Share of open access journal articles published by Berlin authors from 2017: data

Michaela Voigt[1]

January 6, 2019

Published report A. Hübner, M. Voigt, P. Finke, C. Riesenweber: Open-Access-Anteil bei Zeitschriftenartikeln von Wissenschaftlerinnen und Wissenschaftler an Einrichtungen des Landes Berlin: Datenauswertung für das Jahr 2017.
DOI: https://doi.org/10.14279/depositonce-7866

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, normalized and analyzed 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-7867.

This work is distributed under the Creative Commons Public Domain Dedication (CC0). You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.
For more information see https://creativecommons.org/publicdomain/zero/1.0/.

[1]Technische Universität Berlin, michaela.voigt@tu-berlin.de, ORCiD:0000-0001-9486-3189

1 General remarks

The overall goal was to analyze 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 analyzed:

- 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 Berlin (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 ProQuest), PubMed, SciFinder (CA Plus), Scopus, Sport Discus (via EBSCOhost), TEMA and Web of Science Core Collection.

To identify articles in gold 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]. The csv file was saved as tab-delimited file; the file `doaj.txt` constitutes the state of DOAJ metadata as of September 5th, 2018 – listing 12.055 open access journals. An article is considered to be `gold OA` if the journal is published in DOAJ.

To identify open access articles in hybrid journals[5] a combination of data retrieved from the Unpaywall API[6] and the Crossref API[7] was used (September 2018). Unpaywall data were checked for OA status, host type (publisher or repository) for the detected OA version and license in-

[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 Unpaywall data `'Unpaywall[journal_is_oa]'`) for detection of Gold OA articles because 1) it would leave articles without a DOI undetected and 2) samples showed that Unpaywall 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 model with the (fee-based) option to make single articles open access.

[6] http://api.unpaywall.org/v2/

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

formation; Crossref data were checked for license information. An article is considered to be `hybrid OA` if a Creative Commons licensed version is accessible via the publisher website.

To identify green open access articles data from Unpaywall were used. An article is considered to be `green OA` if the article is detected as neither `gold OA` nor `hybrid OA` and Unpaywall 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 5th, 2018 (ISSN + year lookup)
Hybrid OA	according to Unpaywall and Crossref data as of September 20th, 2018 following values must apply: `'Unpaywall[is_oa]' = TRUE`, `'Unpaywall[host_type]' = 'publisher'`, `'license' = 'CC*'` **OR** `'OA status' != gold`, `'license' = 'CC*'`
Green OA	according to Unpaywall data as of September 20th, 2018 entries in `'Unpaywall[OA Repos]'` were searched manually for `'.com'`; if Unpaywall incorrectly assigned the `'host_type' = 'repository'`, the respective entry was corrected (changed to `'OA status' = closed`) following values must apply: `'OA status' != gold OR hybrid`, `'Unpaywall[is_oa]' = TRUE`, at least one entry in `'Unpaywall[OA Repos]'`

Data on APC costs for open access journals were retrieved from DOAJ (as of September 5th, 2018); 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 2017 (January 1st, 2017) and the current rate at the date of analysis (September 22nd, 2018).

2 Bibliographic data

Data were analyzed with regard to the following questions:

- How many journal articles did Berlin-based researchers publish in 2017?
- How many of these articles were published in open access journals (`gold OA`)?
- 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)?
- How many open access articles did researchers from Berlin publish in hybrid journals (`hybrid OA`)?
- How many articles from Berlin-based researchers are available via a repository as green open access (`green OA`)?

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**
OABerlin2017_data.xlsx	script output: list of articles (10.923 items)
OABerlin2017_data.csv	script output: list of articles (10.923 items) in comma-separated format (UTF-8 encoded)
OABerlin2017_data_repositories.xlsx	list of OA versions on repositories (8.542 entries)
OABerlin2017_data_repositories.csv	list of OA versions on repositories (8.542 entries) in comma-separated format (UTF-8 encoded)
OABerlin2017_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	type of OA (`gold`, `hybrid`, `green` or `None`); see tab. 1 for overview on which values were included to determine OA status

Continued on next page

Table 3 – continued from previous page

Field name	Source	Note
DOI	databases, manually	if available; DOIs added and/or corrected manually
doiRA	DOI Foundation	name of the responsible DOI Registry Agency (which agency minted the DOI); the note `DOI does not exist` indicates that the databases included a DOI but it was not registered correctly
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), manually	missing publisher information added manually if retrievable; 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) – `Springer Nature`: Nature Publishing Group; Springer; Springer Healthcare; Springer Heidelberg; Springer International Publishing; Springer Medizin; Springer Nature; Springer New York; Springer Singapore; Springer-VDI-Verlag – `Wolters Kluwer`: 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; where applicable e-mail addresses were anonymized (`xxx@[domain]`)
institution of corr. author	script, manually	script analyzes affiliation data for corresponding author using institution names set up in script; short name for respective (Berlin) institution is given here (otherwise `None`); 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; journals in DOAJ are categorized using the Library of Congress Classification
funding	databases	if available; retrieved from: Web of Science (field code `FN`)
license	DOAJ, Crossref, Unpaywall	if available; consolidated license information (license type); for Unpaywall data license info included for publisher version only (`Unpaywall[bestOA.host_type] = publisher`)
license source	manually	source for license info (Crossref, Unpaywall or DOAJ)
Crossref license	Crossref	data as of Sept. 2018 – 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, manually	various indicators on how article data was processed/ enriched: – `Checked by hand.` = affiliation of corresponding author was checked manually;

Continued on next page

Table 3 – continued from previous page

Field name	Source	Note
notes	script, manually	– `Identified via DOAJ.` = Gold OA status identified via DOAJ; – `Identified via Unpaywall.` = Green/Hybrid OA status identified via Unpaywall; – `Identified via Crossref.` = Creative Commons license URL was found in Crossref metadata (Hybrid OA status identified) – `Unpaywall flags free version (...)` = Unpaywall flags free version on publisher website as `'repository'` version and no other repository version was found (not counted as green) – `DOI added manually.` = DOI not included in database data; added manually – `DOI corrected manually.` = DOI incorrect in database data; corrected manually – `DOI error (not registered).` = DOI not registered correctly at DOI registration agency; could not retrieve data from Unpaywall or Crossref
APC Amount	DOAJ	APC amount for Gold OA articles according to DOAJ (`None or unknown` if DOAJ data included no information)
APC Currency	DOAJ	APC currency for Gold OA articles according to DOAJ (`None or unknown` if DOAJ data included no information)
Unpaywall[is_oa]	Unpaywall	data as of Sept. 2018 – see Unpaywall field documentation at https://unpaywall.org/data-format
Unpaywall[journal_is_oa]		—"—
Unpaywall[bestOA.evidence]		—"—
Unpaywall[bestOA.host_type]		—"—
Unpaywall[bestOA.license]		—"—
Unpaywall[bestOA]		several sub-fields for `best_oa_location` (Unpaywall fields: `Evidence`, `Host-type`, `License`, `URL`, `Version`)

Continued on next page

Table 3 – continued from previous page

Field name	Source	Note
Unpaywall[OA Repos]	Unpaywall, manually	any `oa_locations` with `'host_type' = 'repository'` (Unpaywall sub-fields: `Evidence`, `Host-type`, `Landing page`, `PDF-Link`, `License`, `version`) searched manually for `'.com'`: if Unpaywall incorrectly assigned the `'host_type' = 'repository'`, the respective entry was corrected (changed to `OA status = closed`)
Berlin Inst. Repositories	Unpaywall, manually	repository URL if Unpaywall detected OA copy (in field `Unpaywall[OA Repos]` search for repository URL as listed at http://www.open-access-berlin.de/ressourcen)
xe.com (2017-01-01)	xe.com	exchange rate xe.com as of 01.01.2017
xe.com (2018-09-22)	xe.com	exchange rate xe.com as of 22.09.2018
APC in EUR (2017-01-01)	DOAJ, xe.com	APC amount according to DOAJ as of 01.01.2017 (VAT not included)
APC in EUR (2018-09-22)	DOAJ, xe.com	APC amount according to DOAJ as of 22.09.2018 (VAT not included)

Unpaywall may list several entries for OA-versions of an article (Unpaywall data as of Sept. 2018). These entries were split in a line-based manner to be able to evaluate how often a specific repository was used – i. e. there is one entry per article and repository, there might hence be multiple lines per article. See the article Index number or DOI to evaluate how many repository copies there are per article.

Table 4: Bibliographic data for file `OABerlin2017_data_repositories.xlsx`

Field name	Note
Index	Article ID for articles in file `OABerlin2017_data.xlsx`
authors	see Tab. 3
title	see Tab. 3
OA status	see Tab. 3
DOI	see Tab. 3

Continued on next page

Table 4 – continued from previous page

Field name	note
publisher	see Tab. 3
publisher group	see Tab. 3
institution of corr. author	see Tab. 3
licence	see Tab. 3
Berlin Inst. Repositories	see Tab. 3
OA-Repos_entry-no	index number for *this* repository version
OA-Repos_Evidence	see Unpaywall field documentation at https://unpaywall.org/data-format ("How we found this OA location.")
OA-Repos_Host-type	see Unpaywall field documentation at https://unpaywall.org/data-format ("The type of host that serves this OA location.")
OA-Repos_Landing-page	see Unpaywall field documentation at https://unpaywall.org/data-format ("The URL for a landing page describing this OA copy.")
OA-Repos_PDF-Link	see Unpaywall field documentation at https://unpaywall.org/data-format ("The URL with a PDF version of this OA copy.")
OA-Repos_License	see Unpaywall field documentation at https://unpaywall.org/data-format ("The license under which this copy is published.")
OA-Repos_Version	see Unpaywall field documentation at https://unpaywall.org/data-format ("The content version accessible at this location."); please note: samples showed that the version is not always detected correctly (i. e. Unpaywall says it is the `submittedVersion` (preprint), while it actually is the `acceptedVersion` (postprint) etc.
OA-Repos_PDF-Link_Domain	field `OA-Repos_PDF-Link` was split to include domain only (e.g. `arXiv.org`)
OA-Repos_Landing-page_Domain	field `OA-Repos_Landing-page` was split to include domain only (e.g. `arXiv.org`)
OA-Repos_Domain	fields **OA-Repos_PDF-Link_Domain** and **OA-Repos_Landing-page_Domain** joined; analyze **this** field to get the total number of times a certain repository was used for self-archiving

Tab. 5 lists the pivot tables per sheet for the OA Berlin 2017 evaluation.

Table 5: 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: 2017-01-01) Articles with APC according to DOAJ (exchange rate: 2018-09-22) Articles with Berlin corresponding author – with no APC according to DOAJ (APC = 0 or no record) Articles with Berlin corresponding author – with APC according to DOAJ (exchange rate: 2017-01-01) Articles with Berlin corresponding author – with APC according to DOAJ (exchange rate: 2018-09-22) Articles with Berlin corresponding author – with APC less than 2000 EUR incl. VAT according to DOAJ (exchange rate: 2017-01-01) Articles with Berlin corresponding author – with APC less than 2000 EUR incl. VAT according to DOAJ (exchange rate: 2018-09-22)

Continued on next page

Table 5 – continued from previous page

Tab name	Pivot tables listed
Gold	Distribution of all articles among publishers Distribution of articles with Berlin corresponding author among publishers Distribution of all APC-free articles among publishers Distribution of APC-free articles with Berlin corresponding author among publishers Distribution of APC (according to DOAJ) among publishers
Green	Details Occurrence of copies on repositories Number of articles on Top 3 repositories (arXiv, Europe PubMed Central, PubMed Central) Berlin repositories Distribution of publishers Average number of repositories used per Green OA article (1): all articles Average number of repositories used per Green OA article (2): articles with Berlin corresponding author
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 Unpaywall indication for status “fuzzy gratis access” Distribution of publishers

3 Some remarks on the detailed results

Using 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. Though, it is **not** possible 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.

Unpaywall 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 Unpaywall 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 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 they 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 516 articles fall into this category that we labeled "fuzzy gratis access". Some details on these articles are included in the file `OABerlin2017_data_results.xlsx`.

4 Re-use cases

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

So far data were analyzed on a multi-institutional level. Additionally we analyzed 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 distribution among repositories

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 Unpaywall webservice was used; data reflect the share of Green OA as of September 2018. Since then, authors (or the respective institutions on their behalf) might have self-archived an open access version. Furthermore, Unpaywall 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 2018 data.

While Unpaywall 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 cor-

[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.

relate data with data from (at least) the Berlin institutional repositories (look-up for publisher DOI).

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 special interest to the public)? Are these articles published in ‘gratis’ journals – or are they ‘gratis’ articles in otherwise closed access journals?