Disclosing funding sources for open access publication fees: the Open APC initiative

Najko Jahn[1] **and Marco Tullney**[2]

[1]**Bielefeld University Library, Bielefeld University, Bielefeld, Germany**
[2]**Technische Informationsbibliothek (TIB) - German National Library of Science and Technology, Hannover, Germany**

ABSTRACT

Publication fees in open access publishing hold a prominent place on the agenda of researchers, policymakers, and academic publishers. This paper contributes to the evolving empirical basis on open access funding. It describes the Open APC initiative, in which German universities and research organizations share their expenditures for publication fees. As method, the initiative uses existing open data tools to aggregate and disseminate institutional spending on open access publication fees. In total, 29 German research organizations self-reported funding of 6,279 open access journal articles, which amounted to 8,039,339 €. The average payment for each article was 1,280 €, and the median payment 1,209 €. Our data-set comprises only 53 articles in hybrid journals. With an indexing coverage of 99 %, the findings reveal that the DOI agency CrossRef provides both comprehensive bibliographic coverage of the funded open access journal literature and disambiguated names of journal titles and publishing houses. We show that authority control of these bibliographic information is particularly relevant for the comparative study of the economical effects of open access publishing.

Keywords: Open access, open access journal, scholarly publishing, publication fees, article processing charges, science policy

INTRODUCTION

Publication fees, often paid by funders or universities, are a widely discussed open access business model. Yet, how and to what extent these activities are effective in terms of the number of supported research articles and associated costs remains under debate. This paper describes the Open APC initiative[1], in which German universities and research organization share spendings on open access publication fees, and how it is currently implemented. More specifically, it addresses three problem areas when studying the economical effects of open access publishing: fragmentation of open access funding, variable pricing schemes and the comparison across research institutions. Such an approach extends methods and improves data collection activities for researchers and practitioners, as well as contribute to a better understanding of factors affecting the analysis of publication fees in open access publishing.

The rise of open access journals matches the increasing relevance of publication fees in academic publishing (Davis and Walters, 2011; Laakso and Björk, 2012; Pinfield, 2015). To cover these fees, authors tend to make use of funding that grant agencies or academic institutions provide (Suber, 2012). However, collecting information about what was funded is in most cases difficult. One reason why payments made for open access journal publications are often hard to track is that, on the one hand, they are fragmented across the budgets of funding agencies, research institutions, and libraries, and, on the other hand, taken from personal budgets. Furthermore, open access funding mostly exists in higher income countries, mainly to support research articles in the bio- and physical sciences (Solomon and Björk, 2011). Personal budgets stand in contrast with those support structures and are likely used to cover low price publication fees (Björk, 2015; Solomon and Björk, 2011). Along with the fragmentation of payments, funding for open access publications lacks transparency because the parties involved - authors,

[1] `https://github.com/openapc/openapc-de`

universities, funders, publishers - neither release information on who pays for what nor the costs of publishing (Björk and Solomon, 2014), a situation similar to the lack of transparency regarding journal subscriptions (Lawson and Meghreblian, 2015). It also remains unclear which factors contribute to price formation.[2] While fixed prices for individual articles are common, agreements between publishers and institutions often provide discounts and publishers sometimes waive publication fees for authors from low-income countries (Björk and Solomon, 2012; Lawson, 2015b). Other factors leading to a complex landscape of variable pricing schemes (Pinfield et al., 2015) include submission or page charges (Björk and Solomon, 2012). Hybrid journals substantially add to this complexity, because comprehensive offset systems to avoid paying for the same article twice, through subscription and publication fee, are rare, which, in turn, leads to the phenomena of "double dipping" in scholarly publishing (Pinfield et al., 2015).

This complex situation of fee-based open access publishing creates difficulties for researchers and practitioners alike. Because of fragmented payments, the extent of funding remains unclear. To increase transparency, some research funders have begun collecting and disclosing expenditures for open access journal articles as open data. As per definition, open data is data that "can be freely used, modified, and shared by anyone for any purpose" (Dietrich et al., 2016). Therefore, opening up information about the funding of open access journal articles promises to enhance the discussion about current and future business models in academic publishing. To our knowledge, the first research funders providing such data were the Wellcome Trust (Kiley, 2014) and the Austrian Science Fund FWF (Reckling and Kenzian, 2014), who both released data on publication fees they had funded. The British not-for-profit company Jisc followed by collecting data from higher-education institutions in the UK (Lawson, 2015a). Disclosed as publicly available spreadsheets, these data-sets self-report expenditures along with bibliographic information, including title, journal and publisher, persistent identifier to the publisher's version, and a link to a deposit in a subject repository. Curatorial efforts focused on the disambiguation of publisher and journal titles as well as on detecting duplicates. In the case of the Wellcome Trust, crowd-sourcing data cleaning activities through a Google spreadsheet in combination with checks against bibliographic sources massively improved the spending data (see comments in Kiley (2014)).

The open access landscape in Germany, which is the focus of this paper, shares the general problems of in-transparency regarding funding schemes and costs as discussed above. The Deutsche Forschungsgemeinschaft (DFG), the largest research funder in Germany, has been encouraging open access publishing since years. It launched its "Open-Access Publishing" program in 2009 that has strongly influenced the support of open access publication fees through funds managed by university libraries.[3] With this program, the DFG aims to help universities to establish support structures for publishing in open access journals where authors are requested to pay a publication fee. To reduce administrative burdens, grantees agree not only to reimburse the bills on behalf of the researchers they support, but also to look for ways to improve the handling of those financial transactions. Examples include central invoicing schemes and related agreements between university libraries and publishers. Grantees are also required to report the institutional publication output and their fees paid for open access journal articles to the DFG on a regular basis, and to present the university-wide strategy to sustain the funds when DFG's initial support runs out lately in 2019. The DFG enforces a set of criteria grantees have to comply with, leading to similar implementations for supporting open access publishing across German universities: these criteria exclude sponsoring of articles in hybrid journals, and the funding of articles whose publication fee exceeds 2,000 € (excluding VAT) (Fournier and Weihberg, 2013). Research institutes organized in the Fraunhofer-Gesellschaft, Helmholtz-Gemeinschaft, Leibniz-Gemeinschaft, and Max-Planck-Gesellschaft are not eligible for this funding program, contributing to the diversity of schemes in Germany. In response, some organizations have adopted similar processes to support authors. The Max-Planck-Gesellschaft operates their long-lasting open access activities, including handling spending and publisher agreements centrally, through the Max Planck Digital Library (Schimmer et al., 2013; Sikora and Geschuhn, 2015), while the Leibniz-Gemeinschaft set up a dedicated open access fund in 2016.

The growing share of articles published in fee-based open access journals in recent years has led to calls for an unified approach towards funding of publication fees. The Allianz der Wissenschaftsorganisationen[4], representing all major research organizations in Germany, thus marks transparency as a major means to

[2] These might include article processing, impact, rejection rates, management and investment, and profit margins. See Noorden (2013) for a general discussion and Gumpenberger et al. (2012) and Björk and Solomon (2015) for discussions of journal impact and quality.

[3] Guidelines for the funding program can be found here: `http://www.dfg.de/formulare/12_20/`

[4] `http://www.dfg.de/en/dfg_profile/alliance/index.html`

sustain an “adequate open access publication system” (Bruch et al., 2015). However, there are various ways to achieve this goal. The existing approaches in Austria and the United Kingdom have one institution in charge to collect and analyze the data. The history of the Open APC initiative is rather bottom-up: In May 2014, Bielefeld University Library began to share its expenditures for publication fees. The library put its approach to the working group “Electronic Publishing” of the Deutsche Initiative für Netzwerkinformation (DINI)[5] as a basis for discussion, and invited others to participate. Reflecting the increasing demand for publicly available data, contributions from Universität Regensburg and Universität Hannover followed soon after. As of writing, 29 universities and research institutes voluntary reported their data to the Open APC initiative to be included into a unified data-set of all expenditures.

In this paper, we present the technical workflow of the Open APC initiative. We describe how the data-set is curated and which tools are used in order to produce, disseminate, and preserve the Open APC spending data. Presenting our results, we will particularly discuss how and to what extent the use of the CrossRef index accommodates the demand of disambiguated bibliographic information about journals and publishers when reporting about funding of open access journal articles. CrossRef, a Digital Object Identifiers (DOI) registration agency for scholarly literature, associate these persistent links with metadata CrossRef members such as publishers and research societies contribute.

METHODS AND MATERIALS

The major goal of the Open APC initiative is to gain insights into expenses for publication fees by collecting data directly from the institutions paying on behalf of the authors they support. These institutions can report best on the most important part: How much they have spent for each article. For this aim, the Open APC initiative applies open data methods, designed around the idea of sharing the collected data as permissive as possible. At its core is a data pipeline to collect and to process the contributed data. In this section, we discuss the general idea of this pipeline and describe its stages: data submission (by spending institutions), merging contributions, re-using data, preserving data, and engaging with our communities to increase participation.

General idea

Our approach drew on general open data guidelines that state to “keep things simple” and to “engage early and engage often” with data providers and potential users (Dietrich et al., 2016). We therefore chose a simple data scheme, built our pipeline around the popular social coding platform GitHub, and re-used information from bibliographic indexes.

We followed good practice of other open data projects to guarantee that the data is as open as possible. From the beginning, we emphasized that this also includes information on submissions – date of submission, contributors, etc. – and open file formats. Because institutions should be allowed to self-report data and updates at different times, tracking initial submissions and data-set updates is particularly important.

All data and documentation reside, therefore, on GitHub. Git, a distributed version control system, powers GitHub and allows people to collaborate on software projects. Git keeps a log of changes made in the source code and manages to synchronize local copies of the very same software repository. Because of its distributed and social characteristics, Git in general and GitHub in particular are suitable for researchers to mutually curate and share research artifacts including data-sets, analyses, or visualizations (Ram, 2013).

In addition to GitHub, bibliographic indexes are an essential part of our data pipeline. We use CrossRef to normalize bibliographic metadata. This ensures automatic authority control for journal and publisher titles, which are the most appropriate levels of aggregations when analyzing expenditures for open access articles (Pinfield et al., 2015). We also check automatically the indexing status of each article in Europe PubMed Central, a large bibliographic index for life-science literature, the multidisciplinary database Web of Science, and the open access source Directory of Open Access Journals (DOAJ).

Data Submission

Participants use a template to self-report data (see Table 1). Data structured in this way enables the fetch of additional data from external sources, and ensures that the merge into a single data-set succeeds. By collecting as little data as possible from the institutions, this approach allows for changes and the addition of new fields if required.

[5] http://dini.de/english/ag0/e-pub0/

Table 1. Mandatory and optional data elements for disclosing funding of open access publication fees

Variable	Description	Required
`institution`	Top-level organisation e.g. MPG	mandatory
`period`	Year of APC payment (YYYY)	mandatory
`euro`	The amount paid in EURO (incl. VAT)	mandatory
`doi`	Digital Object Identifier	mandatory
`is_hybrid`	Published in a access journal?	mandatory
`publisher`	Name of publication house	optional
`journal_full_title`	Full name of periodical	optional
`issn`	International Standard Serial Number.	optional

The data scheme reflects how the Wellcome Trust (Kiley, 2014), the Austrian Science Fund FWF (Reckling and Kenzian, 2014) and Jisc (Lawson, 2015a) organized data about expenses for open access publication fees. The project requests items that are already present at the research institutions for internal reporting purposes (Pinfield et al., 2015). Guidance[6] available to the participants furthermore refers to the principles of *tidy data* (Wickham, 2014). This responds to issues encountered in the United Kingdom, when the inexperienced use of spreadsheet software like Excel lead to misaligned tables (Woodward and Henderson, 2014). The principle of tidy data intends to reduce data cleaning efforts before statistical analyses. The three principles of tidy data state that, firstly, each variable must form a column, secondly, each observation must form a row, and, finally, each type of observational unit must form a table. To prevent messy data, it is also crucial that unknown values are not left empty. In our case, we use the R convention `NA` for handling those values.

Data contributors collect the data in different systems. After preparing the data, it must be exported in the csv standard, a plain text file format for tabular data supported by most spreadsheet software that can be easily logged with version control systems like Git.

To illustrate a sample data contribution displaying mandatory information in the csv format:

```
"Hannover U",2013,1241.02,"10.1371/journal.pone.0063501",FALSE
```

Contributors store the csv file into a folder that also contains a README file. Written in plain text or Markdown, the README contains information about the data-set and the contributing institution. The submission itself is facilitated through GitHub's pull request mechanism: contributors fork the initiative's data, add their contributions, upload their modified copy to their repository, and request importing their data into the initiative's repository by alerting the maintainers through sending a "pull request".

Data curation

By using open software, we, as curators, retrieve additional data elements based on the article's DOI. This avoids manual cleaning efforts the other initiatives were faced with. Because the DOI is at the center of the data curation, data normalization starts with a check of the DOI columns for possible duplicates and white space. CrossRef is then queried for article metadata matching the particular DOI.

CrossRef provides several APIs to search for bibliographic information it indexes during DOI registration. In our case, we query the CrossRef REST API[7] by using content negotiation. As resource type, we request the format `application/vnd.crossref.unixsd+xml`, which main function is to support text mining activities.[8] The advantage of this XML format is that it distinguishes full and abbreviated journal titles as well as the media types of ISSNs, the International Standard Serial Number used to identify journals. It also contains license information and disambiguated publisher information, thus avoiding confusion about licensing and naming of publisher houses (Woodward and Henderson, 2014).

As a client, we use the R package `rcrossref` (Chamberlain et al., 2016) developed and maintained by the rOpenSci initiative.[9] With the function `cr_cn`, the client supports all linking types. Table 2 summarizes the data elements we retrieve:

[6] `https://github.com/OpenAPC/openapc-de/wiki/Handreichung-Dateneingabe`

[7] `https://github.com/CrossRef/rest-api-doc/blob/master/rest_api.md`

[8] CrossRed: Text and Data Mining for Researchers: `http://tdmsupport.crossref.org/researchers/`

[9] rOpenSci: `https://ropensci.org/`

Table 2. Sources used to automatically enrich the Open APC data-set

Source	Data element	Description
CrossRef	`publisher`	Title of Publisher
CrossRef	`journal_full_title`	Full Title of Journal
CrossRef	`issn`	International Standard Serial Numbers (collapsed)
CrossRef	`issn_print`	ISSN print
CrossRef	`issn_electronic`	ISSN electronic
CrossRef	`license_ref`	License of the article
CrossRef	`indexed_in_crossref`	Indexed in CrossRef? (logical)
EuropePMC	`pmid`	PubMed ID
EuropePMC	`pmcid`	PubMed Central ID
Web of Science	`ut`	Web of Science record ID
DOAJ	`doaj`	Is the journal indexed in the DOAJ? (logical)

In addition to CrossRef, the indexing status for each article in Europe PubMed Central and the Web of Science is checked. Europe PubMed Central, one of the largest database for life science literature, offers a public RESTful web services to access more than 24 million records and 870,000 deposited open access articles (Europe PMC Consortium, 2014). Information from the Web of Science, which provides indexing of open access journal literature (Walters and Linvill, 2011), is retrieved through the Thomson Reuters Article Match Retrieval Service[10]. This web interface is available to Bielefeld University Library as part of its Web of Science subscription. Finally, an automatic match with the DOAJ, a comprehensive and openly available registry of open access journals, is performed.

After disambiguating and enriching the data, we append the rows into a single csv file that records all expenses. The data additions are logged with Git and pushed to the source code repository on GitHub.

Re-use

We use the main README file of our data repository as general guide to the Open APC data collection and to present sample analyses. The README itself is written in R Markdown, an authoring format that allows for the combination of the R code, results, and text within one document.[11] Compiled with knitr (Xie, 2016), the README reflects new data submissions after every build. Sample results include tables summarizing the number of supported articles and costs per institution as well as figures that illustrate the distribution of publication fees.

A blog hosted on GitHub presents all new data contributions.[12] Since it is technically based on Jekyll,[13] a static site generator, we can use the same development cycle we have for producing the README files. We write blog posts, which include information about the providing institution and the contributed data-set, in R Markdown, build the posts with `knitr` and, then, log and deploy the files to GitHub with Jekyll, making it possible to reproduce the analytical steps.

Preservation

The data-set itself is licensed under the Open Database License (ODbL) v1.0,[14] permitting re-use under conditions of attribution and share-alike. To preserve the collected data, a GitLab installation hosted by Bielefeld University and administrated by Bielefeld University Library, mirrors a copy of the GitHub repository and is accessible via a DOI.[15]

To provide reference points to particular data contributions, unique version names refer to data submissions or other changes in the source code, which will be then distributed as releases through GitHub. This approach allows others to not only to refer, but also to re-use or self-archive particular snapshots of the data.

[10] Thomson Reuters Article Match Retrieval Service: `http://wokinfo.com/directlinks/amrfaq/`

[11] `http://rmarkdown.rstudio.com/`

[12] `http://openapc.github.io`

[13] `https://jekyllrb.com/`

[14] `http://opendatacommons.org/licenses/odbl/1-0/`

[15] `http://dx.doi.org/10.4119/UNIBI/UB.2014.18`

Engagement

Besides technical measures, social aspects are crucial to make an open data initiative successful. For the Open APC initiative, we focus on engaging librarians and collaborating within existing networks. In Germany, the working group “Elektronisches Publizieren” of the Deutsche Initiative für Netzwerkinformation (DINI) coordinates these efforts.[16] The Open APC initiative aims at increasing participation also through the use of social media, workshops, and regular community calls.

A GitHub wiki[17] shares guidance and interim reports. If participants want to report bugs or propose new functionality, GitHub’s “issues” mechanism involves users while keeping track of the discussion. To ensure a constructive environment for all participants, we refer to the code of conduct from Contributor Covenant.[18]

RESULTS

Cost Data

On February 19th 2016,[19] the Open APC initiative covered 6,279 articles, whose publication fees were centrally paid by 29 German universities and research institutions. The number of supported open access journal articles, which were reported to this initiative, grew over the years (see Figure 1). While one institution disclosed 5 payments made in 2005, the majority shared their expenditures from 2013 onwards. With 1,846 articles, the year 2014 was best represented in our data-set. Because of a time lag between payments made and reporting them to the Open APC initiative, only 17 institution were able to partly contribute their cost data for 2015 at the time of this analysis.

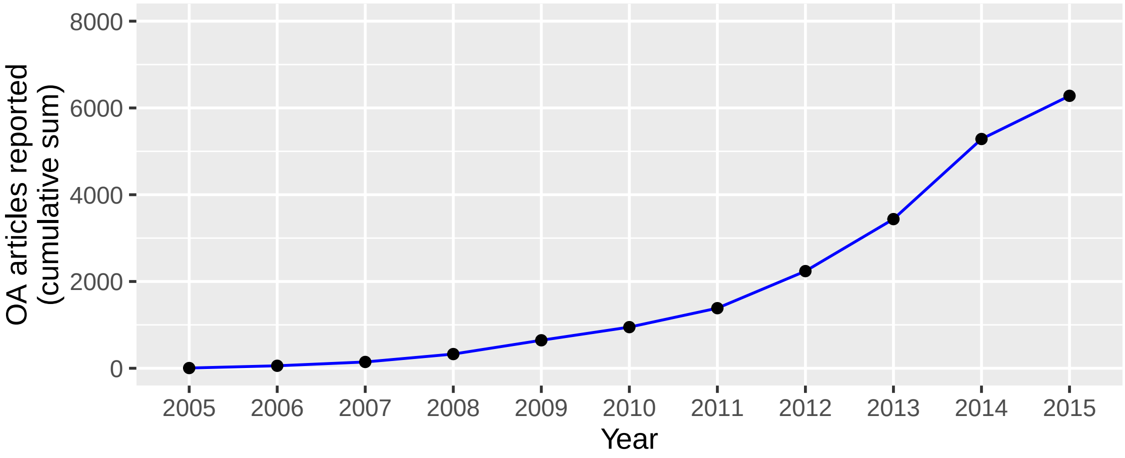

Figure 1. Growth of Open APC Initiative

Among all articles, fees amounted to 8,039,339 € including VAT, the average payment was 1,280 € and the median value 1,209 € . Figure 2 shows the large price variation among the articles. The disclosed publication fees ranged from 40 € to 7,419 €. However, the average price paid varied somewhat during the period 2011 and 2014 (1249 - 1289 €). We also observe that 5,967 (95%) of the publication fees were paid in accordance with the DFG price cap of 2,000 €. Whereas related open data initiatives in Austria and the United Kingdom reported a large share of spending for hybrid journal articles, the situation in Germany is different: only 53 articles in hybrid journals were reported by 3 out of 29 research institutions, accounting for 0.84 % of the overall payments.

The number of APC payments per institutions varied considerably (see Table 3). With 2,796 reported articles, the Max Planck Society contributed 45 % of the overall submissions. In contrast, the two universities of technology, TU Clausthal and TU Ilmenau, who recently begun to set up support structures for fee-based open access journal articles, shared payments made for four articles each.

[16] http://dini.de/english/ag0/e-pub0/

[17] https://github.com/OpenAPC/openapc-de/wiki

[18] http://contributor-covenant.org/

[19] The data is openly available on GitHub. The following analysis is based on version 2.1.13 of the dataset, available at https://github.com/OpenAPC/openapc-de/tree/v2.1.13.

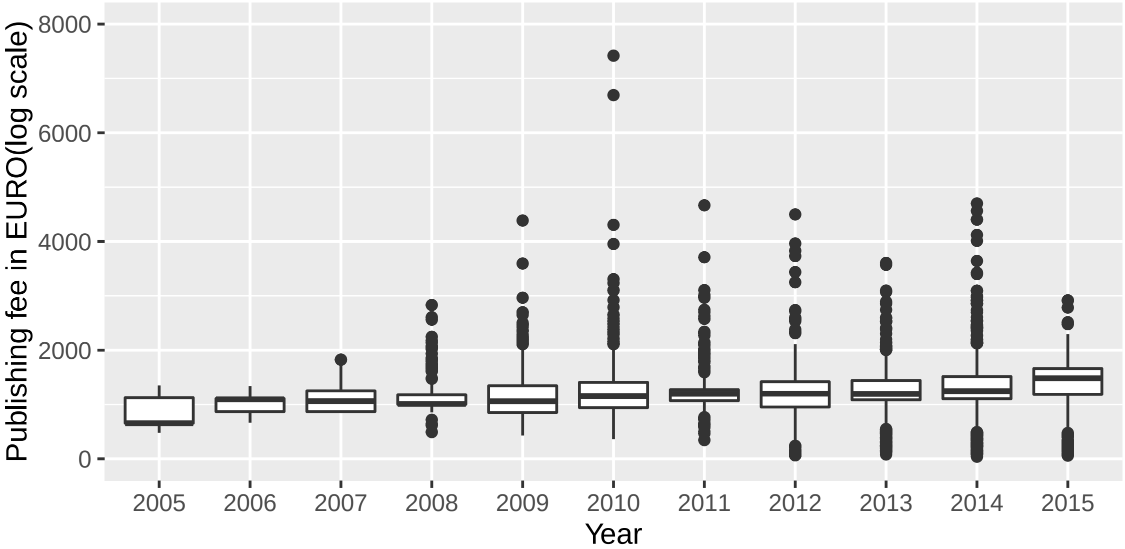

Figure 2. Payments per year

Table 3. Institutions self-reporting expenditures for open access publications (in €)

Institutions	Articles funded	Total	Mean	Median	Minimum	Maximum
MPG	2,796	3,577,537	1,280	1,165	104	7,419
Muenchen LMU	365	463,491	1,270	1,299	496	2,023
Goettingen U	313	409,930	1,310	1,285	180	4,121
KIT	291	344,131	1,183	1,178	69	3,731
Regensburg U	280	331,718	1,185	1,183	77	4,403
Bielefeld U	261	321,475	1,232	1,232	142	2,103
Giessen U	243	326,082	1,342	1,247	81	4,498
Konstanz U	223	304,182	1,364	1,342	40	2,072
Heidelberg U	215	308,348	1,434	1,500	60	2,042
Wuerzburg U	207	286,543	1,384	1,447	105	2,514
Leipzig U	168	236,376	1,407	1,481	341	2,047
Duisburg-Essen U	114	136,911	1,201	1,214	238	1,982
FU Berlin	104	139,284	1,339	1,283	220	2,000
TU Muenchen	103	123,054	1,195	1,269	131	2,046
FZJ - ZB	94	109,701	1,167	1,091	370	2,784
TU Dresden	78	96,046	1,231	1,242	200	1,944
Bochum U	70	91,951	1,314	1,437	100	2,042
Hannover U	69	90,259	1,308	1,241	149	2,159
GFZ-Potsdam	60	69,625	1,160	1,062	438	4,403
Bayreuth U	57	64,519	1,132	1,104	82	1,969
TU Chemnitz	36	37,826	1,051	1,142	78	2,123
Kassel U	35	35,550	1,016	1,142	150	1,861
MDC	34	61,519	1,809	1,348	491	4,700
Hamburg TUHH	24	32,789	1,366	1,466	300	2,027
Bamberg U	16	15,932	996	960	90	2,010
Dortmund TU	9	8,238	915	900	155	1,738
INM - Leibniz-Institut für Neue Materialien	6	8,505	1,418	1,492	237	2,454
TU Clausthal	4	3,771	943	969	460	1,374

Institutions	Articles funded	Total	Mean	Median	Minimum	Maximum
TU Ilmenau	4	4,043	1,011	1,201	178	1,462

CrossRef indexing

Along with the price information, participating institutions were required to identify funded articles by their DOI. They were reported for 6,250 out of 6,279 articles. Of those, 6,228 were indexed in CrossRef, representing 99 % of all funded publications. The reasons why articles identified by a DOI were not registered with CrossRef differed. Some journals were not indexed by CrossRef at the time of our study but by the DOI agencies DataCite (Journal of new frontiers in spatial concepts published by KIT Scientific Publishing) and Medra (DIE ERDE: Journal of the Geographical Society of Berlin). In other cases, either the DOI did not refer to the full text despite the fact that the journal was indexed on a regular basis (compare `http://doi.org/10.1186/1471-2105-13-S19-S7` with `http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-S19-S7`) or the resource type could not be retrieved, although the DOI resolves (`http://doi.org/10.1186/s12885-015-1795-7`).

Cost data by publisher and journal

We used the DOI to automatically fetch publisher and journal names for each article from the CrossRef REST API. Table 4 shows the top ten publishers in terms of payments made that represent 92 % of the spending for publication fees. In total, payments were made to 117 publishing houses. In comparison with data from the UK, full open access publishers have a greater share on total spending. Pinfield et al. (2015), for instance, reported remarkably lower numbers for the open access publishers MPDI AG, Copernicus GmbH, and Hindawi Publishing.

Table 4. Publication fees paid per publisher (in €)

Publisher	Articles funded	Total	Mean	Median	Minimum	Maximum
Springer Science + Business Media	1,751	2,301,460	1,314	1,327	81	2,042
Public Library of Science (PLoS)	1,486	1,972,460	1,327	1,202	556	2,790
Frontiers Media SA	783	1,014,279	1,295	1,106	77	2,380
Copernicus GmbH	676	957,869	1,417	1,299	104	7,419
IOP Publishing	657	674,644	1,027	943	589	1,950
MDPI AG	151	166,948	1,106	1,170	154	2,055
Hindawi Publishing Corporation	91	87,846	965	926	200	2,159
Optical Society of America (OSA)	78	118,798	1,523	1,599	499	3,731
Nature Publishing Group	60	103,174	1,720	1,386	934	4,403
Wiley-Blackwell	57	89,722	1,574	1,457	491	3,000
other	489	552,139	1,129	1,058	40	4,700

Most of the funding of publication fees in Germany went to the publisher Springer Science + Business Media, especially profiting from the merge with the former full open access publisher BioMed Central. In contrast, other established publishing houses such as Elsevier and Wiley-Blackwell rank lower, presumably because they mostly publish hybrid journals, which were not well represented in our data-set at the time of the study. Table 4 also illustrates the variation across and within publishers, which confirms earlier findings (Pinfield et al., 2015).

Table 5. Publication fees paid per journal (in €)

Journal	Articles funded	Total	Mean	Median	Minimum	Maximum
PLOS ONE	1,252	1,500,643	1,199	1,185	749	1,809
New Journal of Physics	653	668,829	1,024	943	589	1,856
Atmospheric Chemistry and Physics Discussions	254	400,183	1,576	1,410	234	7,419
Frontiers in Psychology	241	321,747	1,335	1,142	77	2,123
BMC Genomics	125	164,017	1,312	1,273	920	1,926
Biogeosciences Discussions	115	173,084	1,505	1,331	664	3,641
BMC Bioinformatics	104	128,452	1,235	1,230	655	1,661
Frontiers in Human Neuroscience	100	130,243	1,302	1,106	575	2,000
Frontiers in Plant Science	91	104,067	1,144	1,106	551	2,380
Atmospheric Measurement Techniques Discussions	86	122,927	1,429	1,325	535	3,709
other	3,258	4,325,147	1,328	1,333	40	4,700

Prices also varied within single journals. Based on the number of articles paid for, Table 5 illustrates the top ten out of 640 journals. Payments to these ten journals represent 48 % of all payments. In the case of Atmospheric Chemistry and Physics Discussions, the price range can be explained by the fact that this journal charges per page and also takes the submission's file format into consideration.

The data-set finally confirms the leading role of "mega-journals" in open access publishing, including the multidisciplinary PLOS ONE and the journals New Journal of Physics, Atmospheric Chemistry and Physics Discussions and Frontiers in Psychology, all of which publish contributions from all branches of their respective discipline. In general, an estimated 14 out of more than 10,000 journals registered in DOAJ in 2015 accounted for up to 15–20 % of all articles published in full open access journals (Björk, 2015).

DISCUSSION

The Open APC initiative extends existing methods to disclose spending on open access publication fees. Our workflow benefits from openly available tools and the social coding platforms GitHub, both of which are well established and suited to increase transparency in research (Peng, 2011; Ram, 2013). For 99 % of the articles, CrossRef provided bibliographic information, which substantially contributed to a uniform data-set about formerly fragmented payments made for open access articles.

Although CrossRef disambiguates journal titles and publisher names and is therefore an authority-controlled source for open access journal literature, derivations from CrossRef metadata curation as well as the context of aggregation must be made clear. In particular, problems persist on how to deal with name changes and ongoing mergers. For example, the publisher Public Library of Science (PLOS) has changed its acronym from "PLoS" to "PLOS", which CrossRef metadata reflects from 2015 onward. We therefore normalized all PLOS journal titles in order to secure unique reference to these journals.

Another publisher affected is “The Optical Society,” formerly “Optical Society of America”. Because the ownership of publishing houses can become combined, dealing with mergers is also essential to make cost data comparable. Jisc data, for instance, differentiate between the full open access publisher BioMed Central and the traditional publisher Springer, concluding that “traditional publishing houses” lean on the hybrid model (Pinfield et al., 2015). This stands in stark contrast to our approach, in which CrossRef metadata reflects the merger of BioMed Central and Springer, resulting in Springer Science + Business Media to be the best represented publisher for articles in full open access journals in the Open APC data-set. Another approach for ensuring unique reference, but that we have not evaluated yet, is to use CrossRef’s identifiers for journals and publishers instead.[20]

Because of the dynamic landscape of academic publishing and its representation in CrossRef’s data curation efforts, it is important to consider the time-frame of metadata aggregation. In our case, we re-used metadata shortly after the data submission. However, for some cost analysis – for instance to prepare negotiations with publishers on future schemes to fund open access journal articles – it could be more feasible to re-normalize the complete data. The rossRef API provides incremental metadata updates that can be used to assess the current potential of future funding. While licensing information is incompletely covered in the CrossRef index so far, and therefore not analyzed in our study, the growing importance of facilitating text mining may result in more and more publishers sharing this information with CrossRef in the future.

Participation is voluntary. Therefore, not all institutions in Germany that provide central funding of publication fees contribute cost data to this initiative. In a qualitative survey that also asked why German institutions are reluctant to share their cost data through the Open APC initiative one institution feared that increase in transparency would allow publishers to adjust prices in their favor. Others pointed out that the workload to produce such a data-set could be too extensive (Deppe, 2015).

While there may still be institutions that have no overview of their APC spending, we would like to emphasize that reporting data that is already available within an institution to the Open APC initiative should not lead to much additional work. The central incentive of this initiative is to make it as easy as possible to submit data to it. Being able to combine that data into one standardized data-set increases transparency and comparability, and gives institutions a better understanding of the overall development of open access publishing. In particular, we cannot see the harm of increased transparency; in fact, not knowing how much is spent is undoubtedly a disadvantage in dealing with publishers.

Extending the data template to include information about funders or whether special agreements with publishers applied as suggested by Pinfield et al. (2015), would even increase the efforts needed to participate. However, with the growing demand for action in areas like the large-scale transition of toll-access journals to open access (Schimmer et al., 2015), an updated data template could help institutions to better comply with these policy developments in the future. From our experience, another barrier to participate is the lack of skills in version control: the data submission itself is not always made directly by the institutions. Instead, they sent files to Bielefeld University Library with the request to make them available on GitHub on their behalf.

Future work needs to focus on analyzing the cost data. Of particular interest are questions concerning the coverage of central funding schemes in comparison with open access publication output in general, and the use of other means to cover publication fees in particular. The Open APC initiative does not cover personal budgets or make price reductions explicit, but studies suggest that there is a possible gray area (Björk, 2015; Björk and Solomon, 2012; Lawson, 2015b). Existing study designs could be re-applied to examine the relationship between price on the one hand, and indexing coverage, journal prestige or management costs on the other (Björk and Solomon, 2015, Pinfield et al. (2015); Walters and Linvill, 2011). This, in turn, helps to address the central question of future business models in scholarly publishing from an international perspective (Pinfield et al., 2015).

ACKNOWLEDGMENT

We thank Andrea Hacker and Ada-Charlotte Regelmann for valuable comments on the first draft of this paper. We also thank Christoph Broschinski, Vitali Peil, and Dirk Pieper, the members of the DINI working group “Electronic publishing”, and all data contributors[21] of the Open APC initiative.

[20] We would like to thank Martin Fenner for pointing this out to us. See also the CrossRef API documentation https://github.com/CrossRef/rest-api-doc/blob/master/rest_api.md

[21] https://github.com/OpenAPC/openapc-de#contributors

REFERENCES

Björk, B.-C. (2015). Have the 'mega-journals' reached the limits to growth? *PeerJ* 3, e981. http://doi.org/10.7717/peerj.981.

Björk, B.-C., and Solomon, D. (2012). Pricing principles used by scholarly open access publishers. *Learned Publishing* 25, 132–137. http://doi.org/10.1087/20120207.

Björk, B.-C., and Solomon, D. (2014). How research funders can finance APCs in full OA and hybrid journals. *Learned Publishing* 27, 93–103. http://doi.org/10.1087/20140203.

Björk, B.-C., and Solomon, D. (2015). Article processing charges in OA journals: relationship between price and quality. *Scientometrics* 103, 373–385. http://doi.org/10.1007/s11192-015-1556-z.

Bruch, C., Deinzer, G., Geschuhn, K., Haetscher, P., Hillenkoetter, K., Kress, U., et al. (2015). Positions on creating an Open Access publication market which is scholarly adequate : Positions of the Ad Hoc Working Group Open Access Gold in the priority initiative "Digital Information" of the Alliance of Science Organisations in Germany. Ad-hoc-Arbeitsgruppe Open-Access-Gold der Schwerpunktinitiative "Digitale Information" der Allianz der deutschen Wissenschaftsorganisationen. http://doi.org/10.2312/allianzoa.009.

Chamberlain, S., Boettiger, C., Hart, T., and Ram, K. (2016). *rcrossref: Client for Various 'CrossRef' 'APIs'*. Available at: https://CRAN.R-project.org/package=rcrossref.

Davis, P. M., and Walters, W. H. (2011). The impact of free access to the scientific literature: a review of recent research. *Journal of the Medical Library Association* 99, 208–217. http://doi.org/10.3163/1536-5050.99.3.008.

Deppe, A. (2015). *Ansätze zur Verstetigung von Open-Access-Publikationsfonds.*, ed. K. Umlauf Institut für Bibliotheks- und Informationswissenschaft. Available at: http://nbn-resolving.de/urn:nbn:de:kobv:11-100234262.

Dietrich, D., Gray, J., McNamara, T., Poikola, A., Pollock, R., Tait, J., et al. (2016). *Open Data Guide*. Available at: http://opendatahandbook.org/guide/en/.

Europe PMC Consortium (2014). Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Research* 43, D1042–D1048. http://doi.org/10.1093/nar/gku1061.

Fournier, J., and Weihberg, R. (2013). Das Förderprogramm "Open Access Publizieren" der Deutschen Forschungsgemeinschaft. Zum Aufbau von Publikationsfonds an wissenschaftlichen Hochschulen in Deutschland. *Zeitschrift für Bibliothekswesen und Bibliographie* 60, 236–243. http://doi.org/10.3196/186429501360528.

Gumpenberger, C., Ovalle-Perandones, M.-A., and Gorraiz, J. (2012). On the impact of Gold Open Access journals. *Scientometrics* 96, 221–238. http://doi.org/10.1007/s11192-012-0902-7.

Kiley, R. (2014). *Wellcome Trust APC spend 2012-13: data file*. Figshare. http://doi.org/10.6084/m9.figshare.963054.v1.

Laakso, M., and Björk, B.-C. (2012). Anatomy of open access publishing: a study of longitudinal development and internal structure. *BMC Medicine* 10, 124. http://doi.org/10.1186/1741-7015-10-124.

Lawson, S. (2015a). Article Processing Charges Paid by 25 UK Universities in 2014. *Journal of Open Humanities Data* 1. http://doi.org/10.5334/johd.2.

Lawson, S. (2015b). Fee Waivers for Open Access Journals. *Publications* 3, 155–167. http://doi.org/10.3390/publications3030155.

Lawson, S., and Meghreblian, B. (2015). Journal subscription expenditure of UK higher education institutions. *F1000Research*. http://doi.org/10.12688/f1000research.5706.3.

Noorden, R. V. (2013). Open access: The true cost of science publishing. *Nature* 495, 426–429. http://doi.org/10.1038/495426a.

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science* 334, 1226–1227. http://doi.org/10.1126/science.1213847.

Pinfield, S. (2015). Making Open Access work. *Online Information Review* 39, 604–636. http://doi.org/10.1108/oir-05-2015-0167.

Pinfield, S., Salter, J., and Bath, P. A. (2015). The "total cost of publication" in a hybrid open-access environment: Institutional approaches to funding journal article-processing charges in combination with subscriptions. *Journal of the Association for Information Science and Technology*. http://doi.org/10.1002/asi.23446.

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. *Source Code for Biology and Medicine* 8, 7. http://doi.org/10.1186/1751-0473-8-7.

Reckling, F., and Kenzian, M. (2014). *Austrian Science Fund (FWF) Publication Cost Data 2013*. Figshare. http://doi.org/10.6084/m9.figshare.988754.v4.

Schimmer, R., Geschuhn, K. K., and Vogler, A. (2015). *Disrupting the subscription journals' business model for the necessary large-scale transformation to open access*. Max Planck Digital Library. http://doi.org/10.17617/1.3.

Schimmer, R., Geschuhn, K., and Palzenberger, M. (2013). Open Access in Zahlen: Der Umbruch in der Wissenschaftskommunikation als Herausforderung für Bibliotheken. *Zeitschrift für Bibliothekswesen und Bibliographie* 60, 244–250. http://doi.org/10.3196/186429501360532.

Sikora, A., and Geschuhn, K. K. (2015). Management of article processing charges – challenges for libraries. *Insights: the UKSG journal* 28, 87–92. http://doi.org/10.1629/uksg.229.

Solomon, D. J., and Björk, B.-C. (2011). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the Association for Information Science and Technology* 63, 98–107. http://doi.org/10.1002/asi.21660.

Suber, P. (2012). *Open Access*. MIT Press. Available at: https://mitpress.mit.edu/books/open-access.

Walters, W. H., and Linvill, A. C. (2011). Bibliographic index coverage of open-access journals in six subject areas. *Journal of the Association for Information Science and Technology* 62, 1614–1628. http://doi.org/10.1002/asi.21569.

Wickham, H. (2014). Tidy data. *The Journal of Statistical Software* 59. Available at: http://www.jstatsoft.org/v59/i10/.

Woodward, H. M., and Henderson, H. L. (2014). Report for Jisc Collections on total cost of ownership project: Data capture and process. Information Power Ltd. Available at: https://www.jisc-collections.ac.uk/Global/News%20files%20and%20docs/IPL-Jisc-Total-Cost-of-Ownership-Data-Capture-Report.pdf.

Xie, Y. (2016). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* Available at: https://CRAN.R-project.org/package=knitr.