DeepGreen - ein schnittstellengetriebener Ansatz zur Datenanreicherung und -analyse: Konzeption eines Workflows mit Modulen in RStudio

111 BiblioCon2023 Hannover 24.05.2023 DeepGreen Anwender:innen-Treffen – Austausch und Weiterentwicklung

煎茶 (Sencha) - 八女市 (Yame-shi) 福岡県 (Fukuoka-ken) Juli 2015 T.Wetzenstein

Thorsten Wetzenstein Universitätsbibliothek Kiel Referat Open-Access-Publizieren | Universitätsverlag | https://orcid.org/0000-0001-9589-3136

Inhalte

- Charakteristika der DeepGreen-Sammlung auf MACAU
- Datenschnittstellen
- Anreicherung der DeepGreen-Daten in MACAU
- Wege der Datenanalyse: Harvesten, Transformation, Filtern, Anreichern, Analysieren
- Sichtbarkeit der DeepGreen-Sammlung auf MACAU
- Nutzen einer verteilten Repositorylandschaft mit neuen Werkzeugen und Schnittstellen für DeepGreen-Sammlungen

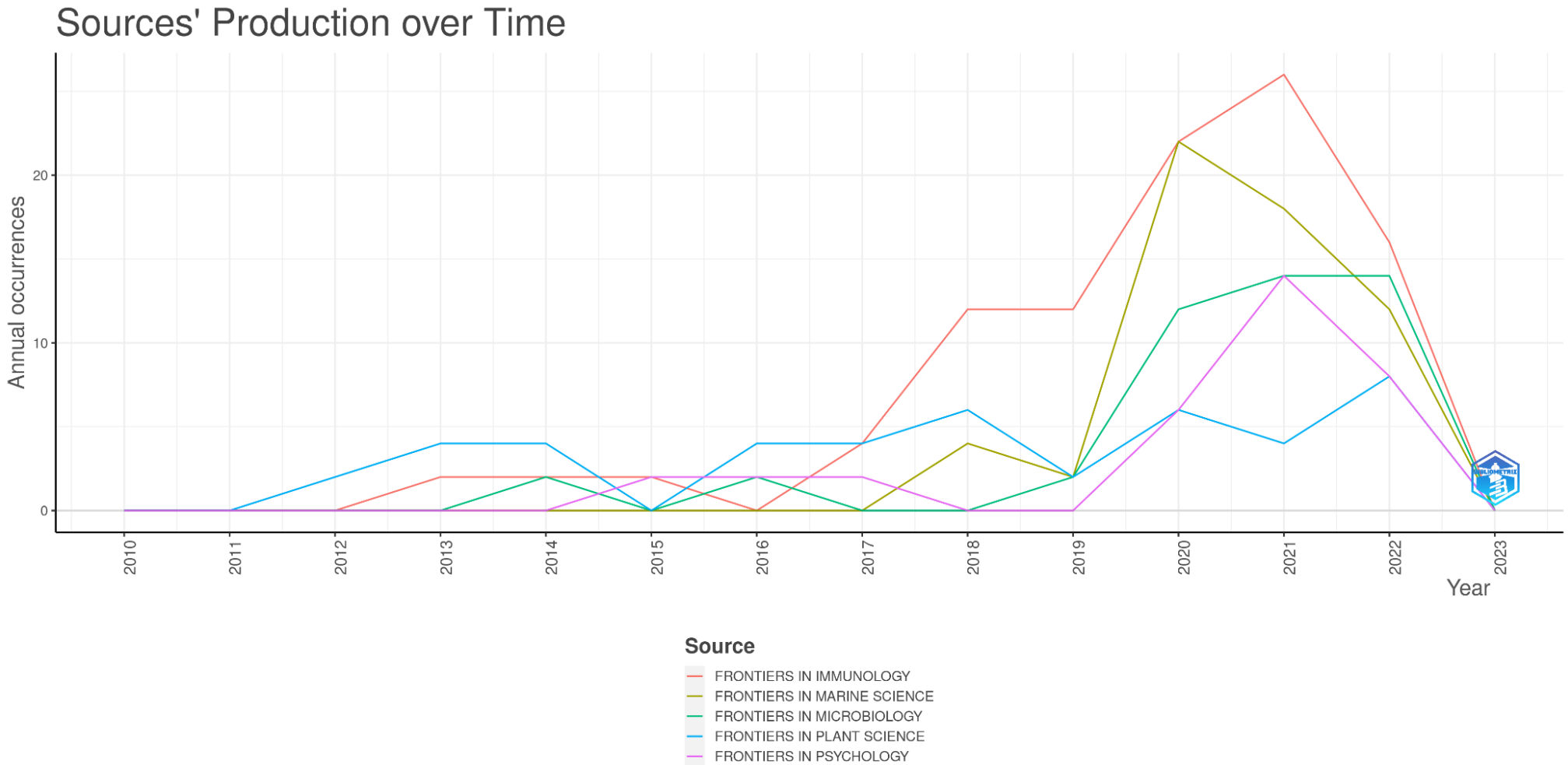
Charakteristika der Sammlung auf MACAU - Workflows

Stand 27.04.2023: **8.080** Dokumente gesamt
5520 Dissertationen
1905 Zweitveröffentlichungen, davon **1011** via DeepGreen
354 Erstveröffentlichungen
304 Verlagsveröffentlichungen (inklusive Kapitel und Artikel)
davon 21 eigenständige Veröffentlichungen (Sammelbände, Zeitschriftenhefte)

DeepGreen

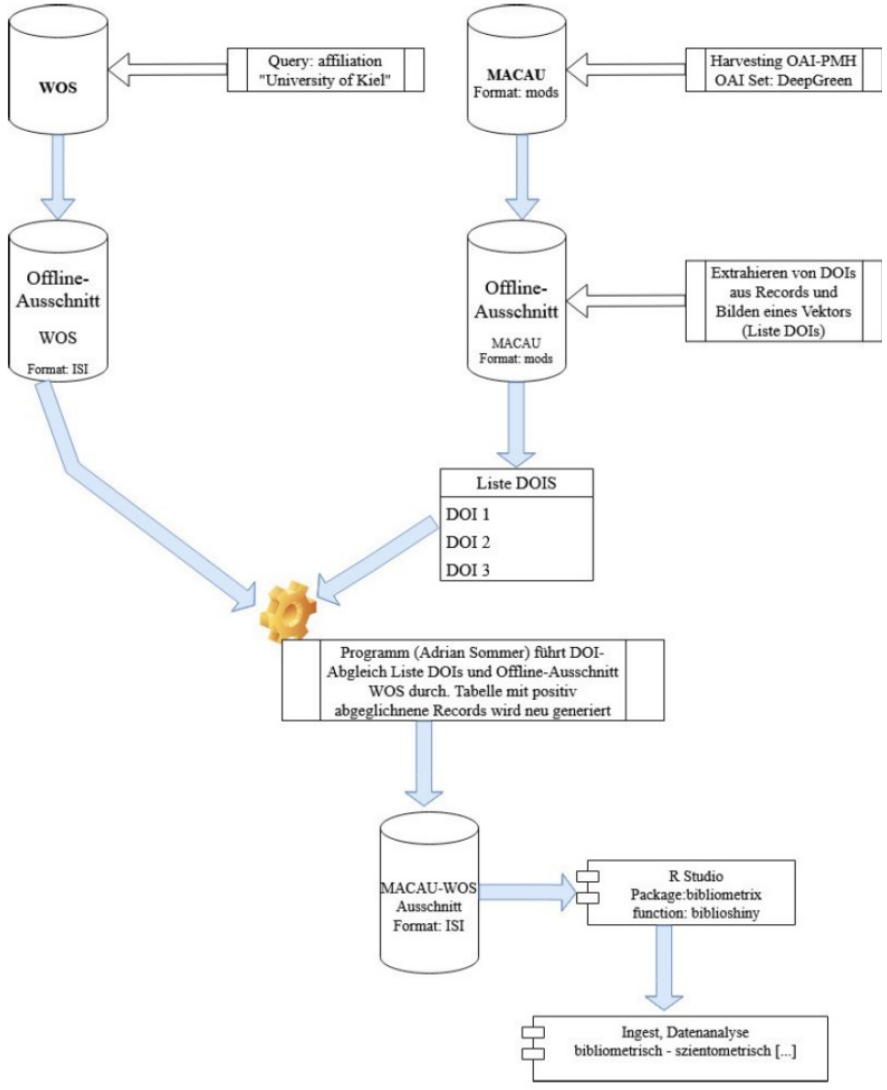
872	Dok.	CC-BY 4.0
35	Dok.	CC-BY-NC 4.0
23	Dok.	CC-BY-NC-ND 4.0
20	Dok.	CC-BY 3.0
3	Dok.	CC-BY 2.0
53	Dok.	Ohne CC-Lizenz

Charakteristika der DeepGreen-Sammlung auf MACAU - Produktivität nach Zeitschriften



Wege der Datenanalyse: Analysieren - Inhalte

Durchführung des Workflows *realisiert*



Datenschnittstellen – Web of Science Daten (WOS) – Open Alex – Crossref – DeepGreen

	JATS (Publ. Router DeepGreen)	Crossref (Verlage / Repo.-Betr. / Bibliotheken)	OpenAlex	WOS (Clarivate)
Funding Information	(ja)	ja	nein	ja
Author ORCID	ja	ja	ja	ja
Author position	(ja)	ja	ja	ja
ror ID	(ja)	ja	ja	nein
weitere Hosts	(nein)	(nein)	ja	nein
zitierte Ref.	(nein)	ja	ja	ja
zitierende Ref.	(nein)	ja	ja	ja
Lizenzinfo CC	ja	ja	ja	nein (OA gold: ja)
Ref. Klassifikationssyst.	(ja)	nein	ja	ja

MACAU als MyCoRe IR verwendet Datenformat MODS. Bezug der Metadaten durch Resolving der DOIs und interner Verwendung des sog. MyCoRe Enrichment Resolvers

OpenAlex integriert unpaywall-Daten: von dort kommen die Angaben über *alternate_host_venue bzw. locations.*

Web of Science API war zum gegebenen Zeitpunkt nicht genutzt worden: Daten von Web GUI

Wege der Datenanalyse: Harvesten - DOI-Liste erstellen

OAI-Harvesting von MACAU vom OAI Set „deepgreen“. Aufruf im Terminal

```
oai-harvest –set deepgreen –metadataPrefix mods https://macau.uni-kiel.de
```

Zählen der Records

```
ls –la | wc -l
```

Anzeigen aller XML Elemente <mods:identifier type=“doi“>

```
cat *.xml | grep doi
```

Anzeigen aller DOI als Textinhalt des XML-Elements als String

```
cat *.xml | grep doi | cut –d “>“ –f2 | cut –d “<“ –f1
```

Schreiben der Liste in eine CSV Datei

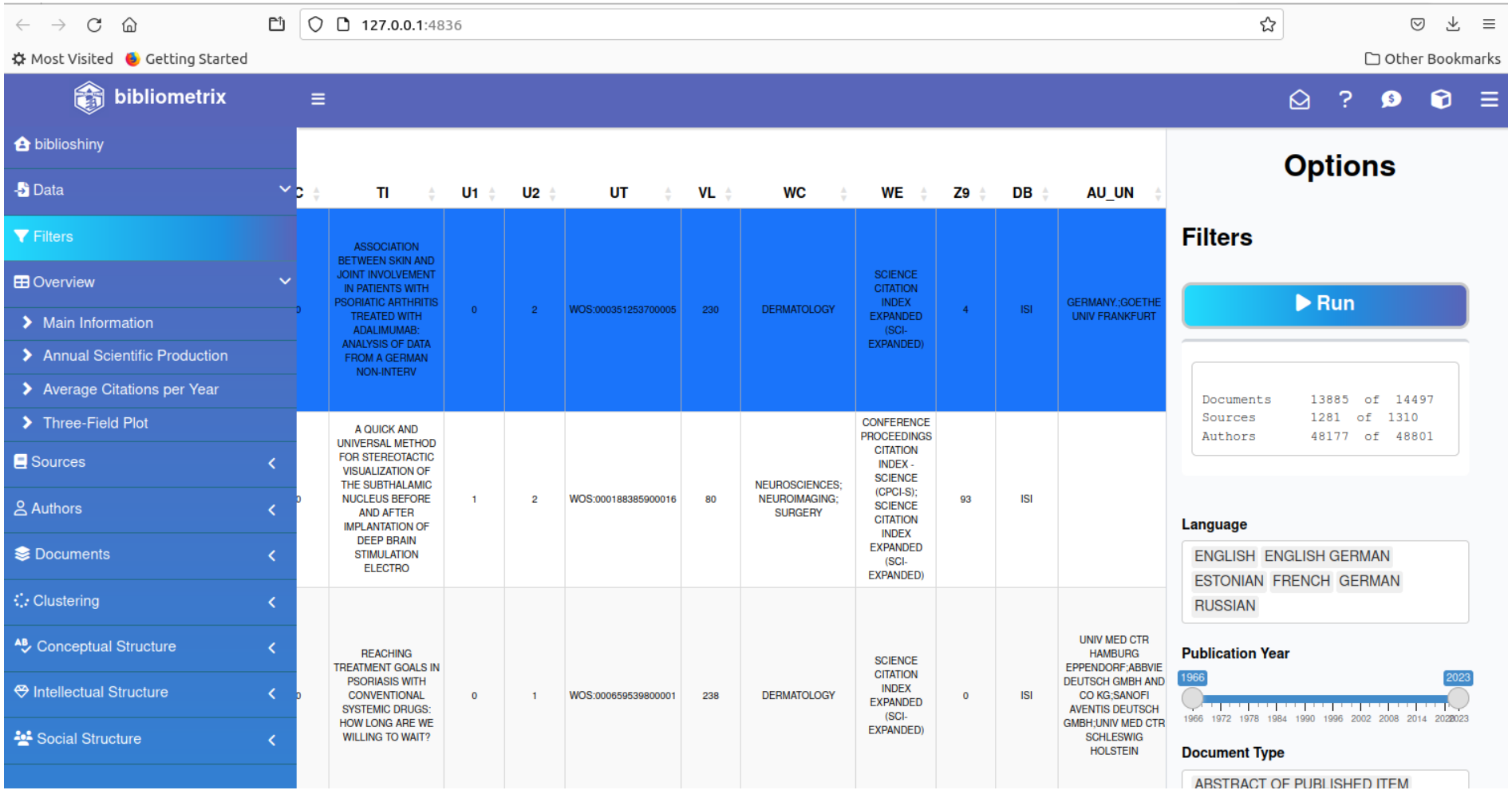
```
cat *.xml | grep doi | cut –d “>“ –f2 | cut –d “<“ –f1 > list_doi.csv
```

Wege der Datenanalyse: Analysieren - Inhalte

- Analysewerkzeuge: CRAN Package bibliometrix
- Data Extraction Tool: CRAN Package openalexR
- Data Extraction Tool: CRAN Package wosr
- Conversion Tool Bibliographische Formate: CRAN Package rbibutils
- Bezug von Daten über Web of Science Webinterface (WOS)

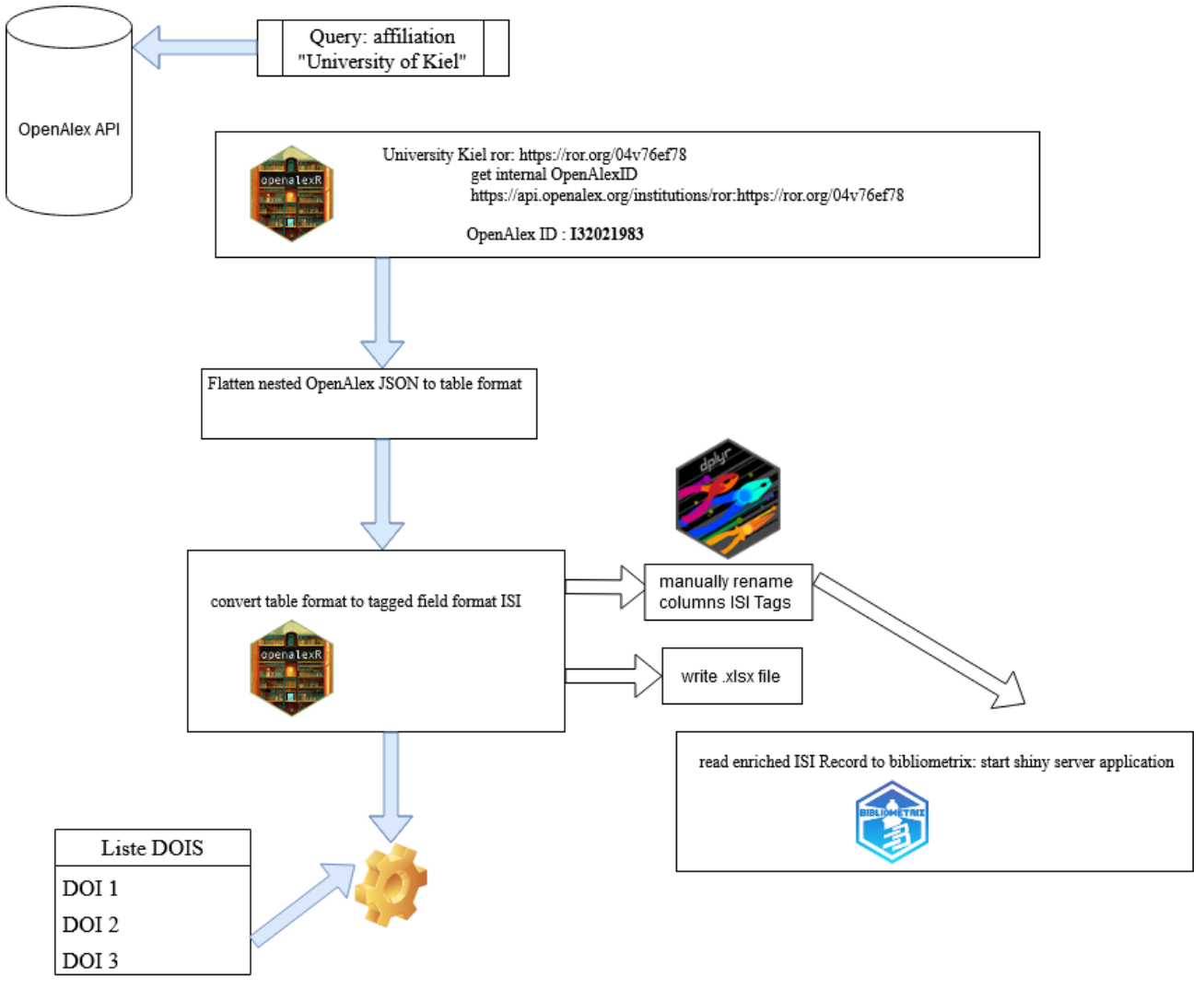
Wege der Datenanalyse: Analysieren - Inhalte

Analysewerkzeuge: CRAN Package bibliometrix



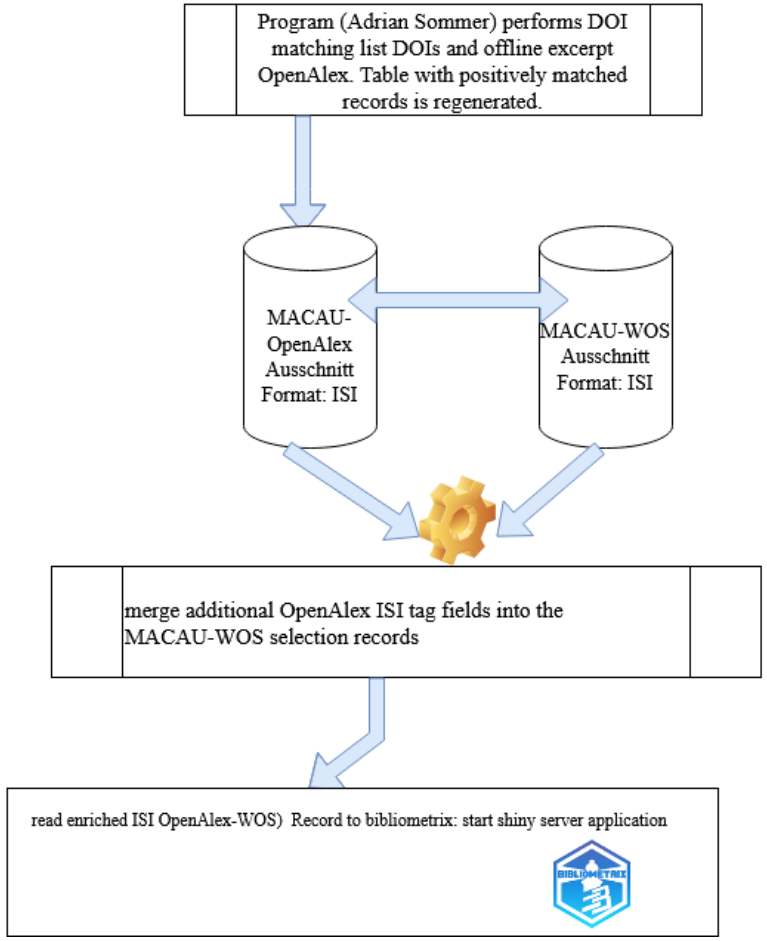
Wege der Datenanalyse: Analysieren - Inhalte

Durchführung des Workflows
Designvorschlag



Wege der Datenanalyse: Analysieren - Inhalte

Durchführung des Workflows
Designvorschlag



Anreicherung der DeepGreen-Daten in MACAU

Rstudio Workflow


```
cau_pub_OA <- openalexR::oa_request(query_url = "https://api.openalex.org/works?filter=institutions.id:I32021983?", verbose = "TRUE")
```

1. Definieren einer Variablen, in der „works“ der Einrichtung CAU Kiel gespeichert wird. Bezug der Daten von der JSON API OpenAlex (tiefgeschachtelte Daten in JSON Codierung)

```
cau_pub_OA_df <- openalexR::oa2df(cau_pub_OA, entity = "works", verbose = "TRUE")
```

2. Definieren einer Variablen, in der der OpenAlex API Abzug in ein flaches Tabellenformat, sog. Dataframe (df) gebracht wird

```
cau_pub_OA_df_bibmet <- openalexR::oa2bibliometrix(df = cau_pub_OA_df)
```

3. Definieren einer Variablen, in der der Dataframe mit ISI Tags in Tabellenspalten gewandelt wird

In der Funktion oa2bibliometrix können Spaltenbezeichnungen teilw. Nicht ISI-gerecht sein, so dass sie in Excel nachbearbeitet werden müssen. Hier liegt evtl. Entwicklungspotenzial für Package-Updates : das erwartete ISI Tag für DOI **DI** ist **„doi“** benannt. Ebenso TAG für Schlagwort **DE** als „concept“ – erweiterte Liste siehe Fachaufsatz

Anreicherung der DeepGreen-Daten in MACAU

Rstudio Workflow

```
WriteXLS::WriteXLS(x = cau_pub_OA_df_bibmet, ExcelFileName = "cau_pub_isi.xlsx", verbose = "TRUE")
```

4. Schreiben des ISI Dataframes in eine Exceldatei

5. Extern oder in Rstudio Umbenennen derjenigen Spalten, für die ISI Tag Bezeichnungen nicht übereinstimmen (Generell sind ISI Tags zum Großteil schon korrekt gesetzt) – Steigerung der Transformationsleistung

```
bibliometrix::biblioshiny()
```

6. Einlesen der Exceldatei in Bibliometrix / Auswertungen

Sichtbarkeit der DeepGreen-Sammlung auf MACAU

OpenAlex

BASE Bielefeld Academic Search Engine

MACAU als Datasource auf OpenAlex : ID S4306401923
https://api.openalex.org/sources/S4306401923

OpenDOAR

Kiel University als Institution auf OpenAlex : ID I32021983
https://explore.openalex.org/institutions/I32021983

OpenAIRE EXPLORE

https://unpaywall.org/sources/repository/ejjyo7aunlatevqeweqj

unpaywall

und andere

Serverseitig:

dini

DeepGreen-Sammlung mit eigenem OAI Set `<setSpec>deepgreen</setSpec>`

In Diskussion: HighWire Press Tag `<meta name="citation_collection_id" content="deepgreen">`

Fazit

- Für Mapping in ISI Format von CRAN Package openalexR wären über spätere Updates Verbesserungen zu erwarten für (Erweiterte Liste siehe Fachaufsatz)
- Eine enge Kooperation in der Entwicklung von openalexR und bibliometrix erscheint wichtig vor dem Hintergrund der Formattransformation nach ISI und die (z.B. in ggplot) vorgesehenen Darstellungen im Package bibliometrix
- Ein offizieller ISI Standard konnte nicht gefunden werden - vorhanden: Web of Science Core Collection Field Tags von Clarivate
- Eine ISI Tag-Erweiterung über das Format von Clarivate hinaus unter gemeinsamer Spezifikation mit openalexR und bibliometrix Entwicklerinnen und Entwicklern scheint weitere Möglichkeiten zu erschließen.
- Für neue Mappings wären auch in bibliometrix neue Auswertungsalgorithmen und –szenarien zu konzipieren

- **Nutzen DeepGreen: Chance zur Nutzung einer noch nie dagewesenen Datenlage durch neue Werkzeuge und Schnittstellen – Reichweitenmessung außerhalb von Verlagen und Contentprovidern durch verteilte Repositorylandschaft – Bildung eigener Metriken auf Artikelebene und teilw. Lösung von kommerzialisierten Metriken. „Open Access auch für Datenanalysen“ Metriken, Sammlungsanalysen, Reichweitenmessung**

Vielen Dank für Ihre Aufmerksamkeit

Vielen Dank für die Unterstützung an die Kollegen der UB Kiel

Dr. Kai Lohsträter
Adrian Sommer
Torsten Krause
Oliver Weiner

Thorsten Wetzenstein
Universitätsbibliothek Kiel Referat Open-Access-Publizieren | Universitätsverlag |
Leibnizstr. 9
24118 Kiel

Telefon: +49 431 880-2740
mail: wetzenstein@ub.uni-kiel.de

https://orcid.org/0000-0001-9589-3136

Ursprünglicher Artikel
Wetzenstein, Thorsten: 1000ste „DeepGreen“-Veröffentlichung an der CAU – ein Zwischenbericht. Feb. 2023

https://oa-info.sh/2023/02/1000ste-deepgreen-veroeffentlichung-an-der-cau-ein-zwischenbericht/