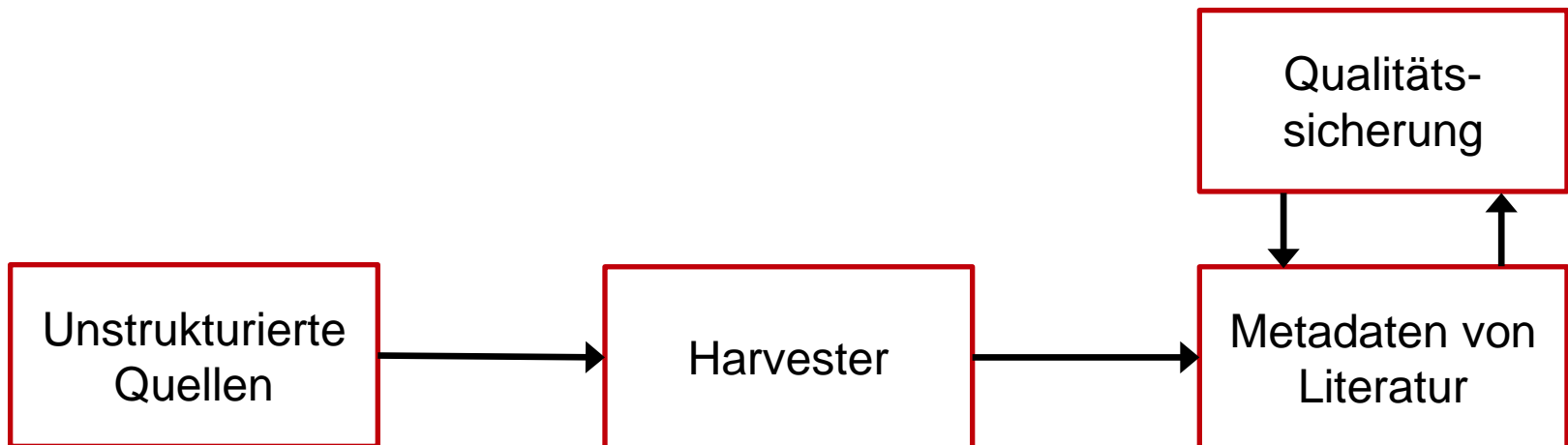
Grundproblem

Wir sind an Quellen interessiert, die

- **nicht durch Schnittstellen**, wie z.B. OAI-PMH, abbildbar sind 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 Ressourcen 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



EDM 2015

**The 8th International Conference
on Educational Data Mining**

**26-29 June 2015
Madrid - Spain**



You are here: [Proceedings](#)

Proceedings

This page holds the proceedings for the 8th International Conference on Educational Data Mining. The conference will be held on June 26 - 29, 2015, in Madrid, Spain.

[Table of contents](#)

Invited Talks (abstracts)

[Behind the Scenes of Duolingo](#)
Luis Von Ahn, Matt Streeter

[Personal Knowledge/Learning Graph](#)
George Siemens, Ryan Baker, Dragan Gasevic

Organized by the
International Educational
Data Mining Society
(IEDMS).

Comercial Sponsors

Gold




- [EDM2015](#)
- [Proceedings](#)
- [Keynotes](#)
- [Panels](#)
- [Workshops & Tutorials](#)
- [Schedule](#)
- [Presenter Instructions](#)
- [Student Information](#)
- [Important Dates](#)

EDM16

The 9th Intl. Conf. on
Educational Data Mining

June 29 - July 2, 2016
Raleigh
North Carolina, USA



You are here: [Proceedings](#)

Proceedings

This page holds the proceedings for the 9th International Conference on Educational Data Mining. The conference will be held on June 29 - July 2, 2016, in Raleigh, North Carolina, USA.

Individual papers

Invited Talks

[Data-Driven Education: Some opportunities and Challenges](#)
Rafael Azeiteiro

[MML: Ways to Strengthen Inquiry-Based Learning](#)
Michael Luo (University of Illinois)

[Enabling people to harness and control EDM for sharing the data learning](#)
John Liu

Organized by the International Educational Data Mining Society (IEDMS)

Sponsors






EDM 2017

THE 10th INTERNATIONAL CONFERENCE
ON EDUCATIONAL DATA MINING

WUHAN, CHINA
JUNE 28 - 30, 2017



You are here: [Proceedings](#)

PROCEEDINGS

This page holds the proceedings for the 10th International Conference on Educational Data Mining. The conference will be held on June 28 - 30, 2017, in Wuhan, Hubei, China.

Main Proceedings

Organized by the International Educational Data Mining Society (IEDMS)

ACADEMIC SPONSORS

Konferenzen – ein besonderes Problem

EDM16

The 9th Intl. Conf. on Educational Data Mining

June 29 – July 2, 2016
Raleigh
North Carolina, USA



EDM2016

Speakers

- Keynotes
- Industry Panel

Proceedings

Awards

Attendees

Proceedings

This page holds the proceedings for the 9th International Conference on Educational Data Mining. The conference will be held on June 29 - July 2, 2016, in Raleigh, North Carolina, USA.

Individual papers

Invited Talks

Data-Driven Education: Some opportunities and Challenges
Rakesh Agrawal

WISE Ways to Strengthen Inquiry Science Learning
Marcia Linn (presentation)

Enabling people to harness and control EDM for lifelong, life-wide learning
Judy Kay

Organized by the International Educational Data Mining Society (IEDMS).

Sponsors









EDM 2015
The 8th International Conference on Educational Data Mining
26-29 June 2015
Madrid - Spain



You are here: Proceedings

Proceedings

This page holds the proceedings for the 8th International Conference on Educational Data Mining. The conference will be held on June 26 - 29, 2015, in Madrid, Spain.

Table of contents

Invited Talks (abstracts)

Behind the Scenes of Owlidge
Luis Van Alen, Matt Shreever

Personal Knowledge/Learning Graph
George Siemens, Ryan Baker, Dragun Geopfert

Organized by the International Educational Data Mining Society (IEDMS).

Commercial Sponsors

Gold






EDM 2017
THE 10th INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING
WUHAN, CHINA
JUNE 25 - 28, 2017



THE 10th INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING

WUHAN, CHINA
JUNE 25 - 28, 2017

SEARCH FORM

SEARCH

This page holds the proceedings for the 10th International Conference on Educational Data Mining. The conference will be held on June 25 - 28, 2017, in Wuhan, Hubei, China.

PROCEEDINGS

Main Proceedings

ACADEMIC SPONSORS

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 XPath - XPath

```

<h3 id="sectionORIGINALARTICLES">
  <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">
            <li class="cit-title-group">Perceptual Hash Function based on
              Scale-Invariant Feature Transform and Singular Value Decomposition
            </li>
            <li class="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">
              </li>
            </div>
          <div class="cit-extra">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
            <li class="cit has-earlier-version from-current-issue toc-cit">
          </ul>
        </li>
      </ul>
    </li>
  </ul>

```

OXFORD UNIVERSITY PRESS | Journals

Journals > Science & Mathematics & Law & Social Sciences > Journal of Cybersecurity > Volume 1, Issue 1

```

<div class="inner-collapsible-content-
  heading-wrapper">
  <h2 id="editorial" class="toc-heading
  editorial wrap-elements-processed inner-
  collapsible-content-
  heading">EDITORIAL</h2>
  <span class="inner-content-toggle"/>
</div>
<div class="inner-collapsible-content-
  wrapper">
  <ul class="toc-section">
    <li class="first last odd toc-item">
      <div class="toc-citation">
        <div id="" class="highwire-article-
        citation highwire-citation-
        type-highwire-article tooltip-enable
        highwire_article_citation_tooltip-
        processed" title="Welcome from the
        Editors-in-Chief" rel="/highwire
        /article_citation_preview
        /60529" data-node-
        nid="60529" data-pisa="cybers;
        1/1/1" data-pisa-master="cybers;
        tyv010" data-apath="/cybers/1/1/1.atom">
          <cite class="highwire-cite highwire-
          cite-highwire-article highwire-citation-
          jnl-oup-toc-one-line clearfix">
            <div class="highwire-cite-title-access">
              <div class="highwire-cite-detail-
              wrapper">
                <span class="highwire-cite-authors
                add-author-link-processed">
                <span class="highwire-cite-jnl-info">
                <span class="highwire-cite-doi">
                <span class="highwire-
                cite-fpub">First published online: 27
                November 2015 (2 pages)</span>
              </div>
            <span class="highwire-cite-extras">
          </cite>
        </div>
      </li>
    </ul>

```

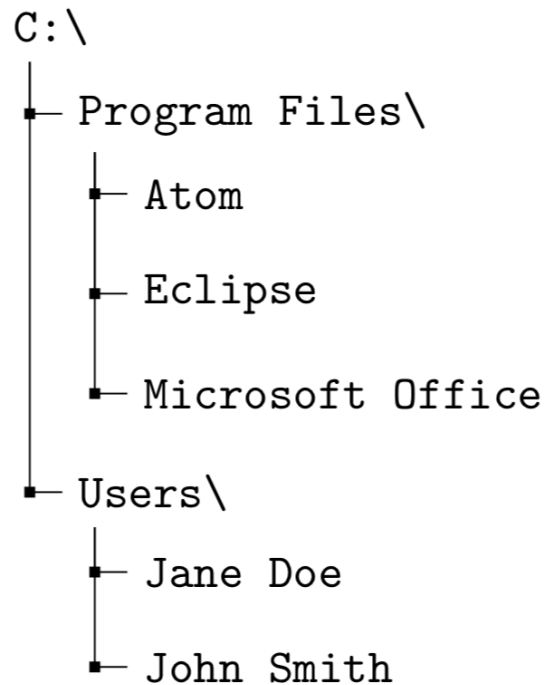
From physical security to cybersecurity

Arunesh Sinha, Thanh H. Nguyen, Debarun Kar, Matthew Brown, Milind Tambe, Albert Xin Jiang
 J Cyber Secur (2015) 1 (1): 19-35 DOI: <http://dx.doi.org/10.1093/cybersec/tyv007> First published online: 17 November 2015 (17 pages)

Abstract Full Text (HTML) Full Text (PDF) Figures & data

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

```

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

OXPath-Beispiel

The screenshot shows a Google Scholar search result for the query "OXPath". The search bar at the top contains "OXPath" and a search icon. Below the search bar, the word "Scholar" is displayed in red, followed by 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 "OXPath: Everyone can Automate the Web! Travel Bursaries. ...". At the bottom, there is a "Create alert" checkbox and a pagination bar with numbers 1, 2, 3, 4 and navigation arrows.

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(.,
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>

```

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

- 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.

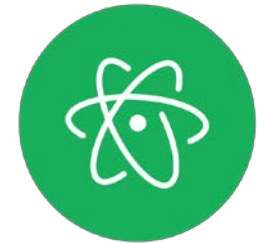


Toolbox rund um XPath

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

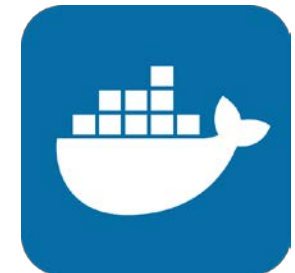
Atom-Modul

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

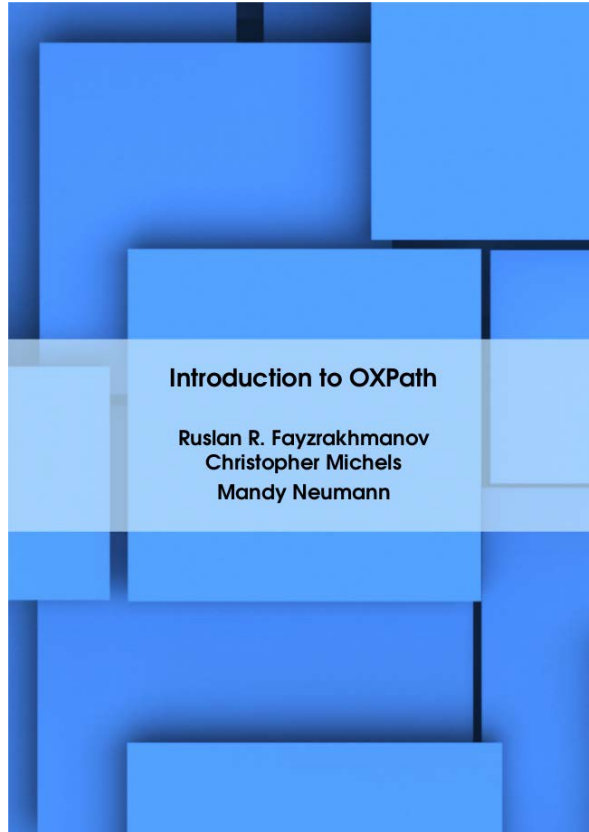


Docker-Container

- Ursprüngliches XPath nur unter Ubuntu
- 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							
ANMO - Austrian Newspapers Online	a0048	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2011-02-01 08:51		1,029	✓
Anabine - Online catalogue for women	a0380	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		185,946	✓
TRIOCATO - Catalogue of the Depart	a0381	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		43,007	✓
Catalogue of the Map Department of t	a0382	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		23,028	✓
Image platform of the Austrian Nation	a0386	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 11:04		14,765	✓
Travel Collection from the National Li	a0429	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		30,928	✓
Picture Archives and Graphics Collec	a0478	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20	2012-08-28 13:30	13,771	✓
Austria	a0479	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 14:50		4,488	✓
Department of Portraits	a0480	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		185,268	✓
Vienna	a0481	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 13:30		8,520	✓
Contemporary History	a0482	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 13:30		89,163	✓
PNB-Virtu	a0482	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 13:30		3,979	✓
Kronprinzenerk	a0488	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		61	✓
Archives of Austrian Folk Music Sock	a0489	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-08-28 09:20		220	✓
EC1914 OAB-Public Material	a0563	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-02-27 12:38		13,858	✓
EC1914 OAB-Picture material (Gibson	a0564	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc	2012-02-27 12:38		1,403	✓
Austrian National Library named after	a0377	oai:10.1001.1001.1001	oai-1.0	OAI-PMH oai_dc			0	✓

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

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 2017

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 2016

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
2016

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 2017?
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

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

M	D	M	D	F	S	S
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

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}
jc dl	3	4	1.88	1.192	0.250	1.029	1.312
tp dl	0	4	1.63	1.577	0.250	1.024	1.352
ic dl	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>

