Share of open access journal articles published by Berlin authors from 2016: data

Michaela Voigt[1], Christian Winterhalter[2]

April 2018

Published report M. Voigt, C. Winterhalter, C. Riesenweber, A. Hübner: Open-Access-Anteil bei Zeitschriftenartikeln von Wissenschaftlerinnen und Wissenschaftler an Einrichtungen des Landes Berlin: Datenauswertung für das Jahr 2016.
DOI: https://doi.org/10.14279/depositonce-6866

Data The data described here were retrieved from multiple bibliographic databases. Due to license terms the database raw data cannot be provided for download. Data were aggregated, normalised and analysed with help of a Python script (https://github.com/tuub/oa-eval, code documentation in English). Search queries and download settings for these databases are documented in the (German) manual that accompanies the script. For a detailed description of the retrieval process and the analysis steps see the report. Data are distributed under the Creative Commons Public Domain Dedication (CC0).
DOI: https://doi.org/10.14279/depositonce-6867.

[1] michaela.voigt@tu-berlin.de, ORCiD:0000-0001-9486-3189
[2] christian.winterhalter@ub.hu-berlin.de, ORCiD:0000-0001-8618-0337

1 General remarks

The overall goal was to analyse the publication output from nine research institutions located in Berlin (Germany) and determine the share of open access journal articles. Journal articles whose authors are affiliated with the following nine institutions were analysed:

- Alice Salomon Hochschule (ASH Berlin)
- Beuth Hochschule für Technik Berlin (Beuth)
- Charité – Universitätsmedizin Berlin (Charité)
- Freie Universität Berlin (FU Berlin)
- Hochschule für Technik und Wirtschaft Berlin (HTW Berlin)
- Hochschule für Wirtschaft und Recht (HWR Berlin)
- Humboldt-Universität zu Berlin (HU Berlin)
- Technische Universität Berlin (TU Berlin)
- Universität der Künste (UdK Berlin)

Data were retrieved from sixteen bibliographic databases: Academic Search Ultimate (EBSCO), Business Source Complete (via EBSCOhost), CAB Abstracts (via OvidSP), CINAHL (via EBSCOhost), Embase (via OvidSP), IEEE Xplore, Inspec, Library and Information Science Abstracts (LISA) (via ProQuest), ProQuest Social Sciences, GeoRef (via EBSCOhost), PubMed, SciFinder (CAPlus), Scopus, Sport Discus (via EBSCOhost), TEMA and Web of Science Core Collection.

To identify open access journals[1] the Directory of Open Access Journals (DOAJ) was used.[2] In order to reduce script run time the API[3] provided by DOAJ was not used. Instead, DOAJ data were downloaded as comma-separated file[4] in September 2017. The csv file was saved as tab-delimited file; the `doaj.txt` constitutes the state of the DOAJ metadata as of September 7th, 2017 – listing 9.888 open access journals.

To identify open access articles in hybrid journals[5] a combination of data retrieved from the oaDOI API[6] and the Crossref API[7] was used (September 2017). oaDOI data were checked for OA status, host type for the detected OA version and license information. An article is considered to be `hybrid OA` if a Creative Commons licensed version is accessible via the publisher website.

[1] An open access journal publishes open access articles, i. e. all published articles are openly available on the publisher's website, without charge or delay.

[2] The analysis does not rely on oaDOI data(`'oaDOI - journal_is_oa' = TRUE`) for detection of Gold OA articles because tests showed that oaDOI data are incomplete on journal OA status.

[3] DOAJ metadata: API https://doaj.org/api/v1/docs

[4] DOAJ metadata: CSV file https://doaj.org/csv

[5] A hybrid (open access) journal publishes both closed access and open access articles. It is operated under a subscription business modell with the (fee-based) option to make single articles open access.

[6] The service oaDOI was renamed to "'Unpaywall data"' in january 2018, the API URL was changed to https://api.unpaywall.org/v2/ (formerly: https://api.oadoi.org/v2/). Data described here were retrieved before this change. Hence, the now outdated name oaDOI is used.

[7] http://api.crossref.org/works/

To identify green open access articles data from oaDOI were used. An article is considered to be `green OA` if the article is not detected as `gold OA` or `hybrid OA` and oaDOI detected at least one OA version in a repository.

Tab. 1 shows which values were included to determine the open access status.

Table 1: Detection of OA status

OA status	**Note**
Gold OA	DOAJ data as of September 7th, 2017 (ISSN + year lookup)
Hybrid OA	according to oaDOI and Crossref data as of September 30th, 2017 following values must apply: `'oaDOI - is_oa' = TRUE` `'oaDOI - journal_is_oa' = FALSE`, `'oaDOI - best_oa_location.host_type' = 'publisher'`, (`'oaDOI - best_oa_location.license' = '$cc$'` OR `'Crossref license' = 'creativecommons.org'`)

Green OA	according to oaDOI data as of September 30th, 2017
	following values must apply:
	`'oaDOI - is_oa' = TRUE`
	`'oaDOI - journal_is_oa' = FALSE`
	at least one `'oa_location'` with `'host_type' = 'repository'`)

Data on APC costs for open access journals were retrieved from DOAJ (as of September 7th, 2017); the costs were not verified manually. Since value-added taxes vary by country publishers usually list costs excluding VAT. APC listed here do not include VAT.

To determine exchange rates we consulted http://www.xe.com: Exchange rates were retrieved for the beginning of 2016 (January 1st, 2016) and the current rate at the date of analysis (October 10th, 2017).

2 Bibliographic data

Data were analysed with regard to the following questions:

- How many journal articles did Berlin-based researchers publish in 2016?
- How many of these articles were published in open access journals?
- How many of these articles have a Berlin-based corresponding author, in other words for how many articles did a Berlin-based author (resp. his/her institution) most likely cover the open access fee (Article Processing Charge, APC)? What were the assumed APC costs for gold open access journals in 2016?
- How many open access articles did researchers from Berlin publish in hybrid journals?
- How many articles from Berlin-based researchers are available as green open access (via a repository)?

For a list of available files see tab. 2. For a list of bibliographic data available in the file containing article data see tab. 3.

Table 2: Overview of files

File name	Note
OABerlin2016_data.xlsx	script output: list of articles (11.005 items), field description, queries for bibliographic databases
OABerlin2016_data.csv	script output: list of articles (11.005 items) in comma-separated format (UTF-8 encoded)
OABerlin2016_data_results.xlsx	detailed results (pivot tables)
DOAJ.txt	script input: DOAJ metadata (tab-delimited file)

Table 3: Bibliographic data

Field name	Source	Note
authors	databases	string trimmed if field length exceeds 200 characters
title	databases	title as indexed as main title in databases; for non-English articles title might be translated to English
OA status	manually	see tab. 1 for overview on which values were included to determine OA status
DOI	databases	if available; DOIs updated if Crossref API returned error

Continued on next page

Table 3 – continued from previous page

Field name	Source	Note
corrupt DOI	manually	if there was a problem with the original DOI (as included in databases): DOI checked → correct DOI added in `DOI`, incorrect DOI added in `corrupt DOI`; error type if no resolvable DOI could be detected
journal	databases	if available; journal names normalized
ISSN	databases	if available; could be ISSN for either print or electronic edition (print ISSN most likely)
eISSN	databases	if available; could be ISSN for either print or electronic edition (electronic ISSN most likely)
publisher	databases, DOAJ, Crossref (data refined)	missing publisher information retrieved from Crossref; publisher names normalized
publisher group	manually (data refined)	cluster publisher syndicates: `publisher group` based on column `publisher` `Wiley`: American Geophysical Union (AGU); International Union of Crystallography (IUCr); John Wiley and Sons; Wiley; Wiley-Blackwell; Wiley-VCH; Econometric Society) `Springer Nature`: Nature Publishing Group; Springer; Springer Berlin Heidelberg; Springer Fachmedien Wiesbaden GmbH; Springer Healthcare; Springer Heidelberg; Springer International Publishing; Springer Nature; Springer Netherlands; Springer New York; Springer Singapore; Springer-VDI Verlag `Wolters Kluwer`: Lippincott Williams and Wilkins; Medknow Publications; Ovid Technologies (Wolters Kluwer Health); Wolters Kluwer `IOP Publishing`: IOP Publishing; American Astronomical Society; Japan Society of Applied Physics
year	databases	PubMed indexes multiple dates: PubMed search covers all date fields while python script uses only one date field (`DP = Date of Publication`)
affiliations	databases	if available; where applicable e-mail addresses were anonymized (`xxx@[domain]`)

Continued on next page

Table 3 – continued from previous page

Field name	Source	Note
corresponding author	databases	if available; retrieved from: Web of Science, PubMed, SciFinder; where applicable e-mail addresses were anonymized (`xxx@[domain]`)
institution of corr. author	script,	script analyses affiliation data for corresponding author using institution names set up in script; short name for respective (Berlin) institution is given here; e-mail addresses anonymized (xxx@[domain]) articles with multiple entries checked manually to confirm multiple corresponding authors; search for Berlin e-mail domains in `e-mail` and manual check for corresponding author
e-mail	databases	if available; e-mail addresses were anonymized (`xxx@[domain]`)
subject	databases	if available; retrieved from: Web of Science, SciFinder
DOAJ subject	DOAJ	available for articles in DOAJ-listed journals; in DOAJ journals are categorised using the Library of Congress Classification
funding	databases	if available; retrieved from: Web of Science (field code `FN`)
license	DOAJ, oaDOI, Crossref	if available; consolidated license information (DOAJ, oaDOI: license type, Crossref: license URL)
oaDOI license	oaDOI	license type according to oaDOI (e. g. `CC BY`)
Crossref license	Crossref	license URL according to Crossref (e. g. `https://creativecommons.org/licenses/by/4.0/`)
database name	Python script	Name of the database from which the article information was extracted
notes	script	various indicators on how article data was processed/ enriched: `Checked by hand.` = affiliation of corresponding author was checked manually; `Identified via DOAJ.` = Gold OA status identified via DOAJ;

Continued on next page

Table 3 – continued from previous page

Field name	Source	Note
		`Identified via oaDOI.` = Green/Hybrid OA status identified via oaDOI; `oaDOI error.` = oaDOI has no record of this DOI; `Identified via Crossref license` = Creative Commons license URL was found in Crossref metadata (Hybrid OA status identified)
APC Amount	DOAJ	APC amount according to DOAJ
APC Currency	DOAJ	APC currency according to DOAJ
oaDOI - is_oa	oaDOI	data as of 30.09.2017 - see oaDOI field documentation at https://unpaywall.org/api/v2 (new name: Unpaywall)
oaDOI - journal_is_oa		—"—
oaDOI - best_oa_location.evidence		—"—
oaDOI - best_oa_location.host_type		—"—
oaDOI - best_oa_location.license		—"—
oaDOI - best_oa_location.url		—"—
oaDOI - best_oa_location.url.domain		—"—
oaDOI - oa_locations		contains JSON data from oaDOI API (to evaluate all OA versions parse JSON)
Berlin Inst. Repositories	oaDOI, manually	repository URL if oaDOI detected OA copy (search for repository URL in field `oaDOI - oa_locations`)
xe.com (2016-01-01)	xe.com	exchange rate xe.com as of 01.01.2016
xe.com (2017-10-02)	xe.com	exchange rate xe.com as of 02.10.2017
APC in EUR (2016-01-01)	DOAJ, xe.com	APC amount according to DOAJ as of 01.01.2016 (VAT not included)
APC in EUR (2017-10-02)	DOAJ, xe.com	APC amount according to DOAJ as of 02.10.2017 (VAT not included)

Table 4: Detailed results (pivot tables)

Tab name	Pivot tables listed
OA	OA status total Articles with CC BY license Distribution OA share per Berlin corresponding authors (1) Distribution OA share per Berlin corresponding authors (2) Gold OA: Distribution of articles with Berlin corresponding authors among Berlin institutions Green OA: Distribution of articles with Berlin corresponding authors among Berlin institutions Hybrid OA: Distribution of articles with Berlin corresponding authors among Berlin institutions Closed: Distribution of articles with Berlin corresponding authors among Berlin institutions
publisher	Stats on publishers Distribution of all articles among publishers (Count, share, cumulative share) Distribution of closed access articles among publishers Distribution of gold OA articles among publishers Distribution of hybrid OA articles among publishers Distribution of green OA articles among publishers
Gold	Details Articles with Creative Commons license Articles with no APC according to DOAJ (APC = 0 or no record) Articles with Berlin corresponding author with no APC according to DOAJ (APC = 0 or no record) Articles with APC according to DOAJ (exchange rate: 2016-01-01) Articles with APC according to DOAJ (exchange rate: 2017-10-02) Articles with Berlin corresponding author – with APC according to DOAJ (exchange rate: 2016-01-01) Articles with Berlin corresponding author – with APC according to DOAJ (exchange rate: 2017-10-02) Articles with Berlin corresponding author – with APC less than 2000 EUR incl. VAT according to DOAJ (exchange rate: 2016-01-01) Articles with Berlin corresponding author – with APC less than 2000 EUR incl. VAT according to DOAJ (exchange rate: 2017-10-02) Distribution of all articles among publishers Distribution of articles with Berlin corresponding author among publishers

Continued on next page

Table 4 – continued from previous page

Tab name	Pivot tables listed
Gold	Distribution of all APC-free articles among publishers Distribution of APC-free articles with Berlin corresponding author among publishers
Green	Details Top 10 repositories Distribution among repositories Berlin repositories Distribution of publishers
Hybrid	Details Articles with Berlin corresponding author Articles with Creative Commons license Distribution of all articles among publishers Distribution of articles with Berlin corresponding author among publishers
fuzzy gratis access	Count of articles and publishers oaDOI/Unpaywall indication for status “fuzzy gratis access” Distribution of publishers

3 Some remarks on the detailed results

Usin the data of this study it is possible to draw conclusions on the number of e. g. gold open access articles with a corresponding author from a certain Berlin institution. It is not possible though to re-use the data to determine the overall share of open access for individual Berlin institutions since data on the total number of articles for individual Berlin institutions are missing.

oaDOI lists more articles to be open access. These articles do not meet the criteria of this study though (see Tab. 1 for the underlying criteria to detect the OA status) and were thus not counted as open access articles: According to oaDOI data some articles are available for free (‘gratis’) on the publisher website – neither Creative Commons licensed nor stored in an independent repository. While other studies count these kind of articles as open access anyway (e. g. labeled as “Bronze Open Access”) they were not included when determining the share of open access articles in this study – unless there are also available via an open access repository and hence counted as Green OA. We do not consider them to be sustainable open access since the gratis access on publisher websites could be of limited time. In total 586 articles fall into this category that we labeled “fuzzy gratis access”. Some details on these articles are included in the file `OABerlin2016_data_results.xlsx`.

4 Re-use cases

We imagine the following re-use cases for this data:

So far data were analysed on a multi-institutional level. Additionally we analysed corresponding authorship for individual Berlin institutions.[8] Since the data basis is comprehensive one could evaluate single institutions:

- breakdown by publisher
- breakdown by APC costs
- breakdown by discipline/subject

A subset of this data might also be of interest for other kinds of studies. As an example, one might take a closer look at aspects of collaboration: How often do authors from Berlin-based institutions collaborate? Are there Berlin-wide collaboration networks? Since affiliation data is not complete, data from other sources should be included.

To determine the share of Gold open access the DOAJ was used. While the DOAJ is the commonly used source to detect Gold OA and the DOAJ index is growing continuously it does not index *all* Gold OA journals – depending on how one defines Gold OA. A research group at the University of Bielefeld has compiled a comprehensive list of open access journals with embargo-free (gratis) access to the publisher version as the main criterion.[9] One could correlate ISSN data to detect further articles that are freely available – probably under a restrictive license, though.

To determine the share of Green open access the oaDOI webservice was used; data reflect the share of Green OA as of September 2017. Since then authors (or the respective institutions on their behalf) might have self-archived an open access version. Furthermore, oaDOI can be considered a progressive service – the underlying technology is being improved, data sources are added. One could thus query the web service again to retrieve a current state of green OA and/or compare the results with the September 2017 data.

While oaDOI seems to be a comprehensive source to detect green OA versions, samples showed it is far from complete. To detect blind spots in finding green versions one could correlate data with data from (at least) the Berlin institutional repositories (publisher DOI look-up).

Furthermore it might be interesting to have a closer look at the category "fuzzy gratis access": Are these articles still available for free at the publisher website? What are possible explanations for the 'gratis' availabilty (e. g. promotion of certain articles, trending research topics, articles of to special interest the public)? Are these articles published in 'gratis' journals – or are they 'gratis' articles in otherwise closed access journals?

[8] It is important to note that it is not possible to determine the share of open access publications for individual Berlin institutions because we have no data on the overall number of publications of each institution.

[9] Rimmert C, Bruns A, Lenke C, Taubert NC. (2017): ISSN-Matching of Gold OA Journals (ISSN-GOLD-OA) 2.0. Bielefeld University. https://doi.org/10.4119/unibi/2913654.