Retrospective and prospective study of the evolution of APC costs and electronic subscriptions for French institutions

Antoine Blanchard, Diane Thierry, Maurits van der Graaf

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE

Liberté
Égalité
Fraternité

Ouvrir la science !

Retrospective and prospective study of the evolution of APC costs and electronic subscriptions for French institutions

December 2022

DATACTIVIST

pleiade

BIBLIOGRAPHIC INFORMATION

Antoine Blanchard, Diane Thierry, Maurits van der Graaf. *Retrospective and prospective study of the evolution of APC costs and electronic subscriptions for French institutions.* Report to the French Ministry for Higher Education and Research. December 2022. doi:10.52949/26

Antoine Blanchard
- 0000-0003-4630-1012
- Datactivist, 7 bis avenue Saint-Jérôme, 13100 Aix-en-Provence, France

Diane Thierry
- 0000-0002-8202-2048
- Datactivist, 7 bis avenue Saint-Jérôme, 13100 Aix-en-Provence, France

Maurits van der Graaf
- 0000-0002-2296-7568
- Pleiade Management and Consultancy, F. Scholtenstraat 45B, 1051 ET Amsterdam, The Netherlands

———

CREDITS

This study was performed by Datactivist (France) and Pleiade Management and Consultancy (the Netherlands) with funding from the French Ministry for Higher Education and Research. It was supervised by a Steering Committee comprised of:

- Odile Contat (Ministry for Higher Education and Research)
- Romane Coutanson (Ministry for Higher Education and Research)
- Marin Dacos (Ministry for Higher Education and Research)
- Thierry Fournier (Couperin; University of Rennes 1)
- Anne-Solweig Gremillet (Ministry for Higher Education and Research)
- Eric Jeangirard (Ministry for Higher Education and Research)
- Anne L'Hôte (Ministry for Higher Education and Research)
- Valérie Larroque (Couperin)
- Françoise Rousseau (Couperin; CEA)
- Sylvie Rousset (CNRS)
- Anne-Sophie Tagliavini (Couperin)
- Didier Torny (CNRS)

The authors acknowledge CNRS-Inist for help with Web of science data, Abes for help with BACON data, and thank the Steering Committee for their illuminating and engaging discussions. Special thanks are due to Valérie Larroque who provided additional analyses in relation to the ERE survey data. We are also grateful to Arjan Schalken (UKBsis, The Netherlands), Najko Jahn (Niedersächsische Staats- und Universitätsbibliothek Göttingen, Germany), Bianca Kramer and Jeroen Bosman (Utrecht University Library, The Netherlands) who shared their experience with regards to OA publishing and APC studies.

The authors take full responsibility for the content of the report.

———

MESSAGE TO READERS

The report was written for an expert audience; however, we did our best to communicate clearly the key insights of the study. The technicalities of the study were kept at a minimum: however,

for the sake of transparency and reproducibility, the interested reader will find along each graph the data and source code that produced it.

———

OPEN DATA

The data used in the study is available under an open license from the national research data repository Recherche Data Gouv: https://doi.org/10.57745/AQ0OPT. The source code is available from Zenodo: https://doi.org/10.5281/zenodo.6940484.

In addition, each graph provides a link to the table of plotted values and the source code that generated it.

———

GRAPHIC DESIGN

opixido

———

LICENSE

TABLE OF CONTENT

TABLE OF FIGURES

TABLE OF TABLES

01 Résumé

CONSTRUIRE LE JEU DE DONNEES DES ARTICLES DE PERIODIQUES

Un jeu de données sur les articles de périodiques a été construit, contenant les métadonnées des articles publiés par des auteurs affiliés en France sur la période 2013-2020, afin de servir de base à une analyse rétrospective et prospective du coût total des APC (que nous appellerons simplement « coût des APC ») pour les institutions françaises. Les APC (*article processing charges*) sont les frais de publication que les chercheurs doivent payer à certaines revues pour que leurs articles soient publiés en libre accès.

Le jeu de données a été construit en deux temps :

- **socle de données du BSO** : le socle du jeu de données est fourni par le Baromètre de la science ouverte (BSO), défini pour cette étude comme l'univers total des articles de périodiques publiés par des auteurs affiliés en France. Compilation et extraction de données issues de la base Unpaywall, enrichies par d'autres sources, le BSO est limité aux publications avec DOI. Il utilise un algorithme pour déterminer si un auteur est affilié en France, et pour déterminer le montant des APC à partir des données OpenAPC
- **enrichissement des données du BSO avec le Web of Science et OpenAlex** : pour évaluer le coût des APC pour les institutions françaises, il convient de connaître le pays d'affiliation de l'auteur correspondant (en principe celui qui paye les APC), de savoir si des APC ont été payés, et de connaître le montant des APC. C'est pourquoi les données du BSO ont été enrichies avec des informations sur les auteurs correspondants déduites du Web of Science. Un autre enrichissement avec les données OpenAlex a permis de déterminer si des APC ont été payés pour la publication en libre accès. Les données ont encore été enrichies avec des données de Couperin et de QOAM.

ANALYSE RETROSPECTIVE DU COUT DES APC POUR LES AUTEURS CORRESPONDANTS AFFILIES EN FRANCE

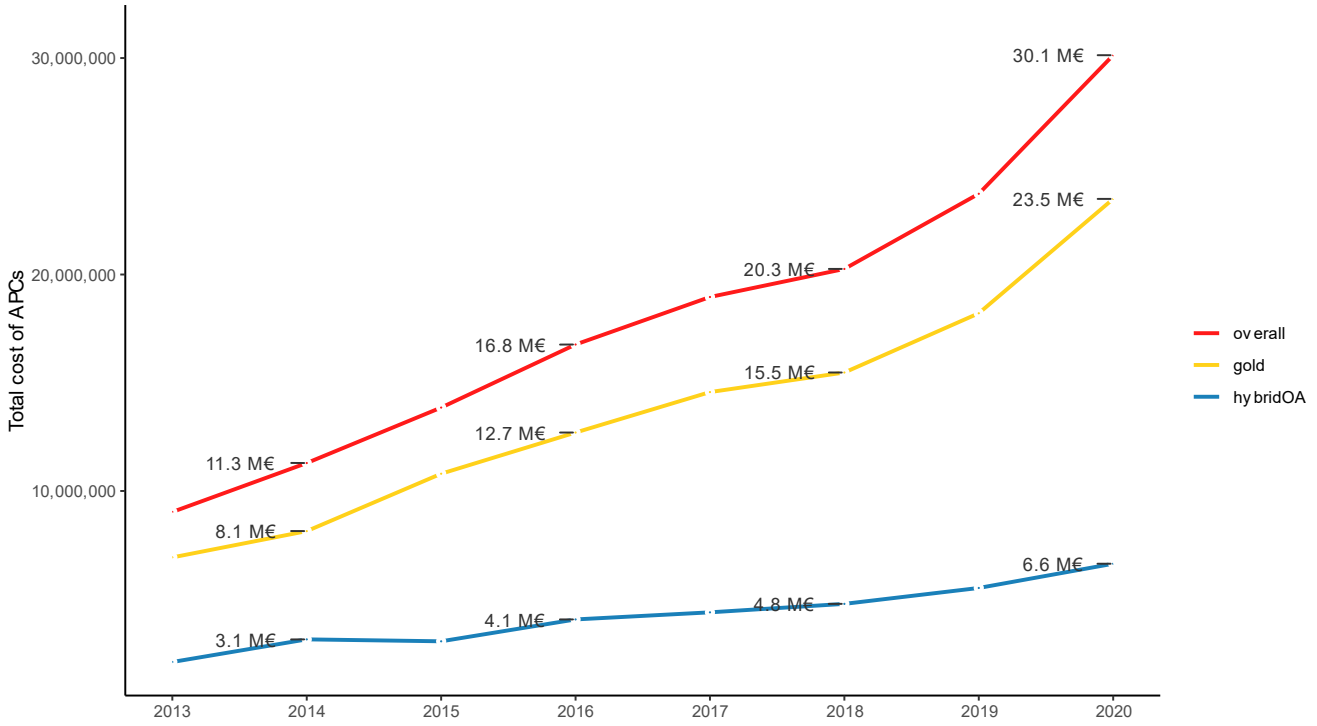
À partir du jeu de données décrit précédemment, l'étude rétrospective du nombre d'articles avec APC et auteur correspondant affilié en France a permis d'obtenir plusieurs résultats principaux :

- le coût des APC triple sur la période 2013-2020, due principalement à la croissance du nombre d'articles *gold open access* (sans cette croissance, le coût aurait été multiplié par 1,69 au lieu d'être multiplié par 3)
- les éditeurs et plateformes de diffusion des périodiques ont été groupés en quatre catégories, représentant chacune entre 20 et 32 % des articles du BSO en 2020 : la classe 1 comprend l'éditeur le plus représenté (Elsevier), la classe 2 comprend trois éditeurs, la classe 3 comprend 16 éditeurs et la classe 4 comprend la longue traîne des éditeurs (n = 1995). Le plus gros taux de croissance des articles avec APC et auteur correspondant affilié en France est observé dans la classe 2 (Springer Nature, Wiley et MDPI)
- plus de trois quarts des articles avec APC et auteur correspondant affilié en France concernent la biologie et la recherche médicale.

L'observation principale concernant l'évolution du prix des APC est que le niveau de prix pour les revues hybrides s'établit en 2013 à un niveau élevé (2 453 € en moyenne) mais est stable au cours du temps pour atteindre une moyenne de 2 488 € en 2020. Le niveau de prix pour les revues *gold*

s'établit en 2013 à un niveau significativement plus faible avec un APC moyen de 1 395 €. Cependant, il croît très vite pour atteindre 1 745 € en 2020.

Nous avons également pu calculer le coût des APC pour les institutions françaises entre 2013 et 2020.



Évolution du coût des APC payés par les auteurs correspondants affiliés en France, au total et par type de libre accès, après reconstruction des données manquantes

ANALYSE DU COUT DES ABONNEMENTS ELECTRONIQUES

Le consortium Couperin, réseau de négociation et d'expertise des ressources documentaires électroniques, a également fourni les données 2019 et 2020 de l'enquête ERE relative au coût des abonnements électroniques pour les institutions françaises membres de Couperin. Ces données ont été traitées avec Microsoft Power BI et croisées avec les catégories des ressources, les catégories d'éditeurs et plateformes de diffusion, et les répondants à chacune des deux éditions de l'enquête. Cette analyse a permis d'estimer à environ 87,5 M€ la dépense d'abonnements aux périodiques en 2020. En outre, une analyse par Couperin des années 2014-2021 de l'enquête ERE a montré que la variation de prix sur la période s'est située entre -1,95% et +7,22% par an, avec une augmentation de prix moyenne de 1,76 % par an.

À partir de ce taux de croissance moyen, nous estimons que les abonnements électroniques s'élèveront à 97,5 M€ en 2030.

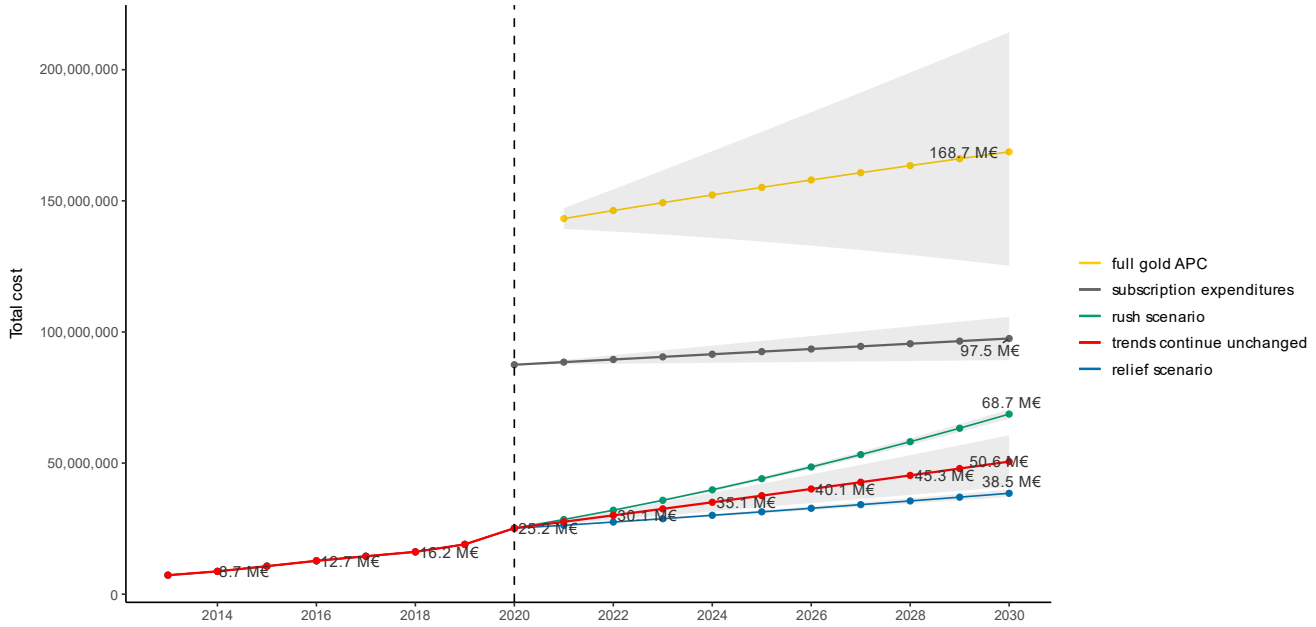
ANALYSE PROSPECTIVE DU COUT DES APC

En utilisant les données de confiance sur les articles (c'est-à-dire en ne retenant pas les articles pour lesquels le pays d'affiliation de l'auteur correspondant n'a pas pu être déterminé, ni ceux dont le montant des APC n'a pas pu être estimé avec suffisamment de certitude), nous avons établi des modèles mixtes du prix des APC en fonction des autres métadonnées des articles.

Ce modèle a permis de prédire l'évolution du coût des APC entre 2021 et 2030 dans plusieurs situations, représentées graphiquement dans la figure ci-dessous :

- sous l'hypothèse d'une évolution à l'identique des tendances observées (courbe rouge nommée « trends continue unchanged »)
- dans un scénario d'accélération vers le *gold open access* (courbe verte nommée « rush scenario »), où la hausse de la part d'articles Gold et la hausse des montants d'APC s'emballent
- et dans un scénario de hausse du libre accès *green* et transition du libre accès hybride vers *gold* (courbe bleue nommée « relief scenario »).

Enfin, nous avons simulé le plafond en prenant une hypothèse d'école (courbe jaune nommée « full gold APC ») : il s'agit du montant d'APC qui serait payé si tous les articles d'auteurs correspondants affiliés en France étaient publiés dans des revues en libre accès (répartis entre 10% de revues diamant et 90% de revues *gold*).



Résumé du coût prédit des APC payés par les auteurs correspondants affiliés en France dans plusieurs situations (2021-2030)

Résultats principaux

Coûts 2020

- Dépenses d'abonnement aux périodiques électroniques en 2020 : 87,5 M€
- Coût des APC en 2020 : 30,1 M€

Coûts prédits sous l'hypothèse d'une évolution à l'identique des tendances observées :

- Dépenses d'abonnement aux périodiques électroniques en 2030 : 97,5 M€
- Coût des APC en 2030 : 50,6 M€

Coût prédit dans un scenario d'accélération vers le *gold* OA :

- Coût des APC en 2030 : 68,7 M€

Coût prédit dans un scénario de hausse du libre accès *green* et transition du libre accès hybride vers *gold* :

- Coût des APC en 2030 : 38,5 M€

Coût prédit pour 90% d'articles d'auteurs correspondants affiliés en France dans des revues *gold* (plafond théorique) :

- Coût des APC en 2030 : 168,7 M€

02 Summary

BUILDING THE DATASET OF JOURNAL ARTICLES

A journal article dataset has been developed with metadata of articles by France-based authors in the period 2013-2020. The purpose of this dataset was to form a basis for the retrospective and prospective analyses of the total costs of APCs paid by French institutions. APCs are the article processing charges (APCs) that researchers must pay to have their articles published in some open access journals.

The dataset has been built as follows:

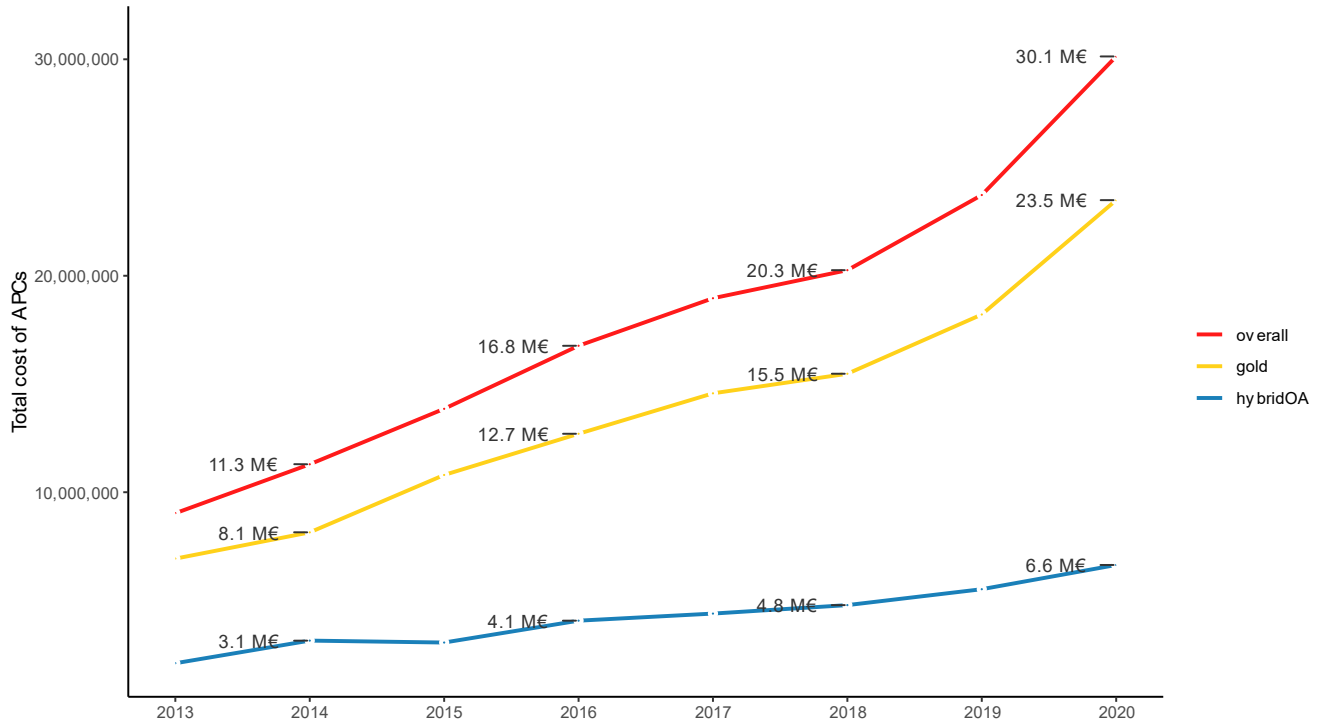
- **BSO data as the basis:** the basis of the dataset is formed by data from the French Open Science Barometer (called BSO as in *Baromètre de la science ouverte*), which we have defined for this study as being the 'total universe of journal articles by France-based authors'. Compiled and extracted from Unpaywall database and enriched by other sources, BSO is focused on publications with a DOI. It uses an algorithm to assess if an author is affiliated with a French institution and collects APC information using an algorithm based on the OpenAPC data
- **enriching the BSO data with data from Web of Science and OpenAlex**: for the assessment of APC costs for French institutions, information on the affiliation country of the corresponding author (who pays the APC), information on whether an APC has been paid, and the amount of APC are crucial elements. Therefore, we enriched the BSO data by adding information on the corresponding author derived from the Web of Science. Another enrichment with OpenAlex data took place in order to ascertain whether the article was Open Access as a result of an APC payment (APC-paid articles). The data were further enriched with data from Couperin and QOAM.

RETROSPECTIVE ANALYSIS OF APC COSTS FOR OPEN ACCESS ARTICLES BY FRANCE-BASED CORRESPONDING AUTHORS

The main results of the retrospective analysis of the above-described dataset regarding the numbers of APC-paid articles with a France-based corresponding author:

- the total cost of APCs has tripled in the period 2013-2020. The major driver is the growth of articles in Gold OA journals, i.e. fully open access journals with APCs (without this growth, the APC cost would have been multiplied by 1,69 instead of 3)
- the journal publishers and dissemination platforms have been categorised in four tiers, each publishing between 20% and 32% of all articles with a French co-author in 2020: tier 1 with the top publisher (Elsevier), tier 2 with three publishers, tier 3 with 16 publishers and tier 4 with the long tail of publishers (n=1995). The highest growth rate of APC-paid articles by France-based corresponding authors is seen in journals published by tier 2 publishers (Springer Nature, Wiley and MDPI)
- more than three quarters of the APC-paid articles by France-based corresponding authors are in the fields of biology and medical research.

The main observation regarding the price evolution of the APCs is that the APC-level for HybridOA articles started in 2013 at a high level (2 453 € in average) but have been stable over the years with 2 488 € in average in 2020. The APC-level for articles in Gold journals started considerably lower with an average APC in 2013 of 1 395 €. However, a rapid increase in the level of APCs for Gold OA articles has been observed, with an average APC of 1 745 € in 2020.
This has led to a calculation of the total cost of APCs paid by French institutions between 2013 and 2020 (see figure below).



Evolution of total cost of APCs paid by France-based corresponding authors, overall and per open access color, after reconstructing missing data

ANALYSIS OF THE SUBSCRIPTION COSTS FOR JOURNAL PACKAGES

In addition to the above-mentioned article dataset, 2019 and 2020 data from the ERE survey by Couperin (the national consortium of research performing organizations that negotiates with publishers the prices and conditions of access to research publications for the benefit of its members) were analysed in order to assess the total subscription costs of journal packages for French institutions. This data was compiled in Microsoft Power BI with additional information about the categories of the products, the publisher tiers, and the respondents to both surveys. This resulted in an estimate of ca. 87,5 M€ for the total expenditure on journal packages by all Couperin members in 2020. An analysis by Couperin of the ERE surveys 2014-2021 showed that the price increases per year in this period varied between -1,95% and +7,22% with an average price increase of 1,76% per year.

The evolution of the total journal subscription costs based on this average annual growth rate results in an estimated 97,5 M€ in 2030.

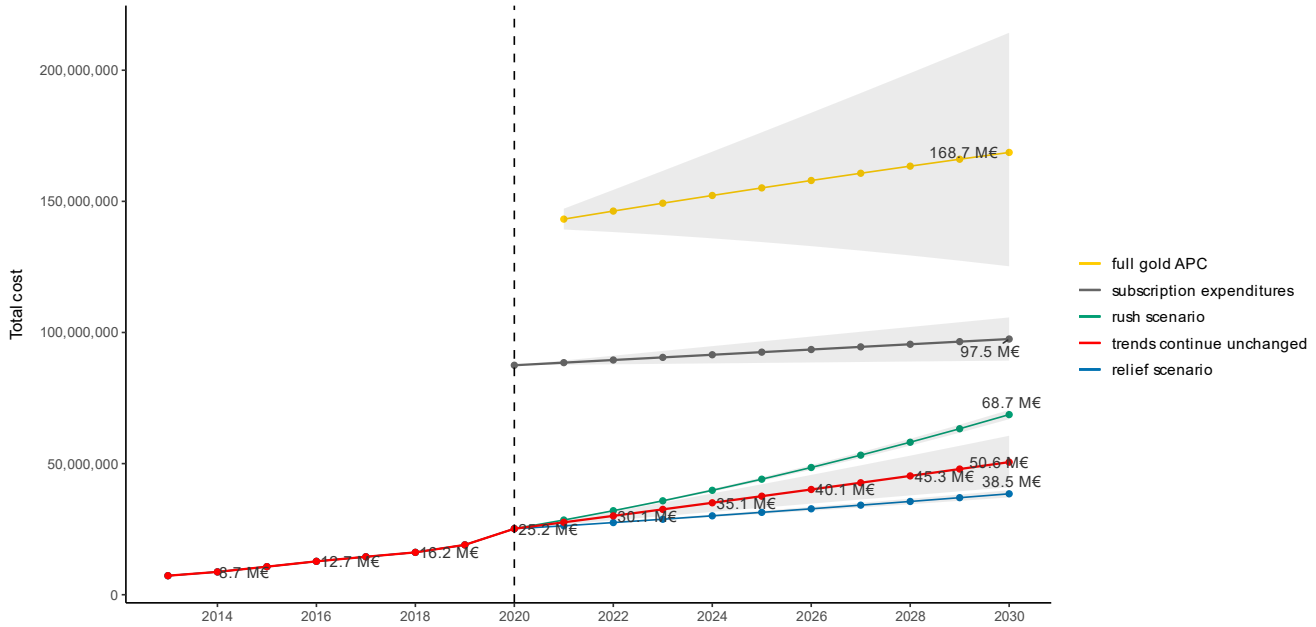
PROSPECTIVE ANALYSIS OF APC COSTS

Using known and trusted data from our dataset (i.e. excluding articles for which the country of the corresponding author could not be determined, or whose APC price could not be precisely estimated), we developed mixed-effects models of APC prices as a function of other articles' features.

These models were used to predict the evolution of the total cost of APCs for the period 2021-2030, in various situations, represented on the graph below:

- under the assumption of continuing and unchanged trends (red line)
- for a scenario with an acceleration towards Gold OA (green line)
- and for a scenario with increase of Green OA and a transition from HybridOA towards Gold (blue line).

Finally, a situation was simulated where 90% of all articles by France-based corresponding authors would be Gold OA (and 10% would be published in Diamond OA journals) in order to assess the cost ceiling of such a 100% open access situation (yellow line).



Summary of predicted cost of APCs in various situations (2021-2030)

Main results

Figures for 2020

- Total expenditure on journal subscriptions in 2020 is 87,5 M€
- Total cost of APCs in 2020 is 30,1 M€

Predictions in case of continuing trends:

- Total expenditure on journal subscriptions in 2030 is 97,5 M€
- Total cost of APCs in 2030 is 50,6 M€

Prediction in case of an acceleration towards Gold OA:

- Total cost of APCs in 2030 is 68,7 M€

Prediction in case of an increase of Green OA and a transition from HybridOA towards Gold:

- Total cost of APCs in 2030 is 38,5 M€

Prediction in hypothetical case where 90% of the articles with a France-based corresponding author are APC-paid (theoretical cost ceiling):

- Total cost of APCs in 2030 is 168,7 M€

03 Introduction

The French Ministry for Higher Education and Research (MESR) has the ambition to develop a publishing ecosystem that avoids the domination of only one road towards open access (OA) and balances the various roads, or colors (Green, Gold, Diamond). While the spiraling cost of Gold OA with article processing charges (APCs) puts enormous pressure on the economical sustainability of OA publishing, the purpose of the study is twofold:

- **a retrospective analysis** of APC costs in order to determine the expenses incurred over the last years
- **a prospective estimation** (over the next 10 years) of the evolution of APC costs.

Said analysis is expected to impact the negotiations with scholarly publishers as well as future open science policies, such as the National Open Science Plan.

The present report documents the methodology of the study as well as the various statistical analysis and their outcomes.

04 Methodology

We developed a methodology that looked at two complementary costs:

- **APC costs**: we built an article-level dataset as a basis for the analyses on the APC costs per article
- **subscription costs**: we obtained data from the ERE survey by Couperin on expenditures for journal subscriptions.

I. Building the article-level dataset

COVERAGE OF THE DATASET

An article-level dataset was developed with data on articles by French authors in the period 2013-2020. This dataset is based on data from the French Open Science Barometer (BSO, see Bracco et al., 2022), and enriched with data from other sources: data provided by Couperin, data from OpenAlex and QOAM databases, and information derived from Web of Science.

We consider the BSO dataset as the **'total universe of journal articles by French authors':**

- BSO is a compilation of data from Unpaywall, enriched by several other sources and using an algorithm to assess if an author is French (Jeangirard, 2019).
- The BSO is limited to publications having a DOI. Publications without a DOI in the French OA repository HAL tend to be in French language and in the field of humanities and social sciences (Jeangirard, 2019).
- Only published journal articles have been included in the dataset for further use (other types of publications e.g. submitted papers are excluded).

The systematic bias of the dataset will be discussed in Section 04.II (p. 19).

CORRESPONDING AUTHOR INFORMATION

It is critical to determine who the corresponding author is, because they will typically pay the APCs (Monaghan, 2020; van der Graaf, 2017).

Corresponding author information in the BSO data: the corresponding author information is collected from the HTML version of the articles available on the publishers' platforms. If there is an email address linked to an author, this author is identified as the corresponding author. However, the scripts used are not perfect. The collected data on authors are probably of better quality for the more recent data in the dataset

Other author information in the BSO data: the field 'author position' reflects the position of the author in case of multiple authors. This field comes from CrossRef.

Affiliation information in the BSO data: affiliation information is also collected and the country per author is determined with an algorithm (affiliation matcher) with high precision (99%) and high recall (97%) (L'Hôte & Jeangirard, 2021). This results in the field 'detected_countries', which contains all countries detected in the address field of the article (see the dictionary of variables in Thierry et al., 2022).

Corresponding author information from the Web of Science: in addition to these BSO data, information about the corresponding author derived from the Web of Science (WoS) has been added to the dataset, for about 60% of the article (the rest of the articles are in journals that are not covered by Web of Science). For each article, we identified the corresponding author(s) and we computed a Boolean variable (which can take the value 'true' or 'false') of whether an article has at least one corresponding author affiliated to France (see the dictionary of variables in Thierry et al., 2022). For obvious intellectual property reasons, the original Web of Science field is not part of the data.

JOURNAL INFORMATION

ISSN: we use the ISSN-L (linking ISSN) provided by BSO in order to catch all media versions of a journal, be it online or print.
Publisher name: BSO provides a publisher_dissemination variable which cleans and maps the different publisher names to the publisher who owns these names. The mapping is a three-step process based on correspondence tables available from https://github.com/dataesr/bso-publications/tree/1eed00eb4d5e524499846975abad82fa5b215583/bso/server/main/publisher:

- **Normalization**: for example, if the name starts with Elsevier, eventual additions such as Masson or B.V. are left out
- **Major mergers and acquisitions** from 2015 onwards
- **Dissemination**: OpenEdition and Cairn, but also Wiley, Elsevier and others, do the dissemination for small publishers. In these instances, the variable takes the value of the dissemination platform instead of the actual publisher.

It should be noted that the publisher_dissemination variable states the situation of the disseminating publisher from one particular starting year. There is no retrospective mapping in case of mergers or acquisitions.

Publisher tier: instead of looking at individual publishers and dissemination platforms (as defined above), we aimed at grouping them in descending order of number of articles published in the dataset in 2020, the most recent year of the dataset (see Table 1). We defined 4 tiers of approximately the same size (number of articles): each tier comprises less publishers or platforms than the next tier, with tier 1 comprising only the top one publisher, Elsevier.

Table 1: List of publishers and publisher tiers by descending order of articles published in 2020

Publisher / Dissemination platform	Number of articles	Percentage	Tier
Elsevier	39521	27,91	1
Tier 1		**27,91**	
Springer-Nature	17343	12,25	2
Wiley	11460	8,09	2
MDPI	6933	4,9	2
Tier 2		**25,24**	
OpenEdition	4747	3,35	3
Cairn	4470	3,16	3
Oxford University Press	4384	3,1	3
Informa	3856	2,72	3
Frontiers	2957	2,09	3
American Chemical Society	2675	1,89	3
Wolters Kluwer Health	2490	1,76	3
EDP Sciences	2395	1,69	3
American Physical Society	2321	1,64	3
IEEE	2147	1,52	3
Royal Society of Chemistry	1686	1,19	3
IOP Publishing	1623	1,15	3
SAGE Publications	1463	1,03	3
British Medical Journal	1104	0,78	3
Public Library of Science	1061	0,75	3
Cambridge University Press	934	0,66	3
Tier 3		**28,48**	
total top-20	115570	81,63	
rest of publishers (n=1995)	26011	18,37	4
Tier 4		**18,37**	
total articles in 2020	141581	100	

Journals covered by Couperin: Couperin is the national consortium of research performing organizations that negotiates with publishers the prices and conditions of access to research publications and other digital resources for the benefit of its members. They provided the journal names and ISSN numbers of all journals covered by Couperin contracts. These data were added to the dataset in the field is_covered_by_couperin (see the dictionary of variables in Thierry et al., 2022). As the list has no historical data, the distinction between journals that are covered or not has limited value for the retrospective analysis.

ARTICLE INFORMATION

Publication year: there are often discrepancies between datasets regarding the publication year. The BSO team recommended to consider the publication year as stated in the BSO field 'year' (this reflects the year the publication was actually published).

OA color of article: BSO determines the OA status of an article, and several colors can coexist when an article is found at the site of the publisher (Gold, Hybrid, Diamond… OA) as well as in a repository (Green OA). We have created several variables to determine the color of an article, based on different rules (see the dictionary of variables in Thierry et al., 2022).
In oa_details.2020.oa_colors_with_priority_to_publisher, BSO determines the OA status of an article by focusing on the business model of the journal. The definitions used by BSO are:

- **Diamond**: OA journal without APC according to the Directory of open access journals (DOAJ)
- (APC-)**Gold**: OA journal with APC according to DOAJ
- **HybridOA**: journal not OA according to DOAJ but article with a known APC
- **Other:** all other cases of OA
- **Closed Access**
- **Green only:** if an article is open access on a repository and not at the site of the publisher.

Across the report, Hybrid stands for a toll-access journal while HybridOA stands for articles published with a fee under an open license in Hybrid journals.

For the purpose of the analysis, we settled the variable oa_color_finale which we derived from existing variables:

- **Diamond**: same definition as for oa_details.2020.oa_colors_with_priority_to_publisher above
- (APC-)**Gold**: same definition as OpenAlex once BSO has determined that the article is OA
- **HybridOA**: same definition as OpenAlex once BSO has determined that the article is OA
- **Other:** all other cases of OA

Discipline: the classification in BSO uses an algorithm based on Pascal and Francis databases from Inist-CNRS that consists of 10 categories.

APC INFORMATION

The amount in Euros of the APC for an article in BSO is based on several sources, listed in Table 2 from the most preferred to the least preferred. The variable apc_source stores which source is used (see the dictionary of variables in Thierry et al., 2022).

Table 2: Sources of APC information by decreasing order of priority (i.e., if A is available then A, is used, if A is not available but B is available then B is used, etc.)

In the APC source field	Description
(A) "openAPC"	APC amount for the DOI available in OpenAPC
(B) "doaj"	APC listed in DOAJ. Note: if there are APC values in OpenAPC, but DOAJ states it is a Diamond journal, then DOAJ has priority
(C) "openAPC_estimation_issn_year"	average APC for the same journal in OpenAPC, for the same year as the article (minimum 10 APCs)
(D) "openAPC_estimation_issn	average APC for the same journal in OpenAPC, for all

	years available
(E) "openAPC_estimation_publisher_year"	average APC for the same publisher in OpenAPC, for the same year as the article (minimum 10 APCs)
(F) "openAPC_estimation_publisher"	average APC for the same publisher in OpenAPC, for all years available

In practice, **A is the exact amount paid by the French institution. B and C are about or precisely the amount paid by the French institution**. D will be in most cases a rather good estimate of the amount that is paid by the French institution.

Regarding E and F, this estimate of the APC could differ considerably from the APC actually paid by the French institution. This is especially true for journals of publishers with a large portfolio of journals with different quality ratings and prices. For smaller publishers with a limited number of journal titles, the APC prices might vary less between the journals. **In the analysis, we will consider E and F as "untrustworthy" sources.**

APC discount under Read & Publish (also called "transformative") agreements: Couperin has signed nine Read & Publish agreements as of May 2022, three of which were active during the period 2013-2020 (see Table 3).

Table 3: List of Read & Publish agreements signed by Couperin

Publisher	Starting year	Effect on APCs paid
EDP Sciences	2017	APC fully paid by the contract
Elsevier	2019	25% discount for the years 2019 and 2020; 33.3% discount for 2021; 35.4% discount in 2022
American Chemical Society	2020	25% discount
Cambridge University Press	2021	No APC in hybrid and OA journals
Karger	2021	No APC
Lippincott	2021	5% or 10% discount on APC
PNAS	2021	20% discount APC (other CC licenses) or 12% discount in case CC-BY
Royal Society of Chemistry (Pack RSC Gold)	2021	Quota of articles free of APCs, above quota 15% discount
Thieme	2021	50% discount on APC

Couperin provided the list of DOIs which benefited from the EDP Sciences and Elsevier Read & Publish agreements.

In the period 2013-2020, there also has been a voucher system for OA articles in journals of the Royal Society of Chemistry. However, this was a rather marginal phenomenon.

II. Preliminary analysis of the article-level dataset

SYSTEMATIC BIAS

Coverage of the article-level dataset: a recent study investigated the coverage of BSO compared with other data sources (Chaignon, 2022). This study used several data sources (Scopus, Web of Science, Microsoft academic Graph, HAL, NASA/ADS and PubMed) to collect publications with

DOIs by French authors published in the year 2019 (Corpus FR-2019): 3,5% of the publications covered by BSO have in reality not a French author (false positives) and BSO misses 9% of publications that are listed in other sources with a French author (false negatives). It must be noted that these numbers include journal articles, but also other types of publications, such as conference proceedings, book chapters and other publications.

For the cost calculations, the articles without a DOI and for a larger part in the HSS domain are most probably published in small journals with a subscription model or a Diamond OA model. So, the effects on the total cost of APCs are considered negligibly small. The effects of the 3,5% false positives and the 9 % false negatives on the APC calculation can be estimated as cancelling each other out, resulting in a net **5,5 % underestimation of the total APC costs calculated**.

However, this underestimation will be canceled out by an overestimation of the APCs paid.

Waivers and discounts: in a study for Knowledge Exchange (van der Graaf, 2017), researchers from Inria who published in 2015 in Gold or Hybrid journals were asked if an APC was paid. The responses showed that 4,2 % of the APCs for articles in APC-Gold journals were waived and in 12,5 % of the cases the APC was discounted (it is not known by how much). For articles in Hybrid journals, 8,3 % were waived and 4,2% had gotten a discount. The study also reports surveys among authors of the University of Helsinki, University of Göttingen, University of Glasgow and the Technical University of Eindhoven, which showed similar percentages of waivers (3 to 5% for APC-Gold journals, less for Hybrid journals).

Despite waiver policies being directed at authors from low- and middle-income countries, many publishers give waivers and discounts for authors from high income countries at their own discretion. Editors and reviewers, but also other authors can be eligible on a case-by-case basis and as such, waivers and discounts can be part of the marketing of the journal. This is primarily the case for APC Gold journals.

Corresponding author and payment of APCs: in the same study, interviews with authors made clear that generally the corresponding author is responsible for the payment of the APC, but there are also cases that the amount of APC is divided among several authors (particularly if they are from different institutions).

In view of the above, **we hypothesize that the overestimation of APCs paid is cancelled out against the 5,5 % underestimation of the number of articles.** Consequently, we do not take waivers and discounts into account.

NUMBER OF ARTICLES

The total number of journal articles in the dataset is presented in Table 4.

Table 4: Number of journal articles per year

Year	2013	2014	2015	2016	2017	2018	2019	2020	total
Nr. journal articles	117 399	119 205	120 229	129 322	130 948	136 506	137 327	141 581	1 032 517

NUMBER OF ARTICLES WITH A FRANCE-BASED CORRESPONDING AUTHOR

We have computed the affiliation country of corresponding authors (CAs) by applying a series of rules. The following tables show the distribution of France-based and non-France-based corresponding authors per year for each step.

Step 1: BSO data already contained 268 750 articles with a France-based corresponding author.

Table 5: Country of corresponding authors per year after step 1

Step		2013	2014	2015	2016	2017	2018	2019	2020	total
1	Nr. of articles with a French CA - BSO	10714	20845	25464	29006	43284	43096	46812	49529	**268750**

Step 2: the incorporation of WoS data led to the identification of:

- [2A] 388 210 articles with a France-based corresponding author, and
- [2B] 273 611 articles with a non-France-based corresponding author.

Table 6: Country of corresponding authors per year after step 2

Step		2013	2014	2015	2016	2017	2018	2019	2020	total
2A	Nr. of articles with a French CA - WOS	41443	41377	45389	50784	51159	51307	51495	55256	**388210**
2B	Nr. of articles with a non-French CA - WOS	23834	24829	27983	33987	36219	38521	40611	47627	**273611**

Step 3: BSO and WOS together (overlap: 63,3% of articles) led to the identification of:

- [3A] 477 479 articles with a France-based corresponding author, according to at least one of the two sources
- [3B] 54 352 articles with a non-France-based corresponding author, according to both WOS and BSO (we only count when there is a consensus between the two sources)
- [3C] out of 733 767 articles with reliable information (non-missing value) on the corresponding author, either from BSO or WOS
- [3D] leaving 298 750 articles without adequate information to establish the country of the corresponding author.

Table 7: Country of corresponding authors per year after step 3

Step		2013	2014	2015	2016	2017	2018	2019	2020	total
3A	Nr. of articles with a French CA - BSO or WOS	48526	50444	54592	59433	62540	62958	66728	72258	**477479**
3B	Nr. of articles with a non-French CA - BSO and WOS	3911	4238	3647	4562	8746	9413	9877	10058	**54352**
3C	Nr. of articles with reliable info on CA - BSO or WOS	72567	75927	81974	90488	95319	97801	103649	116042	**733767**
3D	Nr. of articles	44832	43278	38255	38834	35629	38705	33678	25539	**298750**

	without reliable info on CA – BSO and WOS									

The sum of [3C] and [3D] is equal to the total number of articles [1]. However, the sum of [3A] and [3B] is not equal to [3C] because [3C] also comprises articles for which there is no consensus on whether the CA is non-France-based (one source identifies it so, the other does not have the information).

Step 4: the 298 750 articles without adequate information to establish the country of the corresponding author were treated as follows:

- [4A] firstly, their DOI was checked against OpenAlex to retrieve information about the affiliation countries of the authors. For 37 338 articles whose country codes were all FR, we determined that the corresponding author must be French
- [4B] for the remaining articles, we checked the number of authors: 76 760 articles with a single author must have a France-based corresponding author as per the definition of the BSO data 'containing at least one French co-author'
- [4C] out of the 258 291 remaining articles, 120 614 are written in French and we assume their corresponding author is France-based. This will introduce a slight overestimation when the corresponding author is affiliated to another French-speaking country or affiliated to a non-French-speaking country.

Table 8: Country of corresponding authors per year after step 4

Step		2013	2014	2015	2016	2017	2018	2019	2020	total
4A	Nr. of articles with a French CA - OpenAlex	7518	7116	5885	5259	3661	3694	2304	1901	**37338**
4B	If single author, assumption is CA = FR	11693	10933	10140	9677	9717	10065	9942	4593	**76760**
4C	Not 4A, not 4B, if written in French, assumption is CA = FR	20193	19435	16472	16576	16157	17122	13796	6833	**126584**

Step 5: eventually, the information is consolidated in Table 9 as follows:

- [5A] deduplicated data from BSO, WOS and OpenAlex using the above rules led to the identification of 641 401 articles with a France-based corresponding author, i.e. 62% of the dataset
- [5B] deduplicated data from BSO and WOS led to the identification of 269 787 articles with a non-France-based corresponding author, i.e. 26% of the dataset. Contrary to [3B], one source identifying the articles as having a non-France-based corresponding author was enough
- [5C] leaving us with 121 329 articles, i.e. 12% of the dataset, without adequate information on the country of the corresponding author, which could be France-based or not. Therefore, figures 5A and 5B are underestimated.

Table 9: Country of corresponding authors per year after step 5

Step		2013	2014	2015	2016	2017	2018	2019	2020	total
5A	Nr. of articles by French CA (BSO, WoS. OpenAlex), or single author	76237	76995	76949	81268	82358	83774	82828	80992	**641401**

	or written in French (assumption CA=FR)									
5B	Nr. of articles with non-French CA (BSO or WoS)	25996	27370	29000	32815	34301	36360	38368	45577	**269787**
5C	Nr. of articles without reliable info on country of CA	15166	14840	14280	15239	14289	16372	16131	15012	**121329**
	Total number of articles	**117399**	**119205**	**120229**	**129322**	**130948**	**136506**	**137327**	**141581**	**1032517**
5A	% articles with a French CA or written in French	65%	65%	64%	63%	63%	61%	60%	57%	**62%**
5B	% articles with non-French CA and nor written in French	22%	23%	24%	25%	26%	27%	28%	32%	**26%**
5C	% articles without reliable info on country of CA	13%	12%	12%	12%	11%	12%	12%	11%	**12%**

NUMBER OF ARTICLES WITH APC

The number of articles with APC paid by France-based corresponding authors (CAs) per year is presented in Table 10, which was obtained through the following steps:

- Step 1: BSO has a Boolean variable called has_apc that relies on an analysis of the journals' business models and licenses, as explained in Section 2.1.5.2 of Bracco et al. (2022). We used has_apc to indicate if an article is OA at the site of the publisher and the journal is not diamond. 278 507 articles have this label
- Step 2: we used the DOIs of these articles to collect the information from OpenAlex about the business model. For 165 775 articles whose color is Gold (published in an OA journal that is indexed by the DOAJ) or HybridOA (fee under an open license in a toll-access journal) we assumed that an APC has been paid
- Step 3: breaking down the 165 775 articles for which an APC has been paid:
    - 81 058 articles have a France-based CA
    - 67 130 articles have a non-France-based CA
    - we could not establish the country of the CA for 18 259 articles.

Table 10: Number of articles with APC paid by France-based corresponding authors per year

Step	APCs paid by France-based corresponding authors	2013	2014	2015	2016	2017	2018	2019	2020	total
1	Nr. of articles OA at site of publisher (BSO; has apc=1)	22342	25600	28199	32760	33627	35843	45751	54385	**278507**
2	Nr. of articles OA at site of publisher (BSO; has apc=1) and APC paid - OpenAlex (Gold or HybridOA)	11246	13580	16080	18603	20907	23092	27365	34902	**165775**
3A	Nr. of articles OA at site of publisher (BSO; has apc=1) and APC paid (OpenAlex; Gold or HybridOA) and CA=FR	5774	6988	7983	9342	10484	11102	13037	16384	**81058**
3A1	3A, but only Gold	4840	5563	6651	7529	8561	9016	10478	13306	**65944**
3A2	3A, but only HybridOA	934	1425	1332	1813	1923	2086	2559	3042	**15114**
3B	Nr. of articles OA at site publisher (BSO; has apc=1) and APC paid (OpenAlex; Gold or HybridOA) and CA=nonFR	3840	4815	5924	7151	8384	9677	11691	15648	**67130**
3C	Nr. of articles OA at site publisher (BSO; has apc=1) and APC paid (OpenAlex; Gold or HybridOA) and no info on the country of the CA	1717	1891	2290	2192	2093	2386	2699	2991	**18259**
1	% articles OA at site of publisher (BSO; has_apc=1)	19%	21%	23%	25%	26%	26%	33%	38%	**27%**
2	% articles OA at site of publisher (BSO; has apc=1) and APC paid - OpenAlex (Gold or HybridOA)	10%	11%	13%	14%	16%	17%	20%	25%	**16%**
3A	% articles OA at site publisher (BSO; has apc=1) and APC paid - OpenAlex (Gold or HybridOA) and CA=FR	5%	6%	7%	7%	8%	8%	9%	11%	**8%**
3A1	3A, but only Gold	4%	5%	6%	6%	7%	7%	8%	9%	**6%**
3A2	3A, but only HybridOA	1%	1%	1%	1%	1%	2%	2%	2%	**1%**
3C	% articles OA at site publisher (BSO; has	1%	2%	2%	2%	2%	2%	2%	2%	**2%**

	apc=1) and APC paid (OpenAlex; Gold or HybridOA) and no info on the country of the CA									

NUMBER OF ARTICLES WITH APC PER DISCIPLINE

The proportion of articles with APC paid by France-based corresponding authors varies between disciplines. Figure 1 shows that APCs are most prevalent among Biology articles (30%), even though the absolute number of articles with APC is similar in Medical research. Humanities, Social sciences, and Mathematics have single-digit proportions of articles with APC. All other disciplines are between 12% and 17%.

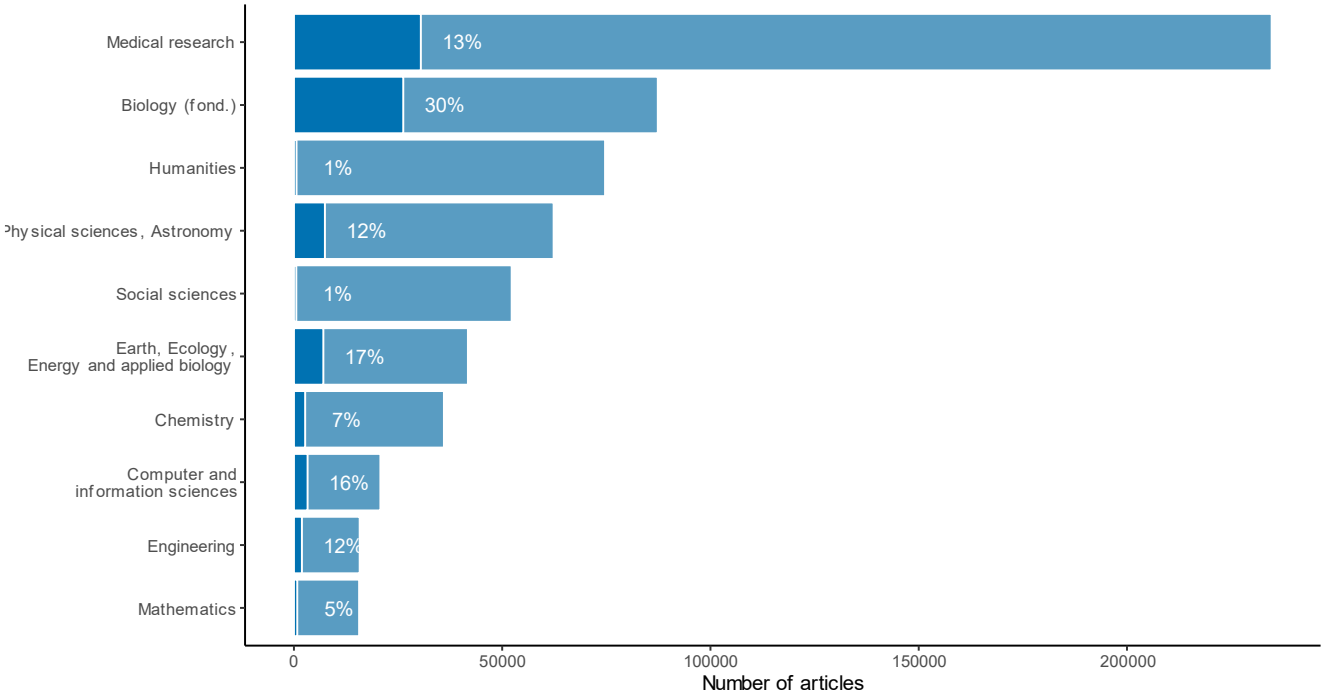


Figure 1: Proportion of articles with APC among all articles by France-based corresponding authors, per discipline[1]

NUMBER OF ARTICLES PER OA COLOR

The distribution of Diamond, Gold and Hybrid articles per year is presented in Table 11, which shows how these were calculated:

1. **Diamond**: the number of articles published in Diamond journals is established by BSO. BSO uses DOAJ for this: the article is published in an OA journal indexed by the DOAJ and without an APC according to DOAJ. 43 183 articles in the dataset have been published in Diamond journals.

[1] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/XPLISK and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4741

2. **Gold**: although BSO has also an identification of Gold articles, we use the OpenAlex data on this for reasons of consistency (we use these data also for the APC calculations). 123 755 articles OA at the site of the publisher are Gold according to OpenAlex
3. **Hybrid**: the journals of the remaining articles were matched with the QOAM dataset. QOAM is a large journal dataset that has 15 532 Hybrid journal titles and their ISSN-e identifiers. As the matching uses the ISSN-e identifier and the words in the journal title, not all articles can be matched (especially because the ISSN-p is missing in the QOAM dataset). Another important remark is that the status of the journal as Hybrid is established by QOAM as recent as possible. As many subscription journals in the past have switched to the Hybrid model and this kind of switching still continues in the more recent years, the number of (purely) toll-access journals is diminishing rapidly. For the retrospective and prospective modelling, this variable will be of lesser importance. The number of articles by France-based corresponding authors in articles covered by Couperin contracts – and thus responsive to potential Read & Publish agreements – will be more important for the modelling exercise.

Table 11: Number of articles, and percentage of articles over the total number of journal articles in the dataset, per OA color (Diamond, Gold, Hybrid) and per year

		2013	2014	2015	2016	2017	2018	2019	2020	total
1	Nr. of articles published in Diamond journals (oa_color_article_BSO)	4363	4805	5196	5658	6298	6452	6336	4075	**43183**
2	Nr. of articles OA at site publisher (BSO;has apc=1) and APC paid (OpenAlex; Gold)	8791	10218	12156	13578	15661	17094	20246	26011	**123755**
3	Nr. of articles published in Hybrid journals (journal_color_qoam)	71413	71317	71880	77932	76986	79569	79003	74077	**602177**
4	Nr. of articles with missing 'journal_color_qoam' field	45986	47888	48349	51390	53962	56937	58324	67504	**430340**
5	subtotal as a check (QOAM Hybrid for all articles not in 1 or 2)	117399	119205	120229	129322	130948	136506	137327	141581	**1032517**
1	% articles published in Diamond journals (oa_color_article_BSO)	4%	4%	4%	4%	5%	5%	5%	3%	**4%**
2	% articles OA at site publisher (BSO;has apc=1) and APC paid (OpenAlex; Gold)	7%	9%	10%	10%	12%	13%	15%	18%	**12%**
3	% articles published in Hybrid journals (journal_color_qoam)	61%	60%	60%	61%	59%	59%	58%	60%	**60%**

4	% articles with missing 'journal_color_qoam' field	39%	40%	40%	40%	41%	42%	43%	48%	**42%**

NUMBER OF ARTICLES PER DISCIPLINE

The distribution of articles per discipline and per year is presented in Table 12, as per the BSO.

Table 12: Number of articles, and percentage of articles over the total number of journal articles in the dataset, per discipline and per year

Disciplines	2013	2014	2015	2016	2017	2018	2019	2020	total
Biology (fond.)	17588	18141	17610	19595	19605	20331	20870	23047	**156787**
Chemistry	6712	6857	6919	7274	6731	6645	6518	7108	**54764**
Computer and information sciences	4339	4351	4725	4755	5256	5452	5449	5755	**40082**
Earth, Ecology, Energy, and applied biology	8624	8952	9019	9589	10295	11080	11256	12094	**80909**
Engineering	3406	3252	3080	3372	3513	3658	3572	3738	**27591**
Humanities	10536	10488	10462	11415	11558	11847	11115	8246	**85667**
Mathematics	2888	2910	3064	3280	3221	3487	3592	3393	**25835**
Medical research	40340	41727	42377	45792	45892	49168	50676	55056	**371028**
Physical sciences, Astronomy	15072	14576	14789	15845	16088	15704	15476	15723	**123273**
Social sciences	7894	7950	8183	8403	8787	9128	8800	7419	**66564**
unknown	0	1	1	2	2	6	3	2	**17**
Subtotal	**117399**	**119205**	**120229**	**129322**	**130948**	**136506**	**137327**	**141581**	**1032517**
% Biology (fond.)	15%	15%	15%	15%	15%	15%	15%	16%	**15%**
% Chemistry	6%	6%	6%	6%	5%	5%	5%	5%	**5%**
% Computer and information sciences	4%	4%	4%	4%	4%	4%	4%	4%	**4%**
% Earth, Ecology, Energy, and applied biology	7%	8%	8%	7%	8%	8%	8%	9%	**8%**
% Engineering	3%	3%	3%	3%	3%	3%	3%	3%	**3%**
% Humanities	9%	9%	9%	9%	9%	9%	8%	6%	**8%**
% Mathematics	2%	2%	3%	3%	2%	3%	3%	2%	**3%**
% Medical research	34%	35%	35%	35%	35%	36%	37%	39%	**36%**
% Physical sciences, Astronomy	13%	12%	12%	12%	12%	12%	11%	11%	**12%**
% Social sciences	7%	7%	7%	6%	7%	7%	6%	5%	**6%**
% HSS	16%	15%	16%	15%	16%	15%	15%	11%	**15%**
% STM	84%	85%	84%	85%	84%	85%	85%	89%	**85%**

NUMBER OF ARTICLES PER PUBLISHER TIER

The distribution of articles per publisher tier and per year is presented in Table 13.

Table 13: Number of articles, and percentage of articles over the total number of journal articles in the dataset, per publisher tier and per year

	2013	2014	2015	2016	2017	2018	2019	2020	total
Tier 1	36525	37139	36273	38749	39044	40340	38733	39521	**306324**
Tier 2	9632	10011	22290	23661	25195	26229	28779	35736	**181533**
Tier 3	40501	39734	40303	43071	42956	43419	43186	40313	**333483**
Tier 4	30741	32321	21363	23841	23753	26518	26629	26011	**211177**
subtotal	117399	119205	120229	129322	130948	136506	137327	141581	**1032517**
% Tier 1	31%	31%	30%	30%	30%	30%	28%	28%	**30%**
% Tier 2	8%	8%	19%	18%	19%	19%	21%	25%	**18%**
% Tier 3	34%	33%	34%	33%	33%	32%	31%	28%	**32%**
% Tier 4	26%	27%	18%	18%	18%	19%	19%	18%	**20%**

NUMBER OF ARTICLES IN JOURNALS COVERED BY COUPERIN CONTRACTS

The annual distribution of articles published in journals covered by 2020 Couperin contracts is presented in Table 14. As explained earlier, we matched the journal names and ISSN numbers of the journals in the article-level dataset with the journal names and ISSN numbers of all journals covered by Couperin contracts. Of course, institutions can subscribe to electronic resources that are not covered by Couperin contracts; however, Couperin contracts are those which are concerned by Read & Publish agreements.

Before 2020, Couperin had only a few Read & Publish agreements, but after 2020 there are a number of these (see Table 3). The impact of this trend depends on the number of articles shown in Table 14.

Table 14: Number of articles in journals covered by Couperin contracts, and percentage of articles over the total number of journal articles in the dataset, per year

	2013	2014	2015	2016	2017	2018	2019	2020	total
Nr. of articles in journals covered by Couperin contracts	73340	73524	72855	76537	76708	77044	74879	74195	**599082**
% articles in journals covered by Couperin contracts	62%	62%	61%	59%	59%	56%	55%	52%	**58%**

III. Building the subscription expenditures dataset

COVERAGE OF THE DATASET

Couperin sends out an annual survey to its members asking to register all the costs for electronic resources they subscribe to. Indeed, while the contracts are negotiated centrally, the subscription and payment are up to each organization. Also, institutions can subscribe to electronic resources with publishers that are not covered by Couperin contracts.

The 2019 and 2020 ERE survey data consists of:

- 283 institutions who responded to the surveys, out of 288 institutes participating in Couperin (as of March 2022); 189 institutions responded in 2019 and 2020
- 1 259 unique product names, the costs in euros paid by the institutions (without VAT) and an indication if the product was part of a Couperin contract
- 644 unique publisher names linked to the products.

Couperin has shared the data from the 2019 and 2020 ERE surveys for this study. The results of the earlier surveys (2014-2018) are available on the national open data portal data.gouv.fr. However, these data have a different set-up and do not list the publishers: they were not included in the dataset.

The 2019 and 2020 data have several limitations:

- the institutions do not always have a comprehensive vision on all electronic resources subscribed by their organizational units. A number of these institutions have only the data on the centrally administered resources while certain organizational units belonging to the institutions might have subscription to other electronic resources
- the definition of the costs for electronic resources regarding journals might differ from institution to institution. Historically, institutions had subscriptions to printed journals. The additional costs for the electronic version of these journals are in some cases administered as just that – additional costs – while the larger part of the costs is administered as costs for the printed versions. This might result in two caveats in the data :
  1. some institutions might see all costs for journals as costs for the electronic versions, others might split the cost and report only the part for the electronic versions in the survey
  2. if an institution cancels all printed journals, the costs for the electronic journals suddenly increases in a year-on-year comparison as the entire cost for the journal package will be seen as costs for electronic resources
- the distinction in the electronic resources between e-books, e-journals, and databases can be difficult, as there are content packages by publishers that give access to a mix of journal articles, books, and other types of publications
- the year of invoicing (as suggested by the survey) might also differ from the subscription year of the electronic resource
- the acquisitions in the framework of the Istex archive are not part of the survey.

DATA ENRICHMENT

Product categories: of the 1259 products listed by the respondents in the 2020 survey, all products with a total cost of over 100 000 € were categorized in order to focus on the journal packages only. This led to the categorization of 200 products; 1 059 products were not categorized, with journal packages among those products (we will introduce a correction for this).

Publisher tiers of journal packages: using the above-mentioned product categories linked to the survey results, the 58 publishers of journal packages were identified. Each publisher was categorized as belonging to Tier 1, Tier 2, Tier 3, or Tier 4 (same tiers as for the article-level dataset).

Organization names: Couperin shared a list of their members of Couperin (as of March 2022) and whether they responded to the surveys of 2019 and 2020. This reference list was used to clean names for 189 organizations. A few dozen of responding organizations could not be linked to the reference list due to name changes, differences in spelling or mergers.

IV. Analysis of the subscription expenditures dataset

The ERE data for 2019 and 2020 and their enrichment, in Excel spreadsheets format, were loaded into Microsoft Power BI for a preliminary analysis, which focuses on the 2020 survey results (the most recent). The 2019 survey results are only used when comparing the two years for the same respondents.

The interactive dashboard[2] consists of 5 screens:

1. Publishers and products
2. Analysis per publisher
3. Analysis per Publisher Tier
4. Analysis per product category
5. Statistics

OVERALL EXPENDITURE ON E-JOURNALS IN 2020

The total expenditure on e-journal packages by Couperin members is an estimated 87,5 M€. This estimate is based on the following (see also Table 15):

- **116 M€ in total**: the total expenditure on electronic resources (e-books, e-journals and databases) by the institutions that have responded to the 2020 survey is 116 M€ (the exact figure is 115 986 749 €).
- **73,4 M€ for e-journals**: 63,26 % of the total amount is spent on categorized electronic journal packages: 73,6 M€ (the exact figure is 73 375 529 €). This relates to 90 categorized journal packages.
- **79,5 M€ for e-journals including the non-categorized e-journal products:** as explained in the methodology, 1 059 products with a total amount of < 100 K euro were not categorized. These products amount to 7,77% of the total costs of the 116 million € (9 014 121 €). For all categorized products, the total expenditure is 106 972 628 €. The proportion of journal packages of all categorized products is 68,6 %. Assuming that the same proportion applies to e-journals in the non-categorized products, then we should add 6 183 039 € to the total expenditures in 2020 on e-journals, thus 79 558 568 €.

[2] https://app.powerbi.com/view?r=eyJrIjoiN2NiYzUxYjItOTdjYy00YzU3LTk1NDQtMDkwNzNjOGQxMDBlIiwidCI6ImIjMmQyOWM3LTUxZGYtNDNlZS1iNDFlLWEwOWI1MjcyMjM5NyIsImMiOjl9&pageName=ReportSection

- **ca. 87,5 M€ is the best estimate of the total expenditure on journal packages by all Couperin members:** Couperin estimates (in its internal document *Descriptif de l'enquête ERE et précautions de lecture des résultats*) that the participants in the survey cover 90% of the total expenses of all Couperin members. The non-responders would in that case have spent 7,9 M€. This would lead to an estimate of 87 514 424 €, to be rounded off to *circa* 87,5 M€. However, it is worthwhile to repeat the most important uncertainties and possible errors of this best estimate: (1) responding institutions might not have a complete overview of the expenses for electronic resources, (2) the above-mentioned extrapolations (from the categorized products to all products; from respondents to the survey to all Couperin members) might have introduced errors.

Table 15: Journal expenditures based on the 2020 results of the ERE survey

2020 expenditure for all e-resources by responding institutions	115 986 749 €
2020 expenditure e-journal packages categorized by responding institutions	73 375 529 €
2020 expenditure for all non-categorized products	9 014 121 €
% non-categorized products of total expenditure	7,77 %
Total 2020 expenditure for all categorized products	106 972 628 €
% expenditure of journal packages of all categorized products	68,6 %
Estimated 2020 expenditure for non-categorized journal packages	6 183 039 €
Estimation total 2020 expenditure E journal packages by responding institutions	79 558 568 €
Estimation the total expenditure by all Couperin member institutions (+10%)	87 514 424 €

ANNUAL PRICE INCREASE

A separate analysis by Couperin using the ERE-survey data from 2014 to 2021 showed that the annual price increases during this period was between -1,95% and +7,22%, with an average price increase of 1,76% per year. This figure – in combination with the above estimation on the total expenditure on e-journals in 2020 – is used for the prospective analysis of the subscription expenditures (see Section 06.I on page 42).

DOMINANT ROLE OF COUPERIN AND DISTRIBUTION OF ORGANIZATIONS

The expenditure on journal packages that are negotiated by Couperin is estimated to be 74,5 % of the total expenditure on journal packages in France, under the assumption that Couperin is not involved in the small journal packages (expenditures under 100 000 €).

18 organizations cover 55 % of the expenditures on journal packages: 18 institutions have spent more than 1 million € in 2020 on (categorized) journal packages. Their total expenditures account for 55% of all expenditures by French organizations on (categorized) journal packages.

PUBLISHERS

The subscription expenditures on categorized journal packages for the four publisher tiers are presented in Table 16. In the last column, the percentage of articles by French authors published by the publisher is added for comparison. Tier 1 and tier 2 (4 publishers in total) account for 69 % of the 2020 expenditures and tier 3 (16 publishers) and tier 4 account for 31 % of the 2020 expenditures of the categorized journal packages.

The comparison with the number of articles by French authors published leads to the following observations: tier 1 takes up 45 % of the expenditures, while publishing a lower percentage (28 %) of the articles by French authors. In contrast, tier 2, tier 3, and tier 4 publishers publish similar proportions of articles by French authors in comparison to their proportions of the subscription expenditures. There are 40 publishers in tier 4 with journal packages over 100 000 €, and 626 tier 4 publishers overall in the ERE survey data (this includes all products, thus also books or databases), while the article-level dataset lists 1995 journal publisher names for Tier 4. These differences are probably partly caused by different publisher imprints of journals from the same publisher, publishers of OA journals (so no subscription needed) but may also reflect that French authors publish in journals that are not subscribed to by their institutions.

Table 16: Subscription expenditures and proportion of articles by French authors per publisher tier

Publisher Tiers	Expenditure	% expenditure on subscriptions	% articles by French authors
Tier 1	36 130 117 €	45,41%	28,91%
Tier 2	14 769 774 €	18,56%	25,24%
Tier 3	14 848 374 €	18,66%	28,48%
Tier 4	13 810 304 €	17,36%	18,37%
total (responding institutions)	**79 558 569 €**		

V. Scoping the analysis

APC costs: the country of the corresponding author is established for 88,2% of the article data. The articles for which an APC has been paid have been established as well. For practically all articles with APC, an APC price is available. In addition, the dataset contains information per article on the discipline, the OA color, the publisher tier of the journal and whether the journal is covered by Couperin contracts. From this article-level dataset will be carried out the retrospective and prospective analysis of the APC costs.

Subscription expenditures: using the dataset from the ERE survey of 2019 and 2020, we analyzed in Microsoft Power Bi the total cost of the subscriptions for journal packages and reported the results and the ensuing estimates for the total cost of e-journal packages. The lack of suitable data of the ERE survey for the years 2015-2018 means that the analysis can only be performed for 2019 and 2020. The annual price increase for e-journal packages that has been determined on the basis of the results of those two years is unexpectedly high and therefore seen as insufficiently trustworthy to use for a prospective analysis.

Coupling of the article-level dataset and the subscription expenditures dataset: BACON, developed by ABES in order to support the management of electronic resources by French libraries, offers data in the KBart format for all content packages from publishers. We tried to link unambiguously the products in the ERE survey and the BACON files which contains ISSNS, and link them again to the article-level dataset. Unfortunately, an ISSN of a journal can be linked to multiple BACON files and a Couperin contract consists of several packages which do not align to BACON files. Therefore, it was not possible to make a connection between the article-level dataset and the subscription costs/Couperin agreements.

Other publication charges: the data available for other publication charges are deemed insufficient for a cost analysis.

05 Retrospective analysis

As already explained in the methodological section, the retrospective analysis is based on known and trusted data:

- known data: we have established the country of the corresponding author for most of the data
- trusted data: we have used the APCs prices and estimates by BSO based on Open APC but excluded the estimations based on the entire publisher's portfolio (sources E and F in Table 2).

This section is set up as follows:

- analysis of the evolution of the number of articles with APC and France-based corresponding authors
- analysis of the evolution of APC prices
- analysis of the cost of diamond articles
- calculation of the total cost of APCs paid, including reconstructed data
- evaluation of the main contributing factors to the total cost of APCs.

I. Evolution of articles with APC and France-based corresponding authors

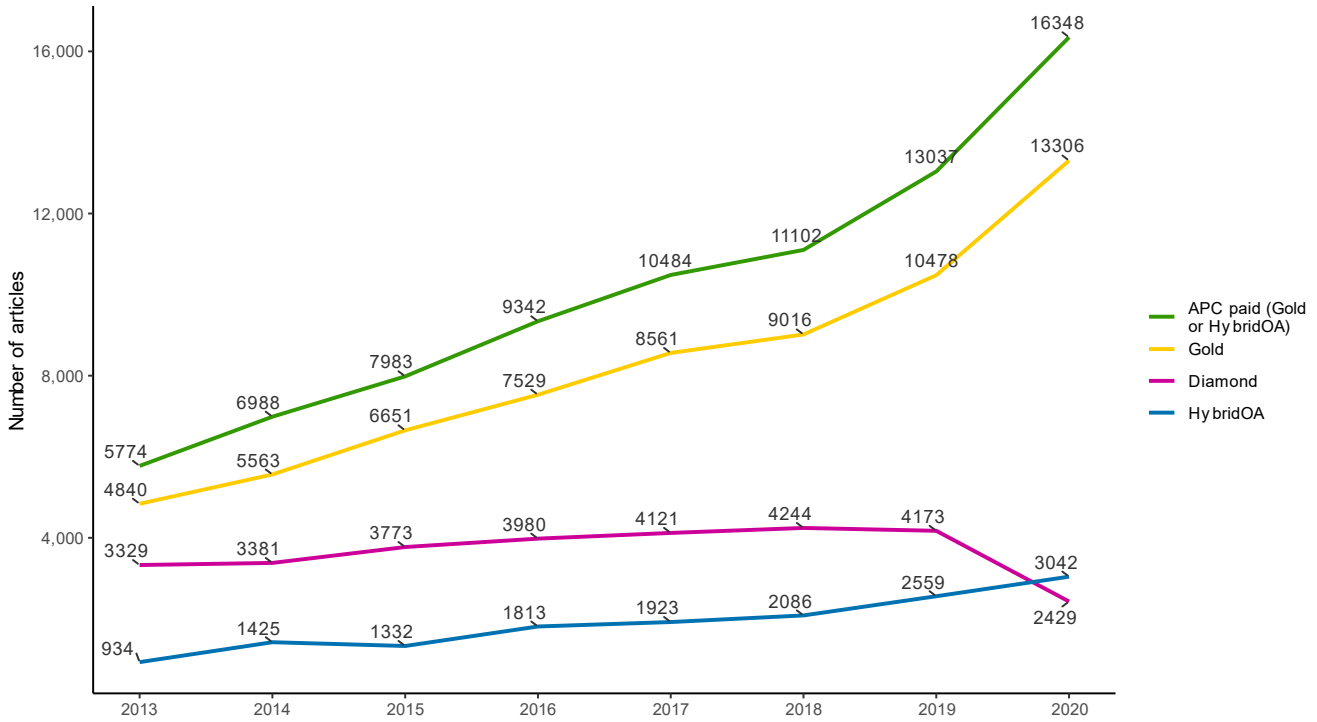


Figure 2: Evolution of OA articles at the site of the publisher by France-based corresponding authors, per OA color[3]

[3] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/LOJI53 and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4650

Figure 2 presents the evolution of OA articles at the site of the publisher by France-based corresponding authors, per OA color. It shows the following:

- the growth of articles with APC accelerates from 2018 onwards, as a result of mix of factors including OA policies, cultural change, international trends, and Read & Publish agreements
- this growth is mainly fueled by the growth of articles in Gold journals. However, it must be noted that Couperin has since 2020 concluded more and more Read & Publish agreements that probably will increase the growth rate of HybridOA articles by France-based corresponding authors in the years after 2020
- there is a slight annual growth in the number of Diamond articles by France-based corresponding authors in the years 2013-2018, while the figure for 2019 and especially for 2020 do seem to indicate a decrease. We cannot think of an explanation for this observation.

The following graphs leave out Diamond OA, to focus on articles with APC.

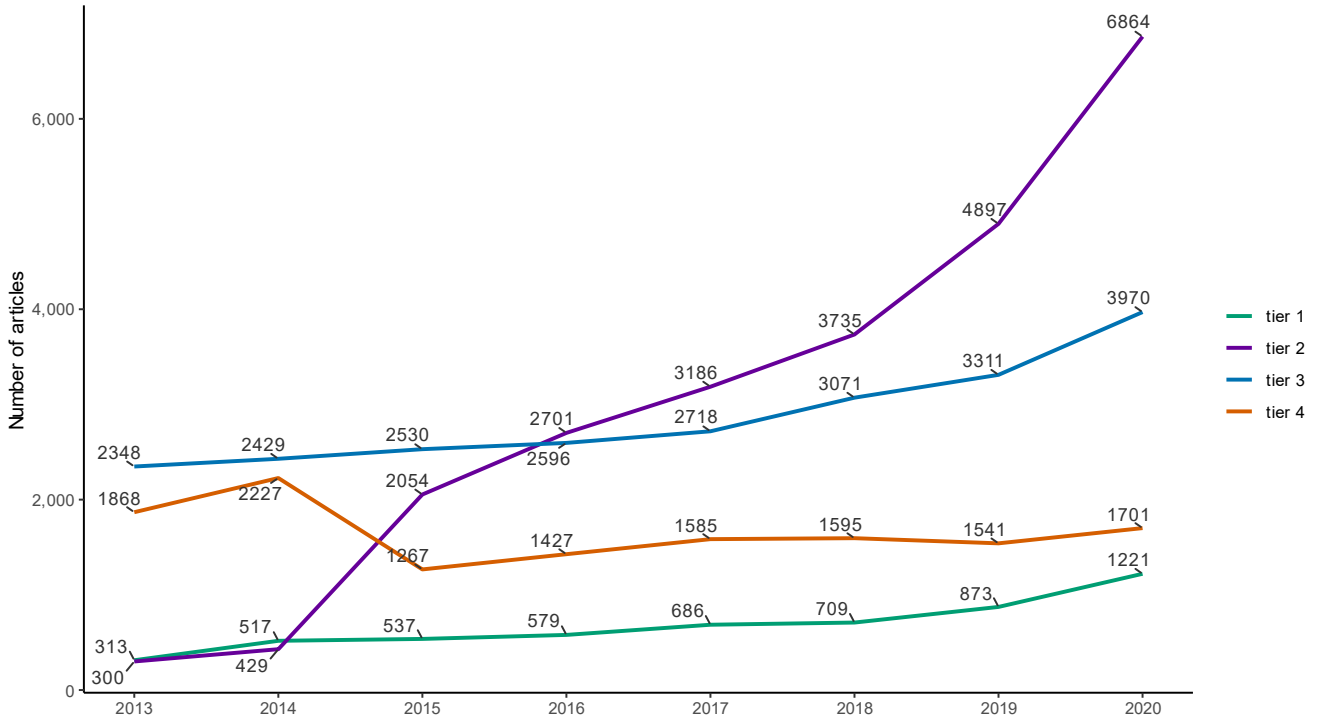


Figure 3: Evolution of articles with APC and France-based corresponding authors, per publisher tier[4]

Figure 3 presents the evolution of articles with APC and France-based corresponding authors, per publisher tier. It shows the following:

- a rapid growth of the number of APC-paid articles in journals published by tier 2 publishers. This includes Springer Nature, Wiley and MDPI
- a considerable growth of the number of APC-paid articles in journals published by tier 3 publishers
- a flat to limited growth rate of the number of APC-paid articles by tier 1 publisher (Elsevier) and by tier 4 publishers (the long tail of smaller publishers). The rise observed in 2020 for tier 1 articles could be explained by the Read & Publish agreement signed by Couperin with Elsevier.

[4] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/NMEPRW and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4779

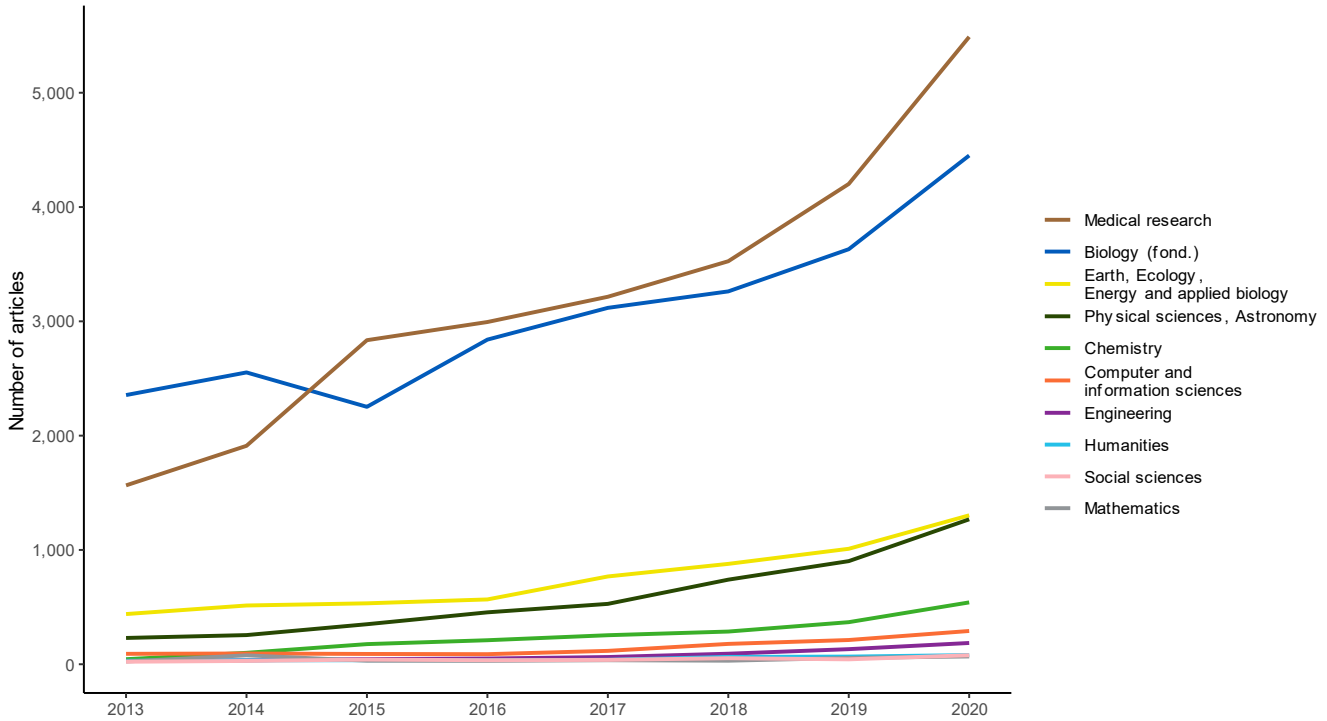


Figure 4: Evolution of articles with APC and France-based corresponding authors, per discipline[5]

Figure 4 presents the evolution of articles with APC and France-based corresponding authors, per discipline. It shows the following:

- two disciplines are dominating: Medical Research and Biology (fond.) (over 76 % of articles combined), while Earth, Ecology, Energy, and applied biology on one hand, and Physical sciences, Astronomy on the other hand, are intermediate (16 % combined)
- the annual growth in the number of these articles in Medicine and Biology is considerable; it is notable for the other two disciplines mentioned
- other disciplines account for dozens or hundreds of articles per year (7% combined).

[5] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/EBA2NK and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4690

II. Evolution of APC prices

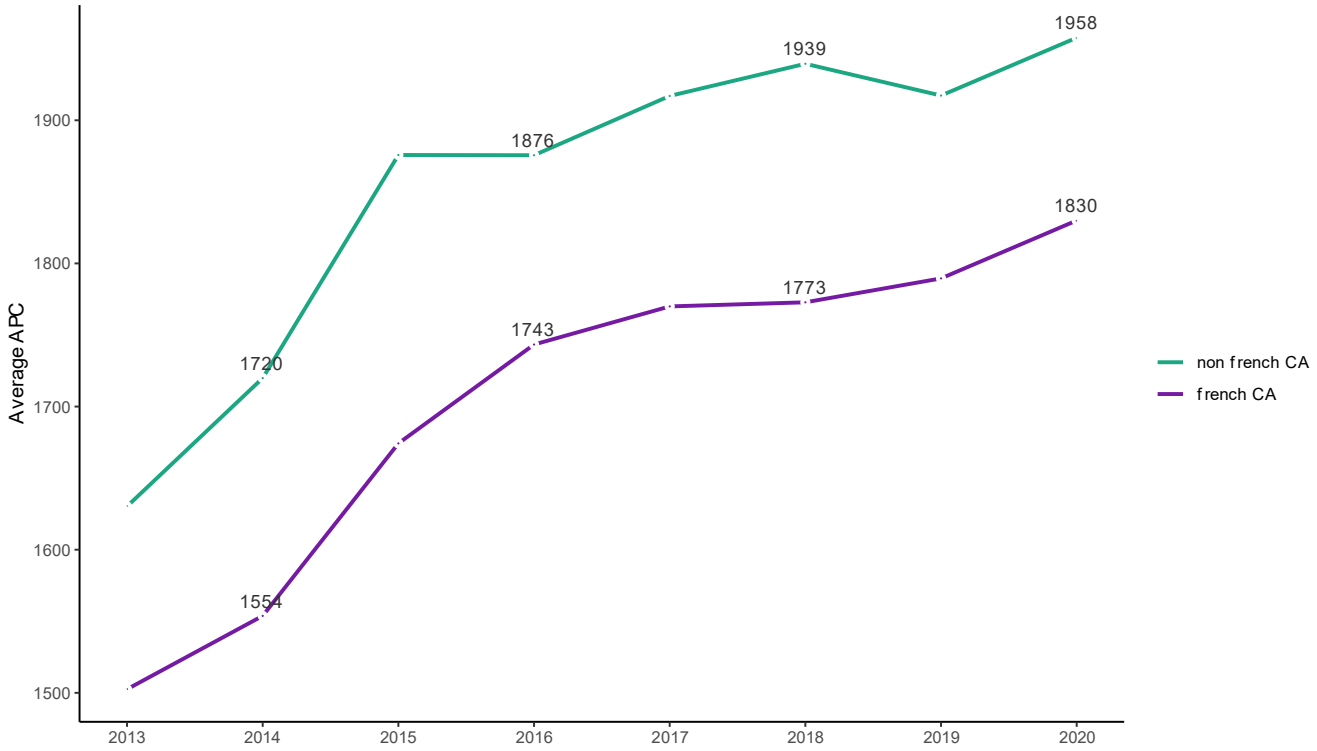


Figure 5: Evolution of average APC paid by France-based corresponding authors and non-France-based corresponding authors[6]

The dataset includes 81 386 articles with APC and France-based corresponding author and 67 130 articles with APC and non-France-based corresponding author (but at least one France-based co-author). Figure 5 presents the average APCs paid by France-based corresponding authors and non-France-based corresponding authors. It shows that France-based corresponding authors on average have paid lower APCs than non-France-based corresponding authors. In 2013, France-based corresponding authors paid on average 1 502 €, while non-France-based corresponding authors paid 1 630 €. In 2020, France-based corresponding authors paid 1 830 € and non-France-based corresponding authors 1 958 €.

We investigated this gap and found that it is most prominent for HybridOA articles, Elsevier articles, and Medical research articles.

[6] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/6SWJD6 and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L369

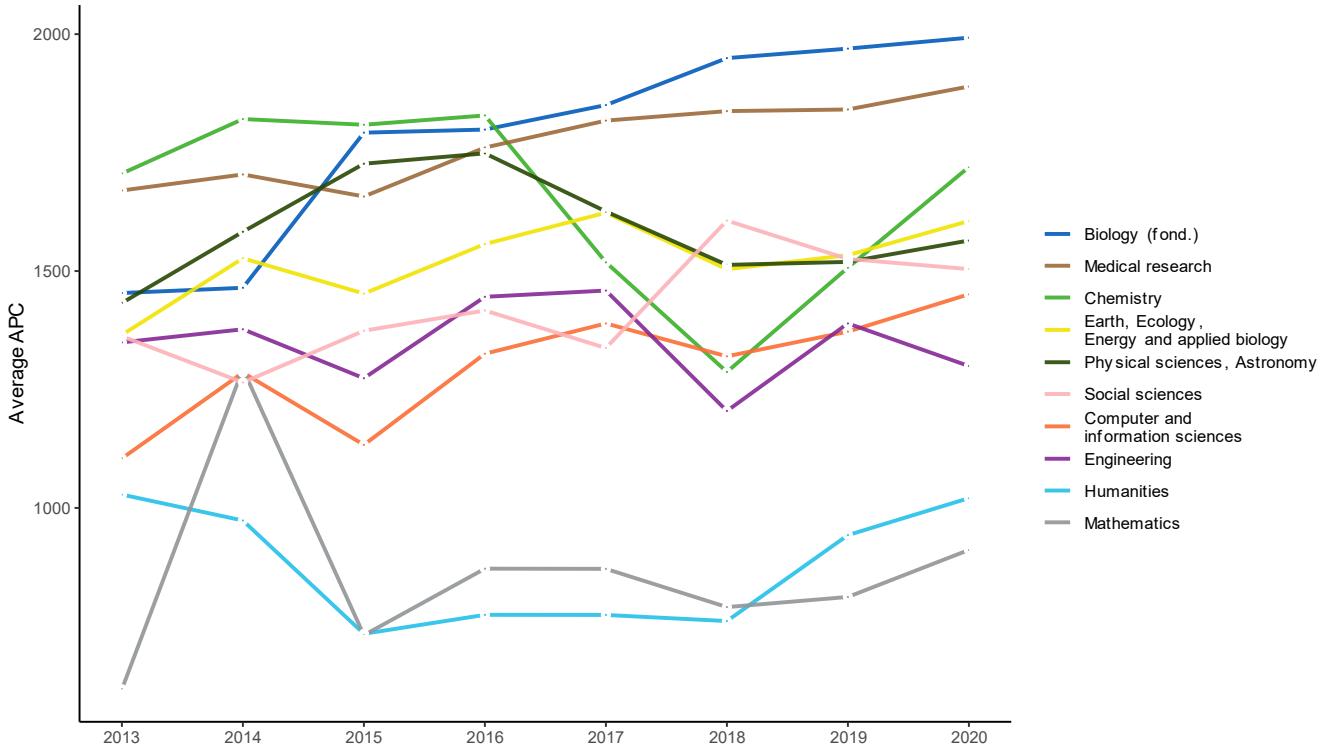


Figure 6: Evolution of average APC paid by France-based corresponding authors per discipline[7]

Figure 6 presents the evolution of the average APC paid by France-based corresponding authors per discipline. It shows the following:

- the wild fluctuations over the years in the APCs paid can be explained by sensitivity of small samples to extreme values
- however, some trends stand out:
    - articles in Mathematics and Humanities pay the lowest APCs, respectively 911 € and 1 020 € in 2020
    - articles in Computer & information sciences and Engineering pay intermediate APCs, respectively 1 451 € and 1 299 € in 2020
    - articles in the disciplines Biology (fond.), Medical research, Physical sciences & astronomy, Chemistry, and Earth, ecology, energy & applied biology have the highest APC levels: the average APC in 2020 for Biology was 1 993 €, for Medical research 1 890 €, for Chemistry 1 720 €, for Earth, ecology, energy & applied biology 1 606 €, for Physical sciences & astronomy 1 564 €.

[7] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/XUGNDA and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L1274

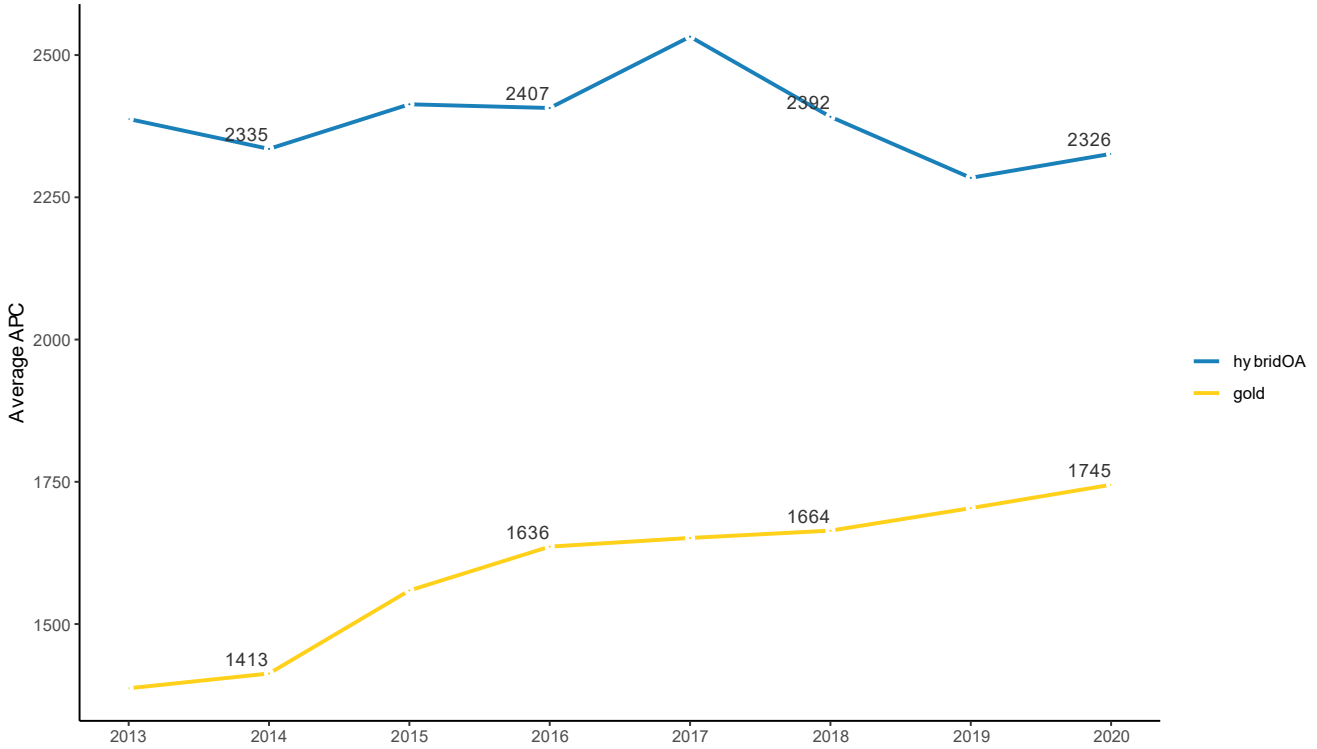


Figure 7: Evolution of average APC paid by France-based corresponding authors per OA color[8]

Figure 7 presents the evolution of the average APC paid by France-based corresponding authors for Gold and HybridOA articles. It shows the following:

- while APCs for articles in Gold journals are considerably lower, their annual increase is quite considerable. In 2013 the average APC for Gold OA article was 1 387 € and in 2020 1 745 €
- the APCs of HybridOA articles are at a much higher level but have been rather stable over the years: 2 388 € in 2013 and 2 326 € in 2020.

[8] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/FZUO5I and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L1332

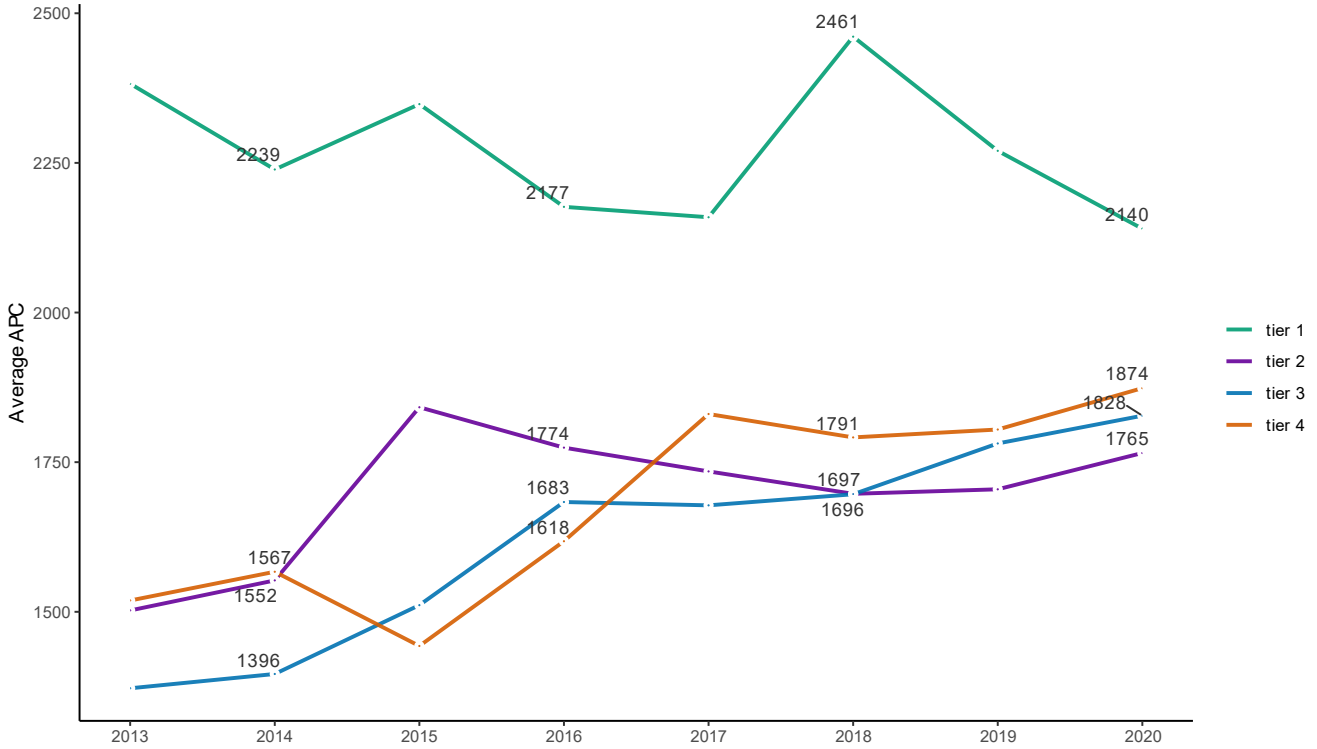


Figure 8: Evolution of average APC paid by France-based corresponding authors per publisher tier[9]

Figure 8 presents the evolution of the average APC paid by France-based corresponding authors per publisher tier. It shows the following:

- the average APC charged by tier 1 publisher (Elsevier) is the highest but is stable over the years. In 2020, the average APC was 2 140 €, while it was 2 382 € in 2013.
- the level of average APCs charged by tier 2, 3 and 4 publishers all follow a similar curve: they start much lower than tier 1 but increase considerably over the years. The average APC of tier 2 was in 2013 1 502 € and in 2020 1 765 €. The average APC of tier 3 started in 2013 at 1 372 € and was in 2020 1 828 €. The average APC of tier 4 publishers started in 2013 with 1 519 €, ending in 2020 with 1 874 €.

III. Total cost of APCs paid in 2013-2020

ESTIMATION OF DIAMOND ARTICLE PUBLICATION EXPENSE

The OA Diamond Journal Study (Bosman et al., 2021) has calculated a median expense per article published in Diamond journals covered by their worldwide survey: 208 €. However, as this is based on Diamond journals that are published over the world, we estimate using their dataset (Bosman et al., 2021) that the expense of publishing a diamond article in France-based journals (a subset of the dataset) is 413 € in average. The estimate is based on dividing the rounded total annual expense including in-kind institutional contributions (Q66) by the rounded average

[9] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/MQOZ6K and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L1302

number of published articles each year during the three last years (Q16), for all journal publishers based in France (Q14).
There is **an average of 3 678 Diamond articles published yearly by France-based corresponding authors, which would have cost 1,519 M€ if published by French journals** (of course not all diamond articles are published in France-based journals, we can think of it as an equivalent expense for the French national research budget).

TOTAL COST OF APCS BASED ON KNOWN AND TRUSTED DATA

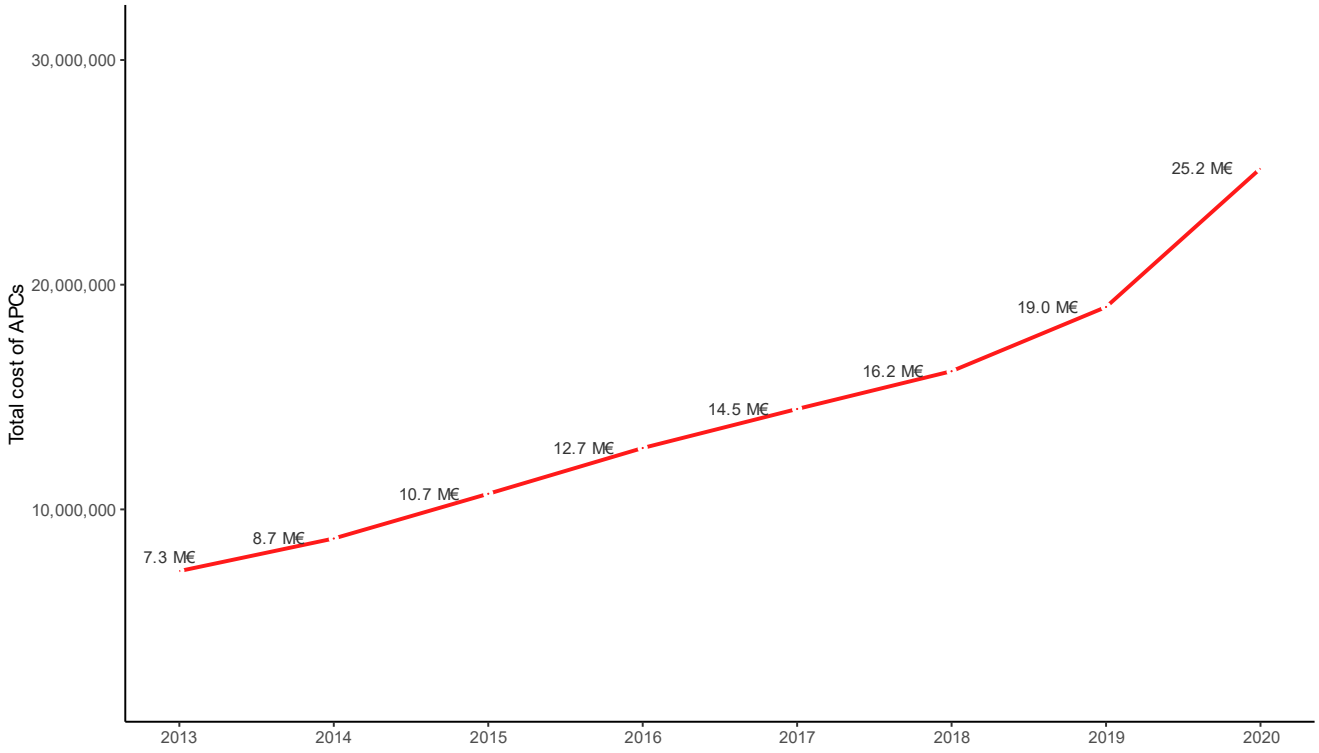


Figure 9: Evolution of total cost of APCs paid by France-based corresponding authors[10]

In Figure 2 we have seen a rapid annual growth in the number of articles with APC by France-based corresponding authors. In Figure 5 we have seen a rapid annual growth in the average level of APCs paid by France-based corresponding authors (albeit at a lower level than paid by non-France-based corresponding authors).
With these numbers, we were able to calculate the total cost of APCs paid 'in the wild' by France-based corresponding authors each year (see Figure 9). 'APCs in the wild' means outside subscriptions, a phrase borrowed from Monaghan et al. (2020).

This calculation is based on known and trusted variables in the dataset. It shows that the total sum of APCs paid by France-based corresponding authors is rapidly growing: it started in 2013 with 7,25 M€ and has grown about 1,8 M€ per year to 16,15 M€ in 2018. After 2018, the growth rate has accelerated at about 4,5 M€ per year to 19,0 M€ in 2019 and 25,17 M€ in 2020.

[10] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/NCPOAA and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L1903

TOTAL COST OF APCS BASED ON RECONSTRUCTED DATA

To move beyond these calculations based on known and trusted data in the dataset, we reconstructed the remaining data based on the following rules:

- correction factor 1: for articles where an APC has been paid and the price is missing or untrustworthy, but it is known that the corresponding author is based in France (15 273 articles), we estimated the APC price based on the model of the prospective analysis (see below).
- correction factor 2: for articles where an APC has been paid and the price is missing or untrustworthy, and the country of the corresponding author is unknown (17 587 articles), we estimated the proportion of France-based CAs per publisher tier and per year and estimated the APC price based on the model.

As per the result in Figure 10, we see that in 2020, French public institutions paid 30,1 M€ of APCs (78% of the amount was spent on Gold vs. 22% on HybridOA).

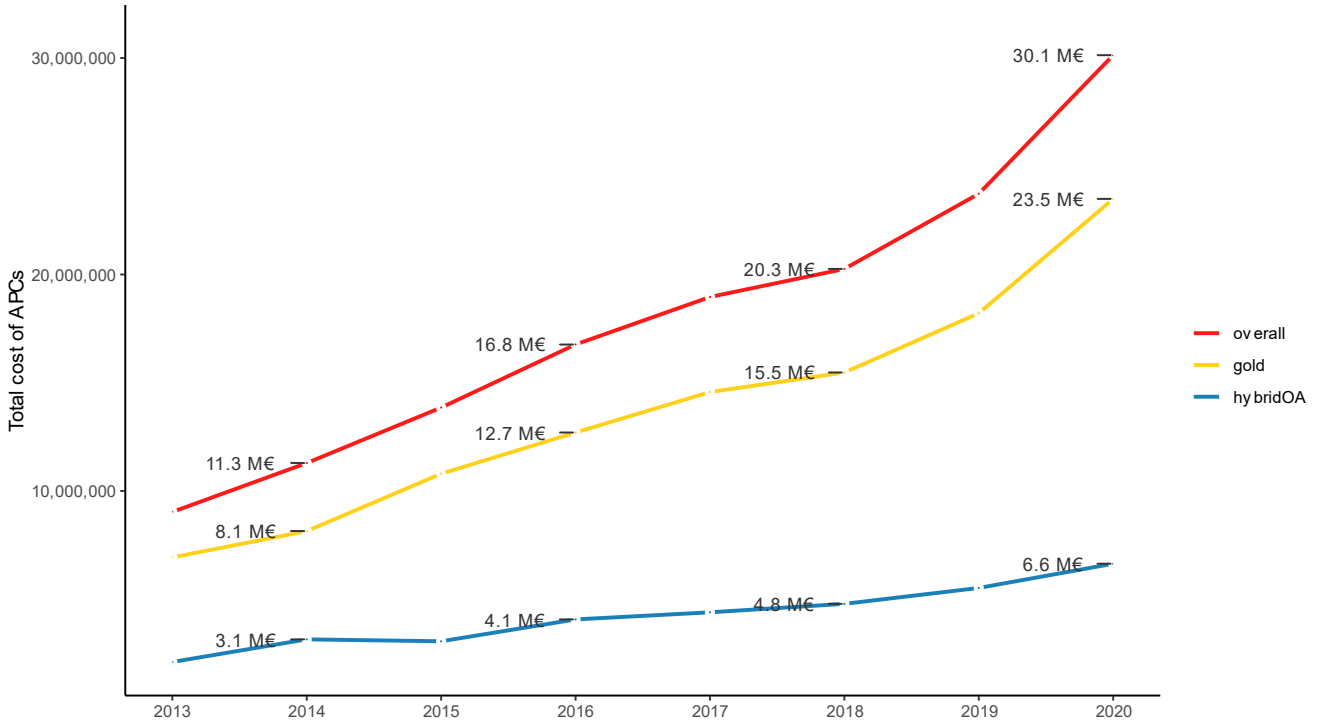


Figure 10: Evolution of total cost of APCs paid by France-based corresponding authors, overall and per OA color, after reconstructing missing data[11]

In conclusion, after reconstruction, the total cost of APCs for the period 2013-2020 is 144 M€.

In 2020, the total costs for reading and publishing journal articles by French public institutions is 117 M€:

- three quarters is spent on journal subscription costs (ca. 87,5 M€)
- one quarter is spent on APCs in the wild (30,1 M€)
- [an equivalent of 1,5 M€ spent on Diamond journals has been discarded]

[11] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/TLFDRS and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L1923

MAIN CONTRIBUTING FACTORS TO THE TOTAL COST OF APCS

The total cost of APCs is made of four variables that varied over the period:

- the number of Gold articles
- the price of Gold APC
- the number of HybridOA articles
- the price of HybridOA APC.

To assess the weight of each variable, we ran four counterfactual scenarios to understand how much France-based corresponding authors would have paid APCs in 2020 if one of the variables had not varied. If the number of Gold articles had been stable, the total cost of APCs paid by France-based corresponding authors in 2020 would have been 15,2 M€ instead of 30,1 M€. If the average price of Gold APC had been stable, the total cost of APCs in 2020 would have been 25,7 M€. If the number of HybridOA articles had been stable, the total cost of APCs in 2020 would have been 25,5 M€. If the average price of HybridOA APC had been stable (instead of decreasing), the total cost of APCs in 2020 would have reached 30,3 M€.

We see that the increasing number of Gold articles (shown in Figure 2) is the major driver of the growth: without this growth, the cost of APCs would have been multiplied by 1,69 instead of 3.

06 Prospective analysis

I. Evolution of subscription expenditures

We forecasted the evolution of subscription expenditures based on past data from Couperin, using a linear regression model. Of course, this prediction doesn't account for events that could change the course of the journal prices.

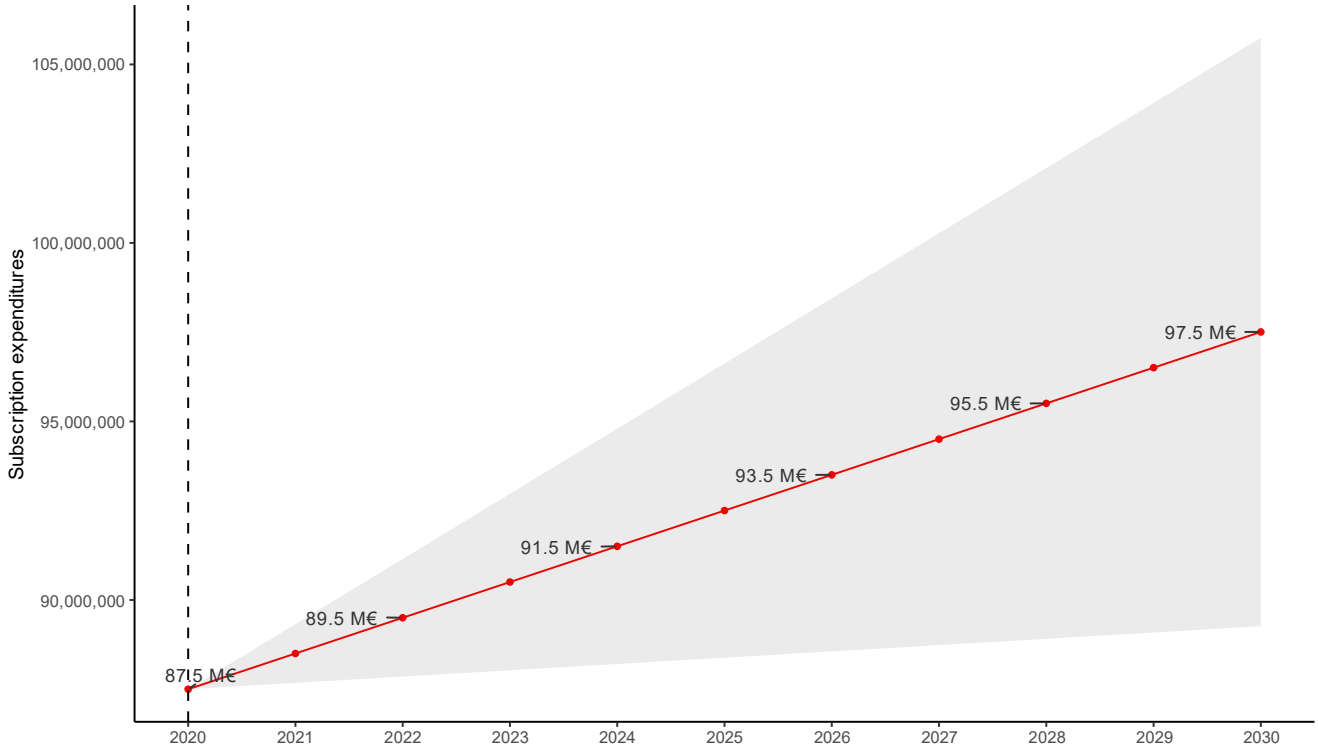


Figure 11: Simulation of subscription expenditures by French institutions (2021-2030)[12]

On Figure 11, the value for 2020 is the actual, observed value; the values from 2021 onwards are predicted.

The grey cone around the line shows the confidence interval, which increases as we attempt to predict further in time. The confidence interval is bound between the predicted value minus two standard deviations, and the predicted value plus two standard deviations. For instance, subscription expenditures are estimated at 97,5 M€ in 2030, with a probability of 95% that it is comprised between 89,3 M€ and 105,7 M€ (the minimum and maximum values of the confidence interval).

II. Building a model to predict APC prices

The basis for the prospective analysis is a model of the APC price (the dependent variable Y) as a function of several variables that might influence the price:

- year
- discipline
- publisher tier
- OA color
- France-based corresponding author or not
- journal covered by Couperin contract or not.

[12] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/MF2ZLS and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4326

The model was built on "trusted data" as explained in the methodological section on APC information. Indeed, the better the quality of the model input, the better its output. Reconstructed data was thus left aside.

We chose a modeling approach called "mixed-effects models", or multilevel models, or varying-intercept varying-slope models. The benefit of mixed-effect models is that they consider groups by combining a non-modeled effect ßX which does not change across groups and a modeled effect α which changes across groups. In mathematical notation, it follows that $Y \sim \alpha + ßX + \sigma$ where Y is the APC price, α is the intercept whose variance across different groups is modeled, ß is the slope, X is a unique independent variable, and σ is the residual error.

Trial and error showed that the best non-modeled effect ßX is the year: each year the APC price varies. We have then created different mixed-effects models looking for the largest source of variation once controlling for the year. All models performed well: Figure 12 shows how the green line (discipline as the source of variation), black line (publisher tier as the source of variation), blue line (journal covered by Couperin as the source of variation), and orange line (OA color as the source of variation) fit the red line of actual, observed data from Figure 9.

The best performing model (black line), which best fits the data, combines a non-modeled effect of the year with a modeled effect of the publisher tier. It means that each year the APC price varies, and it is between publisher tiers that the variation is the largest.

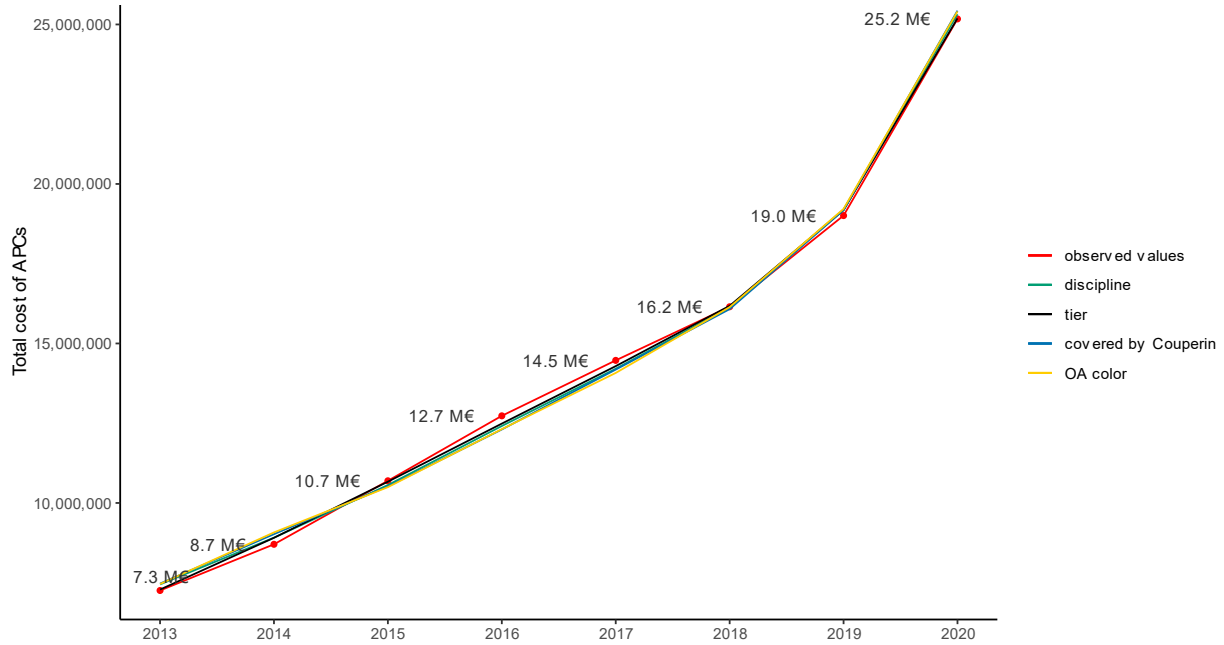


Figure 12: Observed and simulated total cost of APCs paid by France-based corresponding authors per year, per model (2021-2030)[13]

The model was used to forecast the possible evolution of total sum of APCs paid in different scenarios. By default, the best-fit model was used, except when the scenario played on other variables than the publisher tier. For example, when the scenario played on OA colors, we used the alternative model with a non-modeled effect of the year and a modeled effect of the OA color. Although not the best-fit model, it still performed very well.

[13] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/SGSRBQ and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L1634

III. Scenario "trends continue unchanged"

Using the best-fit model (per publisher tier) presented above, we first simulated how the average APC price will evolve in the future, based on its past year-on-year evolution.

Figure 13 shows the growth of the average APC (red line), driven by the dramatic growth of the average APC for tiers 3 and 4, not compensated by the decrease in tier 1. Of course, this scenario based on past trends doesn't account for events that could change the course of the APC prices.

The average APC price is estimated at 2 170 € in 2030, with a confidence interval comprised between 2 109 € and2 231 €.

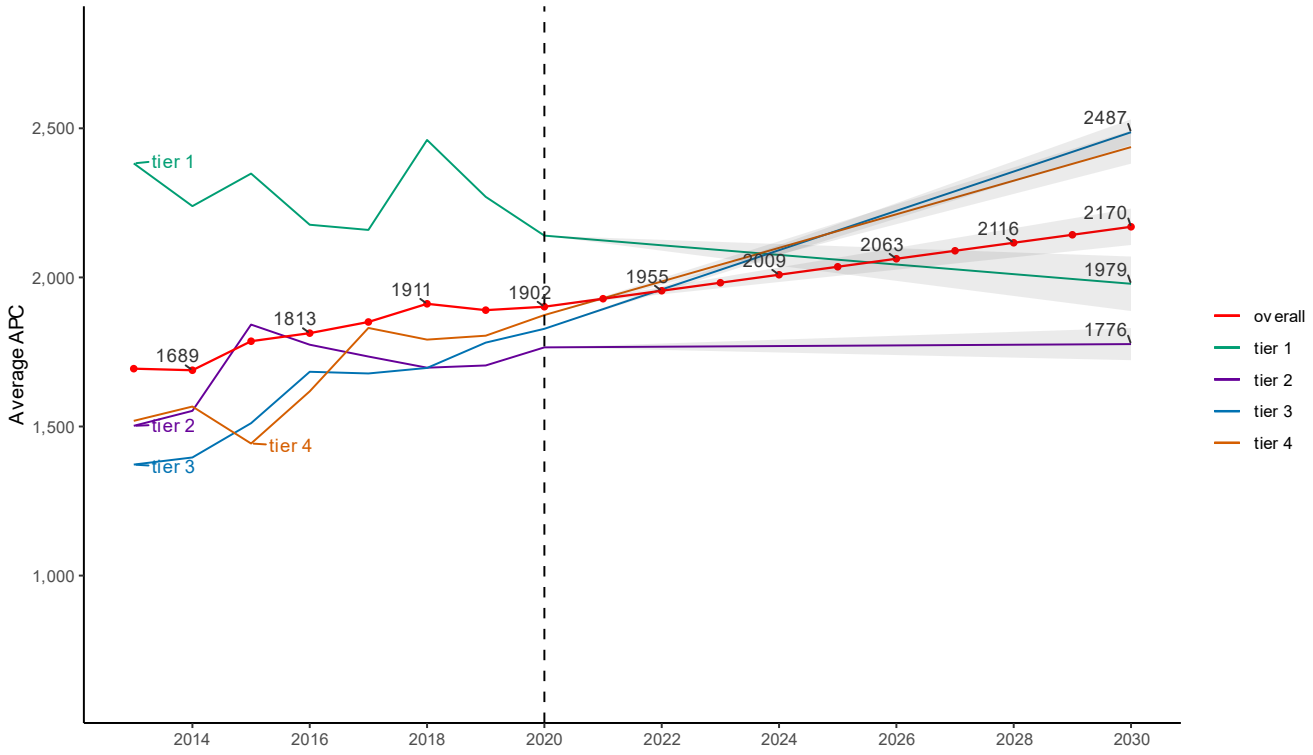


Figure 13: Simulation of average APC paid by France-based corresponding authors, overall and per tier (2021-2030)[14]

The average APC does not reflect the way APCs are distributed within tiers, which does not follow a normal, gaussian distribution. However, the property of the average is that it gives a total sum when multiplied by the number of observations, which is what we will do now.

[14] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/QB3ULH and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L2380

Thus, we estimate, using another mixed-effects model, how the number of articles with APC and France-based corresponding authors per tier will evolve in the future if trends continue unchanged. Figure 14 shows that the overall growth (red line) is driven by the growth of articles in tier 2 journals.

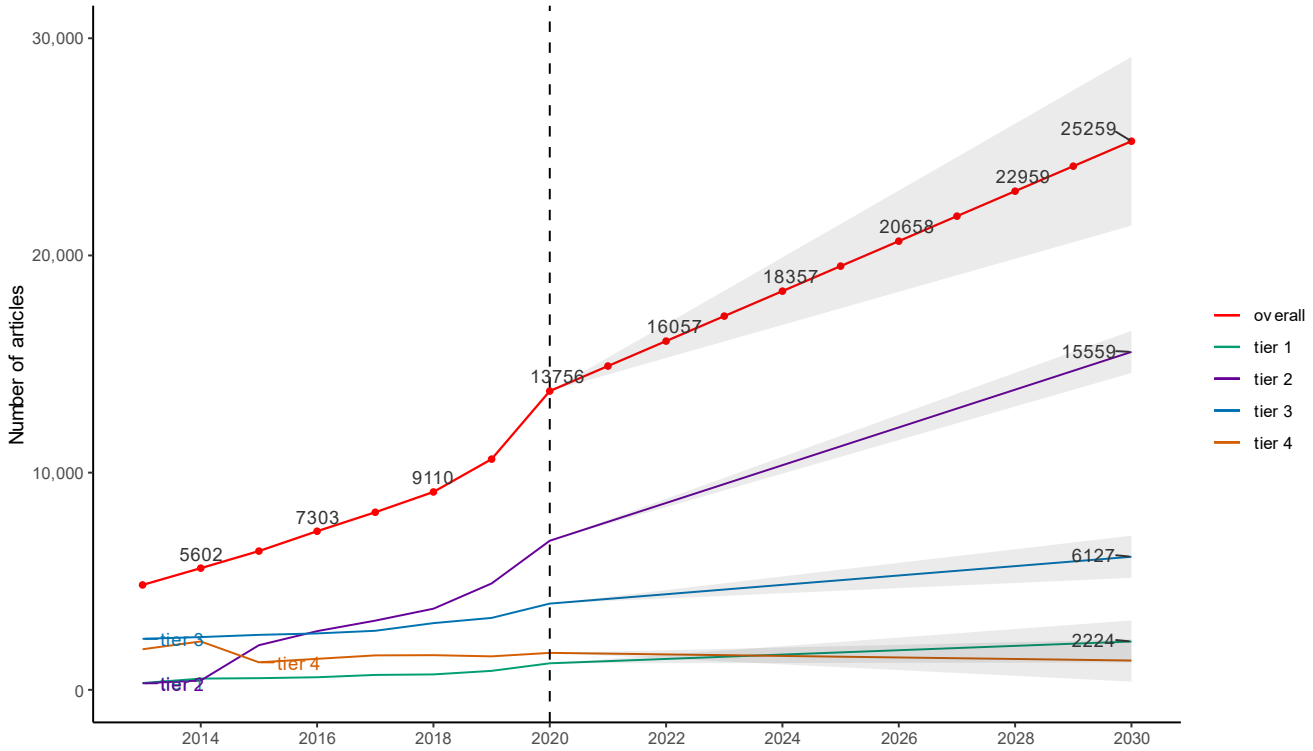


Figure 14: Simulation of the number of articles with APC and France-based corresponding authors, overall and per tier (2021-2030)[15]

[15] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/QB3ULH and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L2392

Finally, for each tier, we multiply the average APC price by the number of articles with APC and with France-based corresponding authors, and add them up. Figure 15 (red line) shows that in ten years, the estimated total cost of APCs would double from around 25 M€ in 2020 to around 50 M€ in 2030. The confidence interval (gray zone) is between 40 M€ and 60 M€ in 2030.

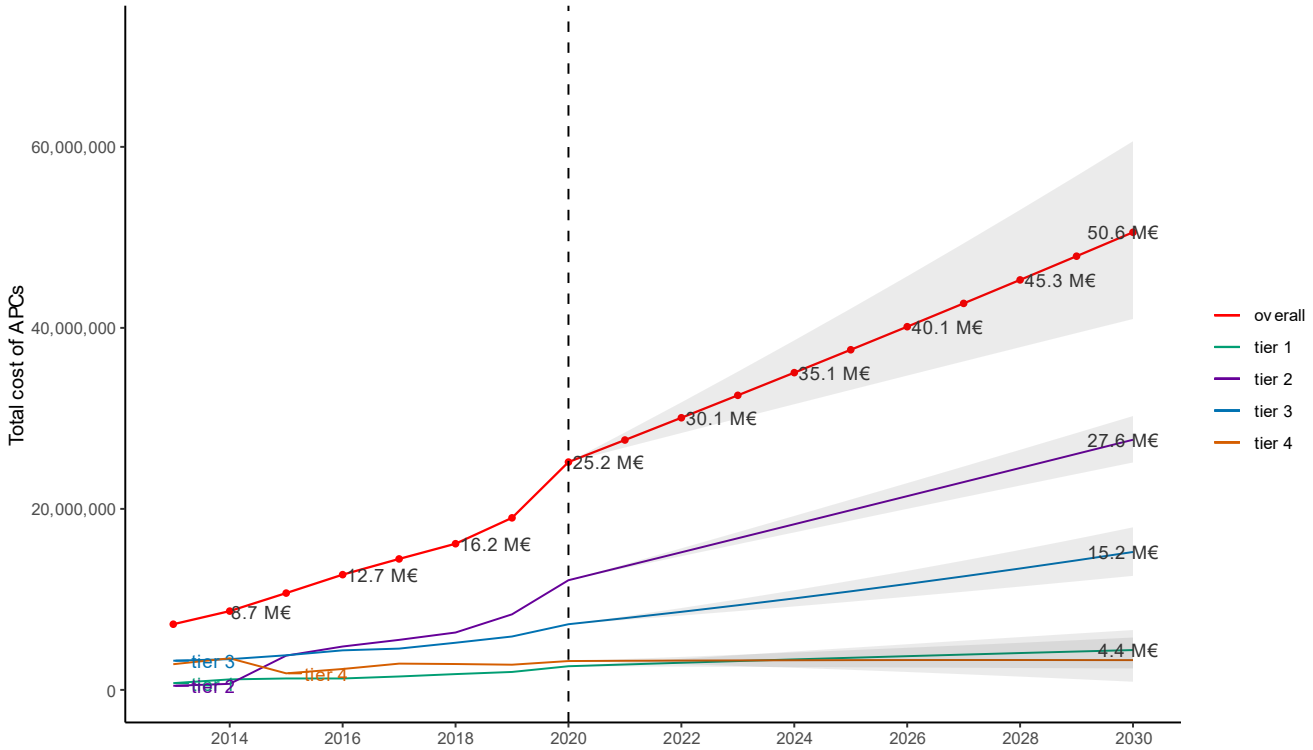


Figure 15: Simulation of the total cost of APCs paid by France-based corresponding authors, overall and per tier (2021-2030)[16]

[16] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/QB3ULH and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L2406

We can also show how OA colors contribute to the simulated trend. To do that, we need to use a different model (black line) than the best-fit model used so far and based on publisher tiers (red line). Figure 16 shows that Gold OA is the main driver to this trend.

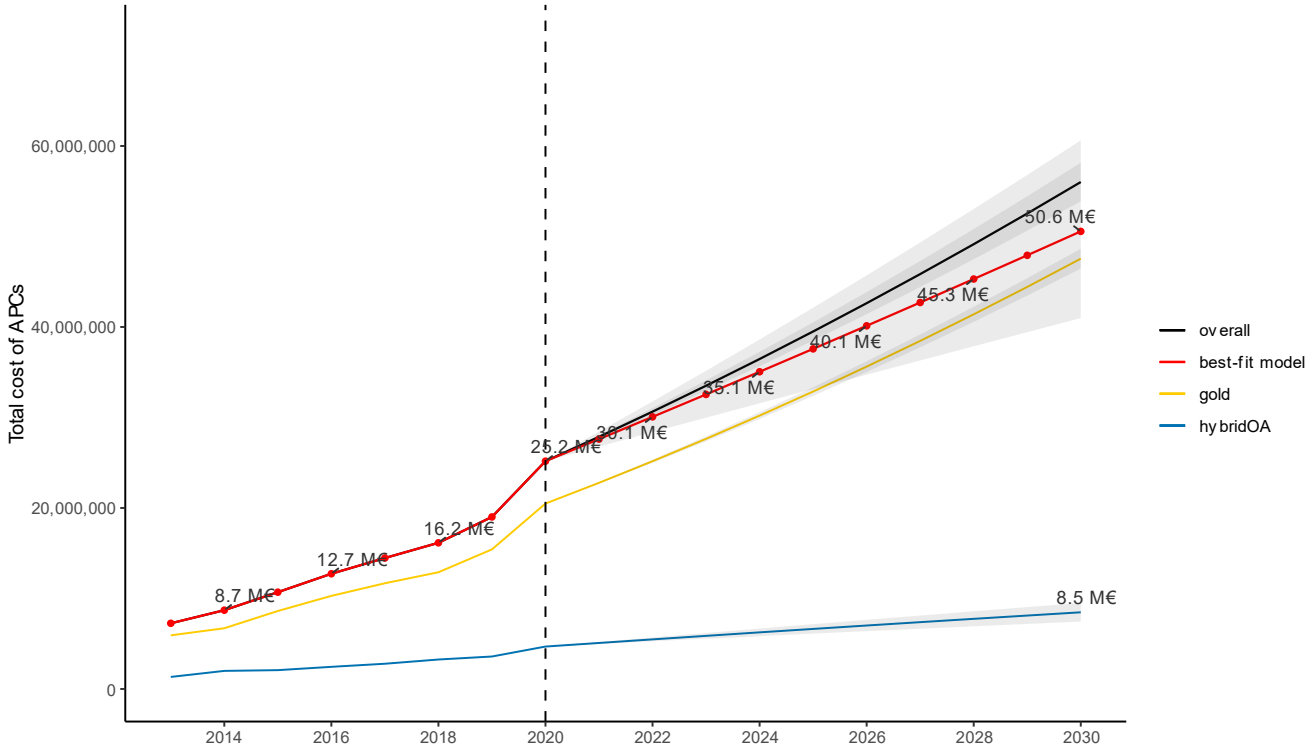


Figure 16: Simulation of the total cost of APCs paid by France-based corresponding authors, overall and per OA color (2021-2030)[17]

[17] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/AIWYQM and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L2593

The number of articles by France-based corresponding authors in articles covered by Couperin contracts in 2020 is responsive to potential Read & Publish agreements. We can also show how these articles contributes to the trend. To do that, we use a different model (black line) than the best-fit model based on publisher tiers (red line). Figure 17 shows that journals not covered by Couperin are the main driver to this trend. Please note that waivers and discounts negotiated by Couperin are not considered.

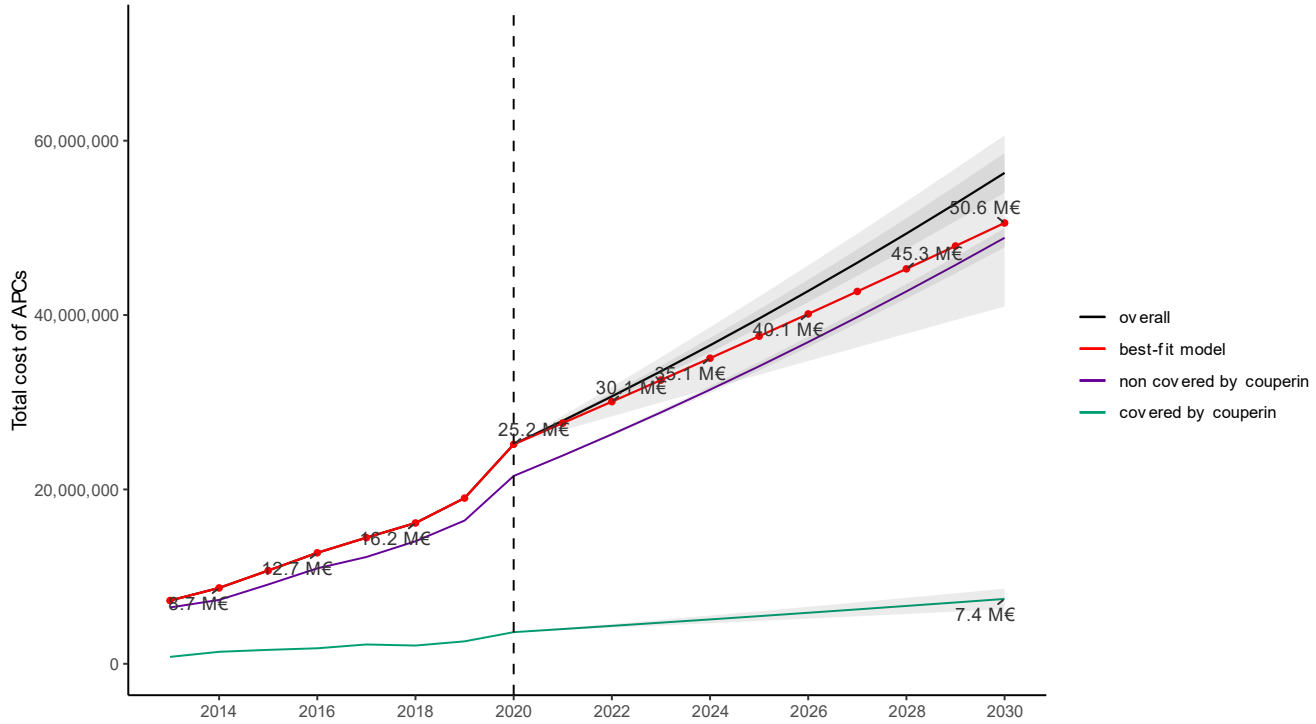


Figure 17: Simulation of the total cost of APCs paid by France-based corresponding authors, overall and for journals covered or not by Couperin contracts (2021-2030)[18]

[18] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/PBVSXJ and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L2744

We can also show how disciplines contribute to the trend. To do that, we use a different model (black line) than the best-fit model based on publisher tiers (red line). Figure 18 shows that articles in medicine and biology are the main drivers to this trend.

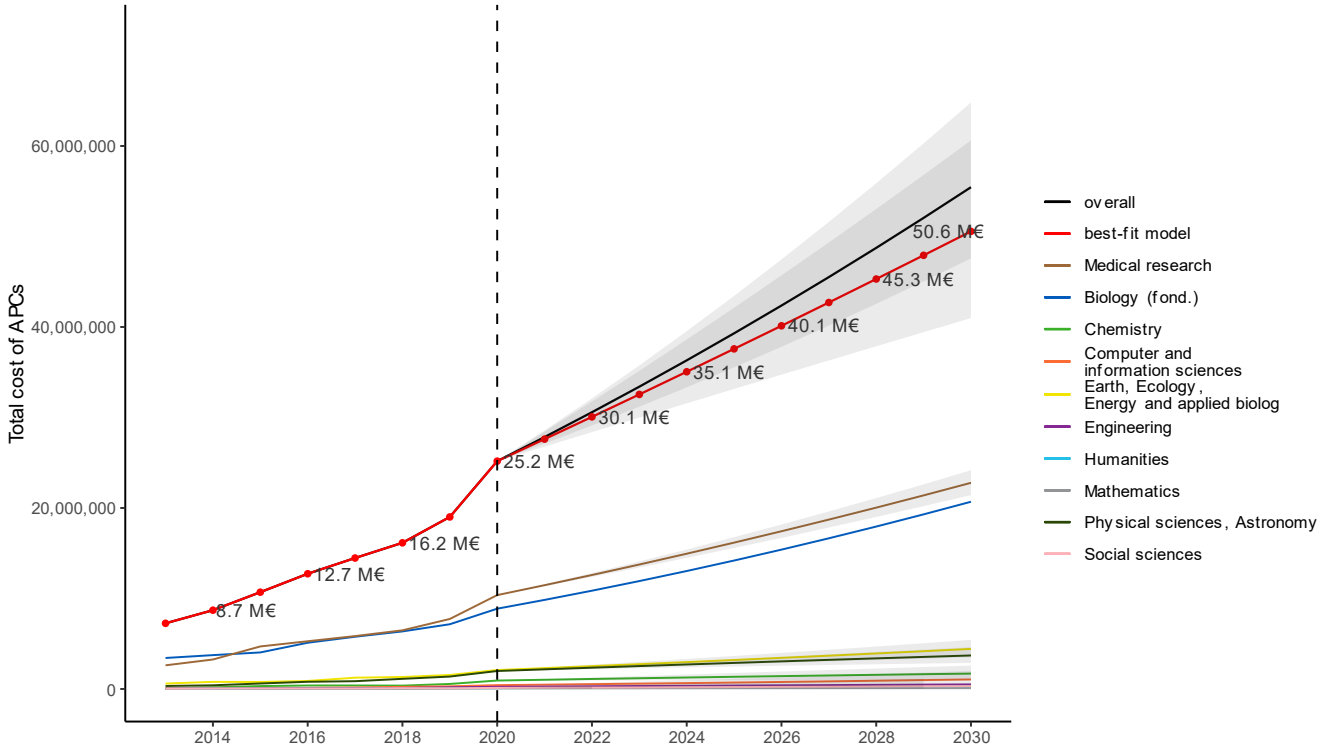


Figure 18: Simulation of the total cost of APCs paid by France-based corresponding authors, overall and per discipline (2021-2030)[19]

[19] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/LZ6KMZ and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L2950

IV. Simulation of theoretical full Gold APC

Assuming that read costs are down to zero, we wanted to simulate the maximum possible APC cost in a total Gold OA environment. In practical terms, we assume that all articles with a France-based CA pay an APC (except 10% of diamond articles). We simulate this theoretical "ceiling" from 2021 onwards.

As expected, this maximum APC cost (143,2 M€ in 2021 to 168,7 M€ in 2030, confidence interval: 125,2 M€ to 214,3 M€) in Figure 19 dwarves the APC cost anticipated in scenario 1.

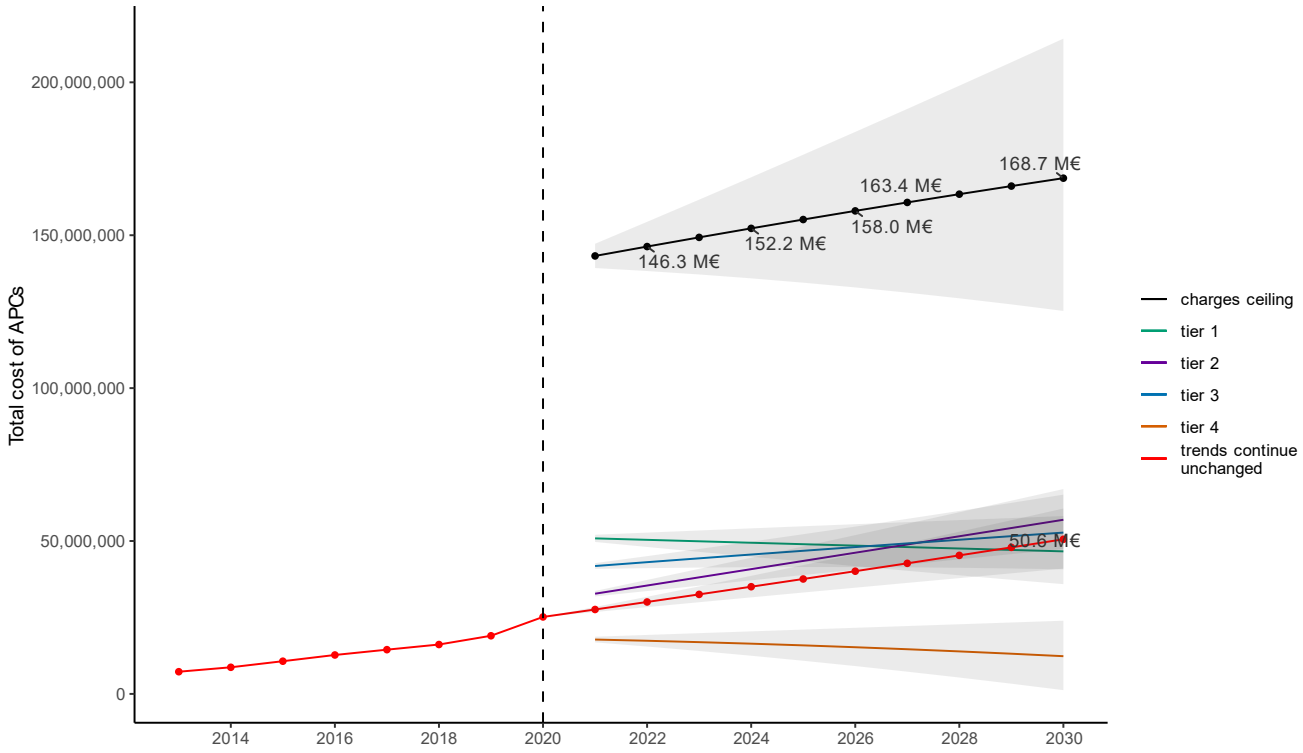


Figure 19: Simulation of the total cost of APCs if all France-based corresponding authors paid APCs except 10% of diamond articles, overall and per publisher tier (2021-2030)[20]

V. Scenario "rush"

In order to explore how APC costs would evolve under certain assumptions, we built a scenario where researchers would go full OA in Gold journals and the APC prices would go up. In practical terms, we input the following assumptions into the model:

- increasing APC prices (twice the trend continuing unchanged)
- increasing proportion of Gold OA articles (1,25 times the trend continuing unchanged)
- decreasing proportion of HybridOA articles (stable absolute number)
- decreasing proportion of closed articles.

[20] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/6VSKQU and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4251

This scenario is called "rush" because it describes a rush on Gold OA publishing and a boom of APC prices. It is, however, conservative in that it smooths out the past trend instead of only extrapolating the 2019-2020 trend which would represent a more radical growth and a much steeper slope. Because the scenario is based on OA colors, it uses the mixed-effects model with a modeled effect of the OA color. Diamond OA is not part of the model.

Figure 20 shows that in the "rush" scenario, the total cost of APCs would go up to 70 M€ in 2030, above the 50M€ predicted (confidence interval: 40 M€ to 60 M€) if trends continue unchanged.

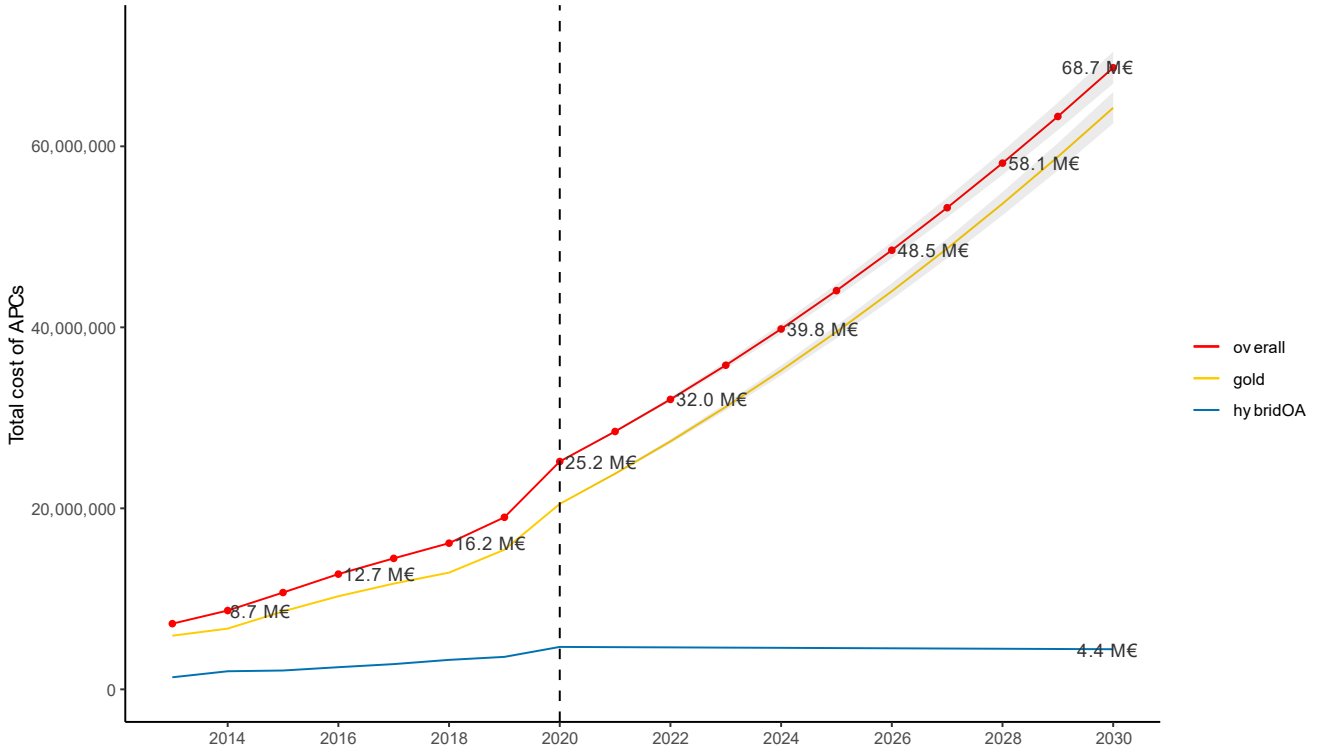


Figure 20: Simulation of total cost of APCs in the "rush" scenario, overall and per OA color (2021-2030)[21]

VI. Scenario "relief"

We built another scenario where the right retention strategy, green, and diamond strategies, as well as research assessment reform would come into full play in favor of bibliodiversity, which would limit the expansion of "star system" journals. In practical terms, we input the following assumptions into the model:

- stable proportion of closed articles at the publisher site, increasingly available as Green OA
- decreasing proportion of HybridOA articles (0,9 times the absolute number from the previous year)
- equivalent increasing proportion of Gold articles, as articles convert from HybridOA to Gold.

[21] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/EGALW9 and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L3228

This scenario is called "relief" because there is not more incentive to publishing open access at the site of the publisher. Because the scenario is based on OA colors, it uses the mixed-effects model with a modeled effect of the OA color. Diamond OA is not part of the model.

Figure 21 shows that in the "relief" scenario, the total cost of APCs would go up to 38,5 M€ in 2030, below the 50M€ predicted (confidence interval: 40 M€ to 60 M€) if trends continue unchanged. Compared to the "rush" scenario, the cost of APCs for HybridOA articles decreases slower and the cost for gold articles (33,4 M€) is half the cost (64,2 M€).

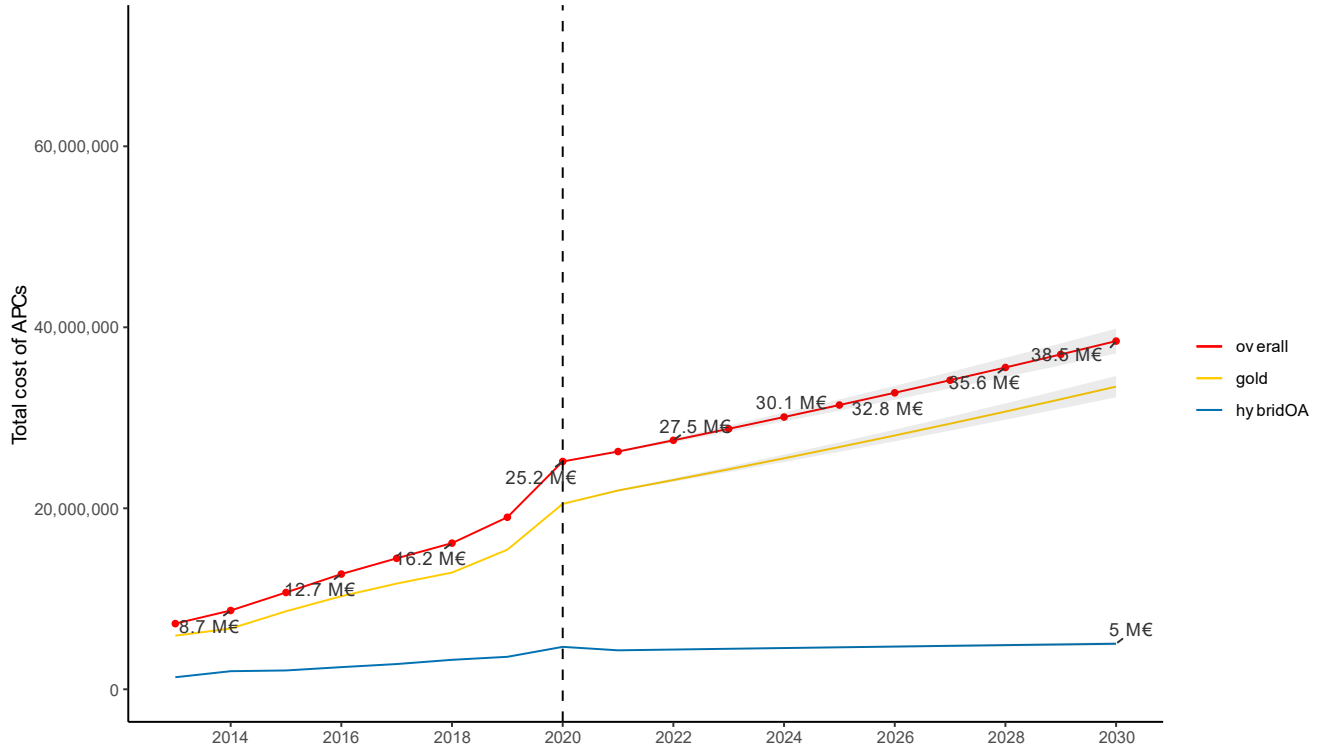


Figure 21: Simulation of total cost of APCs in the "relief" scenario, overall and per OA color (2021-2030)[22]

[22] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/AKBAXW and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L3762

VII. Conclusion

Figure 22 shows that the total cost of APCs in the theoretical situation where all France-based corresponding authors would pay APCs (yellow line) is an order of magnitude above all the other simulations. In the "rush" and "relief" scenarios (green and blue lines), the total cost of APCs increases when Gold OA (in terms of number of articles or average APC price) accelerates, above the simulation of trends continuing unchanged (red line).

These simulations help understand how the cost of publishing could evolve in the future for the French research budget and provides points of reference for policy interventions.

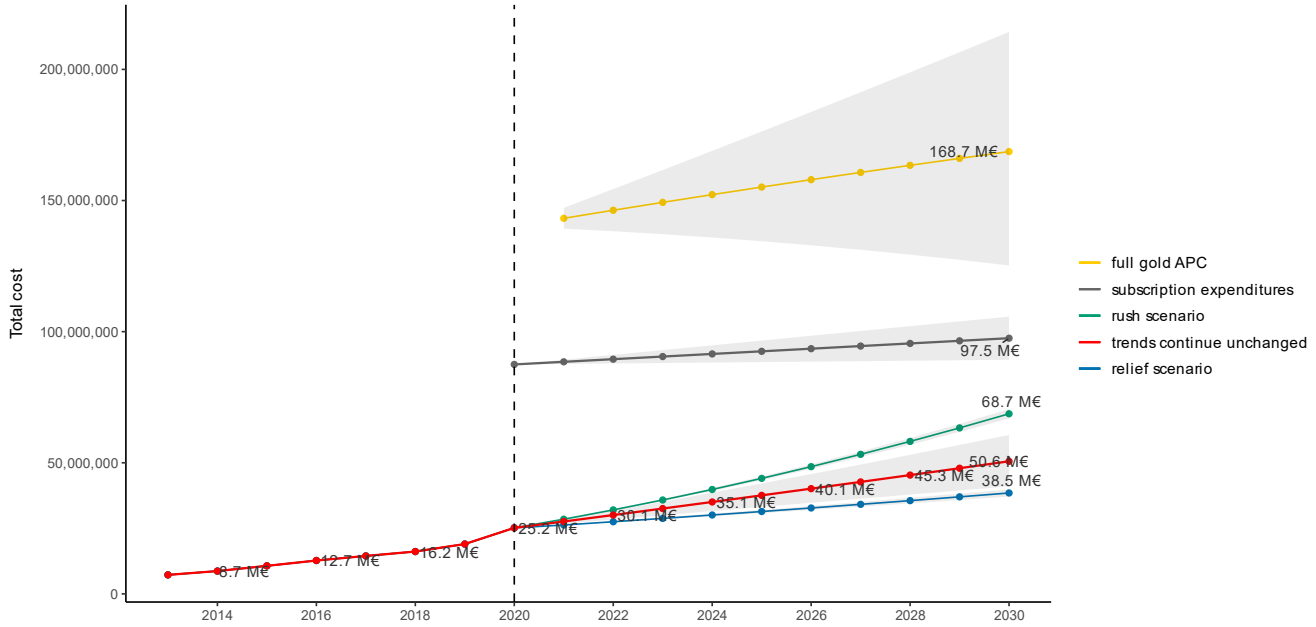


Figure 22: Summary of all simulations for the total cost of APCs as well as subscription expenditures (2021-2030)[23]

[23] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/VNVRMD and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4398

For the "rush" and "relief" scenarios, Figure 23 shows the sensitivity of the model by exploring alternative assumptions where we tweaked the numbers (e.g. 1,20 or 1,30 times the trend continuing unchanged instead of 1,25): the impact on the 2030 estimate is nonexistent for the "relief" scenario, and it is minimal for the "rush" scenario.

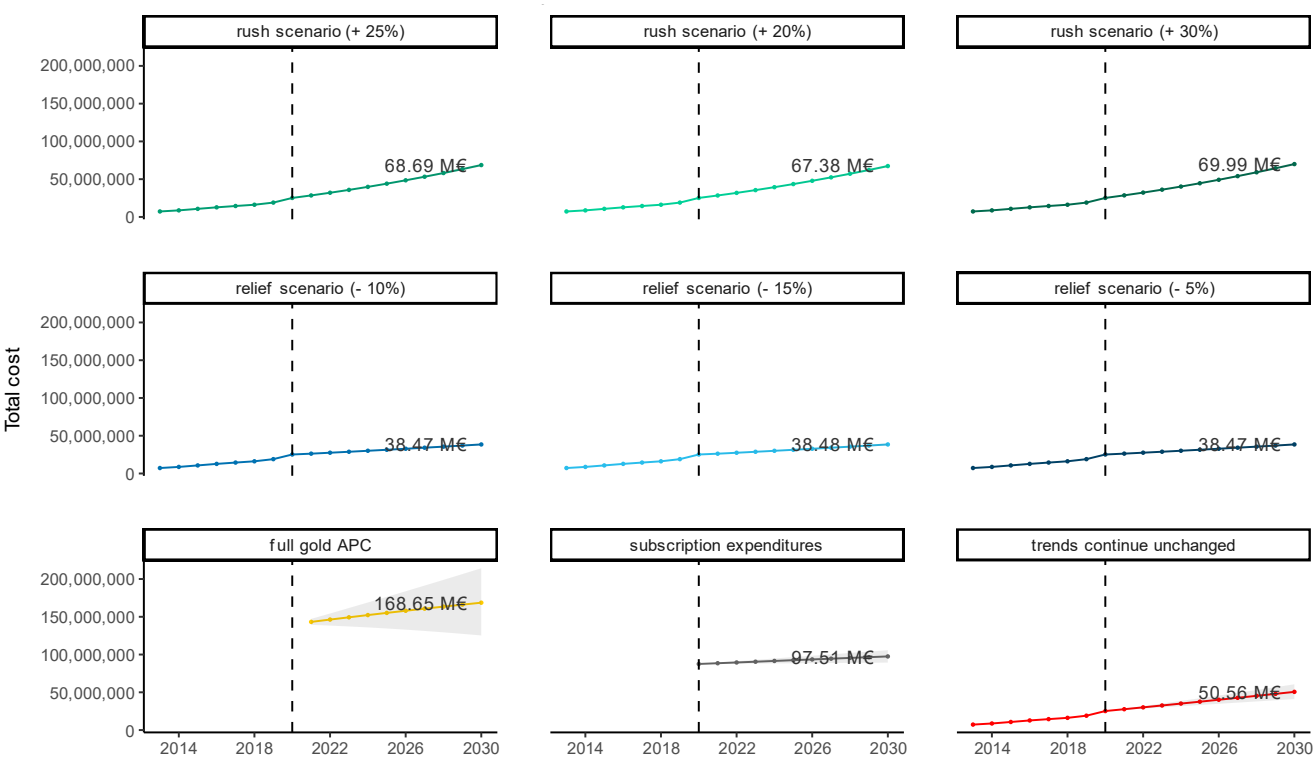


Figure 23: Summary of all simulations, with various quantitative assumptions for "rush" and "relief" scenarios (2021-2030)[24]

We thus demonstrated that the expected cost of APCs in 2030 is in the region of 50 M€ if all trends continue unchanged and could vary between 38 M€ and 70 M€ under several assumptions (funders policy, researchers practice, policy interventions…). In the meantime, the subscription expenditures could go up to 97,5 M€. Therefore, if one assumes no relationship between subscription expenditures and APC-paid articles, the costs of APCs in a full gold scenario would be in the same range as the total sum of APCs ('in the wild') and subscription expenditures in 2030.

[24] Raw data https://entrepot.recherche.data.gouv.fr/file.xhtml?persistentId=doi:10.57745/AQ0OPT/T3CEUD and source code https://github.com/datactivist/etude_APC_public/blob/main/scripts/Modelisations.Rmd#L4492

07 Bibliography

Bosman, J., Frantsvåg, J. E., Kramer, B., Langlais, P.-C., & Proudman, V. (2021). *OA Diamond Journals Study. Part 1: Findings*. Zenodo. doi:10.5281/zenodo.4558704

Bosman, J., Frantsvåg, J. E., & Kramer, B. (2021). *OA Diamond Journals Study Dataset*. doi:10.5281/zenodo.4553103

Bracco, L., L'Hôte, A., Jeangirard, E., & Torny, D. (2022). *Extending the open monitoring of open science. A new framework for the French Open Science Monitor*. https://hal.archives-ouvertes.fr/hal-03651518

Chaignon, L. (2022). Identify scientific publications country-wide and measure their open access: The case of the French Open Science Barometer (BSO). *Quantitative Science Studies*, MIT Press Direct, In press. https://hal.archives-ouvertes.fr/hal-03537679

Jeangirard, E. (2019). Monitoring Open Access at a national level: French case study. *ELPUB 2019 23rd Edition of the International Conference on Electronic Publishing*. https://doi.org/10.4000/proceedings.elpub.2019.20

L'Hôte, A., & Jeangirard, E. (2021). Using Elasticsearch for entity recognition in affiliation disambiguation. *ArXiv:2110.01958*. http://arxiv.org/abs/2110.01958

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

Thierry, D., Blanchard, A., & van der Graaf, M. (2022). Dictionnaire des variables du jeu de données de l'étude APC**.** https://github.com/datactivist/etude_APC_public/blob/main/dictionary.md

Van der Graaf, M. (2017). *Financial and administrative issues around article publication costs for Open Access*. Knowledge Exchange. doi:10.5281/zenodo.438030