Share of open access journal articles published by Berlin authors from 2013 to 2015: data

Michaela Voigt[1], Christian Winterhalter[2]

September 2016

Published report Michaela Voigt, Christian Winterhalter: Open-Access-Anteil bei Zeitschriftenartikeln von Wissenschaftlerinnen und Wissenschaftler an Einrichtungen des Landes Berlin: Datenauswertung für die Jahre 2013–2015, DOI: 10.14279/depositonce-5570

Data The data described here were retrieved from multiple bibliographic databases. Due to license terms raw data from single databases cannot be provided for download. Data were aggregated, normalised and analysed with help of a Python script available at 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: 10.14279/depositonce-5569.

[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 from 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 was retrieved from sixteen bibliographic databases: Academic Search Premier (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, Web of Science Core Collection

To identify open access journals[1] the Directory of Open Access Journals (DOAJ) was used. In order to reduce script run time the API[2] provided by DOAJ was not used. Instead, DOAJ data was downloaded as comma-separated file[3] in July 2016. The file `doaj.txt` constitutes the state of the DOAJ metadata as of that day, listing 8.968 open access journals.

To identify open access articles in hybrid journals[4] the Crossref API[5] was used (July 2016); metadata is checked for open content licenses.

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

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

[4] 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.

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

2 Files and bibliographic data

Data was analysed with regard to the following questions:

- How many journal articles did Berlin-based researchers publish from 2013 to 2015?
- 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 2013–2015?
- How many open access articles did researchers from Berlin publish in hybrid journals? What were the assumed APC costs for hybrid journals in 2013–2015?

For a list of available files see tab. 1. For a list of bibliographic data available in all files see tab. 2.

Table 1: Overview of all files

File name	**Note**
DOAJ.txt	script input: DOAJ metadata (tab-delimited file)
allPubs.xlsx	script output: list of articles (33.172 items)
allOAPubs.xlsx	script output: list of OA articles in open access journals (3.919 items)
allOAPubsWithCorrAuthor.xlsx	script output: list of OA articles in open access journals for which a Berlin author is corresponding author (1.569 items)
hybridPubsWithCorrAuth.xlsx	script output: list of OA articles in hybrid journals for which a Berlin author is corresponding author (204 items)

Table 2: 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
DOI	databases	if available; DOIs updated if Crossref API returned error
journal	databases	if available
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)
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; retrieved from: Web of Science, PubMed, SciFinder; where applicable email addresses were anonymized (`xxx@[domain]`)
corresponding author	databases	if available; retrieved from: Web of Science, PubMed, SciFinder; where applicable email addresses were anonymized (`xxx@[domain]`)
Berlin inst.	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	databases	if available; email addresses were anonymized (`xxx@[domain]`)
subject	databases	if available; retrieved from: Web of Science, SciFinder
DOAJ subject	DOAJ	in DOAJ journals are categorised using the Library of Congress Classification
funding	databases	retrieved from: Web of Science (field code `FN`)
publisher	DOAJ	data available for OA journal articles only
license	DOAJ, Crossref	if available (DOAJ: license type, Crossref: license URL)
notes	script	various indicators on how article data was processed/ enriched `Checked by hand.` = affiliation of corresponding author was checked manually; `License added via CrossRef` = license URL was found in Crossref metadata; `DOAJ=1` = prior missing ISSN retrieved from Crossref and DOAJ API confirmed status of open access journal

3 APC costs

Costs were analysed for articles with Berlin corresponding authors. Data on APC costs for both open access and hybrid journals were collected in July and August 2016:

While 1.569 articles in open access journals were distributed among 144 publishers, about 80 % of the articles were published by 21 publishers (TOP 20 publishers). Due to limited resources data on APC costs for open access journals were collected for the TOP 20 publishers only (see tab. 3).

We could identify 573 open access articles in hybrid journals, of which 204 have a Berlin corresponding author. Data on APC costs for hybrid journals were collected for all identified articles (see tab. 4).

Data on APC costs were collected from publishers' websites: APC represent current APC as of 2016. Publishers often do not charge an exact amount per article. Instead, costs per article can depend on aspects like number of pages, type of article, type of license. Websites then often list price ranges or example calculations. For our analysis we collected all values and calculated costs according to 1) mean, 2) minimum and 3) maximum costs per article.

To determine exchange rates we consulted http://www.xe.com on 12. September 2016. Since value-added taxes vary by country publishers usually list costs excluding VAT. APC listed here do not include VAT.

Table 3: File details: allOAPubsWithCorrAuthor.xlsx

Sheet name	Column name	Note
data		raw script output
eval. by inst.		data analysis: breakdown of OA articles by analysed institutions
publishers TOP20		data analysis: breakdown of OA articles by publishers: cumulated number of articles and APC costs
APC costs		raw script output enriched with data on APC costs for TOP 20 publishers – according to publishers' websites (July/August 2016)
~	APC exact amount	if publisher lists an exact APC the amount is given in the original currency [e. g. PLOS]
~	APC min	if publisher lists a price range the minimum price given in the original currency [e. g. Frontiers]
~	APC max	if publisher lists a price range the maximum price given in the original currency [e. g. Frontiers]
~	APC mean of min/max	if publisher lists a price range the mean price given in the original currency [e. g. Frontiers]
~	APC approx. price	if publisher lists calculation examples (e. g. charge per page) the approximate price is given in the original currency [e. g. Copernicus]
~	APC MEAN in Euro	if no exact price is given mean amount of APC (excl. VAT) in Euro is listed; column sum represents mean APC costs for 2013–2015 (excl. VAT, in Euro)
~	APC MIN in Euro	if no exact price is given the minimum APC (excl. VAT) in Euro is listed for each article; column sum represents minimum APC costs for 2013–2015 (excl. VAT, in Euro)
~	APC MAX in Euro	if no exact price is given the maximum APC (excl. VAT) in Euro is listed for each article; column sum represents maximum APC costs for 2013–2015 (excl. VAT, in Euro)

Note detailed information on open access articles with Berlin corresponding author in open access journals

Table 4: File details: hybridPubsWithCorrAuthors.xlsx

Sheet name	**Column name**	**Note**
data		raw script output
eval. by inst.		data analysis: breakdown of OA articles in hybrid journals by analysed institutions
publisher		data analysis: breakdown of OA articles in hybrid journals by publishers: cumulated number of articles and APC costs
APC costs		reduced article metadata and data on APC costs – according to publishers' websites (July/August 2016)
~	Berlin inst.	affiliation of corresponding author (checked manually)
~	license	license URL extracted from Crossref metadata
~	APC exact amount	if publisher lists an exact APC the amount is given in the original currency [e. g. Elsevier]
~	APC min	if publisher lists a price range the minimum price given in the original currency [e. g. ACS]
~	APC max	if publisher lists a price range the maximum price given in the original currency [e. g. ACS]
~	APC mean of min/max	if publisher lists a price range the mean price given in the original currency [e. g. ACS]
~	APC MEAN in Euro	if no exact price is given mean amount of APC (excl. VAT) in Euro is listed; column sum represents mean APC costs for 2013–2015 (excl. VAT, in Euro)
~	APC MIN in Euro	if no exact price is given the minimum APC (excl. VAT) in Euro is listed for each article; column sum represents minimum APC costs for 2013–2015 (excl. VAT, in Euro)
~	APC MAX in Euro	if no exact price is given the maximum APC (excl. VAT) in Euro is listed for each article; column sum represents maximum APC costs for 2013–2015 (excl. VAT, in Euro)

Note detailed information on open access articles with Berlin corresponding author in hybrid journals

4 Re-use cases

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

So far data was analysed on a multi-instutional level. 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 kind 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.

The method to detect open access articles in hybrid journals is experimental – one might evaluate other data sources than Crossref.

It is rather complex to identify the share of green open access (i. e. self-archiving of articles published in subscription-based journals). Thus, the study focused on the share of gold open access. Considering the objectives of Berlin's open access strategy (e. g. journal articles with 60 % open access share by 2020) a method is needed to identify both the green and gold open access share.