EINE ANALYSE DES LIS OPEN ACCESS PUBLIKATIONSGESCHEHEN ANHAND DER API SCHNITTSTLLE DES DOAJ

Abschlussarbeit

Zertifikatskurs Data Librarian

26.8.2022

Salz, Nadine

salznadine@aol.com

Inhalt

Motivation 1
Hintergrund 1
Beschreibung der Ausarbeitung 2
Auswahl des Themas 2
Analyse des Forschungsstandes 2
Auswahl der geeigneten Quelle 3
Auswahl Auswertwerkzeuge /Softwaretools 3
Analyse der API Schnittstelle & Übertragung der Daten 4
Datenauswahl innerhalb der ausgewählten Datenquelle 4
Ergebnisse 5
Diskussion 7
Projektevaluation 7
Erhebungsprozess 8
Grenzen 8
Empfehlungen für weiterführende Forschung 8
Literaturverzeichnis 9
Anhang Python Code 10

Motivation

Eine Eigenanalyse des LIS (Library and Information Science) Open Access Publikations-geschehen ist im Hinblick auf den Umstand interessant, dass sich insbesondere wissenschaftliche Bibliotheken die Förderung, Unterstützung und Beratung zu Open Access Ihrer jeweiligen Universitätsangehörigen - und bisweilen darüber hinaus – besonders stark auf die Fahnen geschrieben haben. Es stellt sich somit die Frage, wie es in der eigenen Disziplin, dem Bibliothekswesen, Anwendung findet und umgesetzt wird. Betrachtet man Verlage wie de Gruyter fällt auf, dass sie hochpreisig Fachliteratur im Bereich Informationswissenschaft anbieten, es aber gleichzeitig gefühlt wenig qualitative OA Quellen gibt, die sich als Informationsquelle für die fachliche Weiterbildung im Alltag anbieten. Zeitschriften wie beispielsweise das „Journal of Business & Finance Librarianship“ von Taylor & Francis sind weiterhin kostenpflichtig. Die Fragestellung lautet somit, wie qualitativ und wie frei wiederverwendbar und wie kostenfrei die verzeichneten Quellen sind - also kurz wieviel Open Access denn in Wirklichkeit im LIS Bereich, zu finden ist.

Hintergrund

Es gibt bereits einige, auch aktuellere Studien und bibliometrische Analysen, meist aus dem indischen Raum, die sich mit dem Thema bereits beschäftigt haben und dazu ebenfalls in

Teilen auf DOAJ zurückgegriffen haben. Allen ist gemein, dass ein Wachstum von OA Publikationen im LIS Sektor zu verzeichnen sei, oft wird dabei die USA als Hauptakteur genannt.[1] [2] Die Methoden der Studien unterscheiden sich jedoch von dem der hier gewählten Vorgehensweise: während zum Teil Datenbanken wie Scopus herangezogen werden, scheint beim DOJA hauptsächlich die Webbrowser Oberfläche genutzt zu werden, oder die Daten per csv Format extrahiert worden seien, statt die API-Schnittstelle zu nutzen. Diese Ergebnisse werden in der vorliegenden Arbeit verglichen.

Beschreibung der Ausarbeitung

Auswahl des Themas

Die Themenwahl stellt einen aufwändigen Baustein in dieser Arbeit dar. Der zu untersuchende Forschungsgegenstand soll greifbar und die gestellten Anfragen umsetzbar sein. Für eine Themenabgrenzung reicht es nicht, sich einfach eine interessante Forschungsfrage zu überlegen, es braucht auch realistisch zugänglich und händelbare Datensätze, die ein sinnvollen Output ermöglichen.

Analyse des Forschungsstandes

Eine Evaluation des Forschungsstandes wird bei dieser Arbeit dazu genutzt, die in anderen Studien analysierten Fragestellungen als Anregung zu betrachten, um eine Vergleichbarkeit der bereits gemachten Studien im ähnlichen Zeitraum anzustreben. Es bietet sich damit auch die Gelegenheit einen wichtigen Bereich von Data Science zu betrachten nämlich, bereits bestehende Ergebnisse zu hinterfragen, validieren oder zu widerlegen.

[1] Narayan, R., Pati, Pritam K., Sahoo, S. (2021). Growth of Open Access Literature on Library and Information Science during 2011-2020: A Scientometrics Analysis. *Library Philosophy and Practice*

[2] Rajkumar, T., Jeyapragash, B. (2021). Contributions of Open Access Journals in Library and Information Science indexed in SCOPUS Database: A Metric Study. *Library Philosophy and Practice* (e-journal). 6181.

Stand der Forschung

Übersicht aktuellste Studien	Selvam Masilaman (2020)	Narayan, R., Pati, Pritam K., Sahoo, S. (2021)	Rajkumar, T., Jeyapragash, B. (2021).	Sahoo, J., Birtia, T., & Mohanty, B. (2017). wiederholt 2020	Selvam, M. and Amudha, G. (2020).	Chakravarty, R. , Diksha C. (2020).	Sarasu R., Kairali, A., Sayed M.a.J. (2020).
Amount Journals	176 Journals	8.380 Research Paper	61 Journals	158	176		151
Source	DOAJ	Bibliometrical Analysis Scopus	Scopus; Impact (SNIP), Cite Score, h-index ..	DOAJ Website	DOAJ, Excel	DOAJ	DOAJ, Excel
Results	• 96.02% no APC • 68.18% English • Spanish (23.30%) • 58.52% Double Blind Peer Review • 75 journals use CC BY license. • USA Published (15.91%) highest number • 2017 (18.18%) maximum number of journals added	•Expontial grow • 2019 most productive year with 1642 papers • USA highest number of publications with 2166	•USA highest number of publications (10, 16.39%), 7982 •UK (1848) •US highest number of citations (37658)in 2011-2015 •US & UK h-index (73/66) topped the other 29 countries •US highest SNIP (11.147), •UK highest cite score (26.8, 39.70%)	•English dominant language •3 journals published in five language •India fifth Poland and United Kingdom with six	•169 no APC •7 APC •75 CC BY •35 CC BY-NC-ND •29 CC BY-NC •18 CC BY-NC-SA •14 BY SA, •3 CC BY-ND •2 own license. •28 US highest no of pub. •25 Brazil •13 Spain •120 Engl, Spanish, Portuges most Language accepted •103 Double blind •32 Peer Reviewed •29blind •10 ed Review •Most added during the year 2017 (32)	•US highest number published •Double-blind peer reviewing is significantly •high not mentioned their plagiarism policy. •English prominent language	•1 % LIS discipline •64 no APC •32 CC BY-NC-ND •23 journals CC BY-NC •15 journals CC-BY-NC-S •128 CC BY-SA •2 journals CC BY-ND •2publishers own license •US 27(26%) •Brazil 22(22%), •Indonesia 10(10%) Spain •10(10%) •151 double blind peer review, • 88 (59 per cent) peer review, •19 % blind peer review

(Fig 1)

Auswahl der geeigneten Quelle

Als Ausgangspunkt für die Fragestellung bietet sich das Directory of Open Access Journals (DOAJ) an. Das DOAJ verzeichnet zum einen detaillierte Informationen über die Art der OA Lizenz (CC BY-NC etc.) zu den Publikationsgebühren, zum Review Verfahren, zur Sprache und zum Start des OA. Zum anderen deckt es den LIS Bereich ab, und verzeichnet seit 2003, also seit knapp 20 Jahren OA Zeitschriften, sodass es auch die Betrachtung einer Zeitreihe ermöglicht. Dazu kommt, dass Zeitschriften des gesamten DOAJ aus 130 Ländern verzeichnet werden und damit auch die Option geboten ist, einen internationalen Überblick zu erhalten.

Eine weitere Motivation das DOAJ für die Analyse heranzuziehen ist die vom DOAJ angebotene API-Schnittstelle und die Annahme, dass diese kostenfrei und öffentlich zugängliche Schnittstelle keine Probleme hinsichtlich eines Massendownloads verursacht und damit keinen Harvesting Alarm in der Bibliothek auslöst. Das DOAJ verzeichnet aktuell insgesamt 18.141 Zeitschriften und bietet damit grundsätzlich eine valide Masse für Datenanalyse.

Das DOAJ bietet damit die Möglichkeit konkret folgende Fragen zu beantworten:

- Welche der OA Lizenzen werden vornehmlich vergeben?
- In welchen Ländern sind die Zeitschriften überwiegend herausgegeben?
- Wie viele sind peer-reviewed und double blind peer-reviewed?
- Wie sieht es mit Publikationsgebühren aus?
- Sind unter den Herausgebern vielleicht auch namhafte Herausgeber wie de Gruyter?
- In welchen Ländern wird am meisten publiziert?

Auswahl Auswertwerkzeuge /Softwaretools

Aufgrund der Auswahl des Themas und der Datenquelle, die einen Export in JSON Format anbietet, wird der Editor Jupyter und vorwiegend die Pythonbibliothek Pandas genutzt.

Analyse der API Schnittstelle & Übertragung der Daten

Mit der „normalen“ Suche im Browser kann man Titel, Abstract, Keywords, Subject, Author, ORCID, DOI und nach Sprache suchen. Demgegenüber wird vom DOAJ eine sehr differenzierte API angeboten, die ganz konkrete Abfragen der Teilbereiche zulässt, ohne den kompletten Datensatz exportieren zu müssen. Auf diese Weise ist eine komfortable Vorfilterung möglich. Die API-Dokumentation und die gebotenen Möglichkeiten sind im Vergleich zu anderen Anbietern vorbildlich: So gibt es eine Funktion, die es erlaubt, testweise Abfragen an die API zu senden und einen Beispieldatensatz als Antwort in JSON ausgeben zu lassen.

Um die 183 Titel des LCC Bereichs „Bibliography. Library Science. Information Resources“ zu exportieren, muss die genaue Abfrage gefunden werden. Es handelt sich um 3 Begriffe mit Leerzeichen und Punkten, die für die Übergabe kenntlich gemacht werden müssen. Bei der Eingabe im Browser wird es wie folgt wiedergegeben und kann so nicht 1:1 verwendet werden:

Dies muss im Jupyter Notebook übertragen werden zu

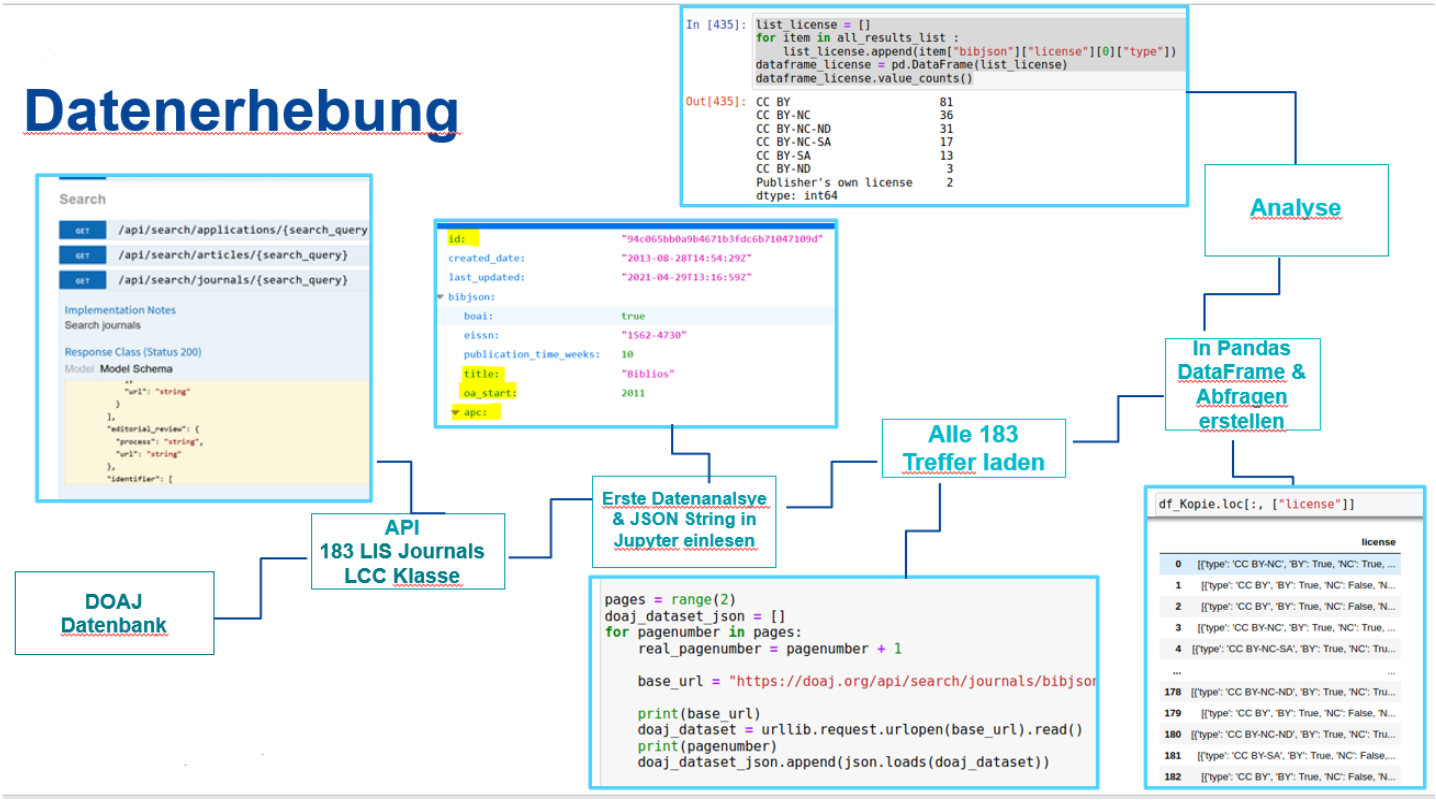
```
base_url = "https://doaj.org/api/search/journals/bibjson.subject.term:/%22Bibliography.
%20Library%20science.%20Information%20resources/%22?page=" + str(real_pagenumber) + "&pageSize=100"
```

Mit %22 werden dabei die Anführungszeichen kenntlich gemacht.
Die API übergibt standardmäßig 10 Treffer und die maximal mögliche Trefferzahl pro Export beträgt 100. Zu importierende Treffer sind es allerdings 183, sodass es 2 Downloads geben muss, die dann zu einem Datensatz über append() verbunden werden. Bei einem Testexport mit wesentlich mehr Datensätzen ist festzustellen, dass die Datenbank doch einen Sicherungsmechanismus enthält, sodass bei zu vielen Anfragen in der Sekunde der Download gestoppt wird. Der Mechanismus kann aber mit einer Verzögerung von 1 Sekunde ausgesetzt werden.

Datenauswahl innerhalb der ausgewählten Datenquelle

Nach einer Betrachtung der bibliographischen Beschreibung auf Artikel- und Zeitschriftebene wird die Zeitschriftebene ausgewählt, da die Informationen dort am besten angehängt sind. Des Weiteren wird die oberste Schlagwortkategorie der LCC (Library of Congress Catalgue) Subject betrachtet, da die beiden Unterkategorien „Bibliography“ oder „Library Science, Information Resources“ alleine zu wenig Treffer ergeben haben. Über die vom DOAJ angebotene API-Schnittstelle werden somit 183 Zeitschriftendaten aus dem Bereich Bibliothekswesen mit Schlagwort LCC Subject Bereich Bibliography, Library Science, Information Resources abgefragt.

Die Beschreibung der Ausarbeitung des ersten Teils lässt sich graphisch damit wie folgt darstellen:



(Fig. 2)

Im Anschluss an den Punkt Analyse folgt noch die Dokumentation, die in dieser Grafik nicht mit aufgenommen ist, da sie der Prüfungsleistung dient und selbst nicht zum Analyseprozess gehört.

Ergebnisse

Konkret lauten die Antworten dieser Datenanalyse wie folgt:

Welche der OA Lizenzen werden vornehmlich vergeben?

```
CC BY                      81
CC BY-NC                   36
CC BY-NC-ND                31
CC BY-NC-SA                17
CC BY-SA                   13
CC BY-ND                    3
Publisher's own license     2
```

In welchen Ländern sind die Zeitschriften überwiegend herausgegeben bzw. welches Land publiziert am meisten OA?

29 USA, 24 Brazil, 13 Indonesien,12 Spanien, [..] 5 Germany

publisher.country	
US	29
BR	24
ID	13
ES	12
PL	9
IR	9
UA	7
IT	6
GB	6
CA	6
DE	5
CH	5
RO	4
EG	3
MX	3
CR	2
RU	2

Wie viele sind peer-reviewed und double blind peer-reviewed?

Double blind peer review	114
Blind peer review	31
Peer review	28
Editorial review	6
Open peer review	3
Committee review	1

Wie sieht es mit Publikationsgebühren aus? Wieviele verlangen Publikationsgebühren?

apc.has_apc	
False	170
True	13

Sind unter den Herausgebern vielleicht auch namhafte Herausgeber wie de Gruyter?

Die Abfrage mit query hat nicht das gewünschte Ergebnis geliefert. Da die Analyse der Verlagslandschaft nicht geglückt ist, konnte nur über die Test API auf der Website ermittelt werden, dass es lediglich 1 de Gruyter Titel gibt.

[Nadine Salz, 26.8.22, Data Librarian 2022]

Request URL

https://doaj.org/api/search/journals/bibjson.publisher%3Ade%20Gruyter

Response Body


```
{
  "total": 0,
  "page": 1,
  "pageSize": 10,
  "timestamp": "2022-08-25T22:51:22.327380Z",
  "query": "bibjson.publisher:de Gruyter",
  "results": [],
  "last": "https://doaj.org/api/search/journals/bibjson.publisher:de%20Gruyter?page=1&pageSize=10"
}
```


Response Code

Folgende beide Fragen wurden zusätzlich beleuchtet

In welchem Jahr wurden die meisten OA Starts registriert?

2016

Betrachtung OA Start

Der OA Start reicht von 1993 -2020

Die Ergebnisse stützen dennoch im Wesentlichen die Ergebnisse der anderen Studien.

Im Hinblick auf die zu Eingang gewählte Fragestellung, wie qualitativ, frei wiederverwendbar und wie kostenfrei die verzeichneten OA Quellen sind, lässt sich feststellen, dass sowohl die große Anzahl an double-blind peer-reviewed Verfahren als auch die große Anzahl an CC BY Lizenzen auf eine hervorzuhebende Qualität sowie die gelungene Übertragung einer „echten“ Open Access „Bewegung“ im LIS Bereichs deuten.

Diskussion

Die folgende Diskussion evaluiert das Projekt, beleuchtet den Erhebungsprozess, geht auf methodische und praktische Limitierungen ein, die sich im Laufe der Arbeit ergeben haben. Abschließend werden Empfehlungen für weitere Forschungsvorhaben gegeben.

Projektevaluation

Die Themenfindung, der Export und die finale Analyse werden in dieser Projektarbeit nicht getrennt nacheinander, sondern gegenseitig bedingend zusammengesetzt: Die Prüfung der Datenauswahl nimmt dabei einen großen Platz ein, um die richtige Basis für die anschließende Analyse zu extrahieren. In diesem Projekt ist das Datenset durch die Vordefinition der API-Schnittstelle bereits sehr gut ausgestaltet, sodass einige Bereinigungsschritte eingespart werden können. Das Vereinfacht objektiv das Handling des Datensatzes, subjektiv ist es dennoch nicht weniger einfach, die richtigen Abfragen zu formulieren und in Code umzusetzen. Bei Ansteuerung der API wird auch sehr deutlich: die Vorarbeit ist entscheidend. Die Auswahl der richtigen Datenquelle in Form des konkreten Teilbereichs hier „Bibliography. Library Science. Information Resources“ innerhalb des DOAJ sowie die Ausgestaltung der API-Anfrage macht deswegen einen Großteil der Vorarbeit aus.

Die Evaluation des Forschungsstandes anderer Studien hat bei diesem Projekt Hinweise darauf gegeben, dass die meisten Abfragen „richtig“ oder zumindest ähnlich gestellt und interpretiert sein könnten.

Erhebungsprozess

Eine Datenanalyse mit Pandas ermöglicht eine Vielzahl von Abfragen. Ohne die Struktur der übertragenden Daten zu verstehen, lassen sich die Daten allerdings weder abfragen noch ansteuern. Es besteht außerdem schnell die Gefahr die Abfrageergebnisse falsch zu deuten. Bei der erfolgten Abfrage in welchen Sprachen publiziert wird, ist eine eindeutige Interpretation daher z.B. nicht möglich und einige Antworten konnten nicht gefunden werden, weil die query() Anfrage nur Fehler produzierte.

Grenzen

Es ist maßgeblich, welche Daten in welcher Form zur Verfügung stehen und wie aufbereitet sie bereits sind. Die Grenzen zeigen sich an der Stelle auf an der z.B. eine API -Schnittstelle schlecht dokumentiert ist. Auch wenn genutzte Datenbibliotheken wie Pandas im Jupyter Notebook bei Nutzung der „import panda as pd – Funktion“ bereits veraltet sind, zeigen sich Grenzen in der Anwendbarkeit auf, an die man zunächst vielleicht nicht denkt.

Eine weiterer Punkt ist die Sorgfalt: Es muss sichergestellt sein, dass nicht nur alle gewünschten Daten übertragen werden, sondern anschließend auch im gleichen Dataframe korrekt vorliegen. Dies unterscheidet sich maßgeblich von einfachen Trefferlisten einer regulären Datenbankabfrage bei denen der Datensatz wesentlich - sicher auch aufgrund von Erfahrungswerten - schneller erfasst werden kann.
Das klingt recht logisch und simpel, wird aber bei der Arbeit mit Python und den entsprechenden Bibliotheken dadurch eminent, dass man den Datensatz auch nicht so ohne Weiteres „sieht“, sondern nur anhand von Abfragebefehlen eigentlich sprichwörtlich mehr „ertastet“ erfordert eine gute Sicherheit in den genutzten Werkzeugen.

Des Weiteren ist der große Vorteil, nämlich der, dass eine größere Freiheit gegeben ist, sich einen beliebigen Datenausschnitt auszuwählen, gleichzeitig auch ein Nachteil, denn die Bereinigung der Daten und Zusammenstellung erfordert dann wesentlich mehr Aufwand. Durch die größere Freiheit der Datenkombination entsteht allerdings gleichsam mehr wissenschaftliche Forschungsfreiheit hinsichtlich der Gegenüberstellung von Daten und damit einhergehendem Erkenntnisgewinn.
Ein Vorteil bei einem bereits recht gut aufbereiteten Datensatz ist sicherlich grundsätzlich etwas mehr Sicherheit eine valide Grundgesamtheit erhalten zu haben, wie in dem vorliegenden Beispiel. gleichzeitig schränkt es jedoch ein, weil der Anbieter der API-Schnittstelle so einen wesentlichen Einfluss auf die Forschungsergebnisse haben kann. Bei Fragen, die der die in dieser Arbeit analysiert wird, fällt das sicherlich nicht so groß ins Gewicht, bei anderen Forschungsfragen sicherlich schon.

Empfehlungen für weiterführende Forschung

Es könnten weiterführende Abhängigkeiten analysiert werden, wenn z.B. noch eine weitere Datenquelle wie Scopus in eine Studie hinzugezogen würde und so bibliometrische Analyseergebnisse mit der Abfrage im DOAJ kombiniert würden. Mögliche Forschungsfragen dabei wären: Werden double-blind peer-reviewed Journals beispielsweise eher zitiert, welche Sprache wird häufiger zitiert, sollte das einen Einfluss auf die Annahme der Manuskripte

haben. Wie produktiv sind die Autoren / Affiliation auch zeitgleich für nicht OA Zeitschriften. Wie verbreitet ist im Bibliothekswesen die Tendenz anderer Wissenschaftsdisziplinen (z.B. Wirtschaftswissenschaften) von und in besonders Namhaften Zeitschriften zitiert werden zu wollen oder wie sind die OA Zeitschriften des DOAJ in Scopus gerankt.

Literaturverzeichnis

Website: DOAJ (Stand 15.7.2022)
DOAJ Schnittstellendokumentation: DOAJ – API Schnittstelle (Stand 15.7.2022)

Narayan, R., Pati, Pritam K., Sahoo, S. (2021). Growth of Open Access Literature on Library and Information Science during 2011-2020: A Scientometrics Analysis. *Library Philosophy and Practice.* (e-journal).

Rajkumar, T., Jeyapragash, B. (2021). Contributions of Open Access Journals in Library and Information Science indexed in SCOPUS Database: A Metric Study. *Library Philosophy and Practice* (e-journal). 6181.

Sahoo, J., Birtia, T., & Mohanty, B. (2017). Open access journals in library information science: A study on DOAJ. *International Journal of Information Dissemination and Technology*, *7*(2), 116 https://doi.org/10.5958/2249-5576.2017.00008.5

Selvam, M. and Amudha, G. (2020). "A Bibliometric Study on open Access Library and Information Science Journals in DOAJ". *Library Philosophy and Practice (e-journal).* 4868. https://digitalcommons.unl.edu/libphilprac/4868

Negi Dheeraj Singh. (2019). Library & Information Science Journals in DOAJ: A Bibliometric Study. *Library Herald.* 57(3).

Chakravarty, R. , Diksha C. (2020). Status of Open Access LIS Journals: An Empirical Study Of DOAJ. *Journal of Indian Library Association, 56 (3).* 88-99. 10.5281/zenodo.4061011

Sarasu R., Kairali, A., Sayed M.a.J. (2020). Open Access Journals in Library and Information Science: A Study based on DOAJ.

Anhang

Genaue Spaltennamen

```
Out[78]: ['id',
          'created_date',
          'last_updated',
          'admin',
          'boai',
          'eissn',
          'publication_time_weeks',
          'title',
          'oa_start',
          'keywords',
          'language',
          'license',
          'subject',
          'apc.has_apc',
          'apc.url',
          'article.license_display_example_url',
          'article.license_display',
          'copyright.author_retains',
          'copyright.url',
          'deposit_policy.has_policy',
          'editorial.review_url',
          'editorial.board_url',
          'editorial.review_process',
          'institution.name',
          'other_charges.has_other_charges',
          'pid_scheme.has_pid_scheme',
          'pid_scheme.scheme',
          'plagiarism.detection',
          'plagiarism.url',
          'preservation.has_preservation',
          'publisher.name',
          'publisher.country',
          'ref.oa_statement',
          'ref.journal',
          'ref.aims_scope',
          'ref.author_instructions',
          'ref.license_terms',
          'waiver.has_waiver',
          'article.orcid',
          'article.i4oc_open_citations',
          'deposit_policy.service',
          'preservation.url',
          'preservation.service',
          'alternative_title',
          'pissn',
          'other_charges.url',
          'preservation.national_library',
          'deposit_policy.url',
          'institution.country',
          'apc.max',
          'waiver.url',
          'replaces']
```

Eine Analyse des LIS Open Access Publikationsgeschehen anhand der API Schnittstelle des Directory of Open Access Journals

Ziel der Arbeit: über die vom DOAJ angebotene API Schnittstelle sollen Daten zu Open Access Publikationen aus dem Bereich Bibliography.Library Science.Information Resources.extrahiert und mit dem Forschungsstand verglichen werden.

Datensatzgröße: 183 Zeitschriften ; Anzahl verzeichnete Zeitschriften Gesamt 18.141 ; Datum :26.8.2022 ; Autor: Nadine Salz ; Abschlussarbeit Zertifkatskurs Data Librarian 2022

Gesamte Trefferliste importieren & Ansprache der API Schnittstelle

Import der benötigen Bibliotheken sowie Abfrage des Datensatzes mit größtmögliche zu ladenden Trefferzahl

urllib .request lädt die gewünschte URL

import json importiert die json Bibliothek

In [382…

```
import urllib.request
import json
```

Die for Loop gibt mit range (2) die Zahlen 0,1 und 2 aus. Die zu ladenen Seiten sind aber 1 und 2, daher +1. Danach werden diese beiden Seiten mit append() zusammengeführt der gewählte DOAJ Datensatz dazu wird dazu aus dem Netz geladen und mit json.loads() als Dictionary ausgegeben.

In [460…

```
pages = range(182)
doaj_dataset_json = []
for pagenumber in pages:
    real_pagenumber = pagenumber + 1

    # Ausgabe der eigentlichen Seitenzahl wäre an dieser Stelle:
    #print(real_pagenumber)
    #base_url = "https://doaj.org/api/search/journals/*?page=" + str(real_

    base_url = "https://doaj.org/api/search/journals/bibjson.subject.term:/
    print(base_url)
    doaj_dataset = urllib.request.urlopen(base_url).read()
    print(pagenumber)
    doaj_dataset_json.append(json.loads(doaj_dataset))
```

```
https://doaj.org/api/search/journals/bibjson.subject.term:/%22Bibliograph
y.%20Library%20science.%20Information%20resources/%22?page=1&pageSize=100
0
https://doaj.org/api/search/journals/bibjson.subject.term:/%22Bibliograph
y.%20Library%20science.%20Information%20resources/%22?page=2&pageSize=100
```

1

Anzeige des Datensatzes zur Analyse des Aussehens mit json(): Die Methode erzeugt ein Javaskript Objekt an der Stelle 1

In [467…

```
doaj_dataset_json[1]["results"]
```

Out[467…

```
[{'id': '94b5aafddf514923802d2372768f30b2',
  'created_date': '2021-08-26T03:21:19Z',
  'last_updated': '2022-03-21T20:00:40Z',
  'bibjson': {'boai': True,
   'eissn': '2447-0198',
   'publication_time_weeks': 12,
   'title': 'Revista Informação na Sociedade Contemporânea',
   'oa_start': 2014,
   'apc': {'has_apc': False,
    'url': 'https://periodicos.ufrn.br/informacao/polacessaberto'},
   'article': {'orcid': False,
    'i4oc_open_citations': False,
    'license_display': ['No']},
   'copyright': {'author_retains': True,
    'url': 'https://periodicos.ufrn.br/informacao/polacessaberto'},
   'deposit_policy': {'has_policy': True,
    'url': 'http://diadorim.ibict.br/handle/1/1611',
    'service': ['Diadorim']},
   'editorial': {'review_url': 'https://periodicos.ufrn.br/informacao/avapo
rpares',
    'board_url': 'https://periodicos.ufrn.br/informacao/about/editorialTeam
',
    'review_process': ['Double blind peer review']},
   'institution': {'name': 'Universidade Federal do Rio Grande do Norte',
    'country': 'BR'},
   'other_charges': {'has_other_charges': False},
   'pid_scheme': {'has_pid_scheme': True, 'scheme': ['DOI']},
   'plagiarism': {'detection': False, 'url': ''},
   'preservation': {'has_preservation': False},
   'publisher': {'name': 'Programa de Pós-graduação em Gestão da Informação
e do Conhecimento/Departamento de Ciência da Informação',
    'country': 'BR'},
   'ref': {'oa_statement': 'https://periodicos.ufrn.br/informacao/polacessa
berto',
    'journal': 'https://periodicos.ufrn.br/informacao/',
    'aims_scope': 'https://periodicos.ufrn.br/informacao/about',
    'author_instructions': 'https://periodicos.ufrn.br/informacao/about/sub
missions',
    'license_terms': 'https://periodicos.ufrn.br/informacao/polacessaberto
'},
   'waiver': {'has_waiver': False},
   'keywords': ['information science',
    'library science',
    'knowledge management',
    'information management',
    'scientific communication'],
   'language': ['PT', 'ES', 'EN'],
   'license': [{'type': 'CC BY-NC',
     'BY': True,
     'NC': True,
     'ND': False,
     'SA': False,
     'url': 'https://creativecommons.org/licenses/by-nc/4.0/'}],
   'subject': [{'code': 'CD1-6471',
     'scheme': 'LCC',
```

```
   'other_charges': {'has_other_charges': False},
   'pid_scheme': {'has_pid_scheme': True, 'scheme': ['DOI']},
   'plagiarism': {'detection': False, 'url': ''},
   'preservation': {'has_preservation': True,
    'url': 'https://portal.issn.org/resource/ISSN/2398-4112',
    'service': ['CLOCKSS', 'LOCKSS', 'PKP PN']},
   'publisher': {'name': 'University of Victoria Libraries', 'country': 'CA
'},
   'ref': {'oa_statement': 'https://kula.uvic.ca/index.php/kula/about',
    'journal': 'https://kula.uvic.ca/',
    'aims_scope': 'https://kula.uvic.ca/index.php/kula/about',
    'author_instructions': 'https://kula.uvic.ca/index.php/kula/about/submi
ssions',
    'license_terms': 'https://kula.uvic.ca/index.php/kula/about'},
   'waiver': {'has_waiver': False},
   'keywords': ['human knowledge processes',
    'knowledge creation',
    'knowledge dissemination',
```

die for Loop gibt eine Liste mit Namen all_results_list mit der eingelesenen Ergebnisstrefferliste als Datensatz aus

In [385…
```
all_results_list = []

for result_page in pages:
    #print(result_page["results"])
    all_results_list += doaj_dataset_json[result_page]["results"]
```

len() gibt die Anzahl der Treffer dieser Liste wieder, angwendet um zu prüfen das alle Zeitschriften importiert wurden. Dies ist mit 183 der Fall

In [386…
```
len(all_results_list)
```

Out[386…
183

Mit der Methode json.dumps wird das Format der erzeugten Liste geändert zu einem JSON string.

In [387…
```
all_results_json = json.dumps(all_results_list)
type(all_results_json[0])
```

Out[387…
str

In [23]:
```
#all_results_json = json.loads(all_results_json)
```

Überprüfung ob das erfolgt ist mit er Methode type() . Die Angabe str bestätigt es.

In [465…
```
type(all_results_json[0])
```

Out[465…
str

Anzeige der Trefferliste an der Stelle 0

In [463…
```
all_results_list[0]
```

Out[463…


```
{'id': '0008a8877b2046b082ef902b2df8647c',
 'created_date': '2020-07-30T21:49:52Z',
 'last_updated': '2022-03-15T17:02:49Z',
 'bibjson': {'boai': True,
  'eissn': '2676-7104',
  'publication_time_weeks': 8,
  'title': 'Frontiers in Health Informatics',
  'oa_start': 2019,
  'apc': {'has_apc': False,
   'url': 'http://ijmi.ir/index.php/IJMI/about/editorialPolicies#custom-3
'},
  'article': {'license_display_example_url': 'http://ijmi.ir/index.php/IJMI
/about/editorialPolicies#custom-4',
   'license_display': ['Embed']},
  'copyright': {'author_retains': True,
   'url': 'http://ijmi.ir/index.php/IJMI/about/submissions#copyrightNotice
'},
  'deposit_policy': {'has_policy': False},
  'editorial': {'review_url': 'http://ijmi.ir/index.php/IJMI/about/editoria
lPolicies#peerReviewProcess',
   'board_url': 'http://ijmi.ir/index.php/IJMI/about/editorialTeam',
   'review_process': ['Double blind peer review']},
  'institution': {'name': 'Iranian Association of Medical Informatics'},
  'other_charges': {'has_other_charges': False},
  'pid_scheme': {'has_pid_scheme': True, 'scheme': ['DOI']},
  'plagiarism': {'detection': True,
   'url': 'http://ijmi.ir/index.php/IJMI/about/editorialPolicies#custom-2
'},
  'preservation': {'has_preservation': False},
  'publisher': {'name': 'Hamara Afzar', 'country': 'IR'},
  'ref': {'oa_statement': 'http://ijmi.ir/index.php/IJMI/about/editorialPol
icies#openAccessPolicy',
   'journal': 'http://www.ijmi.ir',
   'aims_scope': 'http://ijmi.ir/index.php/IJMI/about/editorialPolicies#foc
usAndScope',
   'author_instructions': 'http://ijmi.ir/index.php/IJMI/about/submissions#
authorGuidelines',
   'license_terms': 'http://ijmi.ir/index.php/IJMI/about/editorialPolicies#
custom-4'},
  'waiver': {'has_waiver': False},
  'keywords': ['medical informatics',
   'image processing',
   'decision support systems',
   'artificial neural networks'],
  'language': ['EN'],
  'license': [{'type': 'CC BY-NC',
    'BY': True,
    'NC': True,
    'ND': False,
    'SA': False,
    'url': 'https://creativecommons.org/licenses/by-nc/4.0/'}],
  'subject': [{'code': 'Z',
    'scheme': 'LCC',
    'term': 'Bibliography. Library science. Information resources'},
   {'code': 'R858-859.7',
    'scheme': 'LCC',
    'term': 'Computer applications to medicine. Medical informatics'}]},
 'admin': {'seal': False, 'ticked': True}}
```


For loop zur Ausgabe der Spalte der Schlagworte, um zu sehen das auch die gewünschten Datensätze aus der Kategorie Bibliography. Library Science. Information resources ausgegeben wurden.Dies scheint der Fall.

In [389…

```
for item in all_results_list:
    print(item["bibjson"]["subject"][0]["term"])
```

```
Bibliography. Library science. Information resources
Museums. Collectors and collecting
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Academies and learned societies
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Communication. Mass media
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Arts in general
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Law
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Management information systems
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Information resources (General)
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Information resources (General)
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Information resources (General)
Auxiliary sciences of history
History of scholarship and learning. The humanities
Information resources (General)
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Information resources (General)
Bibliography. Library science. Information resources
Information resources (General)
Bibliography. Library science. Information resources
Communication. Mass media
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
Bibliography. Library science. Information resources
```

Analyse mit Hilfe von Pandas

Import der Pandasbibliothek und lese mit pd.read_json die Trefferliste ein

In [390…

```
import pandas as pd
dataframe = pd.read_json(all_results_json)
```

Zeige mit head() den Kopf des Dataframes Es zeigt sich das bibjson nicht in Spalten anzeigt wird.

In [277…

```
dataframe.head()
```

Out[277…

	id	created_date	last_updated	bibjs
0	0008a8877b2046b082ef902b2df8647c	2020-07-30T21:49:52Z	2022-03-15T17:02:49Z	{'boai': Tr 'eis '2676-71('publica
1	0059b3e34a0d4e5f809b8c1096dd42f5	2019-08-03T23:42:36Z	2021-04-29T13:18:58Z	{'boai': Tr 'eis '2358-07('publica
2	02f38a58853d4a7ea153c1d0383ba9ad	2010-11-09T14:35:09Z	2021-04-29T13:14:05Z	{'alternative_ti 'Journa Library and
3	05847ba5e1ef4386a656a83f4b86999e	2020-01-21T18:11:41Z	2021-04-29T13:19:55Z	{'alternative_ti 'Library Journ 'boa
4	0599ed873dd440e8a66779bca759a448	2018-05-08T23:58:58Z	2021-04-29T13:13:55Z	{'alternative_ti 'Insaniyat', 'bo Tr

Um die Liste in der Spalte bibjson in weitere Spalten aufzuteilen, um den Datensatz dann weiter anzusteuern, wird mit concat und normalize ein bestimmtes Stück des gleichen Dataframes (=bibjson) in die Tabelle nebeneinder eingefügt. pd.concat() verbindet dabei Objekte miteinander und gibt es in einem Datenframe aus. normalize() vereinheitlich die Werte der Zeile für eine saubere Ausgabe. Mit axis =1 werden die beiden Datensätze nicht untereinander zusammengefügt sondern nebeneinander. Die Spalten vom 2. Datensatz erscheinen dadurch neben dem 1. Datensatz. drop()entfernt die ursprüngliche bibjson Spalte.

In [391…

```
from pandas import json_normalize

df = pd.concat([pd.DataFrame(dataframe),
                json_normalize(dataframe['bibjson'])],
                axis=1).drop('bibjson', 1)
```

```
/tmp/ipykernel_102839/397273336.py:3: FutureWarning: In a future version of
pandas all arguments of DataFrame.drop except for the argument 'labels' wil
l be keyword-only.
  df = pd.concat([pd.DataFrame(dataframe),
```

df gibt das bearbeitete Datenframe aus

In [279…

```
df
```

Out[279…

	id	created_date	last_updated	admin	bo
0	0008a8877b2046b082ef902b2df8647c	2020-07-30T21:49:52Z	2022-03-15T17:02:49Z	{'seal': False, 'ticked': True}	Tr
1	0059b3e34a0d4e5f809b8c1096dd42f5	2019-08-03T23:42:36Z	2021-04-29T13:18:58Z	{'seal': False, 'ticked': True}	Tr
2	02f38a58853d4a7ea153c1d0383ba9ad	2010-11-09T14:35:09Z	2021-04-29T13:14:05Z	{'seal': False, 'ticked': True}	Tr
3	05847ba5e1ef4386a656a83f4b86999e	2020-01-21T18:11:41Z	2021-04-29T13:19:55Z	{'seal': False, 'ticked': True}	Tr
4	0599ed873dd440e8a66779bca759a448	2018-05-08T23:58:58Z	2021-04-29T13:13:55Z	{'seal': False, 'ticked': True}	Tr
...	...	...	...	...	
178	f73dfa7769bf417b900143ca4d8951a1	2013-06-23T18:20:28Z	2021-04-29T13:14:36Z	{'seal': False, 'ticked': True}	Tr
179	f7f587617c134cd99447c6ff8aa6c64a	2006-12-21T14:32:17Z	2021-08-16T15:35:08Z	{'seal': False, 'ticked': True}	Tr
180	f936908299e446c1b5c9d8c108cc596f	2019-10-20T03:40:44Z	2021-04-29T13:17:58Z	{'seal': False, 'ticked': True}	Tr
181	fb979e39ff8444eaa7aed77df8ed9169	2017-07-03T13:16:11Z	2021-04-29T13:18:35Z	{'seal': False, 'ticked': True}	Tr
182	fd40919b9baf458f8cff4d11a1f294eb	2019-03-06T16:29:15Z	2021-04-29T13:19:46Z	{'seal': True, 'ticked': True}	Tr

183 rows × 52 columns

list() erzeugt eine Liste der Spalten. Ziel ist zu schauen, wie die Bezeichnung der Objekte lautet von denen, die betrachtet werden sollen.

In [392…
```
list(df.columns)
```

Out[392…
```
['id',
 'created_date',
 'last_updated',
 'admin',
 'boai',
 'eissn',
 'publication_time_weeks',
 'title',
 'oa_start',
 'keywords',
 'language',
 'license',
 'subject',
 'apc.has_apc',
 'apc.url',
 'article.license_display_example_url',
 'article.license_display',
 'copyright.author_retains',
 'copyright.url',
 'deposit_policy.has_policy',
 'editorial.review_url',
 'editorial.board_url',
 'editorial.review_process',
 'institution.name',
 'other_charges.has_other_charges',
 'pid_scheme.has_pid_scheme',
 'pid_scheme.scheme',
 'plagiarism.detection',
 'plagiarism.url',
 'preservation.has_preservation',
 'publisher.name',
 'publisher.country',
 'ref.oa_statement',
 'ref.journal',
 'ref.aims_scope',
 'ref.author_instructions',
 'ref.license_terms',
 'waiver.has_waiver',
 'article.orcid',
 'article.i4oc_open_citations',
 'deposit_policy.service',
 'preservation.url',
 'preservation.service',
 'alternative_title',
 'pissn',
 'other_charges.url',
 'preservation.national_library',
 'deposit_policy.url',
 'institution.country',
 'apc.max',
 'waiver.url',
 'replaces']
```

Anzeige der zu betrachtenden Spalten mit:

In [393…
```
df[["id", "oa_start","language" , "license", "editorial.review_process" ,
```

Out[393…

	id	oa_start	language	license	editorial.review_process	

	id	oa_start	language	license	editorial.review_process
0	0008a8877b2046b082ef902b2df8647c	2019	[EN]	[{'type': 'CC BY-NC', 'BY': True, 'NC': True, ...	[Double blind peer review]
1	0059b3e34a0d4e5f809b8c1096dd42f5	2014	[PT]	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...	[Double blind peer review]
2	02f38a58853d4a7ea153c1d0383ba9ad	1975	[ZH, EN]	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...	[Double blind peer review]
3	05847ba5e1ef4386a656a83f4b86999e	2012	[UK, EN]	[{'type': 'CC BY-NC', 'BY': True, 'NC': True, ...	[Double blind peer review]
4	0599ed873dd440e8a66779bca759a448	2016	[AR, EN]	[{'type': 'CC BY-NC-SA', 'BY': True, 'NC': Tru...	[Blind peer review]
...	...	...	...	...	...
178	f73dfa7769bf417b900143ca4d8951a1	2011	[EN]	[{'type': 'CC BY-NC-ND', 'BY': True, 'NC': Tru...	[Peer review]
179	f7f587617c134cd99447c6ff8aa6c64a	2006	[EN]	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...	[Double blind peer review]
180	f936908299e446c1b5c9d8c108cc596f	2012	[PT]	[{'type': 'CC BY-NC-ND', 'BY': True, 'NC': Tru...	[Double blind peer review]

	id	oa_start	language	license	editorial.review_process
181	fb979e39ff8444eaa7aed77df8ed9169	2003	[EN, ID]	[{'type': 'CC BY-SA', 'BY': True, 'NC': False,...	[Double blind peer review]
182	fd40919b9baf458f8cff4d11a1f294eb	2017	[EN]	[{'type': 'CC BY', 'BY': True, 'NC': False,	[Editorial review, Double blind peer review]

Bereinigung von NAN Werten

Prüfen, ob in den zu betrachteten Spalten NAN Werte vorkommen, um sie ggf zu füllen oder löschen.

mit copy() Kopie erstellt für den Fall, dass Werte gelöscht oder mit unknown gefüllt hätten müssen, um den ursprünglichen Datensatz zur Sicherheit zu erhalten. Da keine NAN Werte vorhanden sind, kann dieser Schritt ausgelassen werden.

In [394…
```
df_Kopie = df.copy()
```

isnull().sum() Summiere die Anzahl der Zeilen, wo keine NAN Werte angegeben werden. Der ISSN sind also 9 Zeilen mit NAN Werten zu finden.

In [395…
```
df_Kopie.isnull().sum()
```

Out[395…
```
id                                     0
created_date                           0
last_updated                           0
admin                                  0
boai                                   0
eissn                                  9
publication_time_weeks                 0
title                                  0
oa_start                               0
keywords                               0
language                               0
license                                0
subject                                0
apc.has_apc                            0
apc.url                                0
article.license_display_example_url    82
article.license_display                0
copyright.author_retains               0
copyright.url                          33
deposit_policy.has_policy              0
editorial.review_url                   0
editorial.board_url                    0
editorial.review_process               0
institution.name                       90
other_charges.has_other_charges        0
pid_scheme.has_pid_scheme              0
pid_scheme.scheme                      54
```

```
plagiarism.detection                  0
plagiarism.url                       61
preservation.has_preservation         0
publisher.name                        0
publisher.country                     0
ref.oa_statement                      0
ref.journal                           0
ref.aims_scope                        0
ref.author_instructions               0
ref.license_terms                     1
waiver.has_waiver                     0
article.orcid                       110
article.i4oc_open_citations         110
deposit_policy.service              100
preservation.url                     87
preservation.service                127
alternative_title                    90
pissn                               101
other_charges.url                    83
preservation.national_library       165
deposit_policy.url                  148
institution.country                 166
apc.max                             170
waiver.url                          177
replaces                            181
dtype: int64
```

Dublettenprüfung

Überprüfung zur Sicherheit ob es ggf. Dubletten geben könnte, die rausgezogen werden müssten. Ist nicht der Fall

dublicated(subset) übergibt angegebene Spalte, in dem Fall id zur Dublettenprüfung. ID gewählt, da die einzigartig ist und eine gesicherte Aussage dazu gibt ob eine Dublette enthalten ist. Es ist keine Dublette vorhanden.

In [311…

```
df_Kopie_Dubletten = df_Kopie[df_Kopie.duplicated(subset = "id")]
df_Kopie_Dubletten
```

Out[311…

id	created_date	last_updated	admin	boai	eissn	publication_time_weeks	title	oa_start	ke

0 rows × 52 columns

Start der Datenanalyse und Beantwortung der Fragen

Betrachtung OA Start: Wann ist der erste und wann der letzte OA Start?

.loc gibt Datenindexwert an. Zeigt aus den "Spalten" 0-183 die Werte aus der Spalte Open Access start an. Die Liste ist noch unsortiert.

In [142…
```
df_Kopie.loc[:, ["oa_start"]]
```

Out[142…

	oa_start
0	2019
1	2014
2	1975
3	2012
4	2016
...	...
178	2011
179	2006
180	2012
181	2003
182	2017

183 rows × 1 columns

sort_values sortiert die Werte der Spalte Open Access start

In [461…
```
df_Kopie.sort_values(by=["oa_start"])
```

Out[461…

	id	created_date	last_updated	admin	b
177	f69de99a68644f7b8aa73b2343cc28a4	2013-10-29T08:51:00Z	2021-08-24T14:27:46Z	{'seal': False, 'ticked': True}	Tı
171	f24d1935251c4802993ef202f1124ede	2006-11-06T14:49:32Z	2021-04-29T13:14:02Z	{'seal': False, 'ticked': True}	Tı
2	02f38a58853d4a7ea153c1d0383ba9ad	2010-11-09T14:35:09Z	2021-04-29T13:14:05Z	{'seal': False, 'ticked': True}	Tı
142	d47bf74be3c44b849d6f73f738b71a52	2015-05-21T11:05:32Z	2021-04-29T13:18:43Z	{'seal': False, 'ticked': True}	Tı
92	8d442f2ade5c41919bf471bbdd7b28d0	2017-09-26T22:33:21Z	2021-04-29T13:17:18Z	{'seal': False, 'ticked': True}	Tı
...	...	...	...	...	
0	0008a8877b2046b082ef902b2df8647c	2020-07-30T21:49:52Z	2022-03-15T17:02:49Z	{'seal': False, 'ticked': True}	Tı

	id	created_date	last_updated	admin	b
123	af3864c617b346578fbc410831f8abb5	2020-10-12T07:42:20Z	2021-04-29T13:19:50Z	{'seal': False, 'ticked': True}	Tı
145	d8fee705be2949e2a84e7cb6d97b87f9	2021-12-15T17:10:01Z	2021-12-15T17:10:01Z	{'seal': False, 'ticked': True}	Tı
21	1a7136a04b8d482da3a650b08d0e5596	2022-07-06T15:02:25Z	2022-07-06T15:02:25Z	{'seal': False, 'ticked': True}	Tı
6	07a8b9e848a7462c91952373e9a10313	2020-07-15T12:57:16Z	2021-04-29T13:14:29Z	{'seal': False, 'ticked':	Tı

Antwort: Der OA Start reicht von 1939 bis 2020

In welchem Jahr wurden die meisten OA Starts registriert?

groupby() und count() Gruppieren und zählen die häufigkeit der Werte der Spalte Open Access und zeigt an, wieviele Zeitschriften im Jahr 1993 zu OA Journals wurden.

In [152…
```
df_Kopie.groupby("oa_start").oa_start.count()
```

Out[152…
```
oa_start
1939     1
1942     1
1975     1
1976     1
1980     1
1981     1
1986     2
1987     1
1989     1
1992     1
1993     1
1995     3
1996     2
1997     1
1998     4
1999     4
2000     1
2001     1
2002     2
2003     6
2004     5
2005     4
2006     5
2007     9
2008    11
2009     5
2010     4
```

```
2011    14
2012    10
2013    12
2014    13
2015    10
2016    15
2017     7
2018    13
2019     6
2020     4
```

Antwort: 2016 wurden die meisten Zeitschriften als OA Journals herausgegeben.

Mit dtypes können die Art der Objekte angezeigt werden. Wenn Fehlerabfragen darauf hinweisen, dass der falsche Typ verwendet wird.Da dies Häufig passiert ist, wird es an dieser Stelle geprüft.

In [167…

```
df_Kopie.dtypes
```

Out[167…

```
id                                    object
created_date                          object
last_updated                          object
admin                                 object
boai                                    bool
eissn                                 object
publication_time_weeks                 int64
title                                 object
oa_start                               int64
keywords                              object
language                              object
license                               object
subject                               object
apc.has_apc                             bool
apc.url                               object
article.license_display_example_url   object
article.license_display               object
copyright.author_retains                bool
copyright.url                         object
deposit_policy.has_policy               bool
editorial.review_url                  object
editorial.board_url                   object
editorial.review_process              object
institution.name                      object
other_charges.has_other_charges         bool
pid_scheme.has_pid_scheme               bool
pid_scheme.scheme                     object
plagiarism.detection                    bool
plagiarism.url                        object
preservation.has_preservation           bool
publisher.name                        object
publisher.country                     object
ref.oa_statement                      object
ref.journal                           object
ref.aims_scope                        object
ref.author_instructions               object
ref.license_terms                     object
```

```
waiver.has_waiver                        bool
article.orcid                          object
article.i4oc_open_citations            object
deposit_policy.service                 object
preservation.url                       object
preservation.service                   object
alternative_title                      object
pissn                                  object
other_charges.url                      object
preservation.national_library          object
deposit_policy.url                     object
institution.country                    object
apc.max                                object
waiver.url                             object
replaces                               object
```

Welche der OA Lizenzen werden vornehmlich vergeben?

loc() Zeigt den Inhalt von der "Spalte" Lizenz an. Es wird deutlich dass die Daten verschachtelt sind. So kann keine Zählung vorgenommen werden.

In [344…
```
df_Kopie.loc[:, ["license"]]
```

Out[344…

	license
0	[{'type': 'CC BY-NC', 'BY': True, 'NC': True, ...
1	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...
2	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...
3	[{'type': 'CC BY-NC', 'BY': True, 'NC': True, ...
4	[{'type': 'CC BY-NC-SA', 'BY': True, 'NC': Tru...
...	...
178	[{'type': 'CC BY-NC-ND', 'BY': True, 'NC': Tru...
179	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...
180	[{'type': 'CC BY-NC-ND', 'BY': True, 'NC': Tru...
181	[{'type': 'CC BY-SA', 'BY': True, 'NC': False,...
182	[{'type': 'CC BY', 'BY': True, 'NC': False, 'N...

183 rows × 1 columns

mit der for loop und append sowie value_counts() wird eine Liste erzeugt, die die gewählten Objekte in die gewünschte Form bringt. Wichtig dabei ist der genaue "Ort": bibjson license 0 type" entspricht daher der Struktur die in der schriftlichen Ausarbeitung beigefügt ist.

In [471…
```
list_license = []
for item in all_results_list :
    list_license.append(item["bibjson"]["license"][0]["type"])
dataframe_license = pd.DataFrame(list_license)
dataframe_license.value_counts()
```

Out[471…
```
CC BY                    81
```

```
CC BY-NC                    36
CC BY-NC-ND                 31
CC BY-NC-SA                 17
CC BY-SA                    13
CC BY-ND                     3
Publisher's own license      2
dtype: int64
```

Antwort: Die am Häufigsten vergebenen Lizenzen sind CC BY Lizenzen

groupby(level=0).max() Groupiert Daten nach Größe nach Angabe

In [470…

Out[470…

	id	created_date	last_updated	admin	bo
0	0008a8877b2046b082ef902b2df8647c	2020-07-30T21:49:52Z	2022-03-15T17:02:49Z	{'seal': False, 'ticked': True}	Tr
1	0059b3e34a0d4e5f809b8c1096dd42f5	2019-08-03T23:42:36Z	2021-04-29T13:18:58Z	{'seal': False, 'ticked': True}	Tr
2	02f38a58853d4a7ea153c1d0383ba9ad	2010-11-09T14:35:09Z	2021-04-29T13:14:05Z	{'seal': False, 'ticked': True}	Tr
3	05847ba5e1ef4386a656a83f4b86999e	2020-01-21T18:11:41Z	2021-04-29T13:19:55Z	{'seal': False, 'ticked': True}	Tr
4	0599ed873dd440e8a66779bca759a448	2018-05-08T23:58:58Z	2021-04-29T13:13:55Z	{'seal': False, 'ticked': True}	Tr
...	...	...	...	...	
178	f73dfa7769bf417b900143ca4d8951a1	2013-06-23T18:20:28Z	2021-04-29T13:14:36Z	{'seal': False, 'ticked': True}	Tr
179	f7f587617c134cd99447c6ff8aa6c64a	2006-12-21T14:32:17Z	2021-08-16T15:35:08Z	{'seal': False, 'ticked': True}	Tr
180	f936908299e446c1b5c9d8c108cc596f	2019-10-20T03:40:44Z	2021-04-29T13:17:58Z	{'seal': False, 'ticked': True}	Tr

	id	created_date	last_updated	admin	bo
181	fb979e39ff8444eaa7aed77df8ed9169	2017-07-03T13:16:11Z	2021-04-29T13:18:35Z	{'seal': False, 'ticked': True}	Tr
				{'seal': True,	

Mit .loc Anzeige des der Länder in einer Liste

In [475…
```
df_Kopie.loc[:, ["publisher.country"]]
```

Out[475…

	publisher.country
0	IR
1	BR
2	TW
3	UA
4	ID
...	...
178	KR
179	GB
180	BR
181	ID
182	CA

183 rows × 1 columns

Mit groupby().size().sort_values(ascending=False) wird die absteigende Sortierung der Häufigkeiten wiedergegeben.

In [476…
```
ser_grouped_publisher_country = df.groupby("publisher.country").size().sor
ser_grouped_publisher_country
```

Out[476…
```
publisher.country
US    29
BR    24
ID    13
ES    12
PL     9
IR     9
UA     7
IT     6
GB     6
CA     6
DE     5
CH     5
RO     4
EG     3
MX     3
CR     2
RU     2
```

```
NL     2
NG     2
LT     2
KR     2
AR     2
CU     2
FR     2
TW     2
CN     2
SE     1
SG     1
TR     1
PT     1
RS     1
QA     1
UY     1
KE     1
PK     1
NO     1
MY     1
AT     1
JP     1
IN     1
HR     1
FI     1
CO     1
BG     1
BA     1
ZA     1
```

Antwort : USA 29 ; BR: 24 ; ID: 13 ; ES: 12 GB: 6 DE:5

In [ ]:
```
mit .loc() und Auswahl des Länderkürzels kann auch eine Einzelabfrage nach
```

In [199…
```
ser_grouped_publisher_country.loc["US"]
```

Out[199…
```
29
```

In [177…
```
ser_grouped_publisher_country.loc["GB"]
```

Out[177…
```
6
```

Wieviele sind peer-reviewed und double blind peer-reviewed?

Mit .loc() Anzeige des Spalteninhalts von editorial.review_process. Auch dieser Auszug zeigt, das die Einträge nicht einzeln sind (siehe Treffer 182)

In [404…
```
df_Kopie.loc[:, ["editorial.review_process"]]
```

Out[404…

	editorial.review_process

	editorial.review_process
0	[Double blind peer review]
1	[Double blind peer review]
2	[Double blind peer review]
3	[Double blind peer review]
4	[Blind peer review]
...	...
178	[Peer review]
179	[Double blind peer review]
180	[Double blind peer review]
181	[Double blind peer review]
182	[Editorial review, Double blind peer review]

Daher wieder mit for Loop wie bereits oben Beschrieben mit append() und value_counts die Anzahl der gewünschten einzelnen Einträge an der "Stelle" review_process" innerhalb des dictionaries ermitteln.

In [437…

```
list_review_process = []
for item in all_results_list :
    list_review_process.append(item["bibjson"]["editorial"]["review_proces
dataframe_review_process = pd.DataFrame(list_review_process )
dataframe_review_process.value_counts()
```

Out[437…

```
Double blind peer review    114
Blind peer review            31
Peer review                  28
Editorial review              6
Open peer review              3
Committee review              1
dtype: int64
```

Antwort: Double blind peer review kommt am Häufigsten vor und deutet damit auf eine Hohe Qualität der gelisteten OA Journals hin.

Wieviele verlangen Publikationsgebühren?

Mit .loc Alle Zeilen der Spalte apc.has_apc in eine Liste übergeben

In [439…

```
df_Kopie.loc[:, ["apc.has_apc"]]
```

Out[439…

	apc.has_apc
0	False
1	False

	apc.has_apc
2	False
3	False
4	False
...	...
178	False
179	False
180	False
181	False
182	False

Wieder mit groupby().size() die Häufigkeit errechnen:

In [398…
```
ser_grouped_apc = df.groupby("apc.has_apc").size()
ser_grouped_apc
```

Out[398…
```
apc.has_apc
False    170
True      13
dtype: int64
```

Antwort: Von 183 verlangen nur 13 Publisher APC Gebühren. 170 dagegen keine.

Sind unter den Herausgebern namenhafte wie de Gruyter?

In [490…
```
list_publisher_name = []
for item in all_results_list:
    list_publisher_name.append(item["bibjson"]["publisher"]["name"])
dataframe_language = pd.DataFrame(list_publisher_name)
dataframe_language
```

Out[490…

	0
0	Hamara Afzar
1	Universidade Federal de Alagoas
2	National Taiwan Normal University
3	Vernadsky National Library of Ukraine
4	Faculty of Adab and Humanities UIN Syarif Hida...
...	...
178	Research Institute for Knowledge Content Devel...
179	University of Edinburgh
180	Universidade de Brasília
181	Universitas Gadjah Mada

	0
182	University of Victoria Libraries

In [498…
```
ser_grouped_list_publisher_name.loc["Hamara Afzar"]
```

Out[498…
```
1
```

In [502…
```
ser_grouped_list_publisher_name.loc["University of Edinburgh"]
```

Out[502…
```
1
```

In [492…
```
ser_grouped_list_publisher_name = df.groupby("publisher.name").size().sort_
ser_grouped_list_publisher_name
```

Out[492…
```
publisher.name
University Library System, University of Pittsburgh       3
Universidad Complutense de Madrid                         3
Levy Library Press                                        2
University of Alberta                                     2
Universidad de Murcia                                     2
                                                         ..
Library Association of the City University of New York    1
Library of the University of Heidelberg                   1
Lodz University Press                                     1
Marketing Libraries Journal                               1
openjournals.nl                                           1
Length: 165, dtype: int64
```

Leider funktioniert keine .query Abfrage oder ähnliches. Ich gehe davon aus in der Falschen "Spalte zu sein. Es sind auch nur noch 165 Datensätze. Anhand der Schnittstellensuche in der API Testsuche konnte ermittelt werden, dass es lediglich ein OA Journal von de Gruyter enthält.