Monitoring OA transformation with APC data analysis final report

SEPT. 2020 | Advancing Open Scholarship

D6.4 – Monitoring OA transformation with APC data analysis final report

Version 1.0 – Final
PUBLIC

This OpenAIRE report is the final document about the analytics, development, and integration of Article Processing Charges (APC). Using the example of the OpenAPC Initiative, one of the largest initiatives for transparent APCs worldwide, the complete integration into OpenAIRE will be demonstrated and thus made possible for subsequent use in the Open Science Observatory, among others.
The report ends with recommendations for repositories, as well as their usage in the OpenAIRE Research Graph, services, and products.

Specifically, it is based on these milestones:

- **MS19 - OpenAPC Workshop for new contributing institutions**
- **MS20 - Technical enhancements of OpenAPC**
- **MS21 - Final Open APC Workshop to promote results**

H2020-EINFRA-2017
Grant Agreement 777541

Document Description

D6.4 – Monitoring OA transformation with APC data analysis final report

WP6 - Towards a Scholarly Commons	
WP participating organizations: UniBI, UGOE	
Contractual Delivery Date: 12/2019	Actual Delivery Date: 07/2020
Nature: Report	Version: 1.0
Public Deliverable	

Preparation Slip

	Name	Organisation	Date
From	Andreas Czerniak Jochen Schirrwagen Najla Rettberg	UNIBI UNIBI UGOE	12/2019
Edited by	Andreas Czerniak	UNIBI	12/2019
Reviewed by	Frank Manista Kathleen Shearer Biljana Kosanovic Camilla Lindelöw Amelie Bäcker	JISC COAR UoB KB UNIBI	08/2020 09/2020
Approved by	Natalia Manola	UoA	10/2020
For delivery	Mike Chatzopoulos	UoA	10/2020

Revision History

Issue	Item	Reason for Change	Author	Organization
V0.1	Draft version	Methodology and initial specifications	Andreas Czerniak *et al.*	UNIBI
V0.2	First Delivery	internal reviewers feedback added		
V1.0	Final version		Andreas Czerniak *et al.*	UNIBI

Table of Contents

Table of Figures

Table of Listing

Disclaimer

This document contains description of the OpenAIRE-Advance project findings, work,and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium head for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the OpenAIRE-Advance consortium and can in no way be taken to reflect the views of the European Union.

OpenAIRE-Advance is a project funded by the European Union (Grant Agreement No 777541).

Acronyms

APC	Article processing charges
API	Application Programming Interface
BPC	Book/monograph processing charges
CSV	comma separated values
DFG	Deutsche Forschungsgemeinschaft (German Research Foundation)
DINI	German Initiative for Networked Information
DOI	Digital Object Identifier
ISO	International Organization for Standardization
NOAD	National Open Access Desk
OA	Open Access
OS	Open Science
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Publishable Summary

In today's scholarly publishing landscape, there are several ways of financing Open Access (OA) publishing. Journal Publishers are looking for income sources to cover the costs of publishing (e.g. for peer-review administration, editorial costs). A variant that has been controversially discussed for many years is shifting costs to the authors' side (or their institutions or funders). This type of publication fee is called Article Processing Charges (APCs).

APCs as a revenue source for open access publishing hold a prominent place on the agendas of researchers, research libraries, universities, policy makers, and academic publishers in recent years. A transparent recording of publication fees provides important indicators in estimating costs for Open Access publishing, provides information on the structure of publication funds and the development of sustainable business models, and can help contain costs.

This report presents the initiative for cost transparency in Open Access - OpenAPC. We analyse the coverage, describe a technical implementation for enriching OpenAIRE metadata with cost information from OpenAPC. We will suggest changes in the OpenAIRE components and supervise them.

Finally, we analyse and discuss how a combination of publication data from OpenAIRE with cost data from OpenAPC can be used to estimate publication costs for different projects and funders in the OpenAIRE context.

The report also incorporates results of webinars and workshops held in the context of OpenAIRE in conjunction with OpenAPC with an international audience from the library and publishing community.

1| INTRODUCTION

There are several ways of financing Open Access (OA) publishing in today's academic publishing landscape. A variant that has been controversially discussed for many years is shifting costs to the authors' side (or their institutions or funders). This type of fee is called Article Processing charges (APC). The transparent recording and reporting of publication fees provides important indicators in the estimation of costs for open access publishing, provides information on the structure of publication funds, the design of sustainable business models, and can help to contain costs. It is a crucial task to help monitor the costs involved in the Open Access and Open Science transition.

As a result, the first national initiatives for the aggregation and monitoring of APC data began to emerge in 2014. These include in particular the Jisc Collections 'Total Cost of Ownership' (TCO) project and the Jisc Monitor for institutions in the UK. In Germany, the OpenAPC initiative was established shortly afterward at Bielefeld University Library with the support of the DINI working group on Electronic Publishing with the aim of making article fees available in a transparent and open manner.

The focus of OpenAPC was initially on APC data from academic institutions in Germany. The institutions, many of which participated in the DFG's program for Open Access publication funds, were made aware of OpenAPC and have since been providing cost data on a voluntary basis. There have been international contributing institutions since 2016 [1,2]. The task and goal in the context of OpenAIRE were to significantly increase the number of institutions in Europe contributing to OpenAPC via the NOADs. The value of OpenAPC increases to the extent that the widest and most complete database possible can be created. It helps to enable a comparison of cost development at the institutional and country-level as well as publisher and journal level. Furthermore, linking OpenAPC with the OpenAIRE research graph would offer additional opportunities to compare costs on the publication level in different disciplines, projects, and research funding agencies.

OpenAPC follows an open science approach by processing, curating, and enriching continuously collected cost data from participating institutions. Open APC publishes the resulting data sets as open data on its GitHub repository[3]. In addition, all scripts used are made available there open source.

This report is based on input from the library and publishing community collected in relation to the OpenAIRE webinars on "*Webinar about transparency of publication fees and the OpenAPC project*"[4] on March 25, 2019, with 103 participants and "*OpenAPC - cost transparency of Open Access publishing*"[5] on October 21, 2019, with 127 participants during the *Open Access Week*

[1] http://openapc.github.io/ (last accessed 31-July-2020)

[2] OpenAPC initiative - https://treemaps.intact-project.org/ (last accessed 31-July-2020)

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

[4] https://www.openaire.eu/item/webinar-about-transparency-of-publication-fees-and-the-openapc-project

[5] https://www.openaire.eu/item/openapc-cost-transparency-of-open-access-publishing

2019, and the workshop "*Monitoring OA publishing costs and revenues of publishing initiatives*"[6] on December 11, 2019, in Paris / France after the General Assembly of OpenAIRE-Advance with 25 participants.

In this report, we first present approaches for cost monitoring using examples of national initiatives and the significance of cost transparency on an international level. In the second section, we present the OpenAPC infrastructure, and, on this basis, in the third section (Methodology) we describe how cost information from OpenAPC can be linked-to publication metadata in OpenAIRE. In the fourth section (Analysis and Results) we present the results on different levels of comparison, discuss the results in the fifth section, and give an outlook on potentials for further development.

1.1 Transparency of Open Access fees

In the last ten years, various business models were established when the transformation of scientific results towards generally open access began. One of these business models is Article Processing Charges (APC). It turned out that APCs can be expensive, decentralized, and opaque. But the need for transparency in APC prices is critical for planning budgets. [1]

Initiatives were formed in a few countries to give priority to these two points. Two of them should be mentioned. On the one hand, the initiative Jisc UK Monitor[7] and the OpenAPC initiative initiated by DINI[8] in cooperation with Bielefeld University Library. Several various APC initiatives had been set up at the national level. In the following, these are briefly presented.

1.1.1 National perspectives

In the United-Kingdom, the report "Article processing charges (APCs) and subscriptions - Monitoring open access costs" [2] provides a detailed analysis of the development of publication fees at higher education institutions in the UK and compares it with the development of subscription fees. They found that the number of article processing charges (APCs) doubled between 2013 and 2014. Growth remained strong in 2015 but slowed in part due to limited room for growth in institutions' internal budgets. And also, the average APC increased by 6% or more over the past years, a rise well above the cost of inflation. This situation should then be examined more in detail.

In Germany, the DINI-EPUB work started the Open APC initiative in 2014 [3] at Bielefeld University. During the project phase of the DFG funded[9] project INTACT – "Transparente

[6] https://www.openaire.eu/blogs/monitoring-oa-publishing-costs-and-revenues-of-publishing-initiatives-workshop-successfully

[7] https://monitor.jisc.ac.uk/uk/about/ (last accessed: 03-August-2020)

[8] DINI-EPUB working group - https://dini.de/ag/e-pub/

[9] https://gepris.dfg.de/gepris/projekt/274983645?context=projekt&task=showDetail&id=274983645&

Infrastruktur für Open-Access-Publikationsgebühren"[10] from 2015 until 2019, the focus was on collecting standardized APCs from institutions and organizations in order to have a simple mechanism for exchange that would be machine-readable, extensible with versioning, and open for re-use. However, it did not remain with the institutions in German-speaking countries but opened itself up in the same way to European and other countries.

In Sweden, the Open APC Sweden [4] project has been initiated by the National Library of Sweden together with Swedish higher education institutions (HEIs) in order to investigate the possibilities of establishing a transparent open access publication cost system. In order to establish Open APC Sweden, there is a great need for a concerted effort by stakeholders. The current project aims to keep the Swedish HEI sector informed about ways in which to monitor the total cost of publication, which includes the collection of APCs. The system is built on the system originally created by the German Open APC project. In 2018, the Swedish government assigned the National Library of Sweden to monitor the costs of scholarly publishing.

1.1.2 Global perspective: OpenAPC Initiative

In the meantime, the OpenAPC Initiative [5,6], originated from the INTACT project, collects data about APCs from different institutions and countries around the world at frequent intervals. The project covers cost data on 104,517 open-access journal articles, amounting to 207,448,024 EUR, contributed by 262 institutions in 16 countries[11]. The updated list of participating institutions is released on GitHub[12].

The APC data from national initiatives like from Sweden and the United Kingdom is included in the OpenAPC dataset.

[10] english: "Transparent infrastructure for Open Access publication fees"

[11] As of June 2020

[12] https://github.com/OpenAPC/openapc-de#participating-german-universities

2| THE OPENAPC APPROACH

However, the initiative of transparent and open handling of publishing fees differs from country to country in the European Union, and so has been the global perspective for the last years.

2.1 The components

The infrastructure is designed to be open and very efficient. In this way, changes in the OpenAPC data can be presented in a correspondingly transparent and comprehensible manner.

Three main areas are covered:

- collecting and enrichments of APC data
- analyzing data and creating reports
- visualization

These three areas are conceptually reflected by three different services:

- GitHub repository (data provisioning; raw and enriched)
- OLAP server (data analysis and creating reports)
- Treemaps (visualization)

2.1.1 APC dataset and GitHub repository

GitHub[13] is a development platform inspired by the way developers work. From **open source** to business, developers can host and review code, manage projects, and build software alongside 50 million other developers.

The GitHub repository[14] holds the complete collection of APC data with the enriched information. The repository is free to use for everyone and from everywhere and is the open sustainability publishing strategy of the project.

2.1.2 OLAP server

The OLAP service is the fast and efficient way of querying the OpenAPC dataset. The service provides an Application Programming Interface (API) in form of a REST interface. OLAP is the backend service for the OpenAPC TreeMaps (Section 2.1.3).

Same as the dataset, the OLAP service is open source software, published at GitHub: https://github.com/OpenAPC/openapc-olap . At this moment, the documentation of the service is also available at GitHub: https://github.com/OpenAPC/openapc-olap/blob/master/HOWTO.md .

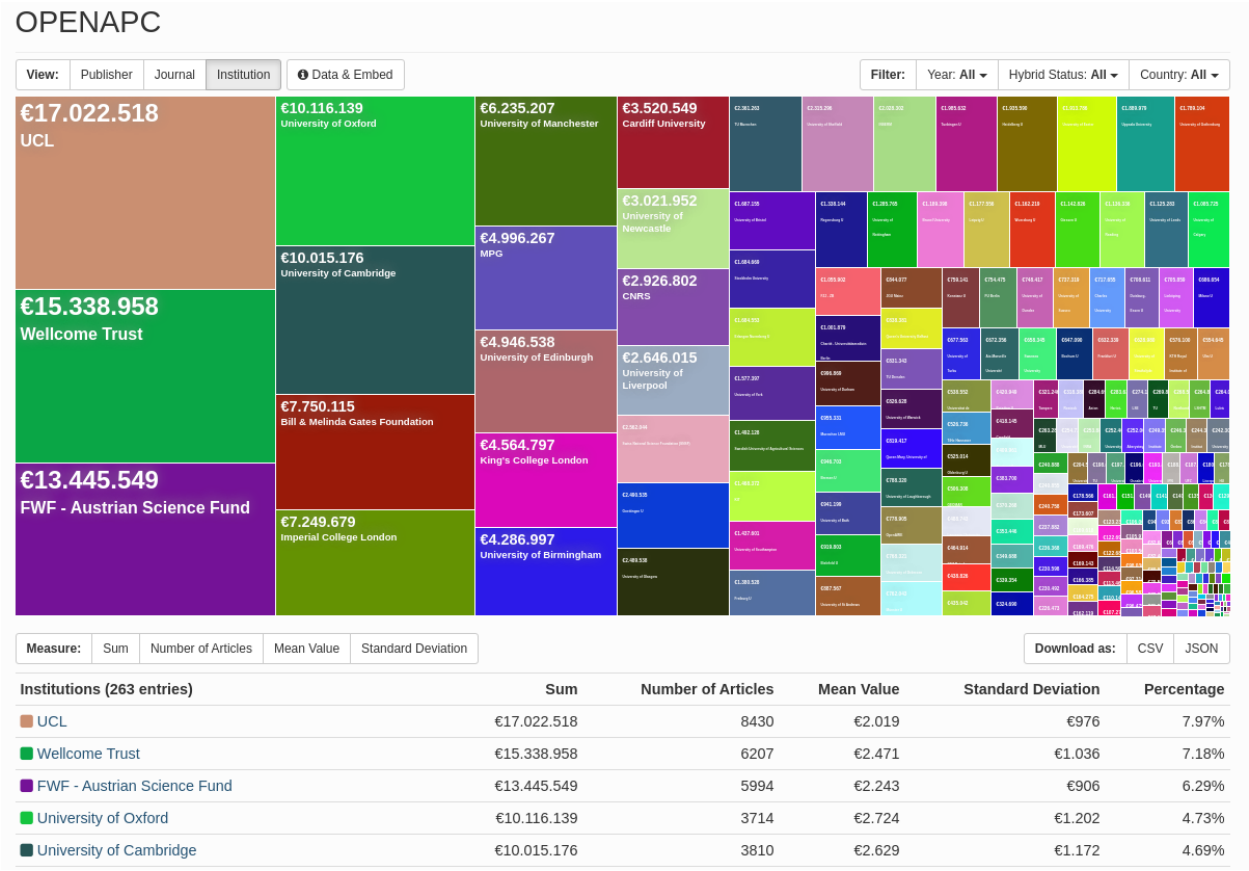
The service is based on the development of flask server, https://flask.pocoo.org/ .

[13] https://github.com

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

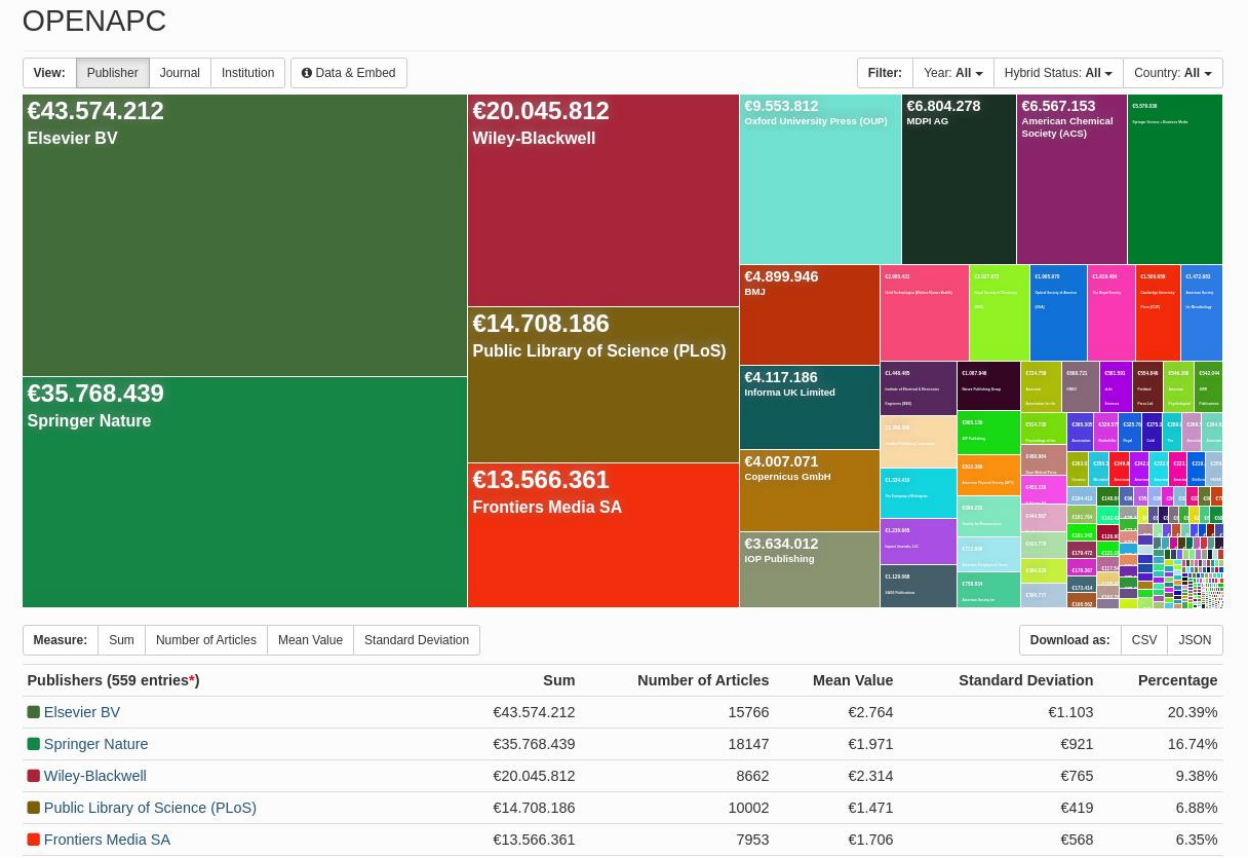
2.1.3 Treemaps

The Treemap service, https://treemaps.intact-project.org/apcdata/openapc/, visualizes the OpenAPC data sets among different perspectives. It uses the OLAP service in the back-end.



Institutions (263 entries)	Sum	Number of Articles	Mean Value	Standard Deviation	Percentage
UCL	€17.022.518	8430	€2.019	€976	7.97%
Wellcome Trust	€15.338.958	6207	€2.471	€1.036	7.18%
FWF - Austrian Science Fund	€13.445.549	5994	€2.243	€906	6.29%
University of Oxford	€10.116.139	3714	€2.724	€1.202	4.73%
University of Cambridge	€10.015.176	3810	€2.629	€1.172	4.69%

FIGURE 1. EXAMPLE OF THE VISUALIZATION OF INSTITUTIONS VIA TREEMAPS. (SOURCE: HTTPS://TREEMAPS.INTACT-PROJECT.ORG/APCDATA/OPENAPC/)



Publishers (559 entries*)	Sum	Number of Articles	Mean Value	Standard Deviation	Percentage
Elsevier BV	€43.574.212	15766	€2.764	€1.103	20.39%
Springer Nature	€35.768.439	18147	€1.971	€921	16.74%
Wiley-Blackwell	€20.045.812	8662	€2.314	€765	9.38%
Public Library of Science (PLoS)	€14.708.186	10002	€1.471	€419	6.88%
Frontiers Media SA	€13.566.361	7953	€1.706	€568	6.35%

FIGURE 2. EXAMPLE OF THE VISUALIZATION OF PUBLISHER VIA TREEMAPS. (SOURCE: HTTPS://TREEMAPS.INTACT-PROJECT.ORG/APCDATA/OPENAPC/)

The visualization can be used to create reports on the level of institutions, publishers, and journals, for example.

In the same way as the datasets and the OLAP service are published, the source code of the treemap service is also open and published at GitHub: https://github.com/OpenAPC/openapc-treemaps

2.2 The Datasets

The actual data sets do not only cover article processing charges, but also Transformative Agreements (like DEAL[15] in Germany), and, more recently, book processing charges (BPCs) have also been added.

This report will focus on APCs and the dataset containing:

- more than 100,000 records
- more than 260 institutions/organizations
- more than 200 Million Euros of aggregated article costs

Each institution or organization can easily provide a minimal set of fields of each article and is as simple as possible.

Only the following, minimal recommendation fields are required for the exchange:

Institution, Period, [APC amount in] Euro, DOI, is_hybrid

Based on the persistent identifier, in this case, the DOI, the record is enriched by further metadata, like

Publisher, Journal_Full_Title, ISSN, ISSN_Paper, ISSN_Electronic, Link_ISSN, License_Reference, In_CrossRef, PubMed-ID, PubMedCentral-ID, …

from different sources. The workflow to enrich and complete the dataset is described in Section 2.3 .

2.3 The Workflow

The three above areas and deployed services work together. Figure 3 shows the workflow from (1) minimum of collected APC data information, (2) enrichment data from CrossRef via DOI, (3) check of persistent identifiers at PubMed Central, (4) enrichment journal information from DOAJ, (5) and related further with ISSN, and (6) get information from the Web of Science.

After this processing, the complete and enriched dataset could be published on GitHub and a new version of the dataset is created there.

[15] https://www.projekt-deal.de/about-deal/

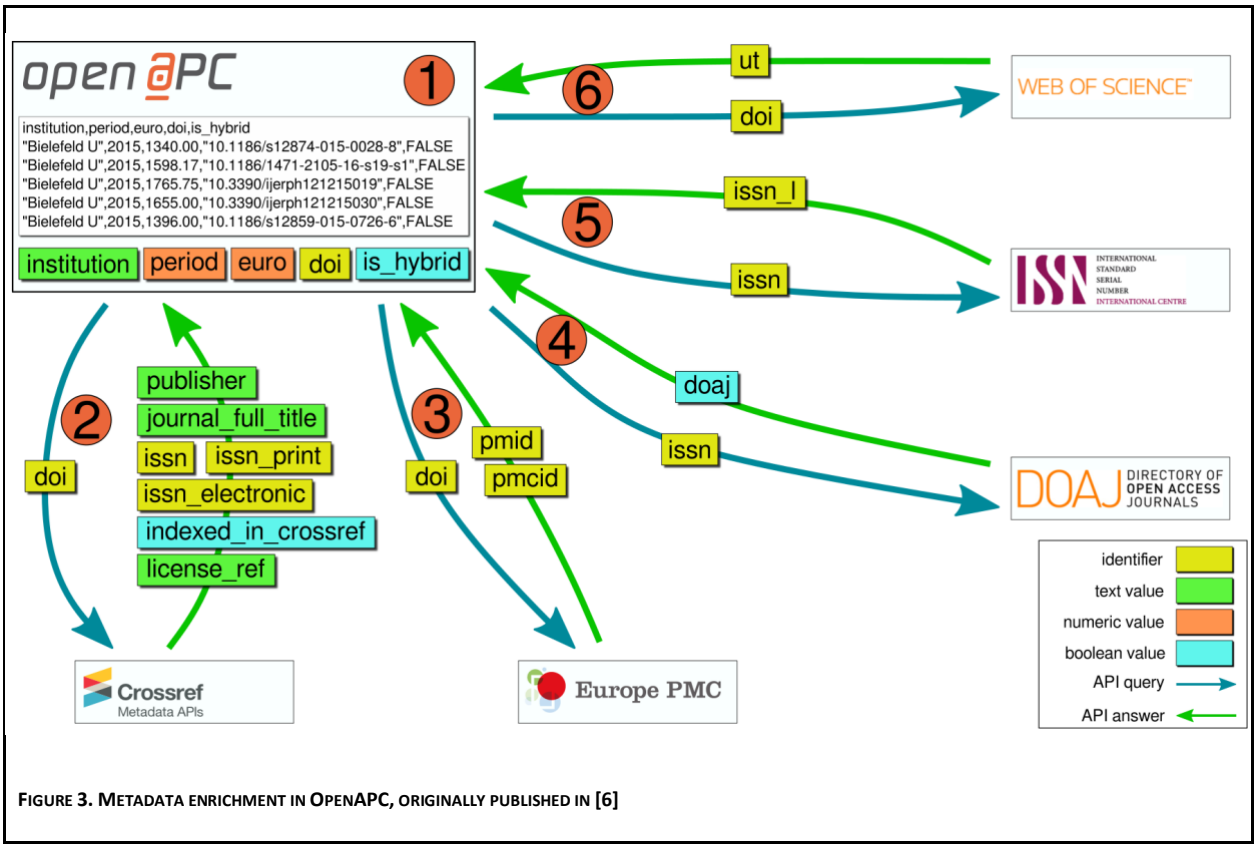


FIGURE 3. METADATA ENRICHMENT IN OPENAPC, ORIGINALLY PUBLISHED IN [6]

3| METHODOLOGY

This section describes the methodology of the analysis between the OpenAPC dataset and the OpenAIRE Research Graph via its API. OpenAIRE produces in regular intervals its Research Graph which is based on the data model described in [7]. To analyze the OpenAIRE Research Graph, it can be downloaded from ZENODO[16] as a dump [8], or retrieved via OpenAIRE's public API[17]. The first analyses during March 2019 were carried out with the OpenAIRE Research Graph in Production with more than 24 million publications, and this offered only a fraction of publications that OpenAIRE collected over the last month. In contrast, the OpenAIRE Research Graph in BETA already had more than 100 million publications in the evaluation period of October 2019 and for all further analyses, the OpenAIRE Research Graph in the BETA environment was used.

3.1.1 Statistic fundamentals

This subsection briefly describes the fundamental mathematics methods, which were used for the evaluation of the finite APC data sets.

The *arithmetic average* is defined as

$$\underline{x^C} = \frac{1}{n}\sum_{i=1}^{v} \quad x_i^C \quad h_i$$

The use of the equation is to calculate the arithmetic averages with $C = \{\text{Projects, Funding Streams, Funders}\}$. The weight value h_i is 1 for every $i \in N$. x_i represents the APC value.

The *median* is defined as

$$\text{median}(x) = \frac{1}{2}(x_{\lfloor (n+1)/2 \rfloor} + x_{\lceil (n+1)/2 \rceil})$$

x is thereby the APC value applied to Projects, Funding Streams, and Funders.

3.1.2 Accessing the OpenAIRE API

This subsection is dedicated to describing access to the publicly accessible OpenAIRE API. It is free to use from everywhere and for everyone, and the license of the metadata is a CC-BY license from Creative Commons.

The public OpenAIRE API is reachable at https://api.openaire.eu . The API is well documented and the documentation is described at https://api.openaire.eu/api.html . The endpoints give

[16] https://zenodo.org

[17] https://api.openaire.eu

access to the different types of research objectssuch as literature, datasets, software, and other publications as well as projects.

The analysis presented here is based only on the literature publication type. This type is used to query the DOIs that are located in the OpenAPC dataset. The endpoint is found at http://api.openaire.eu/search/publications . The DOI is expressed as a simple string like '&doi=...' and also has some default settings, like '?format=xml&model=openaire'.

Each DOI in the OpenAPC dataset can be requested with the procedure described above, to find the complete metadata set from production (or BETA) environment of *OpenAIRE API* as XML.

An example

```
curl -X GET -H "accept: */*"  \

https://api.openaire.eu/search/publications?format=xml&model=openaire&doi=...
```

The OpenAIRE API response to the above request returns an XML structure. Here an example of DOI 10.5194/acp-10-4403-2010 in Listing 1.

```
<?xml version="1.0" encoding="UTF-8"?>
 <response>
  <header>
   <query>(oaftype exact result) and (resulttypeid exact publication) and (pidclassid exact
"doi" and pid exact "10.5194/acp-10-4403-2010")</query>
   <locale>en_US</locale>
   <size>10</size>
   <page>1</page>
   <total>1</total>
   <fields> </fields>
  </header>
<results>
 <result xmlns:dri="http://www.driver-repository.eu/namespace/dri">
 <header xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <dri:objIdentifier>dedup_wf_001::5aa05b17605703ae0181399470e66f45</dri:objIdentifier>
  <dri:dateOfCollection>2017-01-03T14:36:13.185Z</dri:dateOfCollection>
  <dri:dateOfTransformation>2017-01-03T15:22:35.013Z</dri:dateOfTransformation>
  <counters> <counter_similarity_inferred value="14"/>
  <counter_outcome_inferred value="1"/>
  <counter_similarity value="14"/>
  <counter_outcome value="1"/>
  <counter_dedup value="3"/>
  <counter_doi value="1"/> </counters>
 </header>
 <metadata>
  <oaf:entity xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
   xmlns:oaf="http://namespace.openaire.eu/oaf"
   xsi:schemaLocation="http://namespace.openaire.eu/oaf
   https://www.openaire.eu/schema/1.0/oaf-1.0.xsd">
 <oaf:result>
  <source>Atmospheric Chemistry and Physics, Vol 10, Iss 9, Pp 4403-4422 (2010)</source>
  <source>eISSN: 1680-7324</source>
  <source>Atmospheric Chemistry and Physics</source>
  <creator rank="1" name="A." surname="Pozzer">Pozzer, A.</creator>
  <creator rank="2" name="J." surname="Pollmann">Pollmann, J.</creator>
  <creator rank="3" name="D." surname="Taraborrelli">Taraborrelli, D.</creator>
```

<creator rank="4" name="P." surname="Jöckel">Jöckel, P.</creator>
<creator rank="...</creator>
<description>The primary sources and atmospheric chemistry of C₂-C₅ alkanes were incorporated into the atmospheric chemistry ... </description>
<dateofacceptance>2010-01-01</dateofacceptance>
<language classid="eng" classname="English" schemeid="dnet:languages" schemename="dnet:languages"/>
<resulttype classid="publication" classname="publication" schemeid="dnet:result_typologies" schemename="dnet:result_typologies"/>
<journal issn="1680-7316" eissn="1680-7324" lissn="" ep="" iss="" sp="" vol="">Atmospheric Chemistry and Physics</journal>
<subject classid="keyword" classname="keyword" schemeid="dnet:subject_classification_typologies" schemename="dnet:subject_classification_typologies">Chemistry</subject>
<subject classid="keyword" ...</subject>
<title classid="main title" classname="main title" schemeid="dnet:dataCite_title" schemename="dnet:dataCite_title">Observed and simulated global distribution and budget of atmospheric C₂-C₅ alkanes</title>
<format>application/pdf</format>
<fulltext>file:///mnt/uploaded_dumps/copernicus/upload/acp-10-4403-2010.pdf</fulltext>
<publisher>Copernicus Publications</publisher>
…
<tool/>
<pid classid="doi" classname="doi" schemeid="dnet:pid_types" schemename="dnet:pid_types">10.5194/acp-10-4403-2010</pid>
<collectedfrom name="European Research Council (ERC)" id="openaire____::5ecaf0d3af3004219bc6b5907d19b6d9"/>
<collectedfrom name="DOAJ-Articles" id="driver_____::bee53aa31dc2cbb538c10c2b65fa5824"/>
<collectedfrom name="Copernicus Publications" id="openaire____::5a38cb462ac487bf26bdb86009fe3e74"/>
<originalId>oai:doaj.org/article:6ce54a3899504d50a60e7cd3235a4663</originalId>
<originalId>216935</originalId>
<originalId>oai:publications.copernicus.org:acp2539</originalId>
<bestaccessright classid="OPEN" classname="Open Access" schemeid="dnet:access_modes" schemename="dnet:access_modes"/>
<context id="EC" label="European Commission" type="funding">
<category id="EC::FP7" label="SEVENTH FRAMEWORK PROGRAMME">
<concept id="EC::FP7::SP2" label="SP2-Ideas"/>
<concept id="EC::FP7::SP2::ERC" label="ERC"/>
</category>
</context>
<datainfo>
<inferred>true</inferred>
<deletedbyinference>false</deletedbyinference>
<trust>0.9</trust>
<inferenceprovenance>dedup-similarity-result-levenstein</inferenceprovenance>
<provenanceaction classid="sysimport:dedup" classname="sysimport:dedup" schemeid="dnet:provenanceActions" schemename="dnet:provenanceActions"/>
</datainfo>
<rels>
<rel inferred="true" trust="0.9" inferenceprovenance="iis::document_similarities_standard" provenanceaction="iis">
<to class="isAmongTopNSimilarDocuments" scheme="dnet:result_result_relations" type="result">dedup_wf_001::b681e4ce8228b1bac8b7b5830c44127c</to>
<dateofacceptance>2009-02-01</dateofacceptance>
<dateofacceptance>2009-01-01</dateofacceptance>
<pid classid="doi" classname="doi" schemeid="dnet:pid_types" schemename="dnet:pid_types">10.5194/acp-9-1253-2009</pid> <resulttype classid="publication" classname="publication" schemeid="dnet:result_typologies" schemename="dnet:result_typologies"/>
<dateofacceptance>2008-01-01</dateofacceptance>
<dateofacceptance>2009-02-18</dateofacceptance>
<publisher>Copernicus Publications (EGU)</publisher>
<similarity>0.82231545</similarity>
<type>STANDARD</type>
<title classid="main title" classname="main title" schemeid="dnet:dataCite_title" schemename="dnet:dataCite_title">Evaluation of the global oceanic isoprene source and its impacts on marine organic carbon aerosol</title>
</rel>
<rel inferred="true" trust="0.9" inferenceprovenance="iis::document_similarities_standard" provenanceaction="iis">
... </rel>
</rels>
<children>

```
 <result objidentifier="doajarticles::c27dae4630ba5a3a15e4aa59c60ef91a">
  <dateofacceptance>2010-05-01</dateofacceptance>
  <resulttype classid="publication" classname="publication" schemeid="dnet:result_typologies"
schemename="dnet:result_typologies"/>
  <title classid="main title" classname="main title" schemeid="dnet:dataCite_title"
schemename="dnet:dataCite_title">Observed and simulated global distribution and budget of
atmospheric C&lt;sub&gt;2&lt;/sub&gt;-C&lt;sub&gt;5&lt;/sub&gt; alkanes</title>
  <publisher>Copernicus Publications</publisher>
 </result>
 <result objidentifier="copernicuspu::5aa05b17605703ae0181399470e66f45">
  <dateofacceptance>2010-05-12</dateofacceptance>
…
 </result>
…

 <webresource>
  <url>http://dx.doi.org/10.5194/acp-10-4403-2010</url>
 </webresource>
 </instance>
</children>
</oaf:result>
</oaf:entity>
</metadata>
</result>
</results>
</response>
```

LISTING 1. SNIPPET OF RESULT XML RESPONSE FROM OPENAIRE API.

4| Analysis and Results

The OpenAPC initiative's dataset discloses publication fees amounting to more than 207 million EUR paid for 104,517 Open Access articles from more than 260 institutions in over 16 countries. The analysis is based on the methodology described in the previous section.

The complete metadata was requested by the OpenAIRE API and the relevant content was extracted. The OpenAIRE Research Graph API was used to obtain the corresponding enriched metadata for a given Digital Object Identifier (DOI). These include information about the funder, funding stream, and the project the article stems from.

The next subsection explains how the metadata enrichments from OpenAIRE can be used to take a closer look to derive mean and highest APCs differentiated by funders and funding streams. For this purpose, the respective coverage is illustrated using the OpenAIRE Research Graph.

4.1 Coverage of the OpenAPC dataset with article metadata in OpenAIRE

A first analysis was performed in March 2019 with the OpenAPC dataset version 3.55.1 [18].

The analysis shows that around 91% of the articles registered in the OpenAPC dataset are included in the Research Graph of the OpenAIRE production system [9,10]. The dataset of OpenAPC counts 75,287 records that match with 68,430 records in OpenAIRE. The comparison was done via persistent identifiers (DOI), which are present in both datasets.

In the course of the project phase, further analysis was carried out using the OpenAIRE Research Graph in the BETA infrastructure.

The reason for this was that the number of documents in the OpenAIRE Research Graph in BETA (B) grew much faster to more than 110 million metadata records about articles. Access to the data sets was provided via the API of B of the OpenAIRE Research Graph.

The result of this analysis shows that about **98**% [11,12] of the articles in the OpenAPC dataset can be found in the OpenAIRE research graph B.

After analyzing the coverage of the OpenAPC data set in the OpenAIRE Research Graph, we now investigate the distribution of APCs found in OpenAIRE. Figure 4 shows the distribution of APCs from the first analysis in March 2019. The average APC was around 1,958 EUR by a total sum of 134,014,401 EUR. In October 2019[19], we created the same analysis with an updated OpenAPC dataset and the OpenAIRE Research Graph in the BETA environment. The comparison showed that of 84,402 records in OpenAPC, 82,306 records were found in OpenAIRE. Figure 5 shows the distribution and the average APC was around 1,946 EUR by a total sum of 160,192,069 EUR. The figure 6 shows the distribution of APCs in March 2020[20] with a more significant range around

[18] v3.55.1, https://github.com/OpenAPC/openapc-de/releases/tag/v3.55.1

[19] v3.69.4, https://github.com/OpenAPC/openapc-de/releases/tag/v3.69.4

[20] v3.81.5, https://github.com/OpenAPC/openapc-de/releases/tag/v3.81.5

2000 EUR. It covers more than 194,290,612 EUR APC, the average APC was around 1,979 EUR from 98,142 of 100,074 records in total.

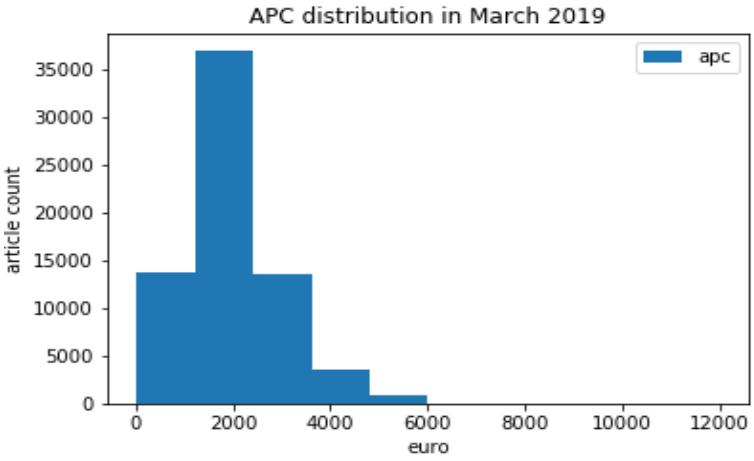


FIGURE 4. DISTRIBUTION OF APCS IN MARCH 2019: COVERAGE ~91% WITH OPENAIRE PRODUCTION RESEARCH GRAPH

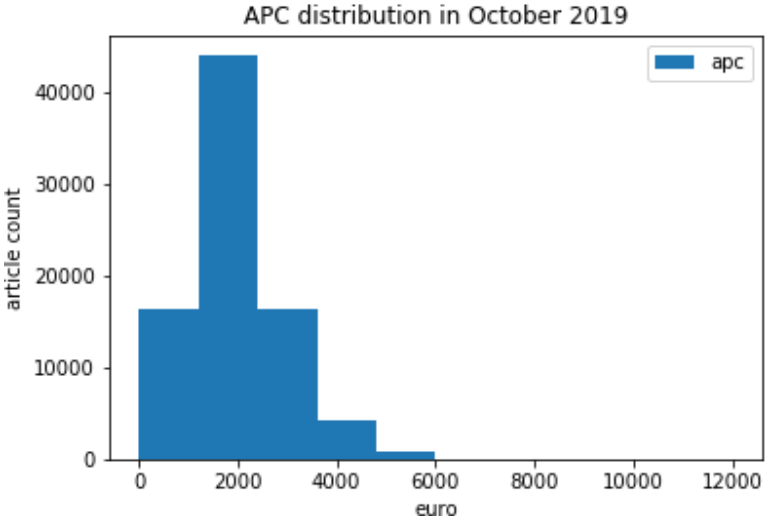


FIGURE 5. DISTRIBUTION OF APCS IN OCTOBER 2019: COVERAGE ~98% WITH OPENAIRE BETA RESEARCH GRAPH

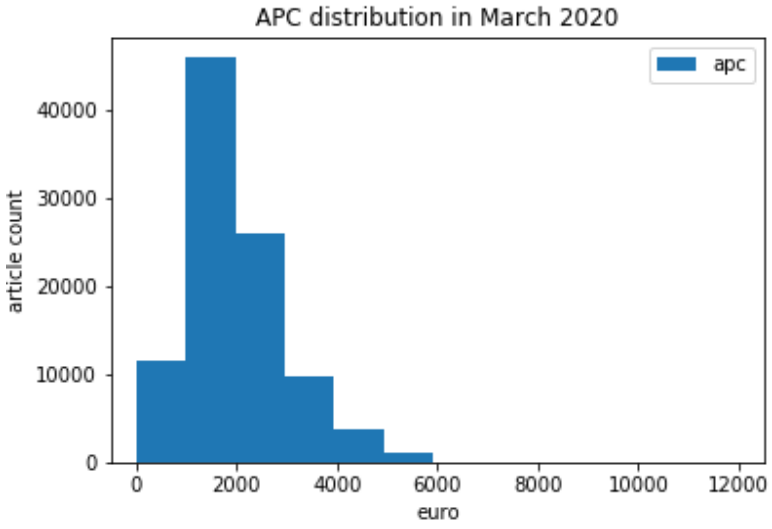


FIGURE 6. DISTRIBUTION OF APCS IN MARCH 2020: COVERAGE ~98% WITH OPENAIRE BETA RESEARCH GRAPH

4.2 Comparison of APC costs at the level of funding agencies

The benefit from OpenAIRE Research Graph is the enrichments in the record like for funders, funding streams, and projects. The section analyses the relation between the APC and the (single) funder.

Figure 7 shows the first 12 top project funders to which publications are assigned, with the highest mean APC from the March 2019 dataset.

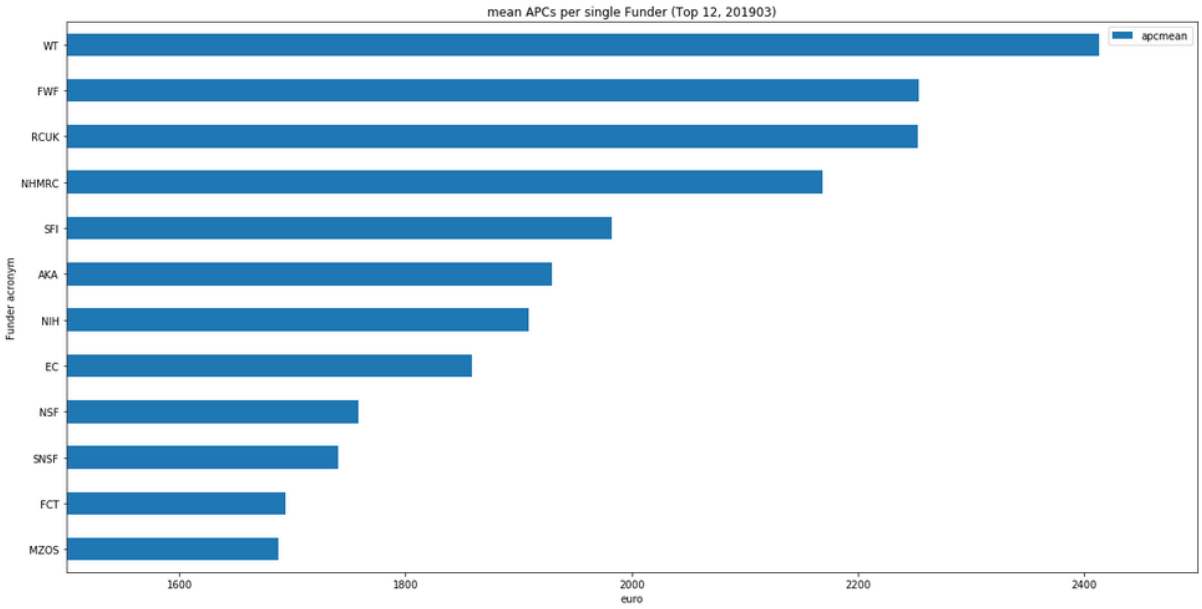


FIGURE 7. TOP 12 SINGLE FUNDER BASED ON THE OPENAPC DATASET FROM MARCH 2019 AND OPENAIRE PRODUCTION RESEARCH GRAPH COVERAGE OF ~91%

Figures 8 and 9 show the first 12 top project funders to which publications are assigned, with the highest mean APC from the October 2019 and March 2020 datasets.

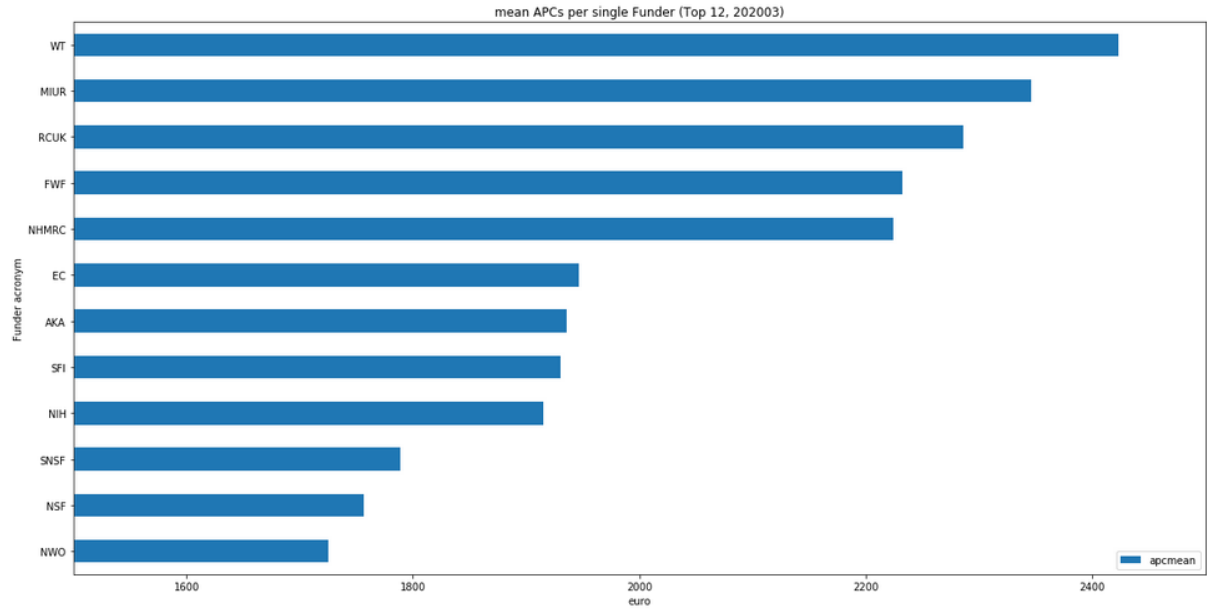


FIGURE 8. TOP 12 SINGLE FUNDER BASED ON THE OPENAPC DATASET FROM MARCH 2020 AND OPENAIRE BETA RESEARCH GRAPH COVERAGE OF ~98%

The y-axis in figures 7 and 8 denotes the funders acronym. The top funders in March 2019 and March 2020 are:

- WT: Wellcome Trust
- FWF: Austrian Science Fund
- RCUK: Research Council United Kingdom
- NHMRC: Australian National Health and Medical Research Council
- SFI: Science Foundation Ireland
- EC: European Commission
- AKA: Academy of Finland
- MIUR: Italian Ministry of Education, University and Research
- NIH: National Institutes of Health
- SNSF: Swiss National Science Foundation
- NSF: National Science Foundation (US)
- NWO: Dutch Research Council

The same analysis was carried out to determine which sponsors shared the funding for a project. Figure 9 shows March 2019 and Figure 10 shows March 2020 with the top 12 results for the highest mean APC values.

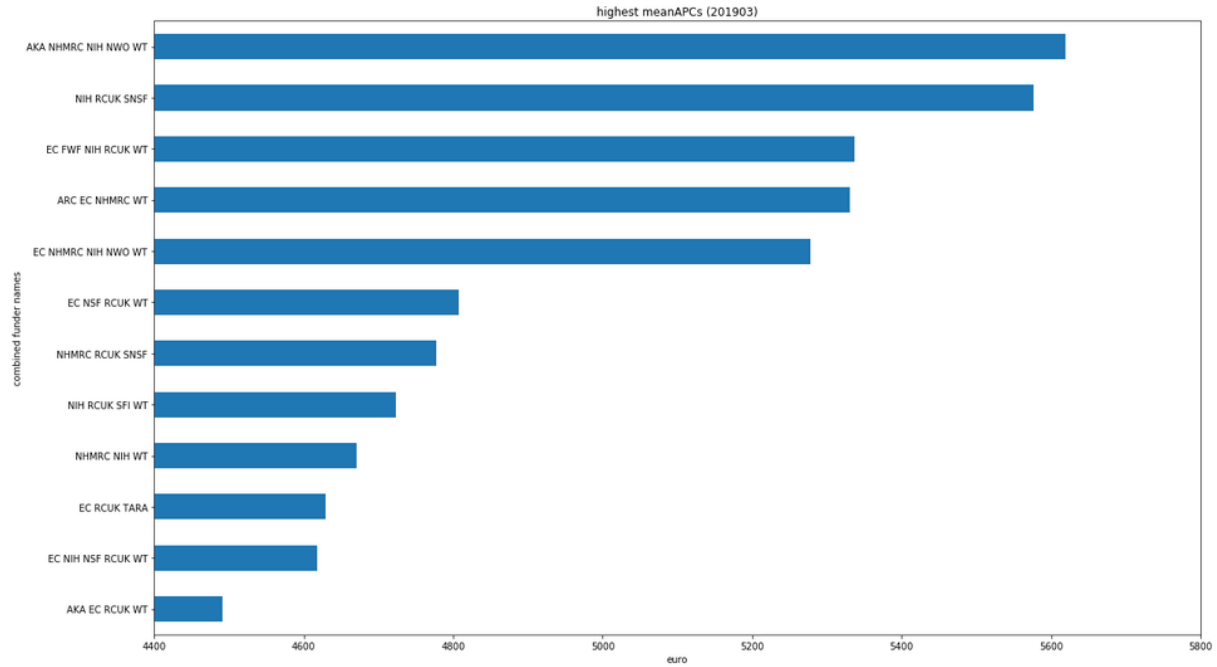


FIGURE 9. TOP 12 MULTI-FUNDED ARTICLES BASED ON THE OPENAPC DATASET FROM MARCH 2019 AND OPENAIRE PRODUCTION RESEARCH GRAPH COVERAGE OF ~91%

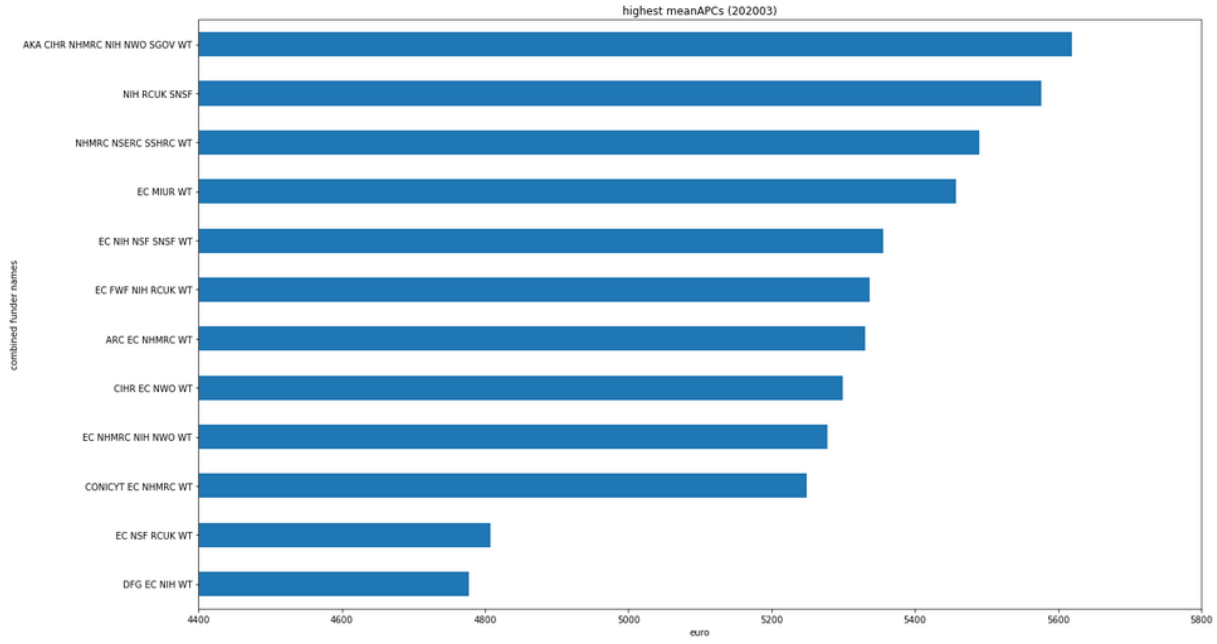


FIGURE 10. TOP 12 MULTI-FUNDED ARTICLES BASED ON THE OPENAPC DATASET FROM MARCH 2020 AND OPENAIRE BETA RESEARCH GRAPH COVERAGE OF ~98%

A differentiated, percentage-based comparison of articles carried by several funders can be made in the future.

4.3 Comparison of APC costs at the level of funding streams

The metadata from OpenAIRE API also provides information about the *Funding Streams* if it is available. The evaluation concerned the element "*funding_level_0*", which can also occur several times. The response in XML from the API provides the information, present in Listing 2.

```
<metadata>
 <oaf:entity xmlns:oaf="http://namespace.openaire.eu/oaf"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://namespace.openaire.eu/oaf https://www.openaire.eu/schema/1.0/oaf-
1.0.xsd">
    <oaf:result>
        …
            <rels>
              <rel inferred="true" trust="0.8145"
inferenceprovenance="iis::document_referencedProjects" provenanceaction="iis">
                <to class="isProducedBy" scheme="dnet:result_project_relations"
type="project">corda_______::79b9c06a91552c2417b5f3ddb7e53923</to>
                <acronym>MANAGED OUTCOMES</acronym>
                 …
                <funding>
                  <funder id="ec__________::EC" shortname="EC" name="European Commission"
jurisdiction="EU"/>
                  <funding_level_0 name="FP7">ec__________::EC::FP7</funding_level_0>
…
```

LISTING 2. SNIPPET OF RESULT XML WITH FUNDING STREAM INFORMATION.

Three selected funder programs were examined in more detail, their article numbers in OpenAIRE, as well as their average APC from OpenAPC[21].

Program name	articles in OpenAIRE	average of APC
Seventh Framework Programme (FP7)	5,669	~2,152 EUR
Medical Research Council UK (MRC)	3,763	~2,523 EUR
Engineering and Physical Science RC	2,657	~2,272 EUR

The above analyses currently only show a part of the opportunities that arise when processing charges are linked to the information on data sets in OpenAIRE.

Another method is to show the development of articles which the OpenAPC dataset contained and indexed in the OpenAIRE Research Graph. Figure 11 shows the example of the European Commission funding stream *Horizon 2020* (H2020) with the article number which can be found in the OpenAPC dataset in the time-line from March 2019 to March 2020. The number of articles has doubled in 2020, one year later.

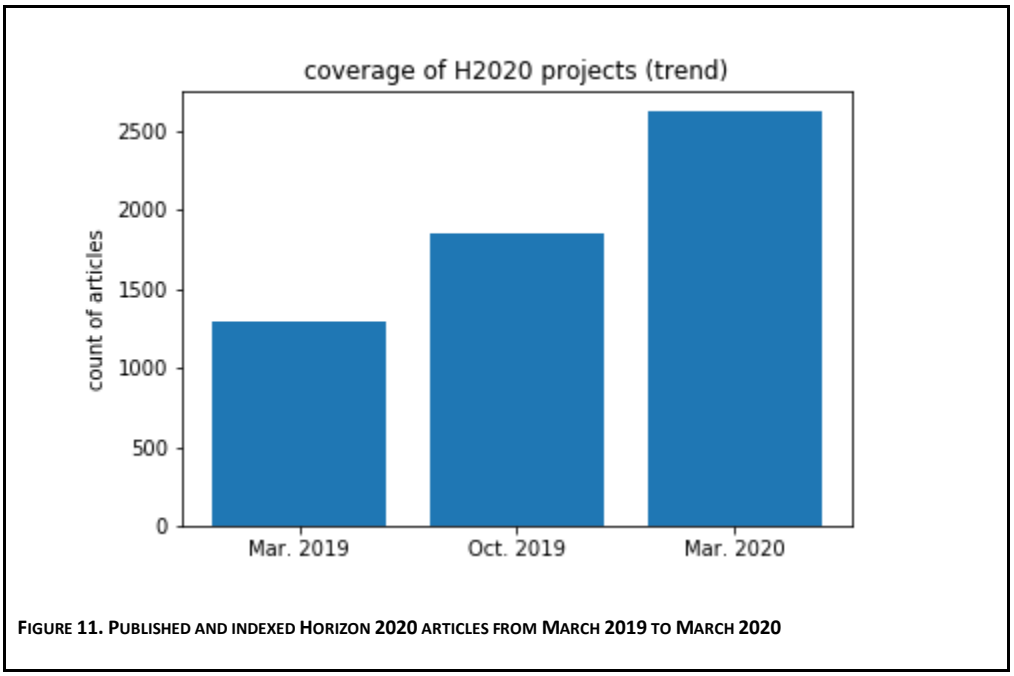


FIGURE 11. PUBLISHED AND INDEXED HORIZON 2020 ARTICLES FROM MARCH 2019 TO MARCH 2020

[21] used data from March 2019.

The long term behavior would be interesting, to each individual project.

Figure 12 shows the combined information of the Horizon 2020 project of articles, average APC, and the sum of APC budget overall articles.

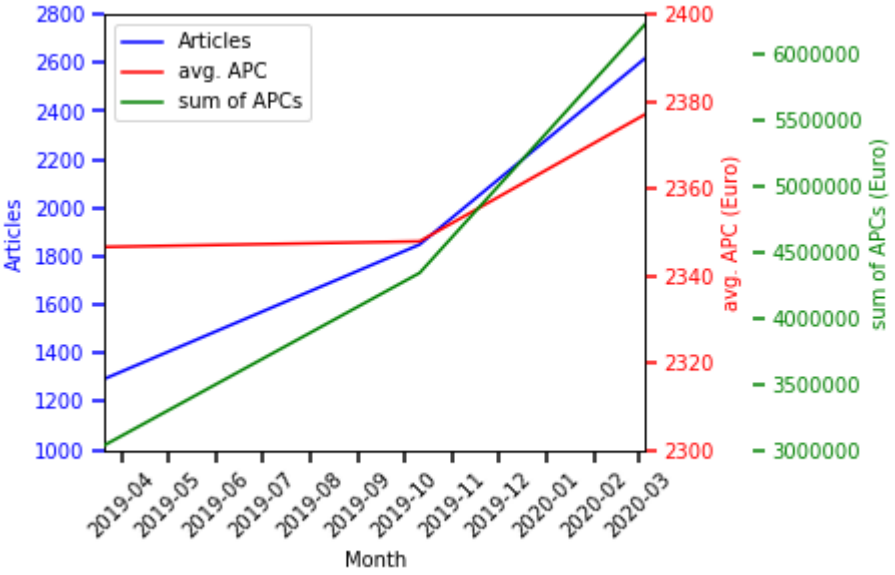


FIGURE 12. ARTICLE COUNT, AVERAGE APC, AND SUM OF APCS IN MARCH 2019, OCTOBER 2019, AND MARCH 2020 OF HORIZON 2020

It was to be expected that the number of articles and the total amount of APCs in the Horizon 2020 funding stream would increase over the period. Surprisingly, the plain average value of APCs also increased from 2,346 EUR to around 2,377 EUR.

5| DISCUSSION

The first analyses of the coverage between the OpenAPC dataset and the OpenAIRE Research Graph, the OpenAPC dataset having far less than 100,000 records, showed coverage of around 91 percent. The coverage in March 2020 was around 98 percent. The result was quite surprising since several million Open Access articles were recorded in OpenAIRE.

Due to the low number of matches between the OpenAPC dataset and the OpenAIRE Research Graph in Production, we decided to perform the comparison with the BETA environment, which will surely be in production again in the near future.

The Community Calls[22] [23] [10,12] that took place and the workshop[24] [13,14] offered the opportunity to get informed about the current status. But the main objective was to exchange information with the community and to encourage OpenAIRE NOADS to join or promote this service in their countries.

The reactions of the participants in the feedback rounds of the virtual calls and the workshop in December 2019 regarding the integration of cost data in the OpenAIRE Research Graph were very positive.

During the final workshop "Monitoring OA publishing costs and revenues of publishing initiatives" in Paris/France, December 11, 2019 [16], there was particular interest in the presentation on "transformative agreements" such as the DEAL-Wiley agreement in Germany [15] and the presentation on the complexity of Open Access book funding models [17]. Both presentations showed potential and approaches how cost information on transformative agreements and OA books can be collected, stored, and transparently presented in a structured way in the future.

6| CONCEPTION AND SCHEMA

6.1.1 Conception

The following assumptions were made for the implementation in the Research Graph, i.e., for the changes in the scheme:

- little complexity (simplicity).

[22] https://www.openaire.eu/blogs/webinar-together-with-the-openapc-project-team-on-the-transparency-of-publications-fees

[23] https://www.openaire.eu/item/openapc-cost-transparency-of-open-access-publishing

[24] https://www.openaire.eu/blogs/monitoring-oa-publishing-costs-and-revenues-of-publishing-initiatives-workshop-successfully

- not only useful for APCs but also for other research areas and resource types. This means that it should be possible to map additional use-cases without changes to the schema.

For this reason, the schema was adjusted;two new fields were added in the Result.proto type definition.

The first added field is named *processing charge amount* that reflects the "number" of a fee and the second added field is named *processing charge currency* that reflects in which currency the number is stated. To standardize the currency field, the field uses the three alphabetic codes which are described first in 1978 from the *International Organization for Standardization* (ISO) in publication 4217[25]. The document was updated the last time in 2015: "Codes for the representation of currencies"[26] known as ISO-4217:2015.

6.1.2 Schema changes

During the work on the conceptual model and with the closed collaboration with our partners, especially our technical partners, the source code's schema extension was implemented as follows described in Listing 3.

```
        // processing charges
        optional StringField processingchargeamount   = 14;
        // currency - alphabetic code describe in ISO-4217
        optional StringField processingchargecurrency = 15;
```

LISTING 3. PART OF THE RESULT.PROTO[27]

The **processing charge amount & currency** are attributes of the research object. Each research object has a resource type. With this combination, the processing charge could be a fee number for an article as "Article Processing Charge" (APC) or for a book as "Book Processing Charge" (BPC) or as well as publishing costs of other research output types.

[25] https://www.iso.org/iso-4217-currency-codes.html

[26] https://www.iso.org/standard/64758.html

[27] https://svn.driver.research-infrastructures.eu/driver/dnet45/modules/dnet-openaire-data-protos/trunk/src/main/resources/eu/dnetlib/data/proto/Result.proto

7| IMPLEMENTATION AND TRANSFORMATION SCRIPT

At the beginning of our work, there were no tools or mechanisms to analyze the data sets between OpenAPC and the OpenAIRE Research Graph. The development of suitable tools and algorithms took place during this phase and should be freely available for re-use and transparency. For this purpose, a separate software repository was created on GitHub to allow collaborative work on these tools.

The open-source and freely available software repository is at:

https://github.com/ACz-UniBi/OpenAIRE-OpenAPC_coverage

In the aggregation environment of OpenAIRE that harvests the data sources, there is no need to adapt the implementation to request the CSV from the GitHub repository directly. The configuration of the collector plugin, the baseURLs, identifier, ID XPath, cronjobs, and some more attributes must be set. The transformation script was developed during this time.

The transformation script uses the periodically updated OpenAPC dataset from GitHub in the OpenAIRE aggregation system. It transforms the OpenAPC data structure into the internal schema for processing and enrichments, where the Research Graph is the result.

In the development of the initial version of the transformation script, the transformation focused on following attributes

doi, euro, issn, jftitle, pmcid, pmid, license

Listing 4 shows the XSLT transformation script.

The result of the transformation is used by the OpenAIRE Research Graph Raw and the steps beyond.

```
<xsl:stylesheet xmlns:dc="http://purl.org/dc/elements/1.1/"
   xmlns:dri="http://www.driver-repository.eu/namespace/dri"
   xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
   xmlns:oai="http://www.openarchives.org/OAI/2.0/"
   xmlns:oaf="http://namespace.openaire.eu/oaf"
   xmlns:date="http://exslt.org/dates-and-times"
   xmlns:dr="http://www.driver-repository.eu/namespace/dr"
   xmlns:datacite="http://datacite.org/schema/kernel-4"
 version="2.0">
   <xsl:output indent="yes"/>
   <xsl:param name="varOfficialName"/>
   <xsl:param name="varDataSourceId"/>
   <xsl:param name="quote">"</xsl:param>
   <xsl:variable name="baseURL" select="string('https://raw.githubusercontent.com/OpenAPC/openapc-
de/master/data/apc_de.csv')"/>
   <xsl:template name="terminate">
      <xsl:message terminate="yes"> record is not compliant, transformation is interrupted.
      </xsl:message>
   </xsl:template>
   <xsl:template match="/">
      <xsl:call-template name="oapcRecord"/>
   </xsl:template>

<xsl:template name="oapcRecord">
      <xsl:variable name="doi" select="//column[./@name = 'doi']/text()"/>
      <xsl:variable name="euro" select="//column[./@name = 'euro']/text()"/>
      <xsl:variable name="issn" select="//column[./@name = 'issn']/text()"/>
      <xsl:variable name="jftitle" select="//column[./@name = 'journal_full_title']/text()"/>
      <xsl:variable name="pmcid" select="//column[./@name = 'pmcid']/text()"/>
      <xsl:variable name="pmid" select="//column[./@name = 'pmid']/text()"/>
      <xsl:variable name="license" select="//column[./@name = 'license_ref']/text()"/>
      <oai:record>
         <oai:header>
            <xsl:copy-of copy-namespaces="no" select="//*[local-name() = 'header']/*"/>
            <dr:dateOfTransformation>
               <xsl:value-of select="date:date-time()"/>
            </dr:dateOfTransformation>
         </oai:header>
         <metadata xmlns="http://namespace.openaire.eu/">
          <oaf:identifier identifierType="doi">
           <xsl:value-of select="$doi" />
          </oaf:identifier>
          <oaf:identifier identifierType="pmcid">
           <xsl:value-of select="$pmcid" />
          </oaf:identifier>
          <oaf:identifier identifierType="pmid">
           <xsl:value-of select="$pmid" />
          </oaf:identifier>

          <oaf:processingchargeamount>
           <xsl:attribute name="currency">EUR</xsl:attribute>
           <xsl:value-of select="$euro" />
          </oaf:processingchargeamount>

          <oaf:journal>
```

```
          <xsl:attribute name="issn">
            <xsl:value-of select="$issn" />
          </xsl:attribute>
          <xsl:value-of select="$jftitle" />
        </oaf:journal>

        <dc:license>
          <xsl:value-of select="$license" />
        </dc:license>

        <dr:CobjCategory type="publication">0004</dr:CobjCategory>
        <oaf:accessrights>OPEN</oaf:accessrights>
        <datacite:rights rightsURI="http://purl.org/coar/access_right/c_abf2">open access</datacite:rights>

            <oaf:hostedBy>
              <xsl:attribute name="name">OpenAPC Initiative</xsl:attribute>
              <xsl:attribute name="id">openaire____::openapc_initiative</xsl:attribute>
            </oaf:hostedBy>
            <oaf:collectedFrom>
              <xsl:attribute name="name">OpenAPC Initiative</xsl:attribute>
              <xsl:attribute name="id">openaire____::openapc_initiative</xsl:attribute>
            </oaf:collectedFrom>
      </metadata>
            <oaf:about xmlns:oai="http://wwww.openarchives.org/OAI/2.0/">
              <oaf:datainfo>
                <oaf:inferred>false</oaf:inferred>
                <oaf:deletedbyinference>false</oaf:deletedbyinference>
                <oaf:trust>0.9</oaf:trust>
                <oaf:inferenceprovenance/>
                <oaf:provenanceaction classid="sysimport:crosswalk:datasetarchive"
                  classname="sysimport:crosswalk:datasetarchive"
                  schemeid="dnet:provenanceActions"
                  schemename="dnet:provenanceActions"/>
              </oaf:datainfo>
            </oaf:about>
    </oai:record>
</xsl:template>
</xsl:stylesheet>
```

LISTING 4. INITIAL VERSION OF XSLT TRANSFORMATION SCRIPT.

8| CONCLUSION

In summary, it can be stated that OpenAPC datasets are continuously released on Github and can be analyzed and visualized via two services of OpenAPC - OLAP and Treemaps. In the OpenAPC project, the APCs are easily manageable using the existing workflows. The extensions regarding book processing charges for book/monographs or the costs of transformative agreements are more recent developments in this field today.

8.1 Recommendations for next steps

Once the full integration of the schema in the OpenAIRE Research Graph and also the provision of cost data via the APIs is completed, the information can be used in portals like *explore* and dashboards like *provide* and the Open Science Observatory. Third-party services can access the Research Graph enriched with cost data for carrying out further analyses thanks to the transparency of the provided data.

Starting from the conceptual basis created during this work, every scientific object in the OpenAIRE Research Graph can have a "Processing Charge" attribute. The implementation is not limited to journal articles. It is also possible to flag fees or costs for books (BPCs) or costs for the creation of datasets, e.g., costs for high-performance computation of a dataset or costs for research vessels to obtain a research object or dataset.

Concerning the integration of repositories into library workflows, the first approaches can already be stated that cost data can be added as an attribute of publication metadata and thus made available e.g., via OAI-PMH. Since 2017, the repository of Regensburg University[28] in Germany has a dedicated OAI set and metadata prefix that allows us to harvest[29] these processing charges.
The enhancement of the Regensburg University repository can also be recommended for other repository platforms and is discussed as an extension in the OpenAIRE Guidelines[30].

Based on the latest developments and the need for transparency in publication fees and the costs of Open Access publishing, it is desirable that OpenAPC continues and finds a smart integration into the OpenAIRE Research Graph and its portals.

[28] https://epub.uni-regensburg.de

[29] https://epub.uni-regensburg.de/cgi/oai2?verb=ListRecords&metadataPrefix=OpenAPC&set=6F615F747970653D676F6C645F70616964

[30] https://guidelines.openaire.eu/

9| REFERENCES

[1] Monaghan, J., Lucraft, M., Allin, K. (2020): 'APCs in the Wild': Could Increased Monitoring and Consolidation of Funding Accelerate the Transition to Open Access?. figshare. Journal contribution. https://doi.org/10.6084/m9.figshare.11988123.v4

[2] Shamash, K. (2016). Article processing charges (APCs) and subscriptions. *Monitoring open access costs.* Online available: https://www.jisc.ac.uk/sites/default/files/apc-and-subscriptions-report.pdf (last accessed 2020-07-29)

[3] Pieper, D. (2015). 1. 2 Open APC Data in Germany. Zenodo. http://doi.org/10.5281/zenodo.3601225

[4] OpenAPC Sweden: https://github.com/Kungbib/openapc-se

[5] OpenAPC Initiative:
https://github.com/OpenAPC/openapc-de
https://www.intact-project.org/openapc/
https://treemaps.intact-project.org/

[6] Pieper, D., & Broschinski, C. (2018). OpenAPC: a contribution to a transparent and reproducible monitoring of fee-based open access publishing across institutions and nations. *Insights the UKSG journal*, *31*. http://doi.org/10.1629/uksg.439

[7] Manghi, P., Bardi, A., Atzori, C., Baglioni, M., Manola, N., Schirrwagen, J., & Principe, P. (2019). The OpenAIRE Research Graph Data Model (Version 1.3). Zenodo. http://doi.org/10.5281/zenodo.2643199

[8] Manghi, P., Atzori, C., Bardi, A., Schirrwagen, J., Dimitropoulos, H., La Bruzzo, S., … Summan, F. (2019). OpenAIRE Research Graph Dump (Version 1.0.0-beta) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3516918

[9] Czerniak, A., & Broschinski, C. (2019). Webinar about transparency of publication fees and the OpenAPC project. http://doi.org/10.13140/RG.2.2.33517.03040

[10] Czerniak, A., & Broschinski, C. (2019). **Recording** of "Webinar about transparency of publication fees and the OpenAPC project". Available online: https://www.youtube.com/watch?v=v323ebK45po (last accessed: 2020-03-27)

[11] Czerniak, A., & Broschinski, C. (2019). OpenAPC - cost transparency of Open Access publishing. Zenodo. http://doi.org/10.5281/zenodo.3517903

[12] Czerniak, A., & Broschinski, C. (2019). **Recording** of "OpenAPC - cost transparency of Open Access publishing". Available online: https://youtu.be/s1opcFWcZg4 (last accessed: 2020-03-27)

[13] Broschinski, C. (2019). The global OpenAPC project - an introduction. Zenodo. http://doi.org/10.5281/zenodo.3634369

[14] Czerniak, A. (2019). Processing charge integration into OpenAIRE Research Graph. Zenodo. http://doi.org/10.5281/zenodo.3634454

[15] Broschinski, C., & Pieper, D. (2019). Cost transparency for transformative agreements (or how to calculate equivalent APCs within the DEAL-Wiley agreement). Zenodo. http://doi.org/10.5281/zenodo.3472208

[16] Pieper, D. (2019). Cost transparency for transformative agreements. Zenodo. http://doi.org/10.5281/zenodo.3634373

[17] Mosterd, M. (2019). Integrating the costs of different OA book models into the ecosystem. Zenodo. http://doi.org/10.5281/zenodo.3634375

[18] Copiello, S. (2020). Business as Usual with Article Processing Charges in the Transition towards OA Publishing: A Case Study Based on Elsevier. Publications, vol. 8, no. 1, p. 3. http://dx.doi.org/10.3390/publications8010003

Special thanks should be given to the OpenAPC team, especially to Christoph Broschinski (OpenAPC Initiative and Bielefeld University Library/Germany), for their contribution, work, and ideas regarding APCs and transformative agreements.