
# Monitoring the open access policy of Horizon 2020

Final Report

**Monitoring the open access policy of Horizon 2020**

European Commission
Directorate-General for Research and Innovation
Directorate A — ERA and Innovation
Unit A.4 — Open Science
Contact Jean-François Dechamp
Email Jean-Francois.DECHAMP@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Manuscript completed in June 2021.



PDF ISBN 978-92-76-40516-0 doi:10.2777/268348 KI-05-21-227-EN-N

Luxembourg: Publications Office of the European Union, 2021

EUROPEAN COMMISSION

# Monitoring the open access policy of Horizon 2020

**Athena Research & Innovation Center, PPMI and UNU-MERIT**


2021 Directorate-General for Research and Innovation

## Table of Contents

## List of figures

## List of tables

# ABSTRACT

*This study is framed within the context of the contract 'Monitoring the open access policy of Horizon 2020 – RTD/2019/SC/021', reporting an authoritative set of metrics for compliance with the European Commission open access mandate within the Framework Programme thus far, and providing advice on how to systematically monitor compliance in the future.*

*Open access requirements for publications under Horizon 2020 are set out in Article 29.2 of the Horizon 2020 Model Grant Agreement (MGA). Regarding open access to research data, the Commission is conducting the Horizon 2020 Open Research Data Pilot (ORDP). The ORDP aims to improve and maximise access to, and reuse of, research data generated by Horizon 2020 projects, balancing the need for openness with the protection of intellectual rights, privacy concerns and security, and commercialisation, as well as questions of data management and preservation.*

*The present study aims to examine, monitor and quantify compliance with the open access requirements of the MGA, for both publications and research data. The study concludes with specific recommendations to improve the monitoring of compliance with the policy under Horizon Europe, together with an assessment of the efficiency and effectiveness of the Horizon 2020 open access policy.*

*The key findings of this study indicate that the European Commission's leadership in the Open Science policy has paid off. Compliance has steadily increased over recent years, achieving a success rate that places the European Commission at the forefront globally (83% open access to scientific publications). What is also apparent from the study is that monitoring – particularly with regard to the specific terms of the policy – cannot be achieved by self-reporting alone, or without the European Commission collaborating closely with other funding agencies across Europe and beyond, to agree on common standards and the common elements of the underlying infrastructure. In particular, the European Open Science Cloud (EOSC) should encompass all such components that are needed to foster a linked ecosystem, in which information is exchanged on demand and which eases the process for both researchers (who only need to deposit once) and funders (who need only record information once).*

# GLOSSARY OF TERMS AND ABBREVIATONS

**Table 1: Glossary of terms**

| Term | Definition |
|---|---|
| **Publications** | **Peer-reviewed** scientific publications, encompassing articles, books, book chapters, monographs, etc. |
| **Repository** | Repository for scientific publications: an online archive that includes the 'payload'/full text of a publication or dataset. **Institutional**, **subject-based** and **centralised** repositories are all acceptable choices under Horizon 2020, while repositories that claim rights over and preclude access to publications deposited within them are not. Interoperability: a repository should allow other systems to use the data it hosts. It should therefore make such data available according to standard metadata exchange formats (e.g. Dublin Core,[1] DataCite[2]), and possibly via standard protocols (e.g. OAI-PMH[3]). |
| **'Green' route to open access** | A scientific publication **(the published version or peer-reviewed, accepted manuscript**) deposited in a repository. |
| **'Gold' route to open access** | A scientific publication with **open access provided by the publisher**. |
| **'Hybrid' route to open access** | An article published immediately under a Creative Commons[4] licence *not in a fully open access journal.* 'Hybrid' open access publications are also 'gold'. |
| **Article Processing Charge** | The fee a publisher charges for providing 'gold' open access to articles or book contributions, (usually) at the time of publication. This is different from 'page charges' or 'colour charges'. |
| **Book Processing Charge** | The fee a publisher charges for providing 'gold' open access to entire books, chapters/Sections of a book, or a monograph, at the time of publication. |
| **Article 29.2** | The article in the Model Grant Agreement of Horizon 2020 projects that specifies the open access mandate for peer-reviewed publications.[5] |
| **Article 29.3**<br><br>*Compliance to Article 29.3*<br><br>*Uptake of Article 29.3* | The article in the grant agreement of Horizon 2020 that specifies the open access mandate for datasets for projects participating in the Open Research Data Pilot.[6]<br><br>Whether the datasets produced in projects that participated and *did not opt out of the ORDP* comply to the rules set out in Article 29.3.<br><br>*Whether the datasets produced in all* European Commission projects follow the rules set out in Article 29.3, i.e. including those that were not obliged to comply. |
| **Creative Commons** | A non-profit organisation that develops, supports, and stewards legal and technical infrastructure to enable sharing of digital outputs, including by the development of a suite of licencing products.[7] |
| **Data** | Facts, measurements, recordings, records, or observations about the world collected by scientists and others, with a minimum of contextual interpretation. Data may be in any format or medium taking the form of writings, notes, numbers, symbols, text, images, films, video, sound recordings, pictorial reproductions, drawings, designs or |

[1] https://dublincore.org/
[2] https://datacite.org/
[3] https://www.openarchives.org/pmh/
[4] https://creativecommons.org/
[5] Details provided in the Annex, Section 7.1.
[6] Details provided in the Annex, Section 7.2.
[7] https://creativecommons.org/

| Term | Definition |
|---|---|
| | other graphical representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing algorithms, or statistical records.[8] |
| **Embargo period** | Publishers permit 'green' open access often only after a certain embargo period. This embargo period can last anywhere between several months and several years. For publications that have been deposited in a repository but are under embargo, usually at least the metadata are openly accessible. |
| **FoS classification** | The classification system built for this study that assigns publications to scientific fields of study using the OECD disciplines/fields of research and development (FORD) classification scheme as developed within the framework of the Frascati Manual.[9] |
| **FAIR principles** | The guiding principles under which publications, datasets and other research output is findable, accessible, interoperable and reusable.[10] |
| **Metadata** | The information that describes an object. In scholarly communication terms the object could be an article, book, dataset, etc. The basic metadata (or bibliographic data) describe the authorship, provenance, publication location, date of publication, object type and so forth. |
| **OpenAIRE Guidelines** | Guidelines set by OpenAIRE to help repository managers expose publications, datasets and CRIS metadata via the OAI-PMH protocol in order to integrate with the OpenAIRE infrastructure.[11] |
| **Pre-print** | A journal article (or book chapter or book) that has not yet undergone peer-review or editorial scrutiny. |

**Table 2: Abbreviations and acronyms**

| Term | Abbreviation/acronym |
|---|---|
| **AAM** | Author accepted manuscript |
| **APC** | Article processing charge |
| **BPC** | Book processing charge |
| **CC** | Creative Commons |
| **DB** | Database |
| **DMP** | Data management plan |
| **DOI** | Digital object identifier |
| **EC-Shared** | Data shared by the client |
| **EOSC** | European Open Science Cloud[12] |
| **FoS** | Fields of study |
| **ID** | Identifier |
| **MAG** | Microsoft Academic Graph[13] |
| **MGA** | Model Grant Agreement[14] |

[8] CASRAI Dictionary – entry for 'Data': http://dictionary.casrai.org/Data
[9] https://www.oecd.org/sti/inno/frascati-manual.htm
[10] https://www.go-fair.org/fair-principles/
[11] https://guidelines.openaire.eu/en/latest/
[12] https://eosc-portal.eu/
[13] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[14] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

| | |
|---|---|
| **OA** | Open Access |
| **ORDP** | Open Research Data Pilot |
| **ORG** | OpenAIRE Research Graph |
| **PID** | Persistent identifier |
| **PMID** | PubMed[15] identifier |
| **R&D** | Research and development |
| **SSH** | Social sciences and humanities |
| **SyGMA** | European Commission's system for grant management |
| **TDM** | Text and data mining |
| **VoR** | Version of record |
| **WoS** | Web of Science[16] |

[15] https://pubmed.ncbi.nlm.nih.gov/
[16] https://webofknowledge.com

# EXECUTIVE SUMMARY

Open access to publications, as well as the Open Research Data Pilot (ORDP), have been key policies throughout Horizon 2020. To further strengthen Open Science and integrate it into all programmes within Horizon Europe, the European Commission has commissioned a study to: (i) measure the compliance of the existing policy under Horizon 2020; (ii) investigate which aspects of the policy have worked and which have not, in order to plan future interventions; and (iii) pilot all aspects of a monitoring mechanism, providing lessons learnt that can be used to potentially optimise the European Commission's internal monitoring platform.

The key findings of this study indicate that the European Commission's leadership in the Open Science policy has paid off. Uptake has steadily increased over the past four years, achieving an average success rate of 83% in Horizon 2020 for open access to scientific publications, which places the European Commission at the forefront globally. What is also apparent from the study is that monitoring – particularly with regard to the specific terms and requirements of the policy – cannot be achieved by the reporting alone, or without the European Commission collaborating closely with other funding agencies across Europe and beyond, to agree on and promote common standards and common elements of the underlying infrastructure. In particular, the European Open Science Cloud (EOSC) should encompass all such components that are needed to foster a linked ecosystem in which information is exchanged on demand and eases the process for both researchers (who only need to deposit once) and funders (who only need to record information once).

## Open methodology

A key objective in the study was to use an open, transparent and re-producible methodology, as a driver to improve the operationalisation of the open access monitoring in the EC, with rules that can be shared with and accepted by the research community. To authoritatively assess the open access compliance among Horizon 2020 peer-reviewed publications and datasets, and to assess the specificities of Article 29.2 and 29.3 of the MGA, this study has explored and combined a number of public and proprietary datasets. **For the first time, open data sources were considered as the primary sources for such monitoring** (OpenAIRE,[17] Unpaywall,[18] CrossRef,[19] OpenAPC[20], DataCite,[21] ORCID,[22] DOAJ,[23] re3data[24] to name a few). These were then **validated against proprietary databases** (Scopus[25] and WoS[26]) **as secondary sources,** when necessary.

Despite the fact that working with and merging open sources often proved to be a painstaking process, it was one that ultimately proved flexible and agile, and allowed the study team to interact with the community and propose changes to the underlying public infrastructure. Moreover, most of the data and metadata contained in the open sources proved to be of good quality, justifying the adoption and curation of open, community-driven standards.

[17] https://www.openaire.eu/
[18] https://unpaywall.org/
[19] https://www.crossref.org/
[20] https://openapc.net/
[21] https://datacite.org/
[22] https://orcid.org/
[23] https://doaj.org/
[24] https://www.re3data.org/
[25] https://www.scopus.com/home.uri
[26] https://webofknowledge.com

Fully in line with Open Science practices, the study produced an authoritative and open access database covering all aspects related to Horizon 2020 publications, accompanied by a data management plan and detailed documentation. [27]

## Efficiency of the Horizon 2020 open access policy

**Overall, the estimated level of compliance to the open access mandate for scientific publications under Horizon 2020 stood at 83%,** which is within the top open access success rates of funders globally. **Compliance and uptake of open access to research data have a success rate of 95%**.[28] This achievement is doubly impressive when considering the context in which the policy is implemented: a decentralised European environment in which Member and Associate countries have different policies and infrastructures (or lack thereof). With a clear upward trend in publications (from 65% in 2014 to 86% in 2019), and a commitment to the policy from projects that participated in the Open Research Data Pilot (ORDP) the potential exists to reach 100% within the early stages or midway through Horizon Europe.

When we compare it with other research funders, **Horizon 2020 is in the top of funders in terms of the level of open access achieved.** In terms of the percentage of publications that are openly accessible, Horizon 2020 came 12th out of the 47 non-discipline specific funders included in the analysis. On average, Horizon 2020 performs better than some of the largest non-discipline specific research funders in Europe (Switzerland, Sweden, Germany, Italy, Spain, Ireland, Portugal) and some of the largest in the US (e.g., the National Science Foundation [NSF]). At the same time, the percentage of publications under Horizon 2020 that were openly accessible was somewhat lower compared with some of the largest research funders in the Netherlands, Hungary, Denmark, Austria and Belgium, which have a similar tradition in open access policies but accompany this with well-established and connected national infrastructures.

In terms of **article processing charges (APCs),** we estimated the average cost of a 'gold' open access article to be around EUR 2,200. 'Hybrid' open access articles, a category that will no longer be reimbursed under Horizon Europe, have a higher average cost of EUR 2,600. Our analysis of six large research funders showed that, on average, **APCs under Horizon 2020 were similar to the average for other funders** in Europe and USA for which the required data was available.

Qualitative evidence also revealed some key sources of inefficiencies, as well as potential areas for improvement in the efficiency of the Horizon 2020 open access policy. To increase open access to research outputs, some beneficiaries expressed **a need to fund the article processing charges (APCs)/book processing charges (BPCs) for post-project publications that resulted from the grant activities**. In many cases, a publication based on Horizon 2020 activities are actually published after the project has formally ended (this is particularly common in the humanities and social sciences, where books and book chapters are common research outputs). In addition, one of the key sources of financial cost-inefficiencies relates to **a lack of awareness and knowledge on the part of beneficiaries with regard to Horizon 2020's open access requirements.** In some cases, project budgets were used to cover APC costs because at the time, beneficiaries were unaware of alternative open access routes. The available evidence also confirms that **excluding APCs for hybrid journals from eligible costs** under Horizon Europe may prove to be a measure to increase the cost-efficiency of the programme's open access policy. This is something that needs to be closely monitored, as even though current data indicates that hybrid options incur considerably higher average APCs compared with fully

[27] All three are deposited in Zenodo under a CC-BY license and are also published in the European Commission's portal. The Zenodo links are as follows: database https://zenodo.org/record/4899767, documentation https://zenodo.org/record/4900100, and data management plan https://zenodo.org/record/4900110.

[28] Compliance refers to adherence to the regulations set out in Article 29.3 for projects that *had to comply* to the article (those in the ORDP), and uptake refers to compliance to Article 29.3 by all projects, whether they had to comply or not.

open access journals, future costs will heavily depend in shifts in the publishing market power brought in by the transition of closed/hybrid journals to Open Access and new publishing platforms such as Open Research Europe.[29]

Our study produced specific indicators to assess the full openness and 'FAIR-ness' of Horizon 2020 results:

- **Licensing: 49% of** Horizon 2020 publications were published **using Creative Commons (CC) licences,** which permit reuse (with various levels of restrictions) while **33% use publisher-specific licences that place restrictions on text and data mining (TDM)**. Another 18% of open access publications (mainly in institutional repositories) come with no licence, which effectively translates into non-legal use for TDM purposes. This calls for further policy action, as real open access should not place any obstacles in the way of both human and machine readability. Concerning research data, things are more straightforward (no publishers in the mix) with a **compliance level to depositing datasets with an open license of 65%** (CC licences).

- **Accessibility and interoperability:** Institutional repositories have responded in a satisfactory manner to the challenge of providing FAIR access to their publications, amending internal processes and metadata to incorporate necessary changes: 95% of deposited publications include in their metadata some type of persistent identifier (PID); a rate of 73% accessibility and interoperability has been observed – i.e., correctly identifying a full text from the metadata (accessibility), and being able to fetch it via a known protocol (interoperability). Datasets in repositories, on the other hand, present a low compliance level as only approximately 39% of Horizon 2020 deposited datasets are findable, (i.e., the metadata includes a PID and URL to the data file), and only around 32% of deposited datasets are accessible (i.e., the data file can be fetched using a URL link in the metadata).

**Effectiveness of the Horizon 2020 open access policy**

On average, the **open access rate among Horizon 2020 publications has increased steadily** over the duration of the programme, from just over 65% of peer-reviewed publications being open-access in 2014, to 86% in 2019. The effectiveness of the policy, however, **differed somewhat between Horizon 2020 programmes.** The highest shares of open access publications were found in the European Research Council (ERC) and 'Science with and for Society' programmes, while the lowest shares were in 'Euratom', 'Industrial Leadership', and 'Spreading Excellence and Widening Participation'. Evidence also confirms that open access under Horizon 2020 **varied according to scientific fields and specific disciplines**. The percentage of open access publications was highest within medical and health sciences, as well as natural sciences, but lower within the agricultural and veterinary sciences, engineering and technology, social sciences, as well as humanities and arts. In some cases, variation under Horizon 2020 also existed at the level of specific disciplines within particular scientific fields.

On the ORDP front, our findings indicate an uptake and compliance success rate **of 95%.** Variations exist in compliance between programmes, although in most cases the level remains well above 90%. The three pillars with the most significant production are Societal Challenges (in proportion to the number of projects, this pillar generates twice as many datasets as the others); Excellent Science; and Industrial Leadership.

Qualitative evidence also reveals that, in general, **Horizon 2020 projects become increasingly compliant with open access requirements over the project's life cycle**. This is mainly due to effective communication, feedback and support provided by

[29] https://open-research-europe.ec.europa.eu/

project officers to beneficiaries, which helps them to meet the open access requirements by the time the project ends.

Study evidence shows that the key result and benefit of the Horizon 2020 open access policy is **wider outreach and dissemination of research work across different fields and to the general public.** Furthermore, the policy led to **learning effects:** fulfilling their open access obligations under Horizon 2020 led to increased awareness and knowledge among beneficiaries with regard to the concepts and principles that underpin Open Science, and improved their related skills. Lastly, at organisational and system level, the Horizon 2020 open access policy has produced spill-over effects by **encouraging other European research funders and institutions to adopt similar open access policies and measures.**

**Monitoring open access**

One of the objectives of the study was to identify the Horizon 2020 open access monitoring workflow, including the key steps, tools and actors involved in the monitoring process. The Horizon 2020 open access monitoring workflow is based on two essential instruments and data sources: automated monitoring and tracking of metadata on research outputs through the OpenAIRE platform, and the continuous self-reporting procedures followed by Horizon 2020 Project beneficiaries using the SyGMa portal.

Our quantitative analysis and interviews with stakeholders identified gaps in the existing Horizon 2020 open access monitoring data, which pose further difficulties in assessing compliance. More specifically: key metadata were not systematically provided by repositories (e.g., peer-review status of publications/ publication release dates/submission history/publication versioning); data displayed to grantees are in many cases of poor quality, mainly due to the **lack of consistent and rigorous data entry practices among many publishers and repositories**; partial coverage of emerging repositories and publishers, particularly in specialised sectors/domains; non-clarity of the different versions of the same publication; delays of appearance of open access publications in OpenAIRE and SyGMa.

**Self-reporting by beneficiaries** also highlighted a number of issues relevant to compliance checking and the assessment of indicators, mainly focusing on the facts that (i) some publications are not reported at all – particularly as beneficiaries do not keep reporting after a project has ended, and (ii) the poor quality of metadata entered by beneficiaries which makes them unreliable and unusable. The latter includes lack of the systematic use of valid digital object identifiers (DOIs) and other valid PIDs; missing links between publications and datasets; data on embargo periods for both publications and datasets being poorly provided or unclear; as well as missing information about the tools and instruments at the disposal of the beneficiaries and necessary for validating the results. One of the main reasons for this is that researchers are very often not fully aware of the semantics and the scope of many open access-related concepts, such as the differences between 'gold' and 'green' open access; embargo periods; DOI; repository links, etc. In many cases, it is impossible for project officers to check if a deposited publication has been made open access within the maximum allowable time limit (at most 6-12 months).

Gaps and challenges relating to **the monitoring of (open access) research data** resulting from Horizon 2020 projects largely resulted from a lack of data management skills and knowledge among beneficiaries. Beneficiaries are often not methodical or meticulous about precisely what type of data to open up (raw vs. annotated vs. processed); what accompanying documentation should be included; and what existing data protection regulations apply. Frequently, data management plans (DMPs) are very rudimentary because researchers do not understand some of the key underlying principles, such as FAIR. In addition, datasets may sometimes be very large and complex. Storing them and

maintaining them in an openly accessible form might require a great deal of storage space and/or qualified staff, which may pose significant financial burdens on the research teams.

In addition to the development of a comprehensive list of open access indicators for both publications and datasets, one of the key inputs required to re-engineer the existing open access monitoring framework was the **identification of key principles and stakeholder expectations regarding the next-generation Horizon Europe open access monitoring framework**. One key expectation is that the Horizon Europe monitoring framework should allow the possibility of checking in real-time the publications resulting from it (including, for example, filtering information by type of publication, discipline, etc.). The scale of the next-generation Horizon Europe open access monitoring framework is also expected to be expanded, incorporating more diverse types of research outputs in addition to publications (e.g., software, prototypes, etc.). Its scope is also expected to expand beyond the direct outputs of the programme, to include medium-term and long-term indicators focusing on the uptake of open access outputs and their impacts on the creation of new research networks.

Based on an analysis of gaps in the previous monitoring framework, the study has prepared **a number of recommendations that address various issues relating to gaps in open access data /monitoring process**. These include recommendations (listed below) on improving the integration of OpenAIRE into the European Commission's SyGMa reporting tool, addressing the processes relating to open access self-reporting by beneficiaries, and regarding the monitoring of open data.

1. Update the OpenAIRE guidelines for repositories and increase the adoption of the OpenAIRE metadata standard among repositories.

2. Streamline internal procedures within OpenAIRE Graph to reduce delays in transferring data to the SyGMa reporting tool.

3. Organise training sessions for beneficiary principal investigators, focusing on the general principles underpinning open access in Horizon Europe, as well as the requirements and reporting process.

4. Prepare a concise 'one-stop source' manual/guidelines for beneficiary principal investigators/project managers/support staff, explaining the key steps in the Horizon Europe open access reporting process.

5. In the case of manual self-reporting by beneficiaries, implement technical safeguards at the data submission stage in the SyGMa reporting tool, to address the issue of beneficiaries incorrectly filling in metadata fields when self-reporting.

6. Deliver regular reminders to the project beneficiaries for several years after the project has ended, calling on them to report the project outputs on the Participant Portal, to increase the level of post-project open access reporting.

7. Improve the quality of open research data management in Horizon Europe projects, by encouraging the inclusion of skilled personnel and by providing guidance and common templates.

8. Disseminating the existing DMP good practice examples to beneficiaries at the beginning of their projects.

9. Develop clear and comprehensive guidelines describing what type of data should be opened up (raw vs. processed), and what documentation should accompany open access research datasets.

# 1 Introduction

This is the final report of the study 'Monitoring the open access policy of Horizon 2020', which sets out to achieve the following:

1. Measure the uptake of open access to publications by Horizon 2020 beneficiaries, and compliance with the requirements set out in Article 29.2 of the Horizon 2020 Model Grant Agreement (MGA).

2. Measure the uptake of (participation to) the Open Research Data Pilot (ORDP) in Horizon 2020, and compliance with the requirements set out in Article 29.3 of the Horizon 2020 MGA.

3. Provide advice to the European Commission, and define a process for monitoring compliance with the policy in the future.

4. Assess the progress achieved to date by the Horizon 2020 open access policy, and provide lessons learned and recommendations for the future.

The study's activities and findings were divided into four tasks. **Tasks 1** and **2** focused on creating the basis for our evidence-based analysis by:

- Creating an authoritative list of Horizon 2020 publications and datasets by gathering and merging data from multiple data sources (both public and commercial).

- Identifying, collecting, linking and integrating the data to be used for the analysis of open access compliance by:

  - performing a thorough gap analysis and quality assessment of the metadata essential for the construction of indicators;

  - filling gaps in metadata by triangulating different data sources and carrying out technical work (e.g., text mining) to produce additional metadata elements (e.g., identify the Fields of Study classification, assess the validity and accessibility of URL's).

- Analysing open access compliance for publications and datasets through a set of indicators that reflect both the overall and the technical implementation of the policy, presenting different aspects such as time, country, programme, discipline.

- Estimating and analysing publication costs for Horizon 2020 'gold' publications.

- Compiling all data into a database (the 'MOAP Horizon 2020 DB'[30]), that was published along with corresponding documentation[31], and a data management plan[32].

**Tasks 3 and 4** included desk research, interviews with key stakeholders in the field, and a validation workshop for the proposed indicators, which effectively allowed us to:

- Model the workflow specification for the Horizon 2020 monitoring process.

- Identify gaps in the monitoring of the Horizon 2020 open access policy.

- Propose interventions, where appropriate, in the monitoring process.

[30] https://zenodo.org/record/4899767
[31] https://zenodo.org/record/4900100
[32] https://zenodo.org/record/4900110

- Test and validate the proposed monitoring workflows.
- Assess the efficiency and effectiveness of the Horizon 2020 open access policy.
- Reconstruct the intervention logic of the Horizon 2020 open access policy.

In line with the above tasks, the report is structured as follows:

- **Section 2** describes the general methodological approach followed in this study.
- **Section 3** presents the data analysis of open access policy compliance and publication costs for Horizon 2020 publications with select observations.
- **Section 4** presents the data analysis of compliance and uptake of Horizon 2020 datasets to Article 29.3 with select observations.
- **Section 5** presents the modelling of the monitoring process workflow, a gap analysis of the framework, and recommendations for the reengineering of the monitoring process in the future.
- **Section 6** provides an overall analysis of the intervention logic for the Horizon 2020 open access policy, as well as an analysis of its efficiency and effectiveness, and the lessons learned from this study.

A set of Annexes present the relevant excerpts from Articles 29.2 and 29.3 of the Horizon 2020 Model Grant Agreement for the assessment of open access compliance and the detailed technical methodology followed in the study.

## 2 General methodological approach

Our methodology followed an integrated approach to all four tasks of this assignment. To be able to meet the challenge put forward by this assignment, we based our work on the following operational quality criteria that drove our whole approach:

- **Openness and transparency**: The end goal of this assignment was to offer an assessment of the performance of the current mandate and future recommendations. For these to be valid, it was essential to openly and clearly identify any potential methodological issues that could skew our findings (e.g., missing data).

- **Coverage and accuracy**: As detailed in our approach we used multiple data sources (triangulation) for *cross-validation* and *coverage to the fullest extent possible*, in order to provide meaningful indicators and recommendations.

- **Clarity and replicability:** Our methodology is described in detail, so that it can be verified and used by the scholarly communication community to create ongoing updates to our proposed statistics and indicators, thus giving the basis for evaluating any changes in the open access mandate.

- **Readiness and timeliness**: We developed our methodology around well-established open databases and already tested knowledge extraction technologies (natural language processing (NLP)/machine-learning (ML) used in operational workflows in OpenAIRE and Data4Impact[33]) to warrant timely results.

- **Trust and robustness**: Our methodology also strived to be reliable, robust, and aligned to other assessment methods so that it can be operationalized, used and reused, in conjunction with other assessment methods.

- **Pragmatism and practicality**: the recommendations for improving the current monitoring and specificities of the open access mandate are guided by incentive (author compliance) and feasibility (data for verification of compliance & cost of collecting it) constraints.

What was important to consider in this study was that we needed to fully understand the dynamics and the actors, what has worked, in which area (region/thematic), what are the factors that have helped or limited the uptake of the policy; to understand the role and the use of underlying infrastructures, evaluate how this has worked and propose changes; to use open and collaborative infrastructures and improve them so as to ensure a continuity, robustness and therefore trust in the monitoring process; to ensure we infuse the right elements in Horizon Europe for a pragmatic approach for implementing Open Science.

Our approach further relied on **three pillars** that were applied throughout our quantitative approach:

- **Using authoritative open data sources** to respond to all analysis requirements for the detailed study of the Horizon 2020 open access policy uptake. Triangulating with bibliographic databases to ensure that we have the most complete coverage of Horizon 2020 outcomes and that we increase the quality of our results (enrichment);

- **Applying big data/NLP/ML technologies** for data cleaning and fusion, for inference of relationships between various entities (e.g., citations, affiliations), and for knowledge extraction (e.g. classifications);

- **Including experts** in order to assess and validate the processes (data, technology and indicators) and consult/recommend on the way forward for a pragmatic intervention for the implementation of Open Science (not only open access) in Horizon Europe.

[33] https://monitor.ilsp.gr/landing

Figure 1 illustrates the conceptual phases of our approach. Starting from an initial set of indicators and statistical figures, we used existing open data and open technologies, we imported-fused-validated-refined, in order to guide our analysts to acquire a 360° view and understanding of the Horizon 2020 open access policy uptake (who, when, what, why) and implementation aspects (what worked, how, where).

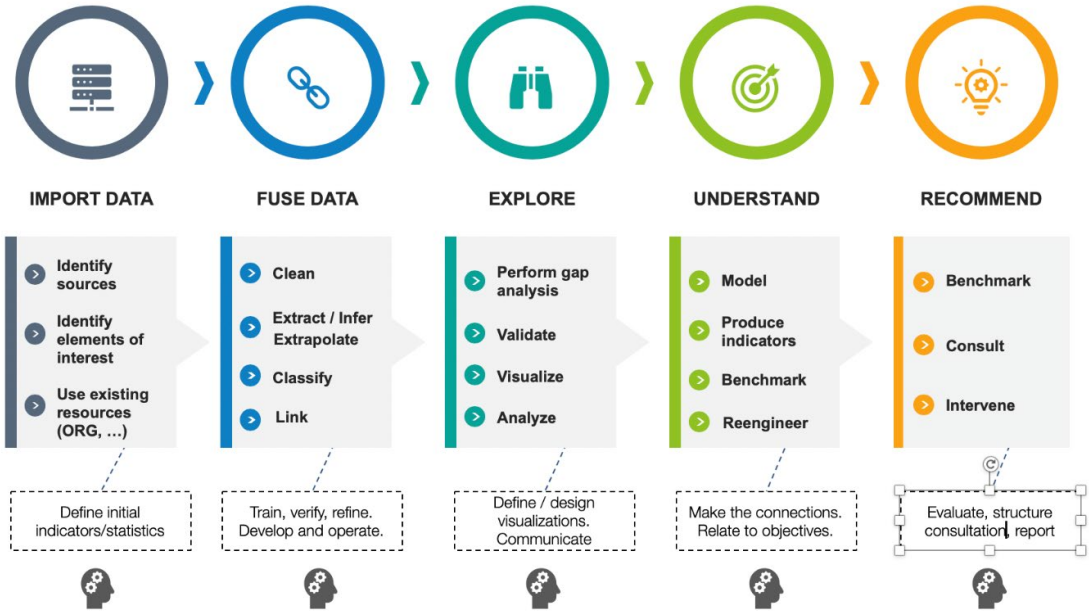



Figure 1: Conceptual methodology overview – human-in-the-loop

Sections 7.3 and 7.4 in the Annex present in detail the methodology followed for Horizon 2020 publications and datasets, respectively.

# 3 Open access to Horizon 2020 scientific publications

This section presents the data analytics and the findings of the compliance with the Horizon 2020 policy on open access to publications, with select observations on patterns of interest.[34] Key aspects of compliance and monitoring are presented with a special focus on those that may require potential interventions under Horizon Europe.

We start by presenting some aggregate figures on the overview of Horizon 2020 production to set the context, continue with an overview of open access uptake reflecting different facets, e.g., country, discipline, programme, and finally examine compliance with respect to specific terms and requirements of Article 29.2. Our work included the definition and calculation of a rich set of indicators. The full list is summarised at the end of this Section (Table 12).

## *3.1 Peer-reviewed publications and compliance to Article 29.2*

### 3.1.1 The context: the overall production of peer-reviewed publications in Horizon 2020

Our first task was to identify and gather a comprehensive and authoritative list of all peer-reviewed Horizon 2020 publications (linked to the grant ID level). Beginning with the European Commission's database, which records outputs from project coordinators and partners, we triangulated using open data sources (OpenAIRE Research Graph, which includes Unpaywall, Crossref, ORCID, DOAJ/DOAB, DataCite and Microsoft Academic Graph) and commercial sources (Scopus/Web of Science). Our methodology is described in detail in Section 7.3.

Through an iterative process that included key quality assessment checkpoints to ensure reproducibility, we identified **218,558 unique publications** (of various types, including 'grey' literature) linked to Horizon 2020, out of which **154,185 are peer-reviewed**.

An initial analysis indicates that:

- The bulk of peer-reviewed Horizon 2020 publications are outcomes of the Excellent Science pillar (Figure 2).
- Social sciences, humanities and arts publications represent a small proportion of Horizon 2020 publications (Figure 2). This may be related to a lower level of funding for SSH-related projects and/or to different publishing procedures (e.g. a greater proportion of books, which involve a time lag in publishing).
- The distribution of authors across Europe (Figure 3) indicates a strong correlation with the size of the country (population) as expected, and a weak correlation with the research and development (R&D) expenditure.[35] For example, countries with the same population such as Austria, Belgium, Czech Republic, Greece, Hungary, Portugal, Sweden, which have R&D expenditure from 1,27%-3,39% of their GDP, produce similar volumes of results.
- Collaboration and synergies with other funders (European, international) mainly occurs within the Excellent Science pillar, followed by Societal Challenges. Other programmes, such as the Industrial Leadership programme, lag behind considerably (Figure 4).

[34] The database (https://zenodo.org/record/4899767) contains a host of metadata elements that can be used for additional analysis.
[35] 2019 R&D expenditure in the EU, https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20201127-1

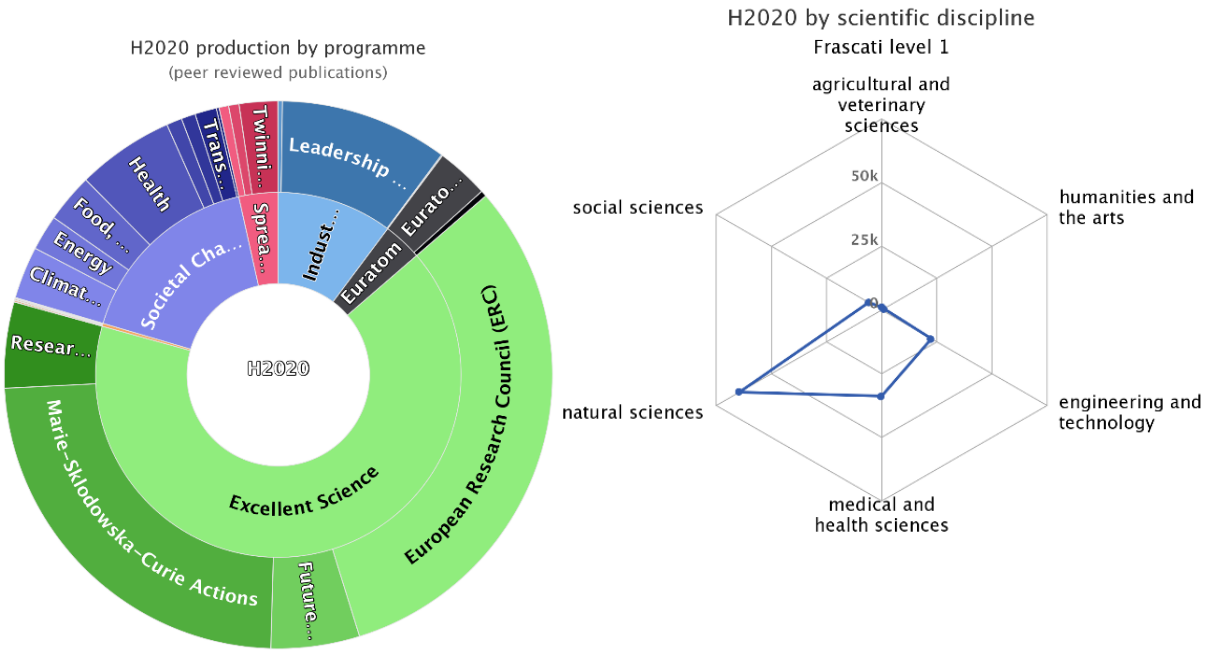



Figure 2. Horizon 2020 peer-reviewed publications, by proramme and scientific discipline

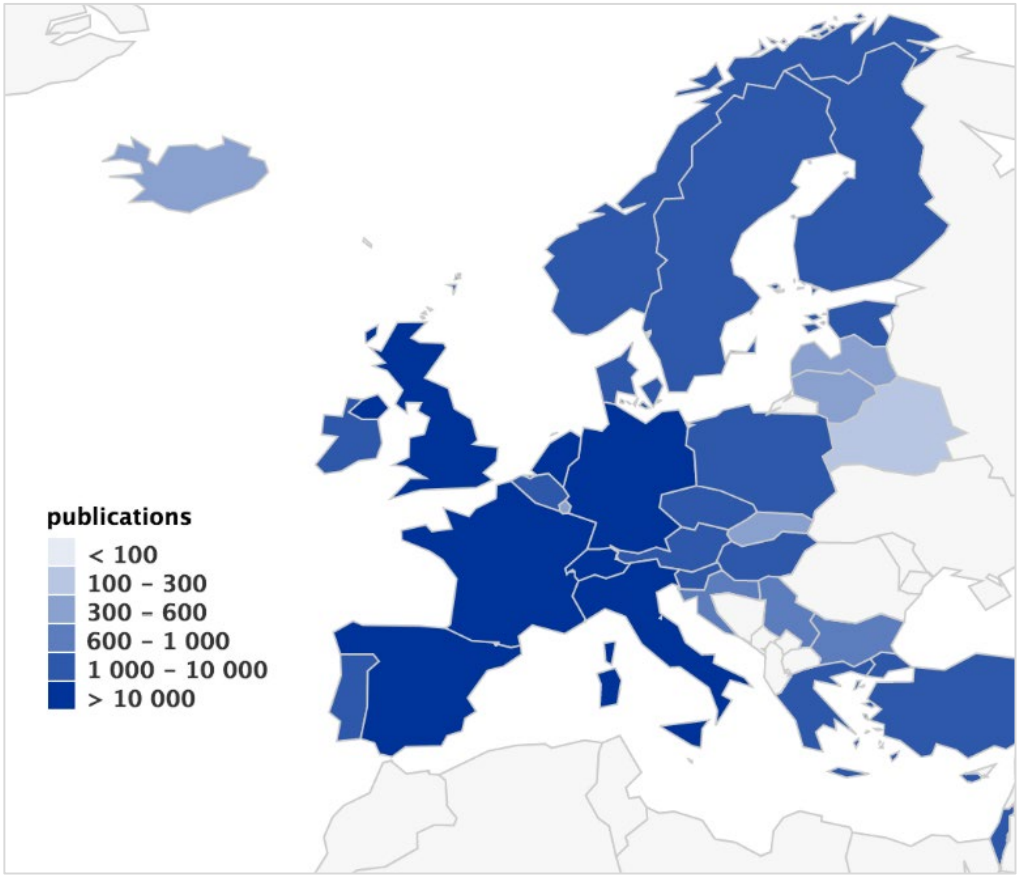



Figure 3. Horizon 2020 peer-reviewed publications, by author country

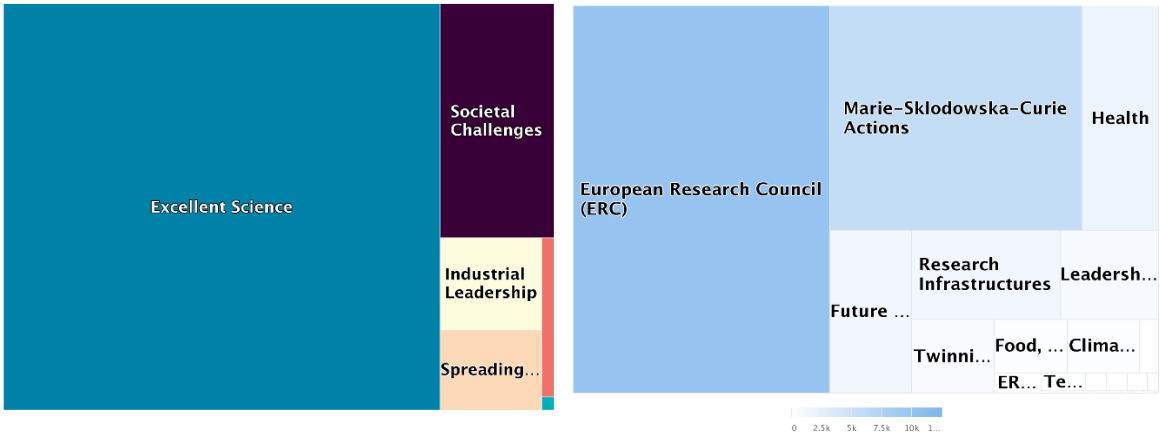


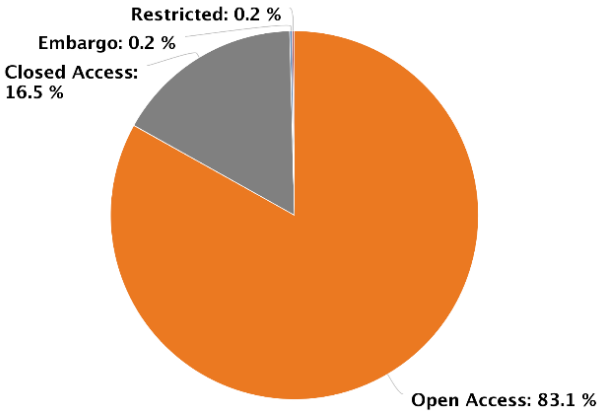

Figure 4. Horizon 2020 collaborations via co-funded publications

### 3.1.2 Compliance with the Horizon 2020 open access policy

By further triangulating the authoritative list of Horizon 2020 publications with the OpenAIRE Research Graph, which includes information on open access repositories and journals, and with Scopus/WoS, which include licencing information, we have come up with the following:[36]

**Horizon 2020 has an open access rate of 83.1% for peer-reviewed publications:**

- Open access: 128,123
- Embargo: 345
- Restricted: 242




As indicated in Table 3 which includes the breakdown of open access by year of publication, we observe an upward trend in the uptake of open access, which increased by more than 20% between 2014 and 2019. This demonstrates the effectiveness of the mechanisms put in place by the European Commission on awareness, infrastructure, tracking and follow up.[37]

[36] Restricted refers to the case where access to the article is not behind a paywall, but access is still restricted to certain users.

[37] The numbers for 2020 are low, probably indicating a lag in open access compliance due to late deposition following the 'green' route or embargoed publications.

**Table 3. Open access rate over time**

| Year | % Open Access |
|---|---|
| *Cumulatively* | 83.1% |
| 2014 | 65.4% |
| 2015 | 79.4% |
| 2016 | 83.2% |
| 2017 | 85.4% |
| 2018 | 86.7% |
| 2019 | 86.3% |
| 2020 | 78.8% |

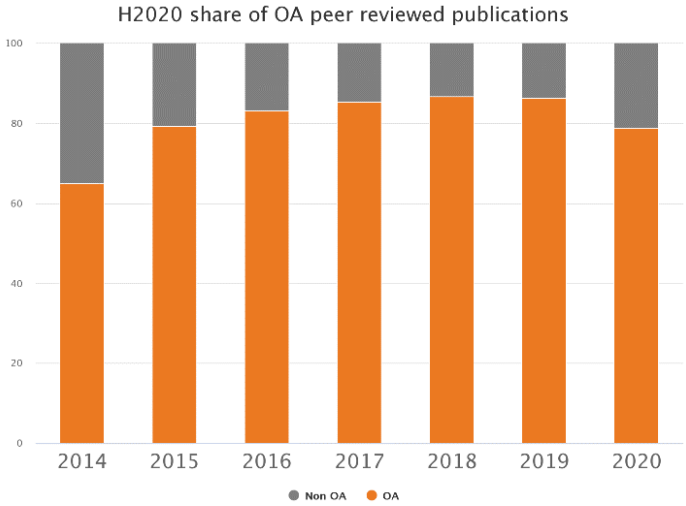



Figure 5: Open access rate over time

- *Publication type, ERC and non-ERC grants*

Figure 6 examines open access rights by different publication types for ERC and non-ERC grants. We distinguish between the two because for *non-ERC* parts of Horizon 2020 the interpretation of the open access mandate was *that books and book chapters were exempted from this obligation*, *but* if made open access, any BPCs would be reimbursed. For ERC grants all types of peer-reviewed publications had to comply to the open access mandate of Article 29.2. In principle, this should imply a higher open access rate for ERC grants for both books and book chapters. This is the case for book chapters (63.5% open access for non-ERC grants and 72.1% for ERC grants), although for both the rate is significantly lower than for all other document types. On other hand, surprisingly, books have a higher open access rate for non-ERC grants (85.9%, against 84.1% for ERC grants).

Overall, however, ERC grants fair better in terms of open access. In particular, for articles and conference objects/proceeding papers, ERC publications exhibit a 6% higher open access rate than those resulting from non-ERC parts of Horizon 2020.

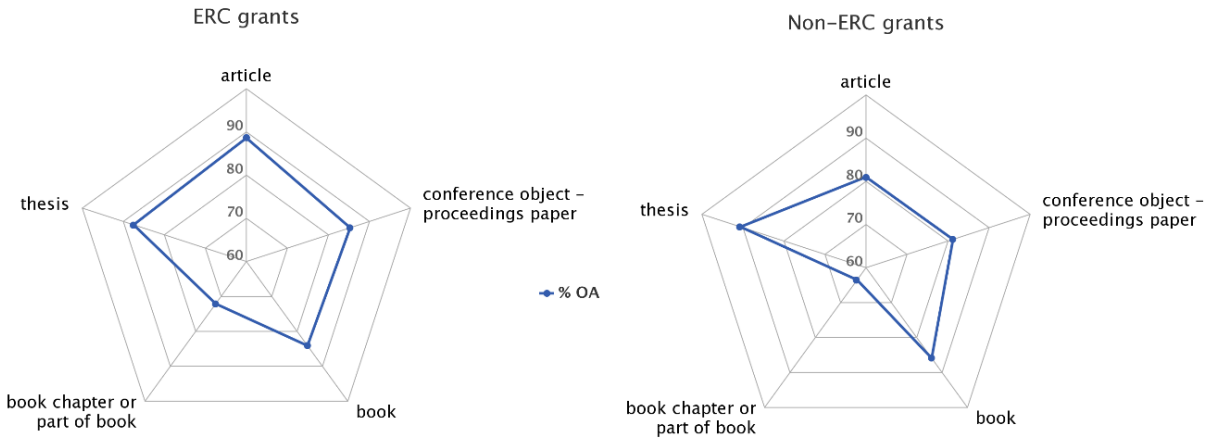



Figure 6: Open access rate, by publication type, ERC and non-ERC grants

Moreover, when analysing specific metadata elements, we find that:

- **Deposited version:** The open access mandate specifies that either the version of record (VoR) or the author-accepted manuscript (AAM) is deposited in a repository. However, identifying the VoR/AAM for open access publications in repositories or in pre/post-print thematic servers is not always possible, as **11.2%** of the deposited publications do not have a valid VoR/AAM attribute in their metadata. This indicates that repositories should either apply stricter rules to ensure that researchers/curators fill in this field, or find automated ways to insert this (e.g., requesting a DOI and resolving it) so that such publications can be captured in any monitoring process.

- **'Green' route:** Compliance in terms of depositing 'gold' open access publications in a repository is relatively high **(81.9%),** indicating that the policy (of depositing publications in repositories) is well understood and implemented by researchers. Out of those, the majority includes the VoR/AAM version when self-depositing **(71.1%)**.

- **Immediate open access:** The 'gold' and 'green' routes are both valid options for researchers and numbers show that they work in parallel. The data from 2020 shows that there is a lag in deposition, which has been observed for the past 10 years (i.e., 'green' catches up with 'gold' in about one year's time), which indicates that immediate 'green' open access is still an issue (Figure 7).

- **Use of repositories and infrastructure:** Horizon 2020 has implemented a 'green' route to open access policy, which supports and promotes the use of repositories. In most cases, particularly where national or institutional policies and established national networks of repositories exist (e.g., in France, Finland, Croatia, Norway, Turkey), 'green' uptake under Horizon 2020 is stronger. This clearly signals that policy implementation cannot be tackled by one organisation alone, however big and important it is, but requires synergies with infrastructure service providers, and primarily with Research Performing Organisations (RPO) (Figure 9).

- **Type of repository:** Deposition ('green' open access, any version of document) is well established in institutional repositories vs. all-purpose repositories: 75,129 publications have been deposited in institutional repositories and 62,037 in thematic repositories or pre-print servers.

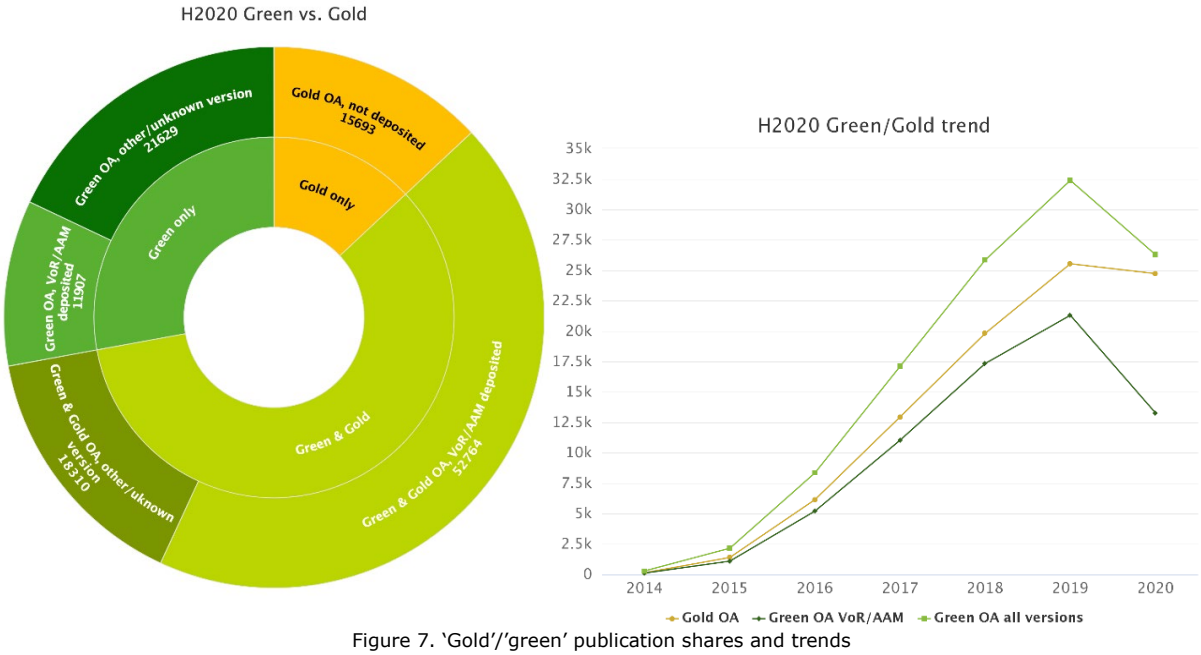


Figure 7. 'Gold'/'green' publication shares and trends

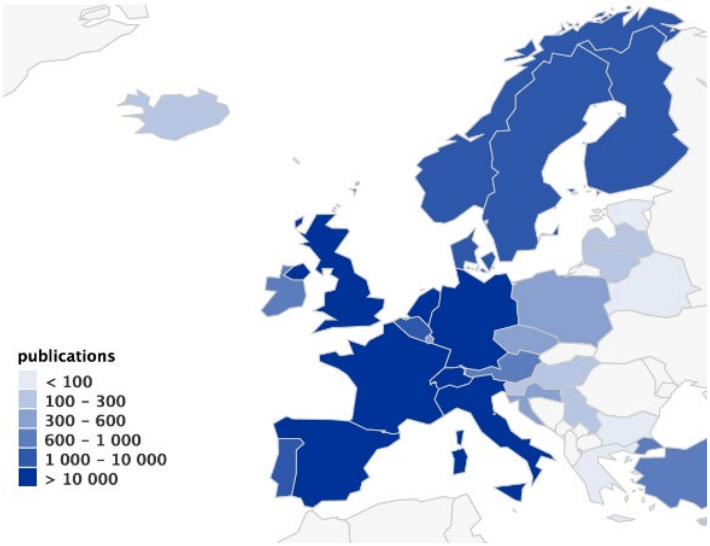



Figure 8. Peer-reviewed publications in institutional repositories, by country

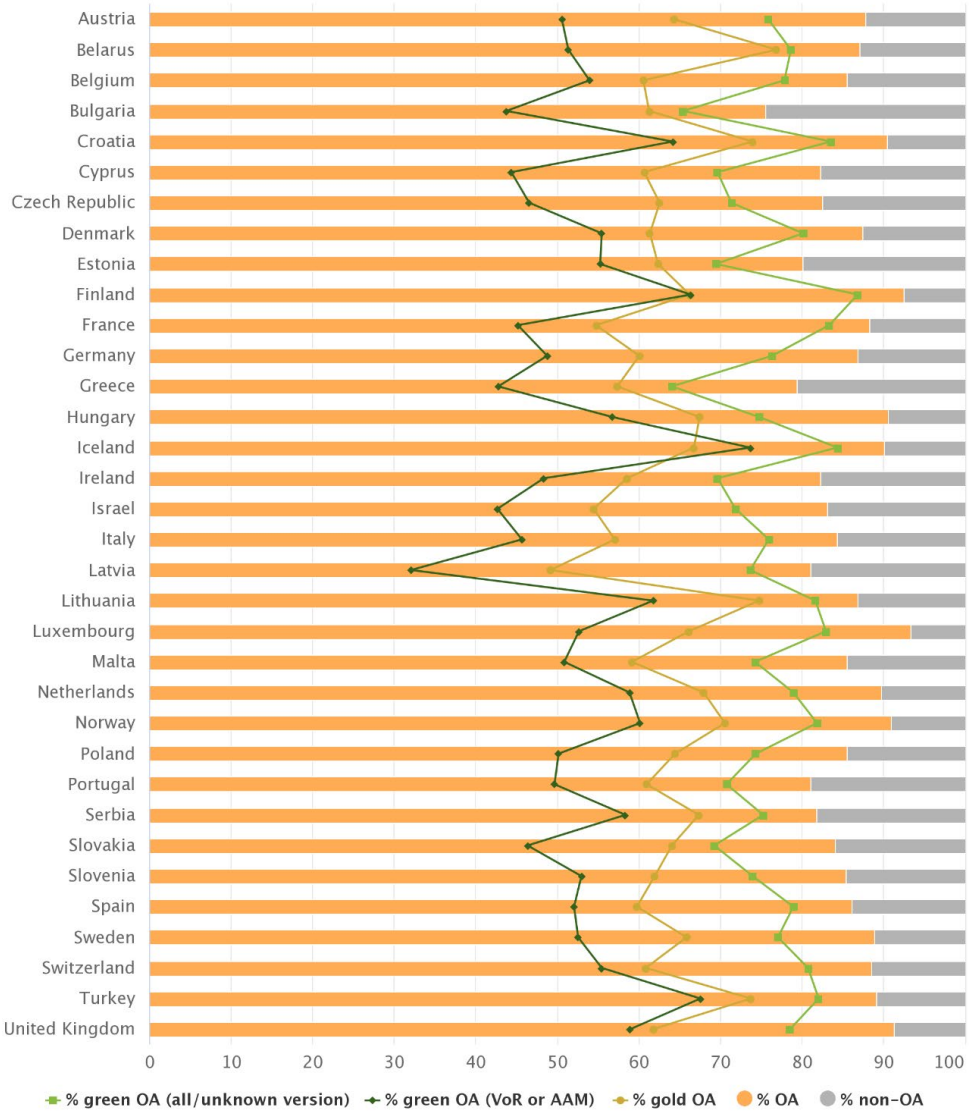



Figure 9: Open access rate and routes, by country of author

### *3.1.2.1 Horizon 2020 programmes*

Table 4 and Figure 10 illustrate the Horizon 2020 open access rate as aggregated to specific programmes (level 2).

The Excellent Science pillar has led the open access success story, with an open access rate of 86%. Of the leaders within this pillar are the European Research Council (ERC) and the Future and Emerging Technologies (FET) programme, with open access rates of over 88%. At the opposite end of the spectrum are the programmes 'Spreading excellence and widening participation' (with an open access rate of 70%) and Industrial Leadership (with 78% open access), both of which are lagging.

Of particular interest is the 'Leadership in enabling and industrial technologies' (LEIT) programme, under which roughly 10% of all peer-reviewed publications are produced, which has an open access rate of 79%. This could be a sign of differences in research environments in academia vs. industry/small and medium enterprises (SMEs), and may potentially require more targeted approaches: researchers in academia work within the broader mechanisms of supporting libraries (repositories and infrastructure), whereas SMEs do not. This should be further explored in terms of awareness-raising and targeted promotion programmes aimed at industry.

**Table 4. Open access rate by programme/sub-programme**

| Programme/sub-programme | Number of peer-reviewed publications | Percentage open access |
|---|---|---|
| **Euratom** | 5,457[38] | 64.7% |
| Euratom – General | 4,909 | 63.9% |
| Indirect actions | 564 | 71.1% |
| **Excellent Science** | 99,253 | 86.3% |
| European Research Council (ERC) | 50,155 | 88.4% |
| Future and Emerging Technologies (FET) | 8,304 | 88.2% |
| Marie Skłodowska-Curie Actions | 37,814 | 84.3% |
| Research infrastructures | 8,061 | 84.2% |
| **Industrial Leadership** | 16,000 | 78.6% |
| Access to Risk Finance | 3 | 66.7% |
| Industrial Leadership General | 110 | 75.5% |
| Innovation in SMEs | 393 | 71.5% |
| Leadership in Enabling and Industrial Technologies (LEIT) | 15,754 | 78.6% |
| **Science with and for Society** | 306 | 83.7% |
| Accessibility and Use of Publicly-funded Research | 18 | 94.4% |

[38] For each, the numbers refer to distinct peer-reviewed publications, therefore publications in more than one sub-programme are *not* counted twice at programme level. Thus, the number of peer-reviewed publications for a programme is not necessarily equal to the sum of the publications in its sub-programmes.

| Programme/sub-programme | Number of peer-reviewed publications | Percentage open access |
|---|---|---|
| Citizens to Engage in Science | 39 | 79.5% |
| Gender Equality in Research | 20 | 65.0% |
| Governance for Responsible Research and Innovation | 136 | 83.8% |
| Improving Knowledge on Science Communication | 3 | 100.0% |
| Integrate Society in Science and Innovation | 110 | 83.6% |
| Potential Environmental, Health and Safety Impacts | 42 | 81.0% |
| Scientific and Technological Careers for Young Students | 55 | 90.9% |
| **Societal Challenges** | 24,849 | 83.2% |
| Climate and Environment | 3,910 | 85.8% |
| Energy | 3,319 | 82.0% |
| Food, Agriculture, Forestry, Marine and Bioeconomy | 4,580 | 84.0% |
| Health | 9,106 | 86.4% |
| Inclusive, Innovative and Reflective Societies | 1,455 | 79.5% |
| Secure Societies | 1,327 | 74.3% |
| Societal Challenges - General | 255 | 72.5% |
| Transport | 2,026 | 76.7% |
| **Spreading excellence and widening participation** | 5,454 | 70.6% |
| Access to International Networks for Excellent Researchers and Innovators | 5 | 100.0% |
| ERA Chairs | 905 | 72.5% |
| Spreading Excellence and Widening Participation – General | 9 | 88.9% |
| Teaming of Research Institutions | 977 | 86.1% |
| Twinning of Research Institutions | 3,605 | 66.0% |

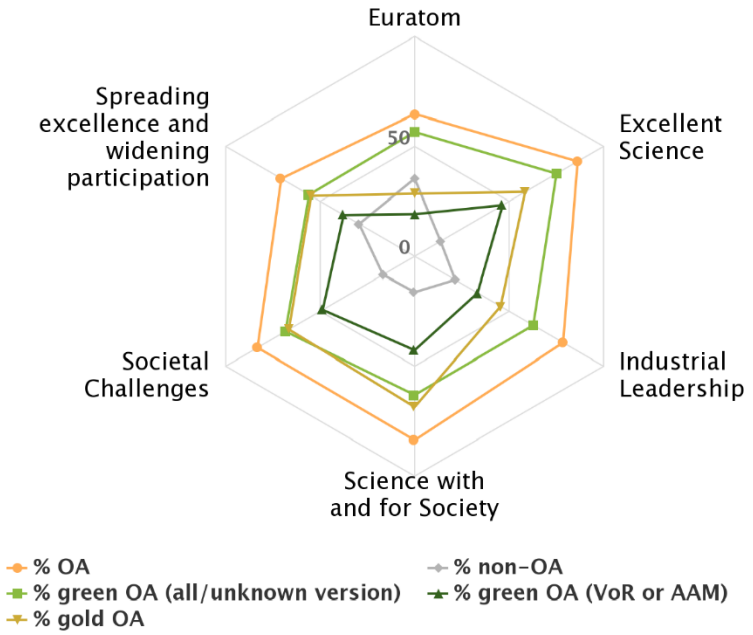



Figure 10. Open access rate and routes, by Horizon 2020 pillar

### *3.1.2.2 Scientific disciplines (FOS classification)*

Using our AI-based classification algorithm (see Section 7.3.2), we classified peer-reviewed publications into Frascati scientific disciplines (FOS levels 1 and 2). As indicated by the data in Table 5, open access rates for level 1 scientific disciplines range from 74.2% (agricultural and veterinary sciences) to 88% (medical and health sciences), indicating variations between different disciplines in terms of both uptake of Open Science and of the open access mandate.

The variations observed may be attributed to two facts: the long tradition of open access policies and the investment and use of infrastructures by the corresponding research communities, as it is the case for the medical and health sciences (88%) and natural sciences (83%). Engineering and technology, as well as social sciences, arts and the humanities (SSH) are at the low end of the spectrum, with 78% open access – a slower uptake, possibly reflecting a lack of community-building structures (e.g., Life sciences have a long history of openness in their practices and infrastructure).

**Table 5: Open access rate per scientific domain (Frascati Level 1)**

| FOS | PUBLICATIONS | PERCENTAGE OPEN ACCESS, HORIZON 2020 |
|---|---|---|
| Agricultural and veterinary sciences | 1140 | 74.2% |
| Engineering and technology | 22547 | 77.9% |
| Humanities and the arts | 1057 | 78.2% |
| Medical and health sciences | 33777 | 88% |

| | | |
|---|---|---|
| Natural sciences | 64519 | 82.8% |
| Social sciences | 5903 | 78% |

Additional differences within level 1 fields can be seen by examining open access rates in level 2, as illustrated in Figure 11.

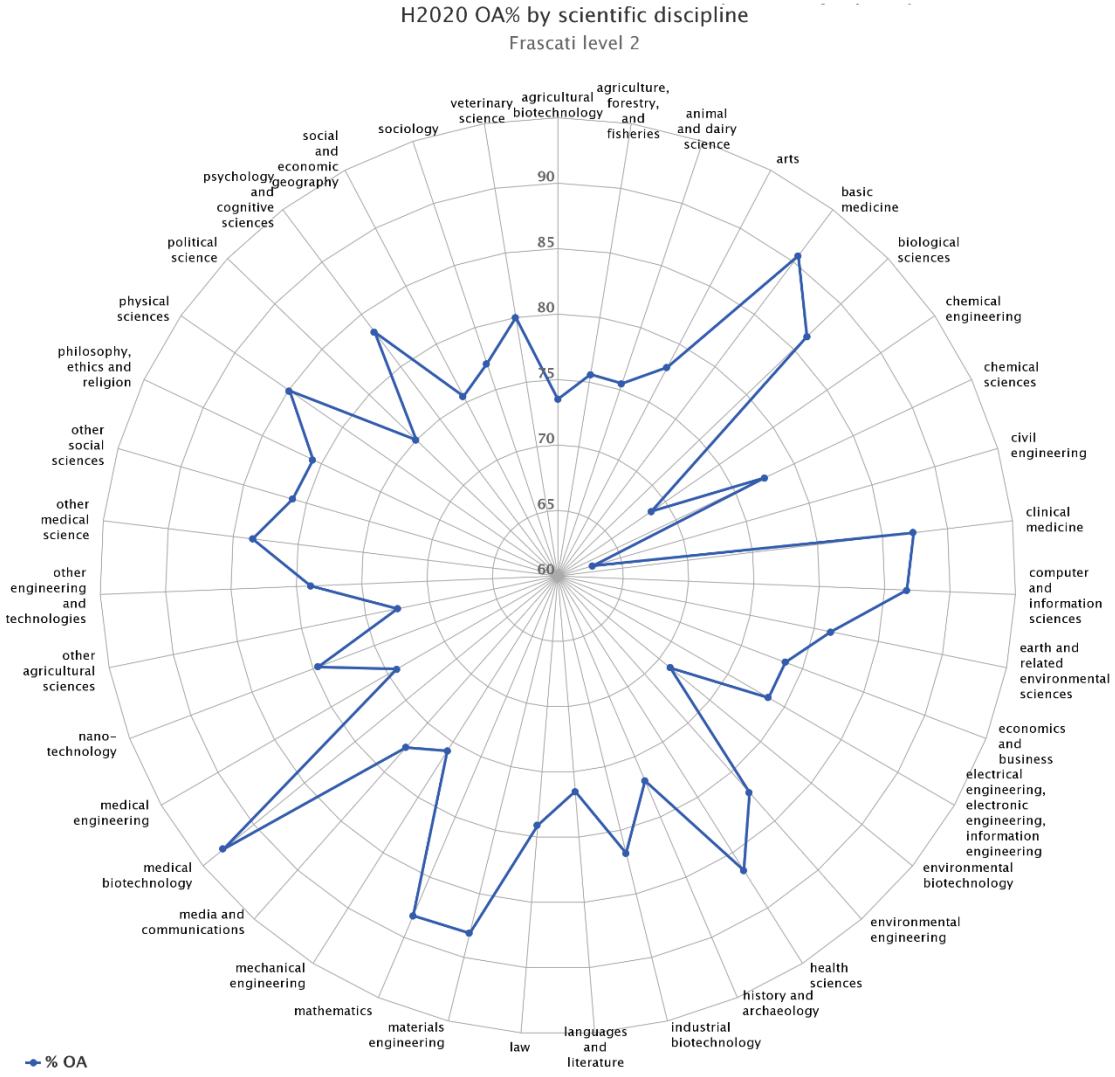



Figure 11: Open access rate, by Frascati level 2 classification

### *3.1.2.3 Publishing venues*

Our analysis yielded no surprises in terms of the publishers chosen by Horizon 2020 recipients. Looking at the 20 dominant publishers/learned societies (globally), 50% of publications are published by the three top publishers (Elsevier, Springer – Nature, Wiley), and 48% by the remaining 17.

**Table 6. Open access rate, by publisher**

| Top publishers (By percentage of Horizon 2020 publications) | Number of peer-reviewed Horizon 2020 publications | Percentage open access ('gold') |
|---|---|---|
| Elsevier | 23,700 | 70% |
| Springer – Nature | 23,392 | 83.5% |
| Wiley | 11,111 | 81.2% |
| MDPI | 7,259 | 100% |
| American Chemical Society | 7,236 | 73.8% |
| Institute of Physics Publishing | 6,570 | 87.3% |
| Royal Society of Chemistry | 4,505 | 81.3% |
| Institute of Electrical and Electronics Engineers (IEEE) | 4,477 | 72.3% |
| Frontiers Media | 3,277 | 100% |
| American Physical Society | 2,510 | 95% |

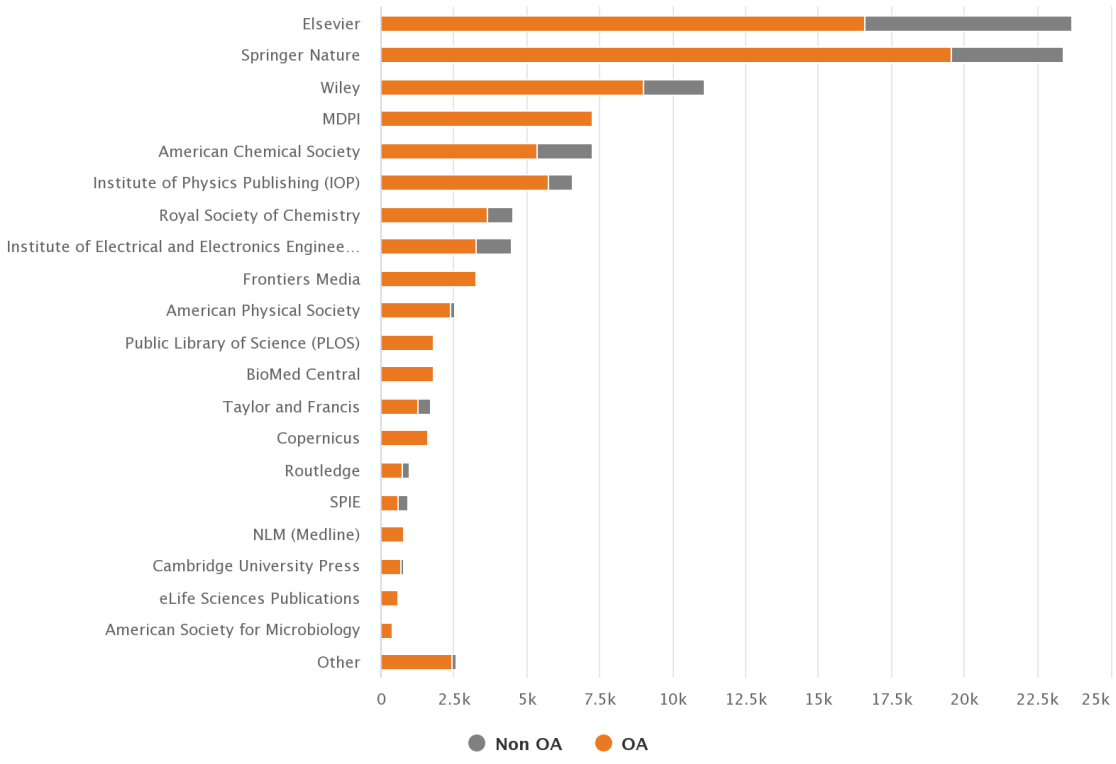



Figure 12: Open access rate for top 20 Horizon 2020 publishers

### *3.1.2.4 Licences*

Our findings show that **125,806 peer-reviewed publications contain a licence** in their metadata (out of a total of 154,185 publications), a rate of **81.6%**.

**Of these, 85% use a Creative Commons (CC) licence or one of the four licences from major publishers,** as illustrated below (see Table 9 for more information and links to the licences). The use of proprietary publisher licences indicates a problem with the mandate, as these licences are all "re-use" licenses that are not considered open. Restrictions include: a paywall, limited API requests per year, limitations on sharing results, among others.

**Table 7: Most common licence types among Horizon 2020 peer-reviewed publications**

| Licence type[39] | NUMBER OF PUBLICATIONS |
|---|---|
| **CC** | **78,086** |
| CC-0[40] | 480 |
| CC-BY | 55,518 |
| CC-BY-SA | 319 |
| CC-BY-NC | 5,149 |
| CC-BY-NC-SA | 1,148 |
| CC-BY-ND | 342 |
| CC-BY-NC-ND | 15,130 |
| **Publisher licences** | **51,758** |
| Elsevier-tdm | 27,726 |
| Springer-tdm | 12,394 |
| IOPscience-tdm | 6,729 |
| Wiley-tdm | 4,909 |
| **No licence** | **28,379** |

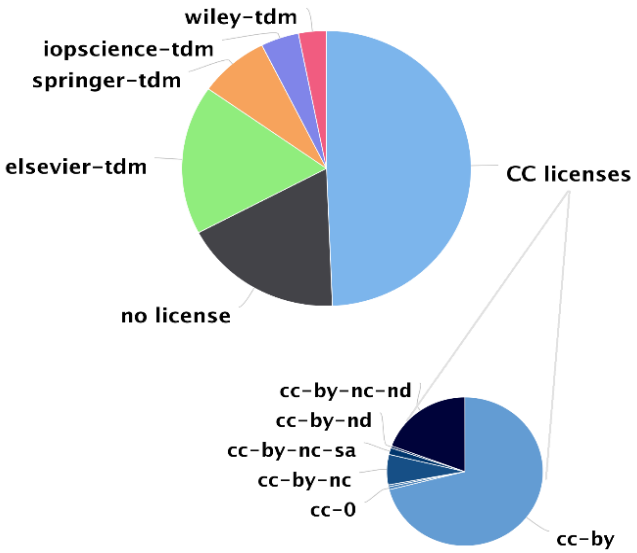



Figure 13: Distribution of license types among Horizon 2020 peer-reviewed publications

The following figures provide insights concerning the distribution of CC licences.

- We observe a **steady increase in the use of the most CC-BY and CC-BY-NC licenses since the start of Horizon 2020,** as Figure 14 shows. The growth in their use reflects the increase in Horizon 2020 open access publications each year. The use

[39] A publication may possess more than one licence.
[40] CC-0 is a public domain dedication and technically not a license, however, it is still of interest in the discussion, we include it and, as is the common practice, refer to it as a license.

of the two other most present CC licenses, CC-BY-NC-ND and CC-BY-NC-SA has been slightly decreasing over time (as a share of total publication with those licenses).

- The more permissive licence CC-BY is used consistently across all FOS level 1 domains (Figure 15), taking into account the different proportions of open access publications produced in different disciplines. A similar relationship exists in CC distribution across Horizon 2020 programmes: **CC-BY is the most consistently used option.**

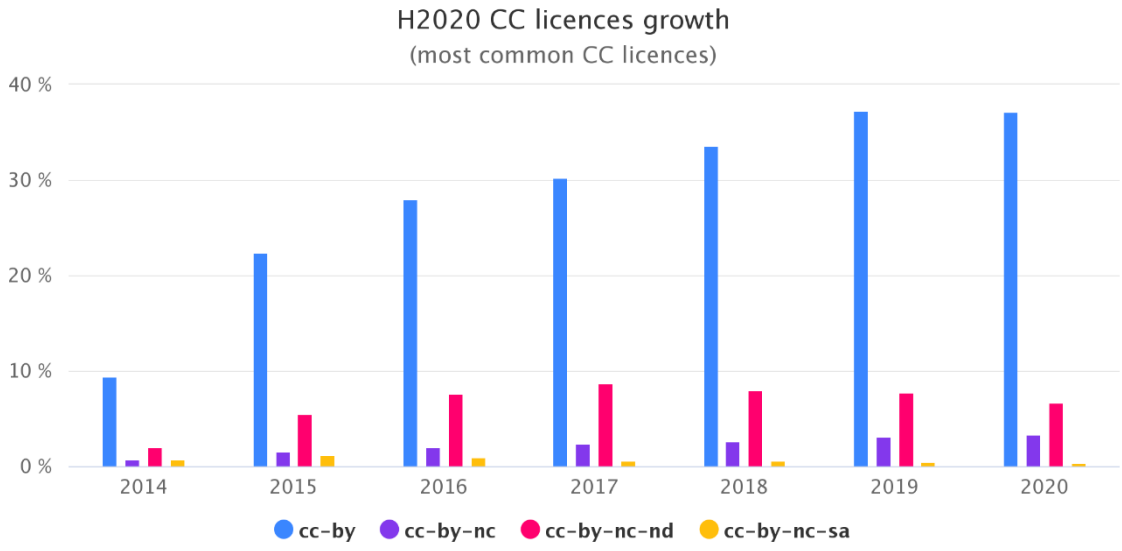



Figure 14. Growth in Creative Commons licences over Horizon 2020 lifespan

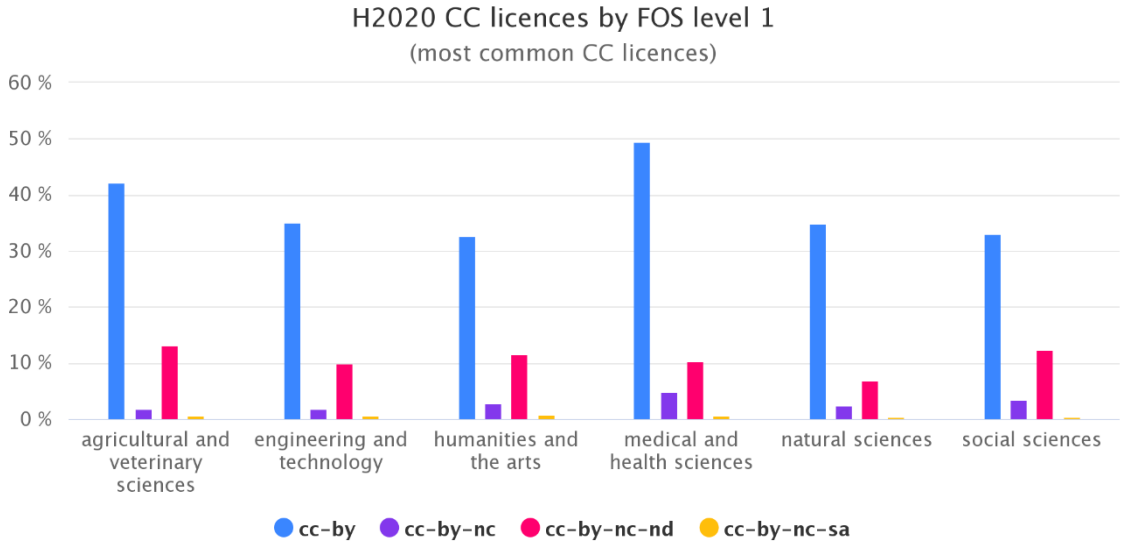



Figure 15. Creative commons Licences, by FOS level 1 classification

Table 8 provides explanations of the various CC licences and their permissiveness/restrictions.

**Table 8. Creative commons licences, by permissiveness**

| CC licence | Permission to produce adapted material | Permission to share licenced material | Permission to mine licenced material for commercial purpose | Permission to share adapted material |
|---|---|---|---|---|
| CC-BY | yes | yes | yes | yes |
| CC-BY-SA | yes | yes | yes | yes |
| CC-BY-NC | yes | yes | no | yes |
| CC-BY-NC-SA | yes | yes | no | yes |
| CC-BY-ND | yes | yes | yes | no |
| CC-BY-NC-ND | yes | yes | no | no |

Source: https://wiki.creativecommons.org/wiki/Content_mining

**Table 9: Licences of most common publishers**

| Licence | URL |
|---|---|
| **Elsevier-tdm** | https://www.elsevier.com/tdm/userlicence/1.0/ |
| **Springer-tdm** | http://www.springer.com/tdm |
| **Wiley-tdm** | http://doi.wiley.com/10.1002/tdm_licence_1.1 |
| **IOPscience-tdm** | http://iopscience.iop.org/info/page/text-and-data-mining |

### *3.1.2.5 Metadata and 'FAIRness'*

This Section addresses: (i) compliance of specific metadata elements with Article 29.2 of the MGA (funding, date, embargo, PID) in repositories; and (ii) metadata openness and completeness, as proposed by the library community. Both of these were assessed in relation to the OpenAIRE Guidelines,[41] which include all metadata elements necessary for the exchange and efficient monitoring of open access, including those that are required by the Horizon 2020 open access policy.

**Openness of metadata:** all repositories ingested in OpenAIRE provide **open access to the bibliographic metadata** that identify deposited publications. Therefore, the corresponding requirement set out in Article 29.2 is satisfied by all peer-reviewed scientific publications deposited in a repository in the MOAP Horizon 2020 database.

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

**Completeness / compliance of metadata:** using the OpenAIRE Validator,[42] a tool that has been developed over the past 10 years (see Section 7.3.2), we devised a scoring mechanism for average metadata completeness, for the purpose of this study. This is only a rough indicator, as the OpenAIRE Guidelines include both mandatory and optional rules. A score of 100 means compliance with all mandatory rules. The current numbers indicate a low level of compliance level among both repositories and publishers: **only 2% of repositories have an average score above 70** (Figure 16). This means that institutions need to invest in their repositories, both in terms of technology and personnel for curation, in order to address the quality of metadata. Even though OpenAIRE uses internal transformations in its aggregation workflows to overcome this issue, the repository community should intensify efforts to comply with the rules set out by funders (Horizon Europe and possibly the cOAlitiion S in their "Plan S Implemeent Guidance"[43]), as well as to the interoperability guidelines proposed by the European Open Science Cloud initiative (EOSC).

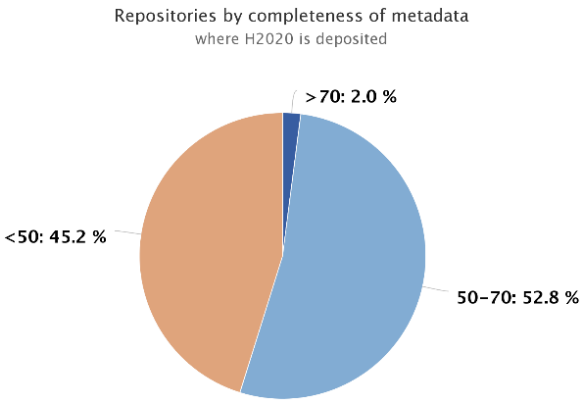



Figure 16. Repositories by completeness of metadata

**Funding information in metadata:** we identified 69,917 'green' open access publications (43,225 of them tagged as VoR/AAM), of which 24,889 include funding information in their metadata (14,556 among those tagged as VoR/AAM).[44] This amounts to a low success rate of 35% in the implementation of the policy, clearly signalling that more work is needed in the future to track the policy. Repositories are key as they must apply global standards such as the OpenAIRE Guidelines and connect to emerging institutional or national current research information systems (CRIS). Figure 17 breaks down this indicator by scientific disciplines. We note that medical and health sciences, that have one of the highest numbers of peer-reviewed publications are the least likely to reference the grant in the deposited publication metadata.

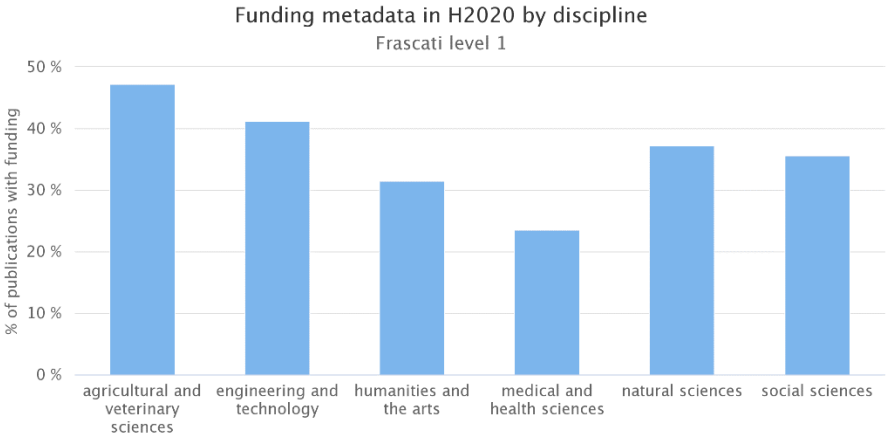



Figure 17. Funding information in metadata, by FoS

**Date information in metadata:** repositories have a good track record when it comes to maintaining the date of publication, as it is present for around 99% of 'green' open access publications).

[42] https://provide.openaire.eu
[43] https://www.coalition-s.org/plan_s_principles/
[44] These numbers do not include publications resulting from ERC grants, as these do not require this policy specificity. The ERC requires only a PID in the repository metadata. All indicators have been adjusted accordingly. However, in Table 12 we provide the indicators for ERC grants as well.

**Embargoes:** the number of publications under embargo has been hard to capture, as: (i) the embargo periods reported in the European Commission database are not reliable; and (ii) repositories update this information over time, so it is unreliable to examine a current snapshot, as original dates and access rights may have changed. Nevertheless, we can still report some values in relation to embargoes (although these are insufficient for a satisfactory assessment of the policy):

- Out of the 43,225 'green' deposits with a VoR/AAM, we observe only 1,809 that contain an embargo end date in their record[45]), a very small share of peer-reviewed scientific publications.

**PID in repository metadata:** As illustrated in Table 10, most repositories have already established a policy of requiring PIDs in publication metadata, which indicates 95% compliance of publications in that respect.[46]

**Table 10. Number of PIDs in institutional and thematic repositories**

| PIDs | Number of publications | Number of publications in repositories |
|---|---|---|
| Digital Object Identifier | 151,527 | 100,367 |
| PubMed ID | 37,648 | 36,067 |
| PubMed Central ID | 36,469 | 35,807 |
| arXiv | 27,512 | 27,466 |
| Handle | 25,291 | 19,820 |
| URN | 17,905 | 6,078 |
| Bibcode | 40 | 40 |
| ORCID workid | 4 | 0 |

### *3.1.2.6 Timeliness for deposition*

The timeliness for deposition has not been assessed in this study, as repositories do not expose metadata on the original deposition date or the original access rights of publications.

### *3.1.2.7 Accessibility and interoperability*

'FAIRness' is equally important in relation to publications as it is for data, since publications constitute corpora for knowledge extraction via natural language processing

[45] Most embargo end dates in the database are in the past (i.e. the embargoes have expired)
[46] The second column of the table refers to the number of publications that have *at least* one instance of that PID type in one of their metadata records. The third column of the table displays the same number *but only for metadata records fetched from repositories.*

(NLP)/machine learning. To define indicators on aspects of 'FAIRness', we used the following definitions:

- a publication is accessible if the text file can be fetched via a valid URL in its metadata,[47] while
- a publication is interoperable if the fetched file is in a machine-readable format.

We verified accessibility by fetching the PDF file of each publication.[48] Table 11 indicates accessibility and interoperability by type of data source available. Unsurprisingly, journals perform better than repositories as, in principle, the former apply tighter guidelines for the (full text) URLs included in metadata. Section 7.3 describes in detail how we assessed the accessibility and interoperability of publications.

**Table 11: Accessibility, by type of data source**

| DATA SOURCE TYPE | NUMBER OF PUBLICATIONS WITH VALID URLs | NUMBER OF ACCESSIBLE PUBLICATIONS | AVERAGE OF SHARE PUBLICATIONS ACCESSIBLE |
|---|---|---|---|
| **CRIS system** | 7,336 | 1,971 | 26.9% |
| **Institutional repository** | 66,660 | 43,229 | 64.8% |
| **Institutional repository Aggregator** | 379 | 121 | 31.9% |
| **Journal** | 128,553 | 87,031 | 67.7% |
| **Journal aggregator/publisher** | 1,695 | 1,121 | 66.1% |
| **Publication catalogue**[49] | 72 | 35 | 48.6% |
| **Publication repository** | 9,563 | 5,325 | 55.7% |
| **Publication repository aggregator** | 24,751 | 16,073 | 64.9% |
| **Thematic repository** | 57,683 | 28,396 | 49.2% |

To conclude these representative findings of our compliance analysis, we present a full list of indicators and their average values in Table 12 below. This list was validated by experts at the Validation Workshop we conducted as part of this study.[50] The last column addresses quality issues/concerns about the indicators.

[47] I.e. if a publication does not include a valid URL, we cannot assess accessibility.
[48] As described also in Section 7.3.2, the current version of our software is able to fetch full texts in PDF file formats only, thus missing the availability of other files in other formats. Thus, although, PDF is the most commonly available file format for publications, the numbers presented depict the lower bound of accessibility and interoperability for Horizon 2020 publications. Moreover, scanned PDF documents although not ideal are still machine-readable with the availability of OCR (optical character recognition) tools on the market.
[49] https://explore.openaire.eu/search/advanced/dataproviders?datasourcetypeuiname=%22Publication%2520Catalogue%22
[50] The workshop took place on 30 March 2021 and is briefly described in Section 5.

**Table 12. Horizon 2020 publication indicators**

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| **CONTEXT** | | | |
| **1. Publications** | Number of peer-reviewed publications linked to Horizon 2020 projects<br>(Number of non-peer-reviewed publications) | 154,185<br>(64,373) | **Peer-Reviewed publications across data sources:**<br>SyGMa 111,343 (72.2%)<br>Scopus: 121,571 (78.9%)<br>WoS: 115,518 (74.9%)<br>OpenAIRE: 152,211 (98.7%) |
| **2. Co-funded publications** | Number of publications with more than one funder *(hereafter, publications are assumed to be peer-reviewed)* | 20,869 | |
| | Share of the total number of publications *w/ valid number of funders* (n=152,211) | 13.7% | |
| | Number of publications linked to more than one project *(of European Commission or other funder)* | 39,331 | |
| | Share of the total number of publications *w/ valid number of projects* (n=152,211) | 25.8% | |
| **3. Co-authored publications** | Number of co-authored publications, by number of authors | 2-4 authors: 52,018<br>5-10 authors: 65,131<br>> 11 authors: 29,202 | |
| | Shares of the total number of publications *w/ valid number of authors* (n=152,211) | 2-4 authors: 34.2%<br>5-10 authors: 42.8%<br>> 11 authors: 19.2% | |
| | Number of publications with at least one author with an ORCID iD | 43,018 | |

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| **4. Publications with at least one ORCID identifier** | Share of the total number of publications (n=154,185) | 27.9% | |
| **OPEN ACCESS** | | | |
| **5. Publications by best available[51] access rights -** *open access, embargo, restricted, closed* | Number of publications by type of access rights | open access: 128,123<br>embargo: 345<br>restricted: 242<br>closed: 25,475 | Restricted is when as access to the article is not behind a paywall, but access is still restricted to certain users.<br><br>**Quality:** content providers do not expose data on the original access rights of a publication. Thus, it is only possible to know the access rights of the *last updated version,* and not if a publication that is open access today was originally embargoed. |
| | Share of the total number of publications *with valid access rights* (n=154,185) | open access: 83.1%<br>embargo: 0.2%<br>restricted: 0.2%<br>closed: 16.5% | |
| **6. 'Green' open access publications (any version of the manuscript)[52]** | Number of 'green' open access publications (open access publications deposited in a repository[53,54]) *without constraint on the version of document* in the repository. | 104,610 | |
| | Share of total number of publications (n=154,185) | 67.8% | |
| **7. 'Green' open access publications[55]** | Number of 'green' open access publications (open access publications w/ VoR or AAM deposited in a repository)[56] | 64,671 | **Quality:** coverage of the version of the manuscript in repository metadata for open access publications (~88.8%). |
| | Share of total number of open access publications deposited in a repository *w/ valid data on the version of the publication* (n=92,892) | 69.6% | |

[51] Across all instances of a publication.
[52] We do not provide here the Unpaywall 'Green' open access numbers, as these give priority to 'gold' publications over 'green' (see Section 7.3.2).
[53] The definition of a repository can be found in the Annex.
[54] Includes embargoed publications (open access after an embargo period is over).
[55] Includes immediate and delayed (embargoed) open access to the publication.
[56] Hereafter, 'green' open access only refers to VoR or AAM deposited in a repository.

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| **8. 'Gold' open access publications** | Number of publications with open access provided by the publisher | 86,767<br>(30,876 of those [35.6%] are hybrid open access) | |
| | Share of total number of publications (n=154,185) | 56.3% | |
| **9. 'Gold' open access publications that are *not* also 'green'** | Number of 'gold' open access publications that are not also 'green' (VoR or AAM). | 34,003 | |
| | Share of total number of 'gold' open access publications (n=86,767) | 39.2% | |
| | Number of 'gold' open access publications that are not also 'green' (*without constraint on version of document*) | 15,693 | |
| | Share of the total number of 'gold' open access publications (n=86,767) | 18.1% | |
| **COMPLIANCE DATES** | | | |
| **10. Publications w/ timely deposition in the repository** | Number of publications deposited in a repository *by the date of publication* | N/A | Almost zero coverage (17 publications) of dates deposited in a repository. It is not a metadata element commonly exposed by repositories. |
| | Share of the total number of 'green' open access publications | N/A | |
| **11a. Publications w/ timely open access in the repository[57] - non-ERC grants** | Number of **Social Sciences and Humanities** 'green' open access publications with embargo end dates **within 12 months of the publication date**[58] | 38 | **Quality:** it is not possible to judge the quality of this indicator, as repositories do not expose metadata on the original access rights of a publication. In other words, we cannot know the full set of |
| | Share of the total number of 'green' open access publications with valid publication date and embargo end date, *without constraint on version of document* (n=43) | 88.4% | |

[57] All *non-embargoed* 'green' open access publications have immediate open access.
[58] ERC grants may have a longer embargo period (see Section 7.1.1 in the Annex for ERC specificities).

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| | Number of 'green' open access publications (**other scientific domains**) with embargo end dates **within 6 months of the publication date** | 1,193 | originally embargoed publications in order to assess whether the availability of embargo end dates is sufficient.<br><br>**Also:** The last two forms of the indicator rely on the date of deposition in a repository, which has almost zero coverage. |
| | Share of the total number of 'green' open access publications with valid publication date and embargo end date (n=1,435) | 83.1% | |
| | Number of 'green' open access publications that are also 'gold' (open access by publisher), and are deposited by the date of publication | N/A | |
| | Share of the total number of 'green' open access publications | N/A | |
| **11b. Publications w/ timely open access in the repository[59] - ERC grants** | Number of **Social Sciences and Humanities** 'green' open access publications with embargo end dates **within 12 months of the publication date**[60] | 24 | |
| | Share of the total number of 'green' open access publications with valid publication date and embargo end date, *without constraint on version of document* (n=24) | 100% | |
| | Number of 'green' open access publications (**other scientific domains**) with embargo end dates **within 6 months of the publication date** | 962 | |
| | Share of the total number of 'green' open access publications with valid publication date and embargo end date (n=1,140) | 84.4% | |
| | Number of 'green' open access publications that are also 'gold' (open access by publisher), and are deposited by the date of publication | N/A | |
| | Share of the total number of 'green' open access publications | N/A | |

[59] All *non-embargoed* 'green' open access publications have immediate open access.
[60] ERC grants may have a longer embargo period (see Section 7.1.1 in the Annex for ERC specificities).

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| **METADATA REQUIREMENTS**[61] | | | |
| **12a. Publications with publication date in the repository metadata - <u>non-ERC grants</u>** | Number of 'green' open access publications with the publication date included in the repository metadata | 42,762 | |
| | Share of the total number of 'green' open access publications (n=43,225) | 98.9% | |
| | Number of 'green' open access publications *without constraint on version of document,* with the publication date included in the repository metadata | 68,594 | |
| | Share of the total number of 'green' open access publications (n=69,917) | 98.1% | |
| **12b. Publications with publication date in the repository metadata - <u>ERC grants</u>** | Number of 'green' open access publications with the publication date included in the repository metadata | 21207 | |
| | Share of the total number of 'green' open access publications (n=21,446) | 98.9% | |
| | Number of 'green' open access publications *without constraint on version of document,* with the publication date included in the repository metadata | 34024 | |
| | Share of the total number of 'green' open access publications (n=34693) | 98.1% | |
| **13a. Publications with embargo period in the repository metadata (out of total number embargoed) - <u>non-ERC grants</u>** | Number of 'green' open access publications with an embargo end date included in the repository metadata | 586 | **Quality:** It is not possible to judge the quality of this indicator, as repositories do not expose metadata on the original access rights of a |
| | Share of the total number of 'green' open access publications with *valid embargo end date in the record* (n=1,809) | 32.4% | |

[61] ERC grants are only required to provide a PID in the repository metadata.

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| | Number of 'green' open access publications *without constraint on version of document* with an embargo end date included in the repository metadata | 653 | publications. In other words, we cannot know the full set of originally embargoed publications in order to assess whether the availability of embargo end dates is sufficient. |
| | Share of the total number of 'green' open access publications *without constraint on version of document* with *valid embargo end date in the record* (n=2,129) | 30.7% | |
| **13b. Publications with embargo period in the repository metadata (out of total number embargoed) - ERC grants** | Number of 'green' open access publications with an embargo end date included in the repository metadata | 371 | |
| | Share of the total number of 'green' open access publications with *valid embargo end date in the record* (n=1,401) | 26.5% | |
| | Number of 'green' open access publications *without constraint on version of document* with an embargo end date included in the repository metadata | 415 | |
| | Share of the total number of 'green' open access publications *without constraint on version of document* with *valid embargo end date in the record* (n=1,789) | 23.2% | |
| **14a. Publications with proper funding reference[62] in the repository metadata - non-ERC grants** | Number of 'green' open access publications with project reference in the repository metadata | 14,556 | **Caveat:** only the grant number and/or acronym are usually found in the **repository metadata,** not the entire funding reference. |
| | Share of the total number of 'green' open access publications (n=43,225) | 33.7% | |
| | Number of 'green' open access publications *without constraint on version of document,* with a project reference in the repository metadata | 24,889 | |
| | Share of the total number of 'green' open access publications *without constraint on version of document* (n=69,917) | 35.6% | |

[62] The action, acronym and grant number; the terms ["European Union (EU)" and "Horizon 2020"];["Euratom" and Euratom research and training programme 2014-2018].

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| **14b. Publications with proper funding reference in the repository metadata - ERC grants** | Number of 'green' open access publications with project reference in the repository metadata | 3,882 | |
| | Share of the total number of 'green' open access publications (n=21,446) | 18.1% | |
| | Number of 'green' open access publications *without constraint on version of document,* with a project reference in the repository metadata | 6,524 | |
| | Share of the total number of 'green' open access publications *without constraint on version of document* (n=34,693) | 18.8% | |
| **15. Publications with PID (to the publication) in the repository metadata** | Number of 'green' open access publications with PID in the repository metadata | 61,538 | **Caveat:** DOIs have been cleaned as far as possible, but other PID types may be dirty. |
| | Share of the total number of 'green' open access publications (n=64,671) | 95.2% | |
| | Number of 'green' open access publications *without constraint on version of document,* with PID in the repository metadata | 99,178 | |
| | Share of the total number of 'green' open access publications *without constraint on version of document* (n=104,610) | 94.8% | |
| **16. Publications providing access to machine-readable copy via the repository** | Number of 'green' open access publications accessible via URL in the repository metadata | 49,355 | • 92,701: number of publications with at least one valid URL |
| | Share of the total number of 'green' open access publications with at least one valid URL in repository metadata (n=58,343) | 84.6% | • 66,835: number of publications w/ full text accessible via URL |
| | Number of 'green' open access publications *without constraint on version of document* accessible via URL in repository metadata) | 66,835 | • 15,459: number of publications w/ full text directly accessible via URL (i.e. link to PDF – for the rest, the site to which the URL linked was crawled for the PDF link) |
| | Share of the total number of 'green' open access publications *without constraint on version of document,* with at least one valid URL in repository metadata (n=92,701) | 72.1% | |
| **17. Publications with standard bibliographic** | Average best score per 'green' open access publication in repository metadata meeting the OpenAIRE guidelines | 48.04 | Description of the OpenAIRE Validation service and the |

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| **metadata** *(following OpenAIRE guidelines[63])* | (Average best score per 'green' open access publication *without constraint on version of document* in metadata, meeting the OpenAIRE guidelines) | (48.29) | resulting score can be found in Section 7.3.2 |
| **FAIR PRINCIPLES** (Here, we consider any record of a publication – not only those in repositories.) | | | |
| **18. (FAIR) findability** | Number of publications with a persistent identifier and a *valid* identifier to the full text (URI to PDF) in their metadata record | 112,731 | |
| | Share of the total number of publications (n=154,185) | 73.1% | |
| **19. (FAIR) accessibility** | Number of publications with the full text accessible via URL in metadata | 112,508 | • 1,626,089 URLs checked (1,529,033 ORG + 97,056 EC-Shared)<br>• 149,467 publications with at least one valid URL<br>• 112,508 publications w/ full text accessible via URL<br>• 54,114 publications w/ full text directly accessible via URL (i.e. link to PDF – for the rest, the site to which the URL linked was crawled for the PDF link) |
| | Share out of total # publications *w/ a valid URL in their metadata* (n=149,467) | 75.3% | |
| | Share out of total # of publications (154,185) | 73% | |
| **20. (FAIR) interoperability** | Minimum number of publications in a machine-readable text format (*These are the ones we were able to verify; we are agnostic as to the rest*.) | 112,508 | Athena RC's software verifies the accessibility of full-text PDFs. Therefore, to the extent that PDFs are machine-readable, the accessible publications are also interoperable. See Section 7.3.2 for details. |
| | Share of the total number of publications (n=154,185) | 73% | |
| **21. Publications with licences** | Number of publications with their **most permissive** licence being:<br>CC-0<br>CC-BY | <br>482 (0.4%)<br>55,787 (44.3%) | We normalised (cleaned and grouped) licences for 94% of those publications with an available licence. See Table 8 |

[63] OpenAIRE guidelines for content providers, to be used by repositories; open access journals; aggregators; CRIS (https://guidelines.openaire.eu)

| INDICATOR FOR HORIZON 2020 *PEER-REVIEWED* PUBLICATIONS | DEFINITION / TYPE OF INDICATOR | INDICATOR VALUE | NOTES AND QUALITY ISSUES |
|---|---|---|---|
| | CC-BY-SA<br>CC-BY-NC<br>CC-BY-NC-SA<br>CC-BY-ND<br>CC-BY-NC-ND | 218 (0.2%)<br>4,535 (3.6%)<br>862 (0.7%)<br>154 (0.1%)<br>12,344 (9.8%) | for a description of the various levels of permissiveness.<br>The share of the total number of publications for which a licence is available is given in parentheses (this may include invalid licences[64]). |
| **22. (FAIR) reusability** | Number of publications with licences: (a) allowing full text and data mining (TDM); and (b) allowing TDM only for non-commercial use | (a) 56,641<br>(b) 17,741 | |
| | Share of the total number of publications with at least one licence | (a) 45%<br>(b) 14.1% | |

[64] That is invalid entries in the license metadata field (e.g. instead of a license, a URL to the PDF file).

### *3.2 Analysis of publication costs*

The cost of open access publishing is an important aspect to consider when formulating the next line of policies and their implementation, particularly given that ongoing initiatives such as Plan S, Open Research Europe (ORE) and transformative agreements are potentially becoming more mainstream. However, identifying the cost of 86,767 **'gold' open access publications** was no trivial task, due mainly to the following factors:

- After close inspection, the European Commission data, which is an authoritative source for Horizon 2020, was found not to provide reliable figures, as they were not comparable with other sources. One explanation is that investigators or project coordinators are not aware of the actual costs, since payments go through an organisation's research office or the library.

- The most comprehensive source for APC costs is OpenAPC[65], the largest database of APCs paid by academic institutions and funders, with 122,999 entries. However, even this is far from comprehensive. One must also take into account that different prices may apply to different countries/institutions. There may be discounts for young researchers or editors; national funding from transformative agreements may also be used.

- Book processing charges not only vary considerably between publishers, but they also vary within the same publisher, based on various parameters including number of pages, curation, etc.

Based on the methodology described in detail in Section 0, we used OpenAPC to calculate Horizon 2020 publication costs by extrapolating costs from the 122,999 APC values (according to criteria that have been found to be significant in determining processing costs), while acknowledging that the aggregate data presented may only be considered an estimate that provides insights for further policy decisions. Our findings indicate that:

- The average APC cost for a Horizon 2020 publication is estimated to be **around 2,200 Euros for full open access journals.** Analysis of six large research funders showed that, on average, **APCs under Horizon 2020 were similar to the average for other funders** in Europe and USA (see Section 6.2).

- The average APC cost for a Horizon 2020 publication in **'hybrid' journals** is estimated to be **around 2,600 Euros**. 'Hybrid' open access articles[66] are, on average, more expensive than 'gold', indicating that removing their reimbursement under Horizon Europe could result in a significant reduction in cost to the European Commission.

- A ballpark figure for the overall cost of 'gold' publications under Horizon 2020 is **187 million EUR**; **however, it is *not* possible to identify who has borne this cost,** due to the existence of transformative agreements, institutional funds, and co-funded publications.

Of the programmes with the highest level of production of Horizon 2020 peer-reviewed publications, Health, Marie Skłodowska-Curie actions and Leadership in Enabling and Industrial Technologies (LEIT), have some of the lowest costs (around EUR 2,000), while Future and Emerging Technologies (FET) has the highest (around EUR 2,200) (Figure 20).

[65] https://openapc.net/
[66] That is articles published in 'hybrid' journals.

Table 13 presents some summary statistics of the Horizon 2020 publications and their overlap with OpenAPC data.

**Table 13. Overlap of Horizon 2020 publications APCs and BPCs with OpenAPC**

| SUMMARY STATISTICS ON HORIZON 2020 PUBLICATIONS APCs | |
|---|---|
| Number of 'gold' publications | 86,767 |
| Number of 'gold' non-book publications (i.e., excluding books and book chapters) – *including 'hybrid' open access* | 85,971 |
| Number of 'gold' non-book publications with APCs provided by OpenAPC | 4,423 (5.1% of 85,963) |
| Number of 'gold' non-book publications with *extrapolated* APCs (see Section 0) | 66,306 (77.1% of 85,963) |
| Average APC (extrapolated) for these **66,306** publications | 2177.5 EUR |
| Number of 'hybrid' non-book publications | 30,609 |
| Number of 'hybrid' non-book publications with APCs provided by OpenAPC | 1,564 (5.1% of 30,609) |
| Number of 'hybrid' non-book publications with extrapolated APCs (see Section 0) | 26,482 (86.5% of 30,609) |
| Average APC (extrapolated) for these **26,482** publications | EUR 2,557.30 |
| Number of 'gold' books | 148 |
| Number of 'gold' books with BPCs provided by OpenAPC BPC database | 5 |
| (Min, max) of BPCs for these **books** | (EUR 1,317.58, EUR 18,000) |
| Number of 'gold' book chapters | 787 |
| Number of 'gold' book chapters with BPCs provided by OpenAPC BPC database | 8 |
| (Min, max) of BPCs for these book chapters | (EUR 527.12, EUR 3,211.87) |

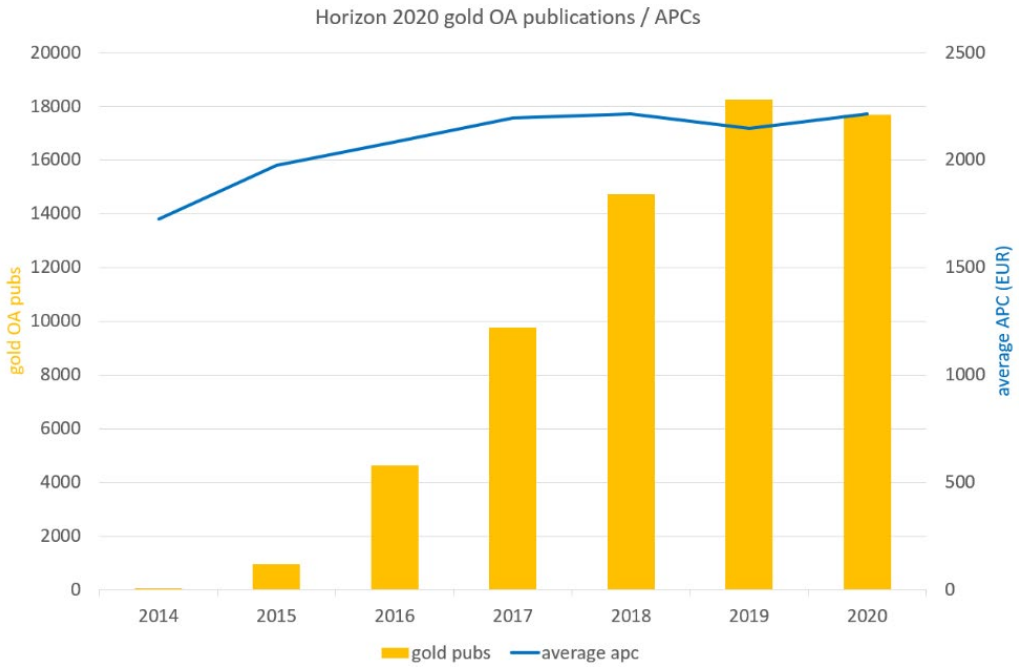



Figure 18. Average APCs by year, over the duration of Horizon 2020

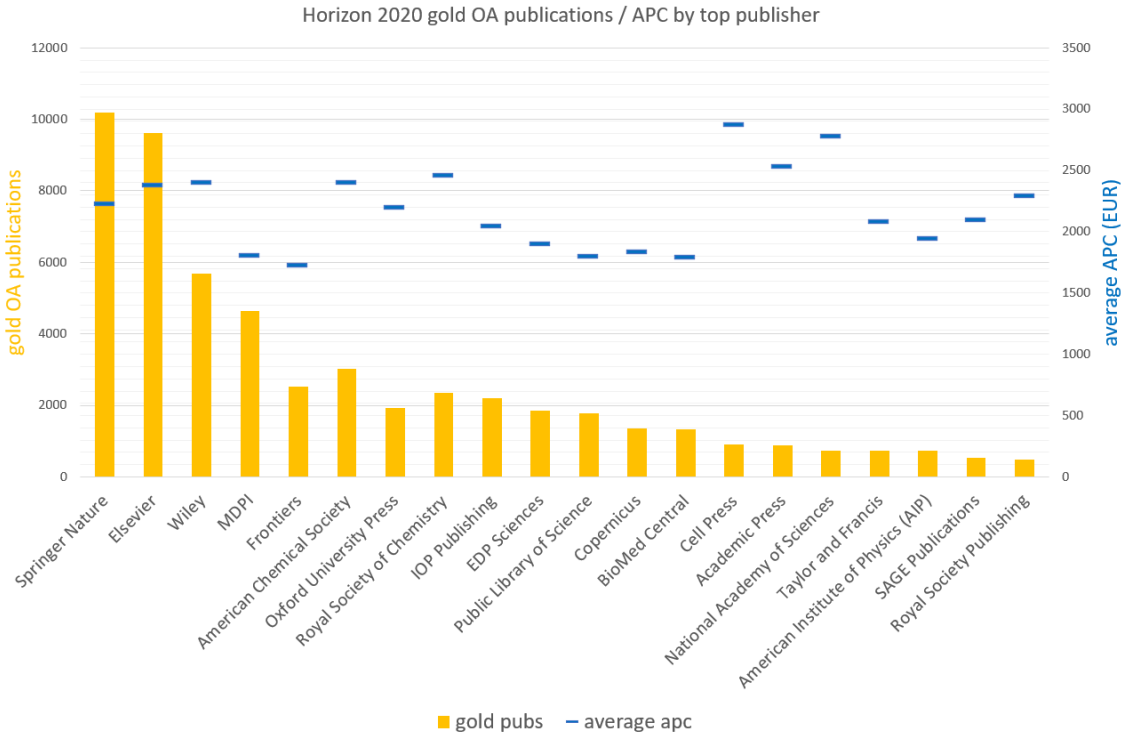



Figure 19. Average APCs under Horizon 2020, by publisher

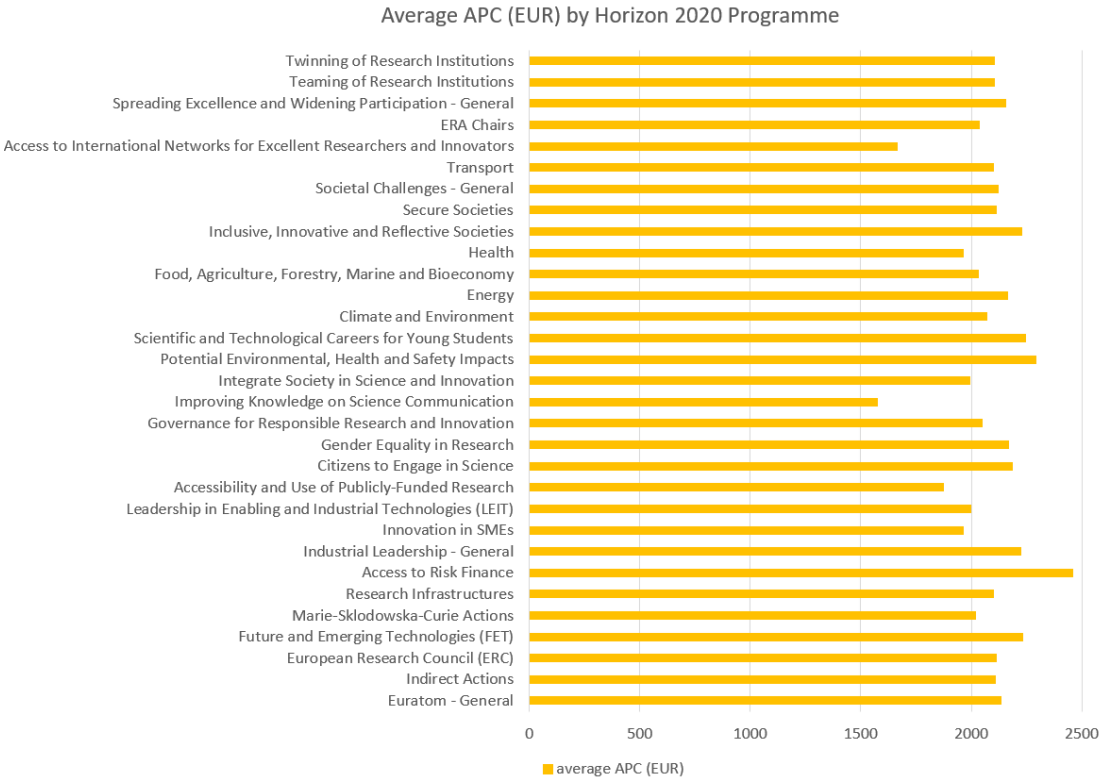



Figure 20. Average APCs per Programme

## ***Book processing charges***

Estimation of BPCs for Horizon 2020 books/book chapters has been a challenging task for various reasons:

- The BPCs reported in SyGMa were not reliable, indicating a lack of knowledge by researchers or project coordinators. Additionally, about 25% of book and book chapters were not reported to the EC.

- The underlying infrastructure (i.e., OAPEN, OpenAPC) does not systematically collect BPCs. Some preliminary efforts exist, but they are not mature enough to be used either as trusted data sources or are sufficient to extrapolate the values for the remainder of the books and book chapters (only 1.4% of Horizon 2020 'gold' books/chapters were in OpenAPC BPC database, see Table 13).

- Our desk research indicated that there is a broad range in the cost of books not only across publisher, but also within the same publisher (detailed in community papers[67,68] and evident by major publisher information pages), so we could not use indicative price lists for our estimation.

- Our communication efforts with major publishers to retrieve an authoritative list of DOI-price pairs, were not successful.

Even so, we considered using the OpenAPC BPC database to extrapolate BPCs based on the publisher alone. Table 14 presents a list of the publishers used by Horizon

[67] Frances Pinter, Why Book Processing Charges (BPCs) Vary So Much, The Journal of Electronic publishing, Volume 21, Issue 1, 2018, DOI: https://doi.org/10.3998/3336451.0021.101

[68] The Costs of Publishing Monographs: Toward a Transparent Methodology, N. Maron, K. Schmelzinger, C. Mulhern, D. Rossman, JEP Volume 19, Issue 1: Economics of Publishing, Summer 2016, DOI: https://doi.org/10.3998/3336451.0019.103

2020 recipients, broken down by the number of available 'gold' open access books or book chapters, and the corresponding number of BPCs available in the OpenAPC BPC database.[69] Springer Nature is by far the most prominent publisher in the list (with 438, of 55% of the 'gold' books or book chapters), however there is not enough data (15 Springer Nature BPCs) in the OpenAPC BPC database to conduct a meaningful extrapolation exercise.

**Table 14. Horizon 2020 books/book chapters by top publishers**

| **Top publishers** *(by number of books and book chapters in MOAP Horizon 2020)* | **Number of 'gold' books and book chapters in MOAP Postgres** | **Number of BPCs available in OpenAPC BPC DB** |
|---|---|---|
| **Springer Nature** | 438 | 15 |
| **Association for Computing Machinery** | 70 | 3 |
| **Elsevier** | 53 | 0 |
| **Wiley** | 29 | 0 |
| **Routledge** | 25 | 24 |
| **Society for Industrial and Applied Mathematics** | 15 | 0 |
| **Taylor and Francis** | 14 | 1 |

[69] Publisher names are cleaned and de-duplicated to the greatest extent possible, to cover almost all publications.

# 4 Open Research Data Pilot and open access to research data

## *4.1 Compliance and uptake analysis of Horizon 2020 datasets*

This section presents key aspects of the data analysis we conducted in relation to the Open Research Data Pilot (ORDP), which includes measures for compliance and uptake, as well as aspects of monitoring.[70]

- **Compliance** with Article 29.3[71] is measured by examining the datasets produced in projects that participated and *did not opt out of the ORDP* (hereafter referred to as 'ORDP projects').
- **Uptake** of the regulations set out in Article 29.3 focuses on compliance with the policy among datasets produced in *all* Horizon 2020 projects, i.e., including those that were not obliged to comply. Uptake is useful for comparing performance across both groups, although we note that projects that did **not** participate in the ORDP *cannot* report the datasets they produced in SyGMa, and may not have an incentive to deposit those datasets. Therefore, when measuring uptake, we are careful to note that it may not capture *all* datasets produced, but only those that are reported and those harvested by OpenAIRE (see Section 7.4).

Table 15 illustrates the reasons given by projects for opting out of the ORDP[72], and the number of projects that cited each reason. It is worth noting that half of all reasons given relate to intellectual property rights (IPR). Figure 20 presents a further breakdown by programme. Protection of results appears to be a valid reason for programmes such as LEIT and the SME instrument, but further exploration should be carried out on its use in relation to more academically targeted programmes such as MCSA (although host actions do aim at industry participation).

**Table 15. ORDP opt-out reasons**

| REASONS FOR OPTING OUT | NUMBER OF PROJECTS |
|---|---|
| To allow the protection of results (e.g. patenting) | *1,536* |
| Incompatibility with privacy/data protection | *455* |
| The project does not generate any data | *421* |
| Other legitimate reasons | *232* |
| Achievement of the project's main aim would be jeopardised | *230* |
| Incompatibility with the need for confidentiality linked to security | *167* |
| **Reason not available** | *11,617* |

[70] The database (https://zenodo.org/record/4899767) contains a host of metadata elements that can be used for additional analysis.
[71] The excerpt of Article 29.3 relevant to this study is presented in Section 7.2 of the Annex, and includes ERC specificities.
[72] Participation in the ORDP became the default with the Work Programme 2017. Since then, projects have been required to explicitly opt out if they do not wish to participate. Prior to the Work Programme 2017, participation in the pilot was the default only in some areas. In most other areas, projects could opt in. ERC projects may opt out of the ORDP at any point without providing a reason.

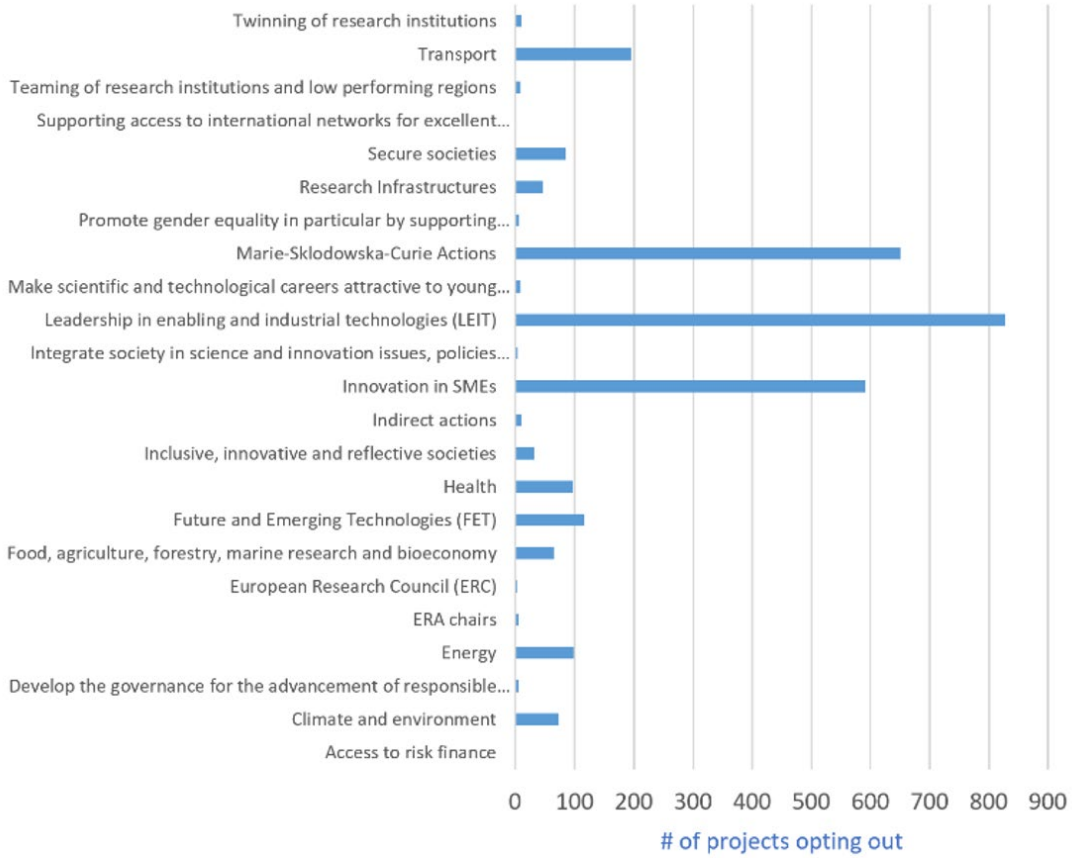



Figure 21. Horizon 2020 opt-outs, by programme (for projects with a recorded opt-out reason)

Based on the methodology described in Section 7.4, we combined the datasets reported to the European Commission and harvested by OpenAIRE to converge them into the MOAP database of Horizon 2020 datasets. The **overall Horizon 2020 open access rate is 94.8%**. A breakdown by years is presented in Table 16, and illustrated in Figure 22. We observe that the production of datasets has steadily increased over time, with the open access rate remaining consistently above 90%.

**Table 16. Open access compliance and uptake per year**

| | COMPLIANCE | | | | UPTAKE | | | |
|---|---|---|---|---|---|---|---|---|
| **Year** | *Number of projects in ORDP* | *Open access datasets* | *All datasets* | ***Per cent open access*** | *Total number of projects* | *Open access datasets* | *All datasets* | ***Per cent open access*** |
| **2015** | 686 | 64 | 64 | **100%** | 4,693 | 64 | 64 | **100%** |
| **2016** | 665 | 121 | 126 | **96%** | 4,930 | 136 | 142 | **95.8%** |
| **2017** | 818 | 333 | 362 | **92%** | 4,958 | 400 | 429 | **93.2%** |
| **2018** | 1,852 | 1,144 | 1,198 | **95.5%** | 5,047 | 1,290 | 1,347 | **95.8%** |
| **2019** | 2,396 | 1,415 | 1,469 | **96.6%** | 5,563 | 1,696 | 1,769 | **95.9%** |
| **2020** | 2,555 | 1,352 | 1,432 | **94.4%** | 4,482 | 1,694 | 1,923 | **92.9%** |

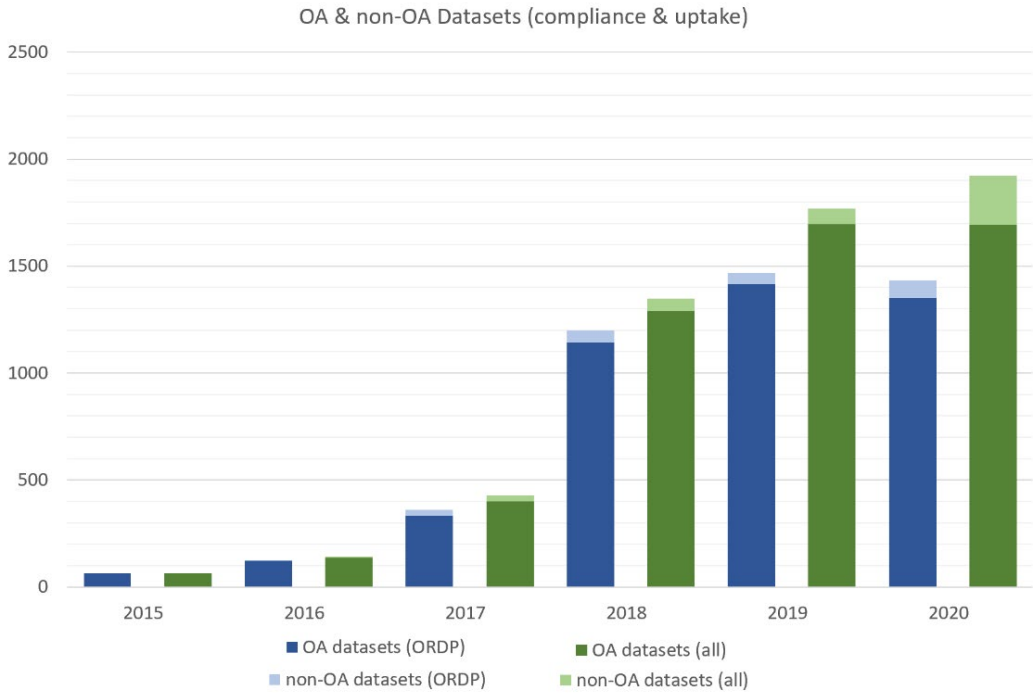



Figure 22. Open access compliance and uptake trends under Horizon 2020

However, these results should be viewed with caution. Since research data comes in many forms, and different research communities employ different practices (each with its own merits), the above figures present a 'slice' of the overall picture, not the whole. Even so, they are indicative of the levels of both compliance and uptake. The challenges and limitations affecting this include:

- The definition of what constitutes a 'dataset' is not the same across disciplines. For example, over recent decades, researchers in life sciences have built infrastructure and collected data in *databases* such as the Protein Data Bank, GenBank, etc. The same is true in earth sciences, where access to large volumes of sensor data is provided via APIs (e.g. SeaDataNet[73]). At the other end of the spectrum, we find depositions in repositories, whether institutional, national, thematic, etc. These repositories store files and make them accessible to end users via standard retrieval APIs. Table 17, below depicts, among others, the (largely) uninformative 'type' metadata element for datasets (97.8% of datasets are of the generic type 'dataset').

- The granularity of a dataset plays an important role when reporting. Packaging the components in the right way is important, so as to promote good practice when reporting.

- The provisions of Article 29.3 require the data underpinning a publication *("…data, including associated metadata, needed to validate the results presented in scientific publications…"*) to be made available in open access, but because data are created in a value chain, it is not always clear what to report.

[73] https://www.seadatanet.org/

- Some confusion exists over the term '*supplementary data*', as some publishers require all figures and tables to be deposited in repositories (e.g. Figshare[74]). This is considered data in machine crawling/harvesting.

- It is possible that in the case of datasets that do not need to be reported, i.e. from projects not in the ORDP, authors do not have other incentives to deposit them. For example, it is common for journals to require authors to submit their datasets to the journal, but they need not be deposited as well. This would also explain the small number of additional datasets from non-ORDP projects.

**Table 17. Open access compliance and uptake per dataset type**

| | COMPLIANCE | | | UPTAKE | | |
|---|---|---|---|---|---|---|
| **Type of dataset** | *Open access datasets* | *All datasets* | ***Per cent open access*** | *Open access datasets* | *All datasets* | ***Per cent open access*** |
| Audio-visual | 42 | 43 | **97.7%** | 45 | 46 | **97.8%** |
| Dataset | 4,438 | 4,668 | **95.1%** | 5,331 | 5,651 | **94.3%** |
| Film | 5 | 5 | **100%** | 9 | 9 | **100%** |
| Image | 96 | 96 | **100%** | 98 | 98 | **100%** |

The table below reveals that out of the large number of datasets we identified, only a small fraction (16.2%) is linked to publications (i.e. listed as citations in the body of the publication, or reported via infrastructures such as ScholExplorer, DataCite, OpenAIRE in the dataset metadata).

**Table 18. Open access compliance and uptake – linked publications**

| | COMPLIANCE | | | UPTAKE | | |
|---|---|---|---|---|---|---|
| | *Open access datasets* | *All datasets* | ***Per cent OA*** | *Open access datasets* | *All datasets* | ***Per cent OA*** |
| Datasets underpinning publications | 719 | 761 | **94.5%** | 870 | 918 | **94.8%** |
| All datasets | 4,538 | 4,769 | **95.2%** | 5,435 | 5,756 | **94.4%** |

[74] https://figshare.com/

#### 4.1.1.1 Open access to research data, by programme

Table 19, below, presents a breakdown according to the top Horizon 2020 programmes (in terms of the production of datasets). It reveals significant differences between them in terms of open access rates. The highest rate of open access for datasets from ORDP projects is found in Health, followed by Research Infrastructures and European Research Council (ERC) projects, under the Excellent Science pillar. At the opposite end of the spectrum, with an open access rate of 87.2%, are ORDP projects from the Inclusive, Innovative and Reflective Societies programme. In Figure 23, within the Societal Challenges pillar, we again see variation in open access rates for datasets from ORDP projects between programmes. Programmes under the Excellent Science pillar, by comparison, have more consistent open access rates throughout.

**Table 19. Open access compliance and uptake per Horizon 2020 programme**

| | COMPLIANCE | | | | UPTAKE | | | |
|---|---|---|---|---|---|---|---|---|
| ***Top Horizon 2020 PROGRAMMES*** *(by number of datasets produced)* | *Number of Projects in ORDP* | *Open access datasets* | *All datasets* | ***Per cent open access*** | *Number of all projects* | *Open access datasets* | *All datasets* | ***Per cent open access*** |
| *Leadership in Enabling and Industrial Technologies (LEIT)* | 1,367 | 1,184 | 1,244 | **95.2%** | 6,221 | 1,271 | 1,337 | **95.1%** |
| *Climate and Environment* | 282 | 954 | 991 | **96.3%** | 658 | 961 | 998 | **96.3%** |
| *Research Infrastructures* | 196 | 665 | 671 | **99.1%** | 309 | 718 | 729 | **98.5%** |
| *Marie Skłodowska-Curie Actions* | 4,275 | 402 | 428 | **93.9%** | 9,819 | 593 | 652 | **91.0%** |
| *European Research Council (ERC)* | 942 | 250 | 254 | **98.4%** | 6,646 | 614 | 627 | **97.9%** |
| *Food, Agriculture, Forestry, Marine and Bioeconomy* | 253 | 350 | 374 | **93.6%** | 840 | 393 | 424 | **92.7%** |
| *Future and Emerging Technologies (FET)* | 347 | 237 | 250 | **94.8%** | 516 | 268 | 283 | **94.7%** |
| *Inclusive, Innovative and Reflective Societies* | 261 | 218 | 250 | **87.2%** | 396 | 221 | 253 | **87.4%** |
| *Energy* | 319 | 110 | 118 | **93.2%** | 1,310 | 167 | 180 | **92.8%** |
| *Health* | 345 | 83 | 83 | **100.0%** | 1,070 | 113 | 132 | **85.6%** |

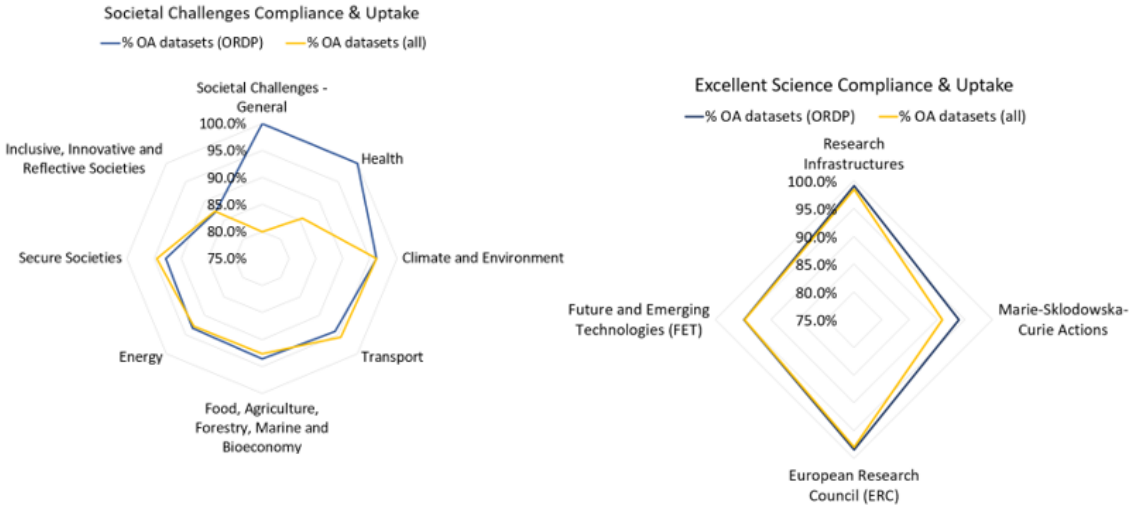



Figure 23. Open access compliance and uptake in the Societal Challenges programme

#### *4.1.1.2 Open access to research data, by scientific discipline*

Research data do not have in their metadata a subject field or other information that would allow us to identify disciplines and map them to known classifications such as Frascati. By extrapolating each publication's scientific discipline through linked publications (as obtained by OpenAIRE and DataCite), we were able to create a subset providing some insights into compliance/uptake within various different domains.[75] While rates of open access compliance and uptake are comparable between disciplines, the production of open access research data is more prominent in natural sciences, followed by engineering and technology, then medical and health sciences. Nevertheless, as the numbers are still fairly low, consistently linking publications to the datasets that underpin them would go along way towards helping to evaluate compliance and uptake across fields.

**Table 20: Open access compliance and uptake per scientific domain**

| SCIENTIFIC DOMAIN | COMPLIANCE | | | | UPTAKE | | | |
|---|---|---|---|---|---|---|---|---|
| *(using FoS of linked publications)* | *No. of linked pubs* | *Open access datasets* | *All datasets* | *Per cent open access* | *No. of linked pubs* | *Open access datasets* | *All datasets* | *Per cent open access* |
| Agricultural and veterinary sciences | 10 | 7 | 7 | 100% | 12 | 9 | 9 | 100.0% |
| Engineering and technology | 69 | 88 | 88 | 100% | 77 | 96 | 96 | 100.0% |

[75] 43% of linked publications (313 publications) fall into the (small) group of peer-reviewed publications that we were not able to classify into scientific domains. See Section 7.3.2 for the FOS classification methodology.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Medical and health sciences | 50 | 49 | 50 | 98% | 73 | 73 | 74 | 98.6% |
| Natural sciences | 202 | 207 | 213 | 97% | 239 | 250 | 260 | 96.2% |
| Social sciences | 17 | 17 | 17 | 100% | 22 | 23 | 23 | 100.0% |

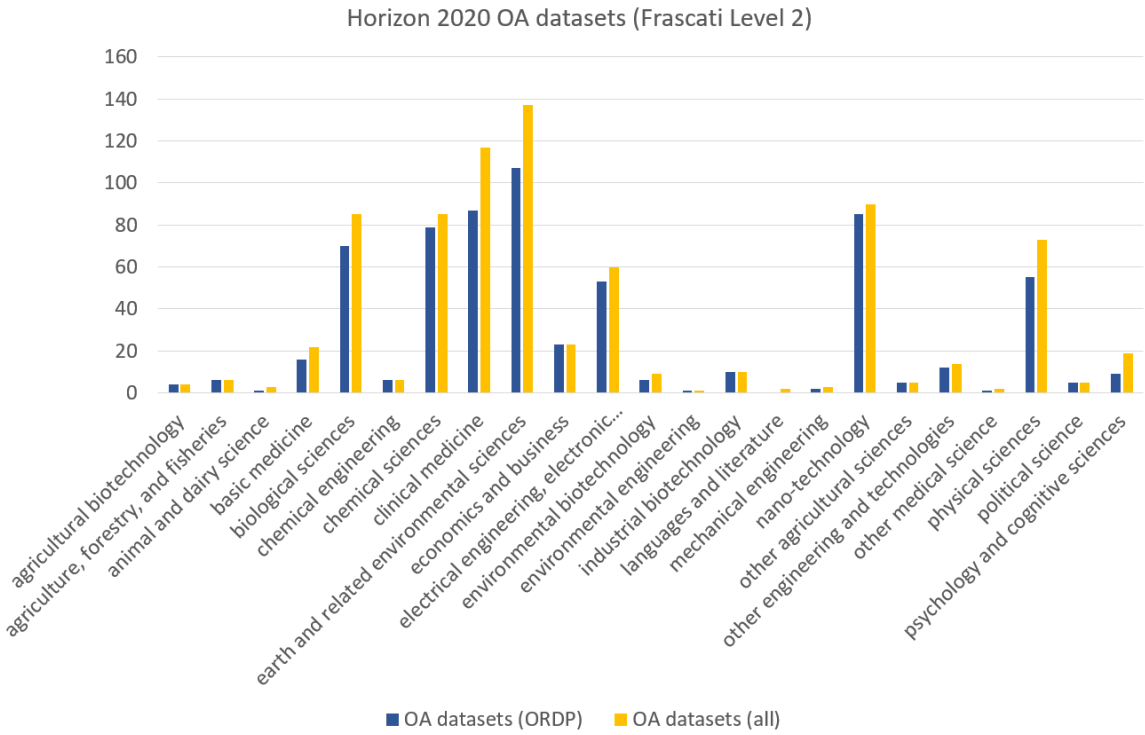



Figure 24. Production of Horizon 2020 open access datasets, by scientific discipline (Frascati Level 2)

#### *4.1.1.3 Analysis of metadata and 'FAIRness'*

We note that all repositories harvested by OpenAIRE provide **open access to the bibliographic metadata** that identify the deposited publications. Therefore, the corresponding requirement of Article 29.3 is satisfied for all datasets deposited in a repository in the MOAP database.

Article 29.3 requires datasets to be FAIR, with the metadata provided at the repository of deposition. As with publications, for the purposes of this study (and given the data available), we estimate here a '*lightweight*' version of FAIR data, constructing the indicators in the following way:

1. A dataset is **findable** if its metadata includes a PID of the dataset and a valid URI to the data file.

2. A dataset is **accessible** if the data file can be accessed (fetched) via a valid URL in its metadata.

3. A dataset is **interoperable** if the data file is in a machine-readable format.

4. A dataset is **reusable** if it has a Creative Commons (CC) licence in its metadata (we distinguish those that allow text and data mining for non-commercial use only).

As is required by Article 29.3, we also examine the metadata standards to which the deposited datasets conform. Our findings are presented below.

First, there is a good coverage of PIDs in repository metadata for Horizon 2020 datasets (87% of deposited datasets), indicating that this is a well-established practice within the community. Table 21 presents the availability of PIDs in the metadata of datasets.

**Table 21. Dataset PID availability**[76]

| | Number of datasets with PID in metadata | Number of datasets with PID in repository metadata (5,370 in repository) |
|---|---|---|
| **Digital Object Identifier** | 6,160 | 4,655 |
| **Handle** | 121 | 33 |
| **URN** | 21 | 9 |
| **Archival Resource Key** | 4 | 0 |
| **Open Archives Initiative** | 3 | 0 |

Overall, however, metadata standards and the provision of valid URLs is still lacking. In particular, the average validation score (i.e. whether a record meets with OpenAIRE guidelines for datasets, Section 7.4.2) is only 41.5 out of 100. Out of all datasets deposited in a repository, the share of datasets with valid URLs in the metadata is just 37.1%.

These findings indicate the following:

- **Findability:** only around 39% of deposited datasets are findable, due to the lack of a valid URI.
- **Accessibility and interoperability:** only around 32% of deposited datasets are accessible, due to a lack of valid URLs. Thus, the share of datasets for which we can assess interoperability is limited to this 32%, as we do not have access to the datafile for the remainder.

Figure 25 shows the number of datasets from ORDP projects that are findable, as well as those that are accessible and interoperable, for the top three repositories of deposition (covering 96.4% of all deposited datasets).

[76] The second column of the table refers to the number of datasets that have *at least* one instance of that PID type in one of their metadata records. The third column of the table displays the same number *but only for metadata records fetched from repositories.* Non-repository data sources for datasets include scholarly communication infrastructures and CRIS (current research information systems).

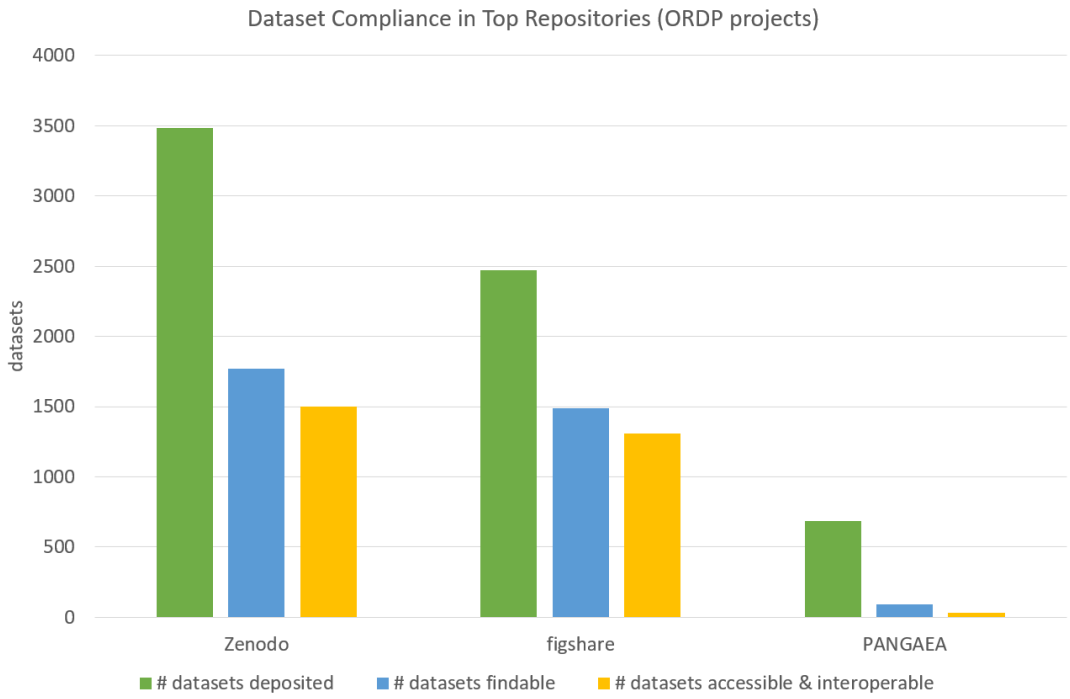



Figure 25. Dataset compliance in top repositories (findability, accessibility, interoperability)

When datasets produced by projects participating in the ORDP (compliance) are compared with datasets from all projects (uptake), the former perform better in all measures (see Table 23), indicating an extra effort to comply with the more technical aspects of Article 29.3.

***Reusability: licences***

Our findings show that **3,591 datasets include a licence** in the repository metadata, which indicates a minimum **compliance level of 66.9%**.[77] After cleaning and grouping, we identified 3,492 datasets with CC licences.[78] As shown in

Table 22, the majority of the licences are CC-BY and CC-BY-SA.

**Table 22. Licences at the repository of deposition**

| LICENCE | COMPLIANCE | UPTAKE |
|---|---|---|
| **CC-0** | 35 (0.8%) | 50 (0.9%) |
| **CC-BY** | 2,049 (46.7%) | 2,622 (48.8%) |
| **CC-BY-SA** | 587 (13.4%) | 615 (11.5%) |
| **CC-BY-NC** | 69 (1.6%) | 86 (1.6%) |
| **CC-BY-NC-SA** | 49 (1.1%) | 79 (1.5%) |
| **CC-BY-ND** | 5 (0.1%) | 7 (0.1%) |
| **CC-BY-NC-ND** | 48 (1.1%) | 57 (1.1%) |

[77] Minimum since these are the datasets that we could identify as re-usable, there are potentially more.
[78] The few remaining licences (just a few per type of licence) appear to be open on first inspection.

Figure 26, below, depicts datasets with and without CC licences in the repository metadata according to Horizon 2020 programme, for ORDP projects. One fact that stands out is that LEIT possesses the highest share of datasets without a CC licence (6.2%). We examine this further by presenting the types of licences for datasets stemming out for LEIT projects in the ORDP (Figure 27). Of these, datasets licenced under (i) a CC non-commercial licence or (ii) another type of licence, together with (iii) datasets without licences, add up to a significant portion of all datasets deposited (although CC-BY and CC-BY-SA are still the most prominent). As discussed previously, this could be the result of SMEs' participation in the projects; other potential causes could be community standards on licencing, as well as the metadata standards of repositories.

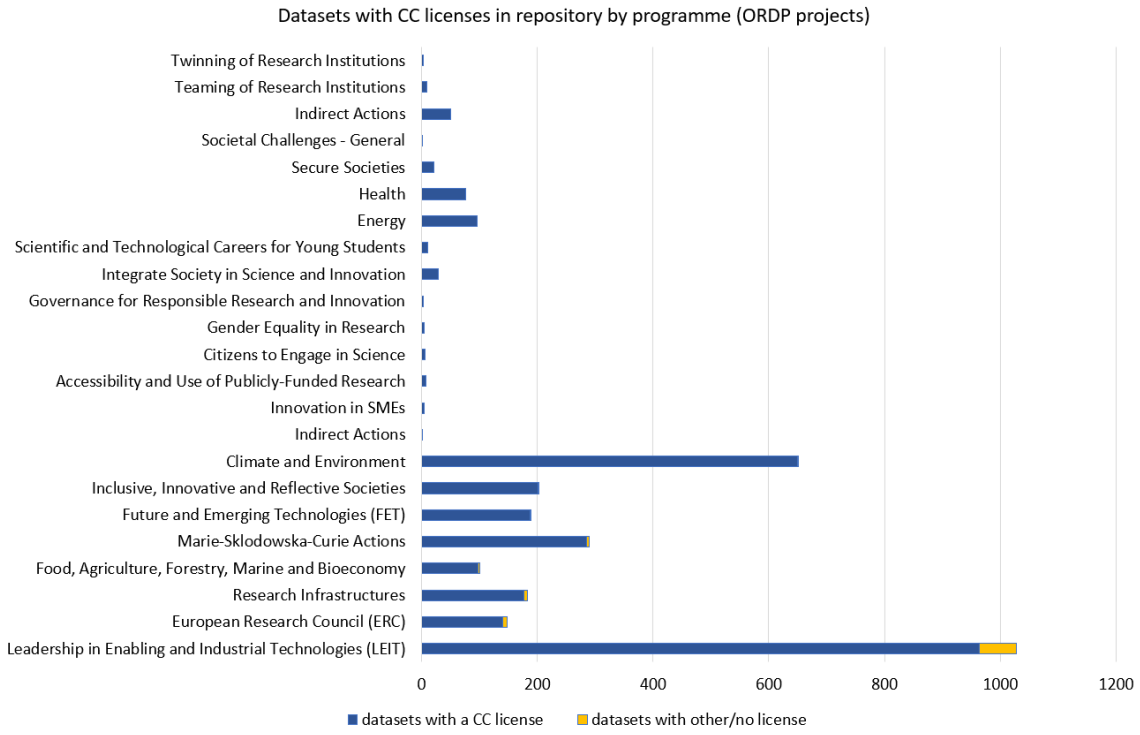



Figure 26. Licence distribution, by programme

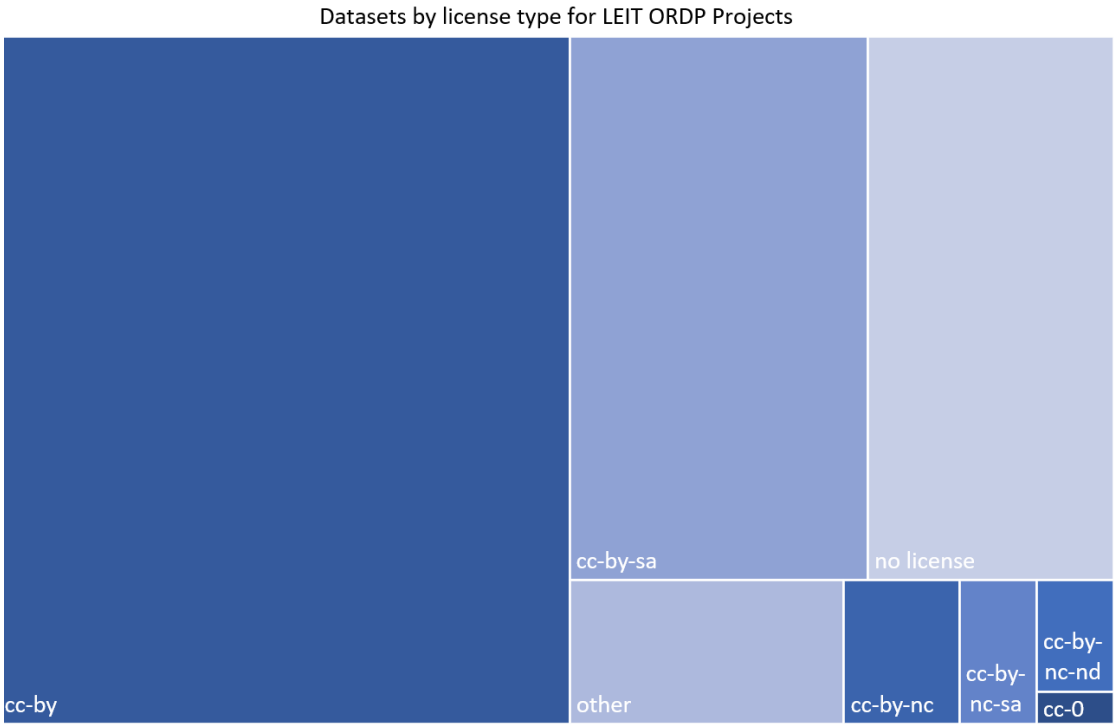



Figure 27. Datasets by license type for LEIT ORDP projects

To conclude these key findings from our compliance analysis, we present the full list of indicators and their average values in Table 23 below. This list was validated by experts during the Validation Workshop we conducted as part of this study. The last column addresses quality issues/concerns with the indicators.

**Table 23. Indicators for Horizon 2020 datasets**

| INDICATOR FOR HORIZON 2020 DATASETS | DEFINITION / FORMS OF INDICATOR | INDICATOR VALUE | | NOTES AND QUALITY ISSUES |
|---|---|---|---|---|
| | | **ORDP projects (compliance)** | **All projects (uptake)** | |
| ***Context*** | | | | |
| **1. Datasets** | Number of datasets linked to Horizon 2020 projects | 5,244 | 6,231 | **Quality:** Datasets cannot be reported for projects that do not participate in the ORDP. |
| **2. Datasets linked to scientific publications** | Number of datasets linked to scientific publications | 851 | 1,008 | |
| | Share of the total number of datasets | 16.2% | 16.2% | |
| **3. Co-funded datasets** | Number of datasets with more than one funder | 22 | 31 | |
| | Share of the total number of datasets *w/ valid number of funders* | 0.4% | 0.5% | |
| | Number of datasets linked to more than one project *(funded by European Commission or another funder)* | 221 | 300 | |
| | Share out of total number of datasets *w/ valid number of projects* | 4.2% | 4.8% | |
| **4. Co-authored datasets** | Number of co-authored datasets, by number of authors | 2-4 authors: 1,955<br>5-10 authors: 1,164<br>> 11 authors: 638 | 2-4 authors: 2,362<br>5-10 authors: 1,453<br>> 11 authors: 697 | |
| | Share of the total number of datasets *w/ valid number of authors* | 2-4 authors: 37.3%<br>5-10 authors: 22.2%<br>> 11 authors: 12.2% | 2-4 authors: 37.9%<br>5-10 authors: 23.3%<br>> 11 authors: 11.1% | |

| INDICATOR FOR HORIZON 2020 DATASETS | DEFINITION / FORMS OF INDICATOR | INDICATOR VALUE | | NOTES AND QUALITY ISSUES |
|---|---|---|---|---|
| **5. Datasets with at least one ORCID identifier** | Number of datasets with at least one author with an ORCID iD | 0 | 0 | |
| | Share of the total number of datasets | 0% | 0% | |
| | Number of datasets linked to a publication with at least one author with an ORCID iD | 125 | 149 | |
| | Share of the total number of datasets *linked to a publication* | 14.7% | 14.8% | |
| ***Open access and timely deposition*** | | | | |
| **6. Datasets by access right -** *Open access, embargoed, restricted, closed* | Number of datasets, by type of access rights | Open access: 4,538<br>Embargo: 66<br>Restricted: 138<br>Closed: 27 | Open access: 5,435<br>Embargo: 79<br>Restricted: 205<br>Closed: 37 | 'Restricted' is defined as access to a dataset being restricted to certain users.<br><br>**Quality:** content providers do not expose data on the original access rights of a dataset. Therefore, it is *only* possible to know the access rights for the last updated version of a dataset. |
| | Share of the total number of datasets *with valid access rights* in their metadata | Open access: 95.2%<br>Embargo: 1.4%<br>Restricted: 2.9%<br>Closed: 0.6% | Open access: 94.4%<br>Embargo: 1.4%<br>Restricted: 3.6%<br>Closed: 0.6% | |
| **7. Datasets with timely deposition into repository** | Number of datasets deposited in repositories by the published date of the linked publication | N/A | | Almost zero dates of deposition were available for Horizon 2020 datasets in repositories. It is not common practice for this metadata element to be exposed by repositories. |
| | Share of the total *number of datasets linked to a scientific publication* | N/A | | |
| ***Metadata requirements and FAIR principles – in REPOSITORY*** | | | | |
| **8. Datasets in repository** | Number of datasets deposited in a repository | 4,384 | 5,370 | |
| | Share of the total number of datasets | 83.6% | 86.2% | |

| INDICATOR FOR HORIZON 2020 DATASETS | DEFINITION / FORMS OF INDICATOR | INDICATOR VALUE | | NOTES AND QUALITY ISSUES |
|---|---|---|---|---|
| **9. Datasets with standard bibliographic metadata in repository** *(following OpenAIRE guidelines[79])* | Average best score per dataset in repository for metadata meeting the OpenAIRE guidelines (out of 100) | 41.42 | 41.58 | |
| **10. (FAIR) findability** | Number of datasets with a persistent identifier and an identifier to the data file (URI) in the repository | 1,888 | 1,899 | |
| | Share of the total number of datasets *deposited in a repository* | 43.1% | 35.3% | |
| **11. (FAIR) accessibility** | Number of datasets with the data file accessible via URL in the repository metadata[80] | 1,549 | 1,555 | • 9,661 URLs checked for 2,447 distinct datasets in repositories |
| | Share of the total number of datasets *w/ a valid URL in their repository metadata* | 78.2% | 78% | • 2,434: number of datasets with at least one valid URL in the repository metadata |
| | Share of the total number of datasets *in repositories* | 35.3% | 29% | • 1,673: number of datasets w/ data file accessible via URL<br>• 5: number of datasets w/ data file directly accessible via URL (e.g. link to CSV – for the remainder, the site to which the URL linked was crawled for the direct link) |
| **12. (FAIR) interoperability** | Minimum number of datasets in a machine-readable format (*This refers to those we were able to verify; we are agnostic as to the rest*.) | 1,549 | 1,555 | Athena RC's software verifies the accessibility of data files by looking for common datafile formats that are also machine-readable. Accessible datasets are therefore also regarded as interoperable (see Section 7.4.2). |
| | Share of total number of datasets *in repositories* | 35.3% | 29% | |

[79] OpenAIRE guidelines for content providers, to be used by repositories, open access journals, aggregators, CRIS (https://guidelines.openaire.eu)
[80] In the MOAP Horizon 2020 database, we provide the accessibility information for all Horizon 2020 dataset URLs available (in OpenAIRE or reported in SyGMa).

| INDICATOR FOR HORIZON 2020 DATASETS | DEFINITION / FORMS OF INDICATOR | INDICATOR VALUE | | NOTES AND QUALITY ISSUES |
|---|---|---|---|---|
| **13. datasets with licences** | Number of datasets with the following licences *deposited* in *repositories*<br>CC-0<br>CC-BY<br>CC-BY-SA<br>CC-BY-NC<br>CC-BY-NC-SA<br>CC-BY-ND<br>CC-BY-NC-ND | <br>35 (0.8%)<br>2,049 (46.7%)<br>587 (13.4%)<br>69 (1.6%)<br>49 (1.1%)<br>5 (0.1%)<br>48 (1.1%) | <br>50 (0.9%)<br>2,622 (48.8%)<br>615 (11.5%)<br>86 (1.6%)<br>79 (1.5%)<br>7 (0.1%)<br>57 (1.1%) | Overall (in merged records)<br>• 3,591: datasets with licences (66.9% of datasets)<br>• 3,492: datasets with CC licences, which we cleaned and grouped (i.e. we identified the types of licences for 97.2% of the licenced datasets)<br>• The remaining licences (just a few per type of licence) also appear to be open on first inspection.<br>In parentheses, we present the share of the total number of datasets in repositories. |
| **14. (FAIR) reusability** | Number of datasets with permissive licences in a repository of deposition: (a) allowing full text and data mining (TDM); and (b) allowing TDM only for non-commercial use | 2,676<br>166 | (c) 3,294<br>(d) 222 | |
| | Share of the total number of datasets in repositories | (a) 61%<br>(b) 3.8% | (c) 61.3%<br>(d) 4.1% | |

# 5 Monitoring open access

## *5.1 Monitoring process modelling and workflow specification*

This section describes the current workflow for the monitoring of open access under Horizon 2020, which is summarised schematically in Figure 28 below. In particular, we explain and elaborate the key steps involved in the process, also highlighting the various tools and actors involved in each of these steps. This overview is based on the findings of our desk research, and on the evidence collected from a number of interviews with key stakeholders in the field, including consultations with OpenAIRE experts.

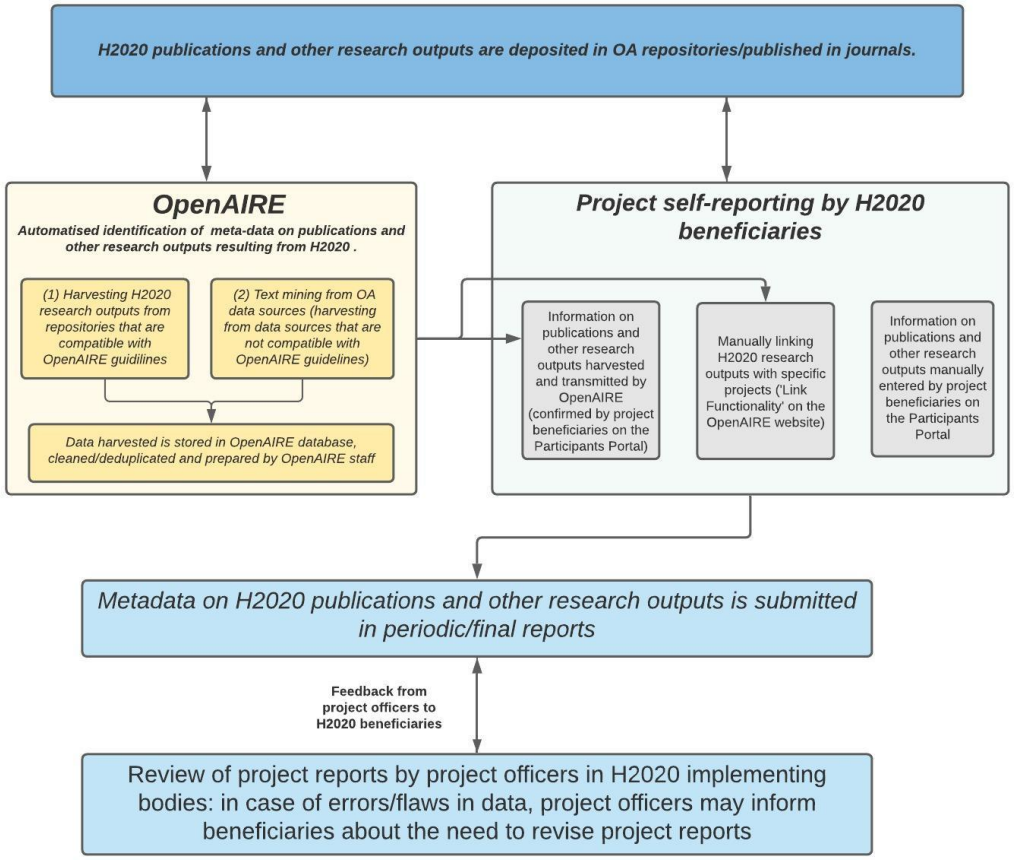



Figure 28. Horizon 2020 open access monitoring workflow. *Source*: desk research and interviews with stakeholders.

**Deposition of Horizon 2020 research outputs into repositories and other archives**

The first step in ensuring open access under Horizon 2020 involves the deposition of each research output (publication, dataset) into a repository, or publishing it in an open access journal. In general, the metadata elements for each research output deposited are entered manually by the researchers themselves or by the administrators of repositories. The quality of the data put in depends on how rigid/methodical the internal protocols and processes of the repository itself are.

**Capturing of Horizon 2020 publications by OpenAIRE**

Once a publication or other research output has been deposited into a repository or published in an open access journal, it can be tracked (together with its metadata) through **OpenAIRE – the network of open access repositories, archives and journals that support open access policies**. It should be noted that there are **two pathways via which OpenAIRE captures Horizon 2020 publications and other research outputs**.

- Currently, OpenAIRE harvests information from around 14,000 data sources. This involves both direct and indirect harvesting (i.e., harvesting from repository aggregators which themselves harvest the data directly from the source). However, **1,600 data sources (repositories, open access journals, etc.) with varying levels of compliance with the OpenAIRE guidelines are aggregated directly by OpenAIRE.**[81] In general, compliance of a repository with the OpenAIRE guidelines means that: a) when a research output has been correctly deposited into this type of repository, the metadata in the repository will also specify a link to a specific Horizon 2020 project from which this output has resulted; and that b) the repository's compliance guarantees better-quality metadata to match the detailed criteria set down for Horizon 2020. Thus, the compliance of a repository with OpenAIRE allows a user to link a research output to a specific project. Once such a link is provided, OpenAIRE can easily collect all of the metadata on a particular research output that is available in the repository.

- In those cases where a data source is not compliant with the OpenAIRE guidelines and no reference to a specific Horizon 2020 project is provided in the metadata, **it is still possible for OpenAIRE to identify the Horizon 2020 project by means of text-mining the full text, as well as the Crossref funding acknowledgement attribute**. In such cases, OpenAIRE uses automated algorithms that search for and try to identify publications linked to Horizon 2020 grants from the publicly available publication repositories, open access journals and other databases. The algorithm captures Horizon 2020 publications by identifying references to Horizon 2020-funded projects in the acknowledgement statements that beneficiaries are obliged to include (see Section 7.3.1.1).

It should be noted that OpenAIRE is integrated into the Horizon 2020 open access monitoring system to the extent that publications (and other outputs) are deposited in repositories that are harvested by OpenAIRE. This includes more than 60% of relevant repositories (excluding, for example, cultural heritage repositories registered in OpenDOAR). However, some more specialised disciplinary repositories might not be harvested by OpenAIRE, which means that the publications deposited there will not be captured by the OpenAIRE system. OpenAIRE's coverage of data repositories

[81] Out of a total of 2,271 European repositories registered in OpenDOAR, 1,130 are *directly* harvested by OpenAIRE, and 246 from a compatible (national) aggregator. Out of the 1,063 repositories used by European researchers registered at re3data.org, 52 are *directly* harvested from OpenAIRE. Source: internal OpenAIRE study Q3 2020 for literature repositories, Q2 for dataset repositories (which have seen a significant (double) increase since the original study).

is expected to increase through its efforts and participation in EOSC developments. It should also be noted that in the case of repositories, OpenAIRE only tracks research outputs that are deposited in openly accessible sources – i.e. it does not pay any licence fees to access publications that are not openly accessible.[82]

In cases where a data source is compatible/harvested, the metadata of publications (and of other types of research outputs) are integrated into OpenAIRE, and are then transferred to Horizon 2020's project reporting system. It should be noted that when a publication is deposited into a repository that is harvested/compatible with OpenAIRE, it is not immediately displayed to a beneficiary on the Participant Portal, since there is some time lag involved in this process:

- In OpenAIRE, harvesting is implemented regularly, approximately twice per month.

- Before displaying information on the Participant Portal, raw (harvested) data must first be cleaned/de-duplicated by the internal staff at OpenAIRE. This data cleaning/de-duplicating process takes some time, which results in a delay before the cleaned information is transferred from OpenAIRE to the SyGMa reporting tool, and suggestions for claims via the portal automatically appear to beneficiaries.

**Integration of OpenAIRE with the Horizon 2020 Participant Portal**

If a Horizon 2020 publication is deposited into an open access repository that is compatible/harvested by OpenAIRE, a grantee who is carrying out continuous reporting should see the publications resulting from their project appear on the European Commission's Participant Portal, even if with some time delays. In such cases, Horizon 2020 beneficiaries are presented with a list of potential publications (tracked by the OpenAIRE system) that are possibly connected with their project. Beneficiaries can then either (a) confirm that the particular publication(s) displayed are linked with their project, or (b) reject the suggestion. It should be noted that even when OpenAIRE suggests some publications/research outputs, not all the metadata fields are automatically pre-filled by the OpenAIRE system. Upon confirming that the research outputs suggested are, indeed, linked with their project, beneficiaries still need to manually enter the relevant information into some of the data fields. Although the information on DOI, author names, repository link and title of publication is pre-filled by OpenAIRE, a beneficiary may still need to key in information such as the publication year, place, type of open access ('green' or 'gold'), APC/BPC, link to publication, embargo length, etc.

In the event that a Horizon 2020 publication identified by OpenAIRE does not automatically appear on the European Commission Participant Portal, a **beneficiary can manually link their project to the publication**[83]. This can be carried out by a user (beneficiary) locating the publication on the OpenAIRE website and using the '**Link functionality**' feature to link that publication to a specific project. The manual 'Link Functionality' tool should only be used in exceptional cases, as in the majority of cases, publications tracked by OpenAIRE should be automatically displayed on the Participant Portal without any additional actions required from project managers/PIs.

OpenAIRE receives notifications and information concerning the acceptance/declining of suggestions relating to specific publications and other research outputs displayed

[82] Having said this, OpenAIRE does harvest metadata from sources such as Microsoft Academic Graph (MAG), and has two ongoing agreements with Springer Nature and Elsevier that provide access to the full text to extract funding information as well as data and software citations.
[83] https://www.openaire.eu/reporting-research-outputs-to-the-ec-using-the-openaire-api

on the SyGMa system. In the event that a suggested publication is rejected by a beneficiary, OpenAIRE experts aim to identify why the suggestion has been rejected and check whether this was done because a publication was not associated with a specific project, or for other reasons (e.g. the beneficiary was unaware of all the research outputs resulting from a specific project). One of the most common reasons why a beneficiary may reject the suggestion by OpenAIRE is when a metadata field suggested (e.g the repository link) is wrong and not editable. In such cases the only option for the beneficiary is to reject the suggestion and manually enter all the necessary data.

**Manual self-reporting by beneficiaries using the SyGMa reporting tool**

**Continuous project self-reporting by Horizon 2020 beneficiaries** is another key instrument in the workflow of Horizon 2020 open access monitoring. In their continuous reporting, beneficiaries can manually report publications that are not found/not displayed to them by the OpenAIRE system. To do so, beneficiaries should input the DOI of their publication[84] into the project reporting tool on the Participant Portal, if it exists. In that case, the reporting system automatically fetches other publication metadata fields from the Crossref database, including information about open access datasets relating to the publication, although some information, e.g. whether 'gold' or only 'green' open access has been provided, information on any associated APC/BPC and on embargo times) must still be added manually by the beneficiaries. In case no DOI exists, even the basic bibliographic metadata have to be entered manually.

Even though beneficiaries can perform *continuous* reporting of their publications and other research outputs on a continuous basis (as the name suggests), they seldom do so. In practice, publications are mainly reported by beneficiaries at the same time as they prepare periodic/final reports (e.g. a mid-term report in the middle of the project, and a final report after the project has ended).

**Data on open access submitted in periodic/final reports**

The data submitted by beneficiaries (accepted suggestions from OpenAIRE and manual entries by beneficiaries) are included in periodic reports. These are verified by **project officers (or scientific officers) in the implementing bodies** (Commission Directorates-General, Executive Agencies) to ensure that the data provided by beneficiaries are correct. If any errors are identified, the project officer may inform beneficiaries and ask them to re-submit the report with the information in the continuous reporting corrected. In the event of a project resulting in hundreds of publications, it is not feasible for each of these outputs to be individually checked by project officers. Therefore, a sample is chosen, and the project officers try to identify patterns of typical errors. It should be noted, however, that some of the data on open access provided by beneficiaries cannot be verified by officers due to timing issues (see below the section on gaps, e.g. the impossibility of verifying the embargo period encoded at the time of deposition).

[84] The European Commission data provided does not reflect this, as we see many records with missing DOIs or very dirty metadata. Please refer to the section on data/monitoring process gaps below.

### 5.2.1 Gap analysis of the Horizon 2020 open access monitoring data

One the basis of the analysis made under Tasks 1 and 2 above, this section summarises the key gaps that exist in terms of the coverage of metadata necessary to calculate/provide a breakdown of open access indicator values, as well as to assess compliance with Articles 29.2 and 29.3 of the Horizon 2020 MGA.

As shown in the previous section and in our analysis under Tasks 1 and 2, in order to collect data necessary for compliance assessment, the OpenAIRE Research Graph (ORG) collects information from various data sources (journals, funders, repositories, etc.). This information is then merged and de-duplicated to produce a set of unique publications with a rich set of metadata. Achieving this, however, requires correct mapping and unified vocabulary across data sources. For instance, a peer-review label must be the same across repositories, journals and conferences. Where information on some metadata element(s) is not commonly provided by a single repository/instance, it is necessary to fetch metadata from multiple sources into a unique record. One example of such a data element is the **peer-review status of a publication.** Information on whether a venue (conference or journal) is peer-reviewed is not a common element of the metadata that accompanies publications deposited in repositories. As a consequence, information on peer-review status needs to be inferred using a host of methods (see Section 7.3.1), or ingested from multiple sources/instances.

**Quality and availability of DOIs in the data shared by the European Commission:** the lack of valid DOIs (by comparison with that in the ORG and with Scopus/WoS) revealed under Task 1 is problematic for tracking publications. Moreover, there is a general **lack of other PIDs in EC-Shared data.** As no PIDs other than DOIs are available in data shared by the European Commission, there is no alternative for matching publications across data sources (aside from using publication titles, which is not ideal in terms of precision).

**Missing links between publications and open access datasets:** only a small fraction of open access datasets produced by Horizon 2020 projects are linked to a publication, even though 97% of datasets come from projects that have also produced publications. Although this may mean that the datasets created simply did not result in a publication, it is also likely that links between publications and datasets are missing.

**Data on embargo period:** both for publications and datasets, data on the embargo period is frequently missing/unclear (see the next section for an explanation of the key reasons for this). It is also very difficult to assess the embargo period because very often, **the repositories do not provide information on the exact publication release dates/submission history.**

**Date of deposition to repository:** this is necessary to assess the timeliness of deposition, which is one element of compliance for both publications and datasets, but is not a metadata element commonly exposed by repositories.

Some other metadata necessary to check compliance with Article 29.3 are currently not collected. Specifically, the Article includes the requirement to *"provide information – via the repository – **about tools and instruments at the disposal of the beneficiaries and necessary for validating the results** (and – where possible – provide the tools and instruments themselves)".* As discussed in Section

7.2 of the Annex, this kind of metadata are neither required from project beneficiaries by SyGMa, nor are they provided as metadata in repositories (to be harvested by OpenAIRE).

Task 1 also revealed gaps in the coverage of metadata necessary for a breakdown of indicators according to different aspects of interest. Specifically, the **European Commission currently does not collect data on a publication's citation band** (uncited, highly cited, etc.). Data on a publication's citation band are currently only available in OpenAIRE beta.

### 5.2.2 Gap analysis of the Horizon 2020 open access process

Qualitative analysis based on consultations and interviews with a number of key stakeholders, as well as on the analysis carried out under Tasks 1 and Task 2, has allowed us to identify a number of existing gaps across various areas of the current process of Horizon 2020 open access monitoring. These are described below.

**Gaps and challenges relating to the integration of OpenAIRE into Horizon 2020 monitoring**

As described in the reconstruction of the current Horizon 2020 open access monitoring workflow above, the OpenAIRE graph is integrated into the overall Horizon 2020 open access monitoring framework, insofar as OpenAIRE tracks Horizon 2020-related research outputs and presents them to the beneficiaries via the SyGMa project reporting tool. Evidence from this study, however, points to a number of challenges and persistent gaps hampering the automation of Horizon 2020 open access monitoring via OpenAIRE.

Although the above framework presupposes the automated identification and tracking of Horizon 2020–related publications, the OpenAIRE system frequently does not transmit all of the data from harvested repositories/databases, due to technical shortcomings and issues in the harvested repositories themselves. Each repository harvested by OpenAIRE has individual standards and internal protocols, which increases the chances of technical problems occurring when OpenAIRE harvests data from them. In practice, this results **in the data transmitted by OpenAIRE and presented to the grantee being of poor quality** and requiring manual corrections that are time- and labour-intensive. One of the key reasons for this is that although OpenAIRE contains a rich set of metadata, as combined from the repositories, CrossRef, MAG, Unpaywall and other sources, the Participant Portal has been configured in such a way that metadata are retrieved only from one instance and not from the merged record. As mentioned earlier, one of the key causes of data noise in the information provided by repositories is the errors/inconsistencies that occur when researchers or the administrators of the repositories/open access journals themselves manually enter metadata on publications/research outputs. These inaccuracies are then reflected in the data harvested from these repositories by OpenAIRE. Evidence from the interviews with key stakeholders indicates that there remains a **lack of consistent and rigorous practices (consistent with the official guidelines of OpenAIRE) within many repositories** with regard to the way metadata on publications and other research outputs are handled.

Only some of the publications and other research outputs that result from Horizon 2020 are captured by OpenAIRE because the latter **does not harvest all journals and repositories, particularly those that are more sector/domain-specific**.

Researchers may publish in highly sector-specific journals that are not listed in engines such as Scopus or Web of Science. One of the key reasons why OpenAIRE may underperform in tracking publications in the field of social sciences and humanities is that many publications in this field are not written in English, but in other languages.

Moreover, in cases where a publication has multiple authors and is deposited in different repositories, possibly with slightly different metadata, it **may not be clear to the grantee which version of the publication is presented via the SyGMa reporting system by OpenAIRE**. Frequently, the same publication is tracked more than once by OpenAIRE (i.e. from different sources) and displayed more than once via the SyGMa reporting tool to a beneficiary. In cases where the publication has already been claimed, the beneficiary rejects the newly displayed copy of the publication, which is then sent back to OpenAIRE as a rejected publication/not linked to a specific project. This implies a need to streamline the internal protocols of OpenAIRE to prevent the same publication being displayed to a beneficiary more than once via the SyGMa tool.

Interviewees also report that a publication or other research output that is tracked by OpenAIRE and displayed to a beneficiary is sometimes rejected by the latter due to their lack of awareness (e.g. reporting on publications in a large project may be carried out by a person who is not completely familiar with all research outputs produced by various members of the team).

Finally, even when OpenAIRE captures publications that result from Horizon 2020, their **presentation to the beneficiary via the SyGMa reporting tool may occur after a delay of several months** (see the previous section on the integration of OpenAIRE and the SyGMa tool).

**Gaps in processes related to Horizon 2020 open access self-reporting by beneficiaries**

A number of stakeholders confirmed that despite attempts at automation, the **overall process of open access reporting in Horizon 2020 is rather burdensome and time-consuming for both project officers and beneficiaries**. The workload related to reporting and ensuring open access to publications often surpasses the workload required by all other aspects of project monitoring.

Our analysis under Tasks 1 and 2, which matched data from OpenAIRE with data from the European Commission, revealed that some publications were not reported to the European Commission via the Participant Portal at all. Some of these publications were published after the end of the project, implying that **project beneficiaries do not keep reporting after the end of the project**.

At the same time, the metadata retrieved from the **European Commission reporting tool, which is largely based on manual entries by beneficiaries, is often of poor quality and unreliable**. The main reason for this is that beneficiaries are very often unaware of the specific open access reporting requirements because they do not understand the technical terms relating to open access (e.g. DOI, repository link, embargo period, 'green' vs. 'gold' open access, version of publication, etc.). For example, the Horizon 2020 participant reporting system contains a field that has to be filled in, labelled 'URL link to the repository' in which the open access publication is deposited. One of the most common issues is that the link to the publication presented to the grantee by OpenAIRE is not the repository link, but the DOI of the published version. If the grantee accepts OpenAIRE's suggestion and it

turns out to be the wrong link (e.g. a DOI instead of a repository link), this field is filled incorrectly and the project officers must undertake a lot of work to correct this mistake, given that the repository link coming from OpenAIRE is not editable by the beneficiary.

Previously, **the mixing up of the DOI with the repository link** was largely caused by technical bugs in the SyGMa reporting system itself – the information entered in the DOI field was automatically transmitted to the repository link field. This technical bug has already been solved, so the key factor behind these mix-ups is a lack of awareness among those beneficiaries who manually fill in the reports. In other words, a large proportion of them are still unaware of the difference between the two, and therefore enter DOIs or links to a publisher's website or other platforms (e.g. Researchgate.net) or research-specific web search engines (https://inspirehep.net/) instead of the correct link. In some cases, researchers wrongly enter the repository link instead of the link to the publication.

In general, **researchers are very often not fully aware of the meaning behind many other terms relating to open access, such as the differences between 'gold' and 'green' open access, embargo period, etc**. Project officers at the implementing bodies must be efficient, and very often do not have sufficient time to check whether a specific publication is 'green' or 'gold' open access. In practice, as long as a publication is uploaded to a repository, project officers do not check its type of open access ('green' vs. 'gold'). Bearing in mind that this data field is filled out manually by beneficiaries and the validity of their input is very rarely checked by project officers, data on the type of open access that is collected from manual reporting by beneficiaries is somewhat unreliable.

**There are differences in open access requirements between different Horizon 2020 programmes**. Such variations in open access requirements across programmes may create confusion among stakeholders and institutions. Differences are particularly pronounced with regard to long-form publications: although Article 29.2 of the MGA applies in the same way to all programmes, interpretation of the Article differs. The general interpretation of Article 29.2 does not focus on books and book chapters, because most Horizon 2020 programmes managed by the Commission produce relatively few books and book chapters. In contrast, books and book chapters are very important for ERC grants, especially among humanities grantees, where books are one of the main outputs.

The **timing of reporting by beneficiaries** was also identified as a cause of difficulties in the current open access monitoring workflow. Under FP7, beneficiaries were formally obliged to report their publications and other research outputs within a certain time after publication. In contrast, under Horizon 2020 beneficiaries usually report only at the time of the periodic/final reports. This means that beneficiaries very often have to report on a large number of publications within a short period of time (when the periodic/final report is due), which increases the likelihood of errors. Early reporting of publications would allow the project officers to flag issues requiring correction, such as EU funding acknowledgments or non-compliant publishing options (although in practice, project officers often cannot check publications as soon as they have been reported, due to time/resource constraints). If reporting occurs late, many of these issues cannot be corrected anymore.

**Gaps and challenges related to the monitoring of open access research data resulting from Horizon 2020**

Based on the evidence gathered from interviews with stakeholders (project officers, beneficiaries and others), a number of difficulties were identified in relation to the monitoring of open access to research data resulting from Horizon 2020:

- In many cases, research data cannot be opened up because in some research fields (e.g. particle physics), **datasets are not owned by individual researchers but by large collaboration groups** of researchers that have a policy of not sharing data. Due to ongoing competition between different experiments, the data cannot be shared (though in practice, they can still be reported without being openly accessible);

- Beneficiaries are often unsure of exactly **what types of data must be opened up**. Significant differences exist between raw data (e.g. recorded by detectors) and data that have undergone several steps of processing, which can be analysed and reused more easily;

- Both project officers and beneficiaries also reported that researchers sometimes face challenges in making their research data openly accessible due to a **lack of knowledge concerning existing data protection regulations**, i.e. it is not always clear to them what types of data containing personal details can be published in open access, and under what conditions;

- In some cases, datasets may be very large, and storing them in repositories might require a large amount of storage space and would constitute **very significant financial costs**. In disciplines such as film studies, research often results in very large amounts of data (i.e. tens of terabytes), and requires complex and expensive software and staff to process and store them (e.g. cloud computing specialists). Maintaining open access to such large datasets and ensuring their operability constitutes significant costs, which are seldom fully covered by the project budget, especially after the project has ended;

- The data themselves are also very often of **no use without the accompanying documentation explaining** how they should be read and reused. Such information is required by the FAIR principles; however, in practice it is very challenging to implement because preparing this type of documentation often requires significant amounts of researchers' time, taking them away from direct research work. Traditions within research communities have not yet fully crystallised with regard to what types of data are shared, and what documentation researchers attach to these openly accessible data. It will take some time for such traditions to develop.

- **Data management plans (DMPs) are often very rudimentary** because researchers do not understand some of the key underlying principles such as FAIR and others (e.g. what a licence for data is, which licences can be used). For instance, researchers may propose licences intended for software, despite the fact that research data are usually not protected by copyright. This is frequently a result of the lack/absence of trained data management specialists within project teams.

### *5.3 Re-engineering the monitoring process*

This chapter of the report focuses on re-engineering the open access monitoring framework for the next-generation Horizon Europe programme. The key requirement for this new open access monitoring framework is the development of a comprehensive list of open access indicators for both the publications and the datasets that result from Horizon Europe projects. We propose that the next-generation monitoring framework for the Horizon Europe programme should collect data on Open access **indicators** for Horizon Europe publications (see Table 12 for a

detailed list of indicators) and indicators for Horizon Europe datasets (see Table 23 Indicators for Horizon 2020 datasets).

This study has also identified several **metadata elements required to compile indicators** for open access publications and open access datasets. For both open access publications and datasets, the study has identified a required list of **metadata elements for the breakdown of indicators** by different aspects of interest.

The following sections also provide a summary of the key expectations/guiding principles for the updated monitoring Framework. These overarching expectations and guiding principles were identified on the basis of our consultation of key stakeholders. Lastly, Section 5.3.2 presents a list of recommendations to address the key data/process-related gaps in the Horizon 2020 open access monitoring framework.

The instruments proposed by the study team to address key gaps in the current open access monitoring framework are tailored to the types of problems identified. Of the problems and gaps identified during the study, many were of a non-technical nature (e.g. lack of awareness among beneficiaries with regard to open access concepts and skills). As a consequence, the study team chose to address these problems by providing specific, dedicated recommendations, instead of proposing a completely new open access monitoring workflow. In the present context, therefore, re-engineering the open access monitoring framework primarily means providing specific recommendations/actions addressing the key challenges, gaps and weaknesses identified in the study.

### 5.3.1 Key expectations and requirements for the updated open access monitoring Framework

Evidence stemming from our consultations of stakeholders allowed us to identify a number of expectations regarding the key principles that should guide the shaping of the next-generation Horizon Europe open access monitoring framework.

One of these is that, in the future, the **Horizon Europe open access monitoring system should allow the checking in real time** of the publications that are produced as a result of the Horizon Europe programme (in addition to information received from periodic reports). This should include the ability to filter these data by scientific domains/areas, the status of publication (open access vs. non open access), region/country, publishing venue, etc.

It is also expected that the next-generation Horizon Europe open access monitoring framework should **expand its scale by including more diverse types of open access research outputs**:

- Currently, only a small percentage of datasets resulting from Horizon 2020 projects are made open access (or reported as such). In the future, this share must increase and, similarly to publications, the general expectations are that it should be possible to access these datasets in real time, as well as filtering them by scientific domain, geographical area, the repository in which the dataset is deposited and other criteria.

- In the Horizon Europe programme, open access monitoring should also aim to encompass not only open access publications and datasets, but also other research outputs including software, trademarks, registered designs, utility models, software protocols, workflows, prototypes, and so on.

**Flexibility and sensitivity to the specificities of particular scientific domains/research fields** is another key criterion emphasised by a number of stakeholders in relation to shaping the next-generation open access monitoring framework. The re-engineered monitoring framework should not be too rigid and should not force all grantees to submit to a narrow set of reporting/publishing rules (e.g. mandatory publishing in specified journals/depositing in specific repositories) that do not take into account the diversity of scientific cultures and existing communities within specific research fields/disciplines.

Lastly, a general expectation exists that the next-generation Horizon Europe open access monitoring framework will **expand its scope beyond short-term indicators** relating to open access to research outputs (which are the focus of the present study). Instead, the next-generation Horizon Europe open access monitoring framework should also allow the inclusion of **medium- and long-term indicators focusing on the uptake of open access outputs and their impact on the creation of new networks respectively**:

- In the medium term, it is expected that the next-generation Horizon Europe monitoring framework will collect and systematise data not only about the numbers of research outputs, but also **usage statistics** – i.e. the indicators should show how many users are picking up and using these outputs for one or another purpose (if a research output is deposited and made open access).

- It is also expected that the next-generation monitoring system will encompass long-term indicators, such as measuring to what extent the opening up of research outputs that result from Horizon Europe projects contributes to the **development of new networks**, including actors that were not directly involved in Horizon Europe projects. In other words, these long-term indicators should measure the extent to which ensuring open access to research outputs contributes to the creation of new communities/research networks.

All indicators relating to open access monitoring will be aligned under 'Key Impact Pathway 3: Fostering diffusion of knowledge and Open Science' of the Horizon Europe programme. It should be also noted that both medium-term and long-term indicators for open access should not be based on self-reported data, i.e. the monitoring of such indicators should be based on automated processes, as far as possible using external databases such as OpenAIRE and commercial databases.

### 5.3.2 Recommendations for the re-engineering of the Horizon Europe open access monitoring process, addressing recurrent issues in current open access monitoring

Based on the qualitative and quantitative analysis conducted in the previous Tasks, this section presents a number of recommendations to address various issues relating to gaps in open access data and the monitoring process. These recommendations to improve open access monitoring in Horizon 2020 and Horizon Europe are grouped thematically, and cover the following areas:

- Recommendations regarding OpenAIRE and its link to the European Commission Reporting Tool;

- Recommendations regarding processes related to open access self-reporting by beneficiaries;

- Recommendations regarding the monitoring of open data.

Each of these recommendations addresses relevant issues/gaps in data and processes in the open access monitoring framework, as identified during our earlier analysis. Non-technical recommendations were subjected to validation and approved by experts during the Validation seminar, which involved a number of stakeholders (including policy makers and experts) in the field of Open Science. The recommendations were revised and complemented on the basis of feedback and comments provided by the stakeholders during the Validation seminar.

**Table 24. Recommendations regarding OpenAIRE and its link to the European Commission reporting tool**

| |
|---|
| **Recommendation 1: Updating the OpenAIRE guidelines for repositories, and increasing the adoption of the OpenAIRE metadata standard among repositories.** |
| Some data elements (e.g. peer-review status, date of deposition) are not common data elements in repositories. This hinders full monitoring of open access compliance. In addition, there is a lack of consistent and rigorous data management/entry practices (consistent with the official guidelines of OpenAIRE) among many repositories.<br><br>We recommend:<br><br>Updating the OpenAIRE guidelines for repositories, in accordance with the list of open access indicators and the list of metadata elements identified in the present study.<br><br>Disseminating and encouraging maximum adoption of the OpenAIRE metadata standards among repositories, using the available channels (including the European Commission). Specifically, it should be ensured that repositories collect all the metadata elements that are necessary to assess compliance, and expose these elements to OpenAIRE. |
| **Recommendation 2: Streamlining internal procedures within OpenAIRE Graph to reduce delays in transferring data to the SyGMa reporting tool.** |
| Currently, the time lag from depositing a publication into a repository and it being displayed to a beneficiary via the SyGMa tool often ranges from several weeks to several months.<br><br>We recommend:<br><br>Streamlining OpenAIRE processes to ensure repositories are harvested at more frequent time intervals;<br><br>Shortening the amount of time dedicated to cleaning/de-duplicating the harvested data and to processing data in the OpenAIRE database;<br><br>The SyGMa reporting tool should be updated to send an automatic alert to a beneficiary once information on publications has been transferred from the OpenAIRE database to the SyGMa system, and is awaiting confirmation by the beneficiary;<br><br>Streamlining OpenAIRE's internal protocols to prevent the same publication being displayed more than once to the beneficiary by the SyGMa tool, while at the same time displaying the version of the publication with the most complete coverage of metadata elements. |

**Table 25. Recommendations regarding processes relating to Horizon Europe open access self-reporting by beneficiaries**

| |
|---|
| **Recommendation 3: Organising training sessions for beneficiary principal investigators, focusing on the general principles underpinning open access in Horizon Europe, as well as the requirements and reporting process.** |
| Often, beneficiaries do not fully understand specific technical terms relating to open access, and encounter difficulties when reporting open access on the Participant Portal.<br><br>We recommend supporting specific training for beneficiary principal investigators, focusing on open access reporting within the Horizon Europe programme. Such training should focus on:<br><br>Summarising the key principles of the Horizon Europe open access policy and its requirements/obligations for beneficiaries (including an explanation of the different routes to open access);<br><br>Explaining the step-by-step process used to report open access outputs on the Participant Portal;<br><br>Explaining key technical terms relating to open access, as well as concepts and the most common errors/misconceptions(e.g. mixing up DOIs and repository links);<br><br>Q&A session. |
| **Recommendation 4: Preparing a concise 'one-stop source' manual/guidelines for beneficiary principal investigators/project managers/support staff, explaining the key steps in the Horizon Europe open access reporting process.** |
| The 'one-stop source' practical guidelines on open access reporting in Horizon Europe for principal investigators/project managers and support staff should include information on the following:<br><br>The step-by-step process used to report open access outputs on the Participant Portal;<br><br>Key technical terms relating to open access, and the most common errors/misconceptions (e.g. mixing up DOIs and repository links);<br><br>Practical step-by-step guidelines should be disseminated to principal investigators during the project inception phase, together with an updated version of the 'Guidelines on Open Access to Scientific Publications and Open Access to Research Data'.<br><br>Different versions of these guidelines could be adapted, taking into account the differences and specificities of different programmes and stakeholders within Horizon Europe. |

**Recommendation 5: In the case of manual self-reporting by beneficiaries, implementing technical safeguards at the data submission stage in the SyGMa reporting tool, to address the issue of beneficiaries incorrectly filling in metadata fields when self-reporting.**

Researchers often mix up DOIs and repository links, or provide incorrect repository links.

For cases when metadata fields are manually filled in by beneficiaries, technical checks using IT algorithms should be integrated into the SyGMa reporting tool that automatically check the validity of information entered by beneficiaries. Such technical checks should verify:

That the repository link is not broken;

Whether it is in fact a repository link and not the DOI or a link to the version on the publisher's website;

If beneficiaries have provided links to common platforms, such as Researchgate.net or Academia.edu, instead of genuine repository links.

Before the report is submitted, these checks should automatically flag errors in repository links or other data fields and ask the beneficiary to correct the relevant data fields.

**Recommendation 6: Delivering regular reminders to the project beneficiaries for several years after the project has ended, calling on them to report the project outputs on the Participant Portal, to increase the level of post-project open access reporting.**

Evidence shows that project beneficiaries often do not keep reporting after the end of a project. As a consequence, a share of publications and other research outputs is not included in the monitoring/open access compliance check.

To address this problem, we recommend ensuring that PIs receive regular reminders (e.g. via email) for at least several years after the formal project end. These reminders should inform the beneficiaries about the need to, for instance, report on the Participant Portal any as-yet-unreported publications that have resulted from their Horizon 2020/Horizon Europe grant.

**Table 26. Recommendations Regarding the Monitoring of Open Data in the Horizon Europe Programme**

| |
|---|
| **Recommendation 7: Improving the quality of open research data management in Horizon Europe projects.** |
| DMPs are often rudimentary or do not follow a fixed/streamlined format or vocabulary.<br><br>We recommend that the quality of research data management in Horizon Europe research projects should be improved through the following means:<br><br>Encouraging project teams to include more personnel professionally trained in research data management (RDM), as well as providing support to project teams in the form of professional training in data management;<br><br>Providing more guidance to beneficiaries and project officers on the available and recommended repositories for depositing research datasets. These guidelines should also explain some of the key principles behind ORD (e.g. the FAIR principle, what a licence for open data is, etc.).<br><br>Disseminating the template for the Data Management Plan, and encouraging its use among principal investigators.<br><br>Disseminating existing DMP good practice examples to beneficiaries at the beginning of their projects. |
| **Recommendation 8: Developing clear and comprehensive guidelines describing what type of data should be opened up (raw vs. processed), and what documentation should accompany open access research datasets.** |
| It is not always clear to beneficiaries what type of data is to be opened up (raw vs. processed data) and what type of accompanying documentation should be provided.<br><br>To address the above problem, we recommend:<br><br>Disseminating the Commission's guidelines on FAIR Data Management more actively among beneficiaries;<br><br>Disseminating the FAIR Guiding Principles for scientific data management to beneficiaries at the beginning of their project, explaining what they mean, and what the underlying data quality standards are[85];<br><br>Disseminating and encouraging the use of the Metadata Standards Directory[86], which can be searched for discipline-specific standards and associated tools (including standards for data documentation);<br><br>Updating the guidelines on FAIR Data Management by addressing questions relating to ethics/personal data protection in research data. |

[85] e.g. Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[86] https://rd-alliance.github.io/metadata-directory/

# 6 Lessons Learned

## *6.1 Intervention logic of the Horizon 2020 open access policy*

In line with the Technical Specifications, the study team has prepared a diagram of the intervention logic for the Horizon 2020 open access policy (see Figure 29). The main sources of evidence used to prepare this were: a review of the main Horizon 2020 open access policy documents; the Horizon 2020 intervention logic[87]; reports and academic articles; as well as the evidence collected during our interviews with key stakeholders. The intervention logic represents the links between causes and effects within the overall Horizon 2020 open access policy framework, including inputs, outputs, short-term results, medium-term results and the long-term impacts of the policy. These are:

- **Inputs to the Horizon 2020 open access policy** (i.e., the main policy interventions and activities funded under Horizon 2020, and aimed at enhancing open access in research). The key inputs into the Horizon 2020 open access policy are the open access requirements and obligations described in Articles 29.2 and 29.3 of the Horizon 2020 MGA (including the Horizon 2020 Open Research Data Pilot).[88] Other inputs into this policy also include Horizon 2020 funding dedicated to the uptake of the open access policy/requirements (e.g., to support open access publication costs via project budgets). Lastly, inputs in the model also include investments via Horizon 2020 grants in the development of open access infrastructure through the European Research Infrastructure Work Programme[89], as well as investments in the European Open Science Cloud (EOSC)[90].

- The key **outputs of the Horizon 2020 open access policy** primarily include the increased compliance of Horizon 2020 beneficiaries with the open access rules and obligations set down in the Horizon 2020 MGA (Articles 29.2 and 29.3). The intervention logic also includes other outputs of the open access policy such as the outputs of Horizon 2020 investments in open access infrastructure: new and improved tools, standards, processes, specifications for interoperability and the sharing of open access research outputs (e.g. the OpenAIRE Research Graph, Zenodo, etc.). Lastly, investments in open access infrastructure are also intended to result in the development of a fit-for-purpose, pan-European governance structure to federate scientific data infrastructures and overcome fragmentation in open access infrastructure.[91]

- The **short term/immediate results** of the Horizon 2020 open access policy in this model include free of charge, open access to Horizon 2020 research outputs for researchers and the general public. In addition, interviews with key stakeholders also reveal that the Horizon 2020 open access policy often has significant learning effects on beneficiary researchers in terms of increased awareness, knowledge and skills in Open Science/data management. The results of this policy also include the increased findability, accessibility, interoperability and reusability of Horizon 2020 research outputs (i.e. implementation of the FAIR

[87] Horizon 2020 Intervention Logic, https://www.kowi.de/Portaldata/2/Resources/horizon2020/Horizon 2020-intervention-logic.pdf
[88] Horizon 2020 Programme AGA – Annotated Model Grant Agreement, Version 5.2, 26 June 2019 https://ec.europa.eu/research/participants/data/ref/Horizon 2020/grants_manual/amga/Horizon 2020-amga_en.pdf
[89] Horizon 2020 Work Programme 2018-2020. European research infrastructures (including e-Infrastructures).
[90] European Commission (2018), Implementation Roadmap for the European Open Science Cloud, Brussels, 14.3.2018 SWD(2018) 83 final.
[91] European Cloud Initiative – Building a competitive data and knowledge economy in Europe. Brussels, 19.4.2016 COM(2016) 178 final

guiding principles[92]). Lastly, the available evidence also shows that the Horizon 2020 open access policy has the positive effect of encouraging other funding bodies and institutions across Europe to develop their own open access policies (spill-over effects).[93]

- It is expected that the immediate results described previously (free and open access to research outputs; Open Science learning effects; open access spill-over effects to other funding bodies and organisations; implementation of FAIR principles) will produce **longer-term results in terms of knowledge diffusion**: improved outreach, sharing and uptake of knowledge, innovation, services and products across different disciplines, sectors, research communities and countries/regions.

- The intervention logic envisages that the Horizon 2020 open access policy will have a number of long-term impacts resulting from knowledge diffusion, across various areas of the R&D system, the economy, and society as a whole:

  - An increase in scientific excellence: pushing the frontiers of knowledge;

  - New transdisciplinary, international and intersectoral networks, new research communities, and new research fields;

  - Boosting innovation and business-research cooperation, as well as the commercialisation of research results;

  - Knowledge diffusion should also contribute to better public decision making and, therefore, help to address EU policy priorities and global challenges through research and innovation (R&I);

  - More and better jobs, economic growth;

  - Strengthening the uptake and general awareness of R&I developments in society, popularising science within society.

[92] European Commission, Turning Fair into Reality, 2018. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
[93] Guedj, D & Ramjoué, C., European Commission Policy on Open-Access to Scientific Publications and Research Data in Horizon 2020 Biomed *Data J.* 2015; 1(1): 11-14, https://doi.org/10.11610/bmdj.01102.

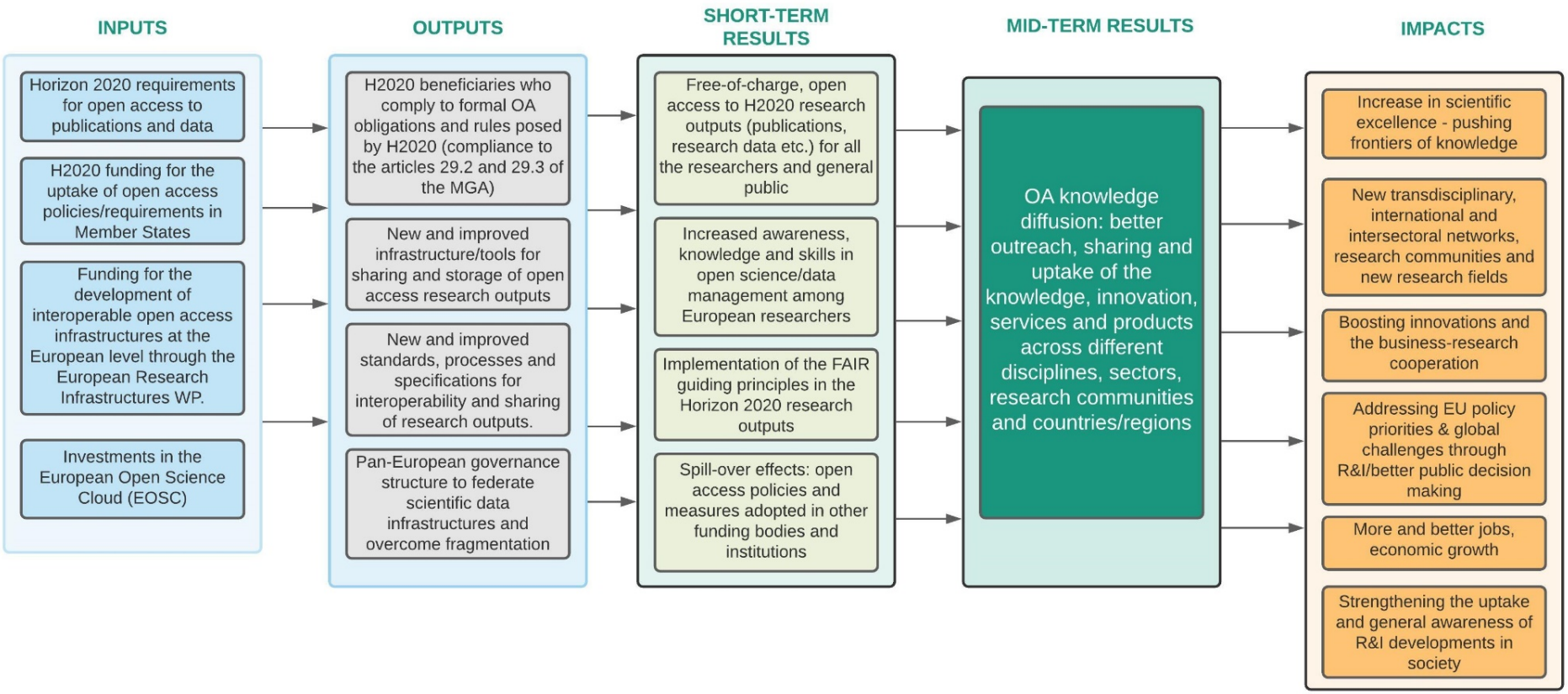



Figure 29. Horizon 2020 open access policy intervention logic. *Source*: based on desk research and interviews with stakeholders.

### *6.2 Efficiency of the Horizon 2020 open access policy*

Some of the key instruments for assessing the efficiency of the Horizon 2020 open access policy include relevant benchmarks: that is, comparing open access compliance rates under Horizon 2020 to those for other R&D funding programmes (in Europe or worldwide) that have similar open access policies. To carry out these comparisons, we collected and systematised data on open access compliance rates for various funders in Europe and around the world. The main sources of evidence were the publicly available databases (e.g., Google Scholar), recent studies and academic articles analysing open access compliance for different countries/funders, as well as data directly provided by research funding organisations at the request of the study team.

**Efficiency of open access under Horizon 2020 – comparison with other funders**

Before comparing the open access compliance rates across different funders, it must be noted that while similar, the open access policies and requirements of these funders are not identical to those of Horizon 2020. For example, unlike Horizon 2020, the Gates Foundation in the US requires that "all funded research, including articles accepted for publication, shall be available immediately on publication, without any embargo period".[94] Some funders also permit a longer embargo period than Horizon 2020: for example, the National Science Foundation (NSF) allows an embargo or administrative delay to access of up to 12 months from the date of publication for journal articles or juried conference papers. In addition, some funders compile extensive lists of exceptions to their open access requirements (see, for instance, the UK Research Council's open access policy[95]).

A comprehensive source of information for comparing publication open access rates for different research funders is Google Scholar's public access metrics.[96] These provide information on the percentage share of publications that are open access for each funder for 2015-2019. It should be noted that this percentage includes both peer-reviewed and non-peer-reviewed publications. In total, the metrics cover 184 research funders from all countries, regions and disciplines. According to Google Scholar's metrics, Horizon 2020 is in the top half of funders in terms of the rates of open access achieved. With an average open access rate of 84% (2015-2019), Horizon 2020 comes between 70th and 90th place[97] out of a total of 184 funders included in the public access metrics. One should note, however, that the vast majority of those funders with a higher percentage of open access publications than Horizon 2020 are discipline-specific funders which mostly focus on life sciences/biomedical research. For a more meaningful comparison, the study team eliminated discipline-specific funders from its analysis and compared Horizon 2020's percentage of open access publications with those of other non-discipline-specific funders. This comparison placed Horizon 2020 in 12th place out of 47 non-discipline-specific funders. This indicates that Horizon 2020 performs better on average than some of the largest research funders in Europe (Switzerland, Sweden, Germany, Italy, Spain, Ireland, Portugal), as well as some of the largest non-discipline-specific funders in the US such as the NSF. The percentage of open access publications in Horizon 2020 is the same as for the main research funders in the UK and France. At the same time, however, the percentage share of open access publications in Horizon

[94] Bill and Melinda Gates Foundation open access policy, https://www.gatesfoundation.org/about/policies-and-resources/open-access-policy.
[95] REF 2021: Overview of open access policy and guidance, November 2019 https://www.ref.ac.uk/media/1228/open_access_summary__v1_0.pdf
[96] Google Scholar, https://scholar.google.com/citations?view_op=mandates_leaderboard
[97] The range means that out of a total of 184 funders included in the analysis, 20 funders (including Horizon 2020) had the same rate of open access publication (84%).

2020 was somewhat lower than those for some of the largest research funders in the Netherlands, Hungary, Denmark, Austria and Belgium (see Table 27).

When comparing Horizon 2020 with other funders, in addition to the range of disciplines funded, other important aspects and differences between funders must be taken into account. As previously mentioned, funders implement different policies with regard to open access, which might significantly influence their performance in terms of the percentage of publications that are openly accessible. For instance, unlike Horizon 2020, the German Research Foundation (DFG) does not mandate open access, but merely recommends that its beneficiaries should ensure open access to their publications. In addition, the scale of funders must be taken into account; many funders with a higher percentage of open access publications are national-level funding agencies that are significantly smaller than Horizon 2020.

**Table 27: Percentage of open access publications (included non-peer-reviewed publications), by funder and by year (non-discipline specific funders)**

| COUNTRY /REGION | FUNDER | 2017 | 2018 | 2019 | TOTAL (2015-2019) |
|---|---|---|---|---|---|
| **NL** | Netherlands Organisation for Scientific Research | 89% | 90% | 89% | 89% |
| **NL** | Royal Netherlands Academy of Arts and Sciences | 89% | 93% | 91% | 88% |
| **US** | Doris Duke Charitable Foundation | 90% | 89% | 84% | 88% |
| **HU** | National Office for Research, Development and Innovation | 90% | 88% | 87% | 88% |
| **FR** | AXA Research Fund | 89% | 86% | 88% | 88% |
| **AT** | Austrian Science Fund | 90% | 91% | 90% | 87% |
| **BE** | National Fund for Scientific Research | 87% | 88% | 89% | 87% |
| **HU** | Hungarian Academy of Sciences | 86% | 85% | 87% | 86% |
| **DK** | Danish National Research Foundation | 88% | 88% | 87% | 86% |
| **US** | Smithsonian Institution | 86% | 86% | 84% | 86% |
| **US** | Hewlett Foundation | 88% | 90% | 86% | 86% |
| **EU** | **Horizon 2020 - EU Research and Innovation Programme** | 86% | 85% | 85% | 84% |
| **FI** | Academy of Finland | 84% | 88% | 89% | 84% |
| **FR** | Agence Nationale de la Recherche | - | - | 86% | 84% |
| **UK** | UK Research & Innovation | 86% | 85% | 85% | 84% |
| **LU** | Luxembourg National Research Fund | 84% | 82% | 85% | 84% |
| **DE** | Leibniz Association | 85% | 81% | 87% | 84% |

| | | | | | |
|---|---|---|---|---|---|
| **CH** | Swiss National Science Foundation | 85% | 86% | 85% | 83% |
| **SE** | Swedish Research Council | 85% | 85% | 85% | 83% |
| **NO** | Research Council of Norway | 83% | 85% | 85% | 83% |
| **US** | US National Science Foundation | 81% | 83% | 84% | 82% |
| **IE** | Science Foundation Ireland | 81% | 81% | 80% | 80% |
| **DK** | Danish Council for Independent Research | 82% | 85% | 81% | 80% |
| **BE** | Research Foundation (Flanders) | 83% | 85% | 84% | 79% |
| **HU** | Hungarian Scientific Research Fund | 82% | 81% | 82% | 79% |
| **US** | State of California | - | - | 81% | 79% |
| **IT** | Government of Italy | 77% | 77% | 75% | 77% |
| **DE** | Volkswagen Foundation | 80% | 76% | 76% | 77% |
| **SE** | Bank of Sweden Tercentenary Foundation | 78% | 84% | 84% | 77% |
| **ES** | Government of Spain | 78% | 78% | 78% | 76% |
| **AU** | Australian Research Council | 77% | 77% | 75% | 76% |
| **DK** | Danish Council for Strategic Research | 77% | 80% | 77% | 76% |
| **PT** | Fundação para a Ciência e a Tecnologia | 76% | 75% | 73% | 75% |
| **DE** | Federal Ministry of Education and Research | 74% | 74% | 74% | 75% |
| **SI** | Slovenian Research Agency | 72% | 75% | 77% | 75% |
| **IE** | Irish Research Council | 76% | 75% | 75% | 75% |
| **IS** | Icelandic Centre for Research | 76% | 72% | 76% | 75% |
| **ET** | Ministry of Science and Higher Education, Ethiopia | - | - | 64% | 75% |
| **RS** | Ministry of Education, Science and Technological Development of the Republic of Serbia | - | 69% | 74% | 74% |
| **IE** | Higher Education Authority | 68% | 81% | 73% | 74% |
| **SA** | National Research Foundation, South Africa | 74% | 74% | 72% | 73% |
| **DE** | German Research Foundation | 75% | 74% | 76% | 72% |
| **LT** | Lithuanian Research Council | 71% | 68% | 72% | 71% |
| **SG** | National Research Foundation, Singapore | 69% | 70% | 70% | 69% |
| **CN** | Chinese Academy of Sciences | 65% | 63% | 61% | 64% |

| | | | | | |
|---|---|---|---|---|---|
| **HK** | Research Grants Council, Hong Kong | 62% | 61% | 59% | 61% |
| **IN** | Department of Science & Technology, India | 57% | 55% | 52% | 57% |

*Source*: Google Scholar, https://scholar.google.com/citations?view_op=mandates_leaderboard.

The data required to compare different **research funders *by route of open access*** (i.e., 'green' or 'gold') are much scarcer, and are largely based on previous studies carried out in this area. Often, the data available do not allow systematic comparison of Horizon 2020 against other funders in terms of open access by route, because of the ***differing methodologies*** used by funders to calculate the percentage shares of open access publications by route. For example, according to data provided by the Swiss National Science Foundation (SNSF),[98] out of the total number of peer-reviewed publications published between 2015 and 2019, 19% were pure 'gold' open access; 20% were in hybrid open access journals; and 14% were 'green' open access. However, another 21% of publications was ascribed to "other open access."[99] Similarly, according to a report from the Austrian Science Fund (FWF), out of the total of 9,353 publications listed in final project reports submitted in 2019, the vast majority (89%) were open access. The open access option most frequently chosen in Austria was hybrid open access (40%), with the share of pure 'gold' open access being 19%, and 'green' being 22%; "other" open access was 8%.[100] In their monitoring reports, both the SNSF and FWF regarded 'gold' and 'green' as mutually exclusive, and therefore did not provide data on publications that were both 'gold' *and* 'green'.

The most comprehensive analysis so far conducted on open access by routes for each funder was by Larivière and Sugimoto, who analysed 12 funders (mostly in the US, Canada and the UK) to determine the percentage share of open access publications by route up to the year 2016.[101] This study relied on analysing the publications indexed in the Web of Science database, and identified the funding sources of papers using the published acknowledgements (mandated by most funders). The study provides a reference point for comparing the percentage share of open access by route for various funders.[102] Comparing Horizon 2020 with other funders reveals that on average, as of 2016, the percentage share of 'gold' open access publications (out of the total number of open access publications) in Horizon 2020 was similar to those for some of the largest funding bodies in the US (the NIH and NSF).[103] Around 53% of all open access publications in Horizon 2020 were 'gold', compared with 53% for the NIH and 54% for the NSF. Compared with most funders in the UK and Canada, however, the percentage share of 'gold' open access publications in Horizon 2020 was significantly lower (see Table 28[104]). This suggests that on average, Horizon 2020 (together with some of the largest US funding bodies) was more likely to choose 'green' open access than other large funders in the UK, Canada and Australia. The percentage share of 'gold' (as opposed to 'green') open access allows us to estimate

[98] Data directly received from the SNSF. Analysis done in March 2020.
[99] According to SNSF methodology, if it was not possible to automatically verify the status of open access, instead of assigning closed or open, a publication is assigned to 'Other open access.' 'Gold', Hybrid, 'Green' or closed is only assigned if it is possible to automatically verify the publication is either of those.
[100] Austrian Science Fund (FWF) open access Compliance Monitoring 2019, Kunzmann M (2020); Zenodo; https://doi.org10.5281/zenodo.3931234.
[101] Larivière, V. & Sugimoto, C.R. (2018). Do authors comply with mandates for open access? *Nature, 562*(7728), 483-486. https://doi.org/10.1038/d41586-018-07101-w
[102] Because the study in question provided data on the percentage share of open access publications by route for different funders as of 2016, we have also used the same time period for our comparison with Horizon 2020 (i.e. those publications under Horizon 2020 published up to 2016).
[103] In 2016 there were 6,149 'gold' open access Horizon 2020 peer-reviewed publications.
[104] It must be noted that out of the 10 funders (not including H2020) included in this comparative analysis, 7 are from the biomedical/life sciences area - in contrast to Horizon 2020, which is a non-discipline-specific funder.

the potential costs of the open access policy, since 'gold' open access mostly transfers the cost of open access to authors, who often need to allocate funds from their research budgets to cover publishing.[105,106]

**Table 28. Percentage open access publications by open access route and by funder**

| Country /region | Funder | Percentage of 'gold' open access publications that are *not* also 'green' (%) | Percentage of gold' open access publications that ARE *ALSO* 'green' (%) | 'Gold' as a share of all publications (%) | Percentage of publications that are open access (%) | 'Gold' as a share of all open access |
|---|---|---|---|---|---|---|
| **EU** | Horizon 2020 (2014-2020) | 6.4 | 46.4 | 52.9 | 84.2 | 63 |
| **EU** | **Horizon 2020 (2014-2016)** | **-** | **-** | **43.34** | **81.2** | **53** |
| **USA** | National Institutes of Health (NIH) | 2.1 | 46 | 48.1 | 91.4 | **53** |
| **UK** | Wellcome Trust | 6.5 | 70.5 | 77 | 86.6 | **89** |
| **USA** | Gates Foundation | 10.3 | 54 | 64.3 | 78.3 | **82** |
| **UK** | Medical Research Council | 11.6 | 56.4 | 68 | 77.4 | **88** |
| **UK** | Biotechnology and Biological Sciences Research Council | 12.5 | 52.3 | 64.8 | 72.5 | **89** |
| **AU** | National Health and Medical Research Council | 56.2 | | 56.2 | 67.3 | **84** |
| **UK** | Economic and Social Research Council | 10.6 | 35.5 | 46.1 | 63.5 | **73** |
| **CA** | Canadian Institutes of Health Research | 18.5 | 31 | 49.5 | 57.5 | **86** |
| **UK** | Engineering and Physical Sciences Research Council | 9.5 | 18.6 | 28.1 | 49.9 | **56** |
| **USA** | National Science Foundation | 8 | 17.9 | 25.9 | 47.8 | **54** |

*Sources:* (1) Larivière, V. & Sugimoto, C.R. (2018). Do authors comply with mandates for open access? Nature, 562(7728), 483-486. https://doi.org/10.1038/d41586-018-07101-w; (2) Kirkman, N. & Haddow, G. (2020). Compliance with the first funder open access policy in Australia. Information Research, 24(4), paper 857. http://InformationR.net/ir/25-2/paper857.html (Archived by the Internet Archive at https://bit.ly/36Dj4fx).

[105] Larivière, V. & Sugimoto, C.R. (2018). loc.cit.
[106] Data for Horizon 2020 refer to the period 2014-2020/2014-2016; for other funders, data refer to the situation in 2016.

Lastly, the **average APC covered by a funding body** is another potential indicator to assess the efficiency of open access policies between different funders. A common assumption is that the level of the APC is an indicator of an open access journal's prestige/level of impact. However, previous analyses have shown that this is not the case, and high APCs do not always reflect the high impact of specific articles. Empirical studies have shown that the high fees charged by certain journals to publish articles do not correlate with greater numbers of citations – in other words, prices charged are not a good indication of a journal's prestige.[107] The analysis of six large research funders (see Table 29) shows that, **on average, APCs under Horizon 2020 were similar to average APCs for those other funders in Europe and the US** for which average APC data were available. On the basis of our extrapolations (see Section 0), the average APC for Horizon 2020 was EUR 2,145 in 2019. The average for the whole programme period (2014-2020) was EUR 2,178. The average APC for Horizon 2020 in 2019 was lower than that for the Gates Foundation (EUR 2,392 for the period 2016-2019) and lower by an even greater margin compared with the Wellcome Trust (EUR 2,777 in 2018-2019). At the same time, average APCs for Horizon 2020 were somewhat higher than those for Austria's FWF (EUR 1,979 in 2019), and very similar to those of the Swiss SNSF and the UK Research Councils. It must be noted, however, that the average APCs covered by Horizon 2020 have increased significantly over the course of the programme, from an average of EUR 1,724 in 2014 to EUR 2,214 in 2020.

**Table 29. Average APC per Funder**

| COUNTRY/REGION | FUNDER | TIME PERIOD | AVERAGE APC COVERED (EUR) |
|---|---|---|---|
| **EU** | **Horizon 2020** | **2014-2020** | **2,178** |
| **EU** | **Horizon 2020** | **2019** | **2,145** |
| **Switzerland** | SNSF | 2019-2020 | 2,220 |
| **Austria** | FWF | 2019 | 1,979 |
| **USA** | Gates Foundation | 2016-2019 | 2,392 |
| **UK** | Wellcome Trust | 2018-2019 | 2,777 |
| **UK** | UKRI (UK Research Councils) | 2017-2018 | 2,118 |

*Sources*: (1) Broschinski, C. (2020). SNSF provides more APC data for 2019 and 2020. https://openapc.github.io/general/openapc/2020/11/26/snsf/.
(2) Broschinski, C. (2020). FWF reports expenditures on APCs and BPCs. https://openapc.github.io/general/openapc/2020/12/17/fwf/
(3)https://openapc.github.io/general/openapc/2020/03/09/gates_foundation/
(4) https://wellcome.org/funding/wellcome-and-coaf-open-access-spend-201819
(5) http://www.open.ac.uk/blogs/the_orb/?p=3038

[107] Mizera, K. (2013), Cost Effectiveness for open access Journals, https://openscience.com/cost-effectiveness-for-open-access-journals/; Eigenfactor Index of open access Fees. http://www.eigenfactor.org/openaccess/

**Potential areas for improvement**

The qualitative analysis that stemmed from our interviews with key stakeholders revealed several inefficiencies related to covering APCs. In some cases, the process of **covering these fees, which are necessary to ensure open access to Horizon 2020 publications, was burdensome and lengthy due to a number of administrative restrictions at research institutions/universities**. For example, some publishers (particularly those outside Europe) expect immediate payment of APCs/BPCs using a credit card, whereas many European higher education institutions do not allow credit card payments due to local legal regulations governing public institutions. As a consequence, in such cases the process of paying APCs/BPCs might take as much as several months, which may in some cases give rise to the risk that a researcher will not be able to claim primacy and intellectual property rights to the results and ideas contained within a paper (because in the meantime somebody else has published similar results). One of the most feasible solutions to help beneficiaries avoid these situations would be to conduct initial training on open access at the beginning of projects, at which time the researchers would be informed about all of the possible challenges relating to open access (including administrative difficulties when processing APC/BPC payments), as well as the measures to mitigate these risks.

Qualitative evidence based on our interviews with beneficiaries reveals that **one of the key sources of inefficiencies in terms of the financial costs of open access relates to a lack of awareness and knowledge among beneficiaries with regard to Horizon 2020 open access requirements**. In some cases, this lack of knowledge about different routes to open access led to inefficiencies that might otherwise have been avoided: some beneficiaries reported that they spent thousands of euros out of their project budget to cover APCs because, for instance, at the time they were unaware that the programme's open access requirements could be fulfilled by depositing a peer-reviewed manuscript into an online repository. As a consequence, beneficiaries chose 'gold' open access to provide immediate open access to their publications, and thus had to bear the expenses related to covering APCs.

**The available evidence also confirms that excluding hybrid APCs from eligible costs in the future Horizon Europe programme is a significant measure to increase the cost-efficiency of the programme's open access policy**. The available data shows that 'hybrid' options (subscription journals that also offer open access to individual articles on payment of an APC) have considerably higher average APCs than fully open access titles. One study carried out in 2016 of APCs covered by UK institutions showed that between August 2014 and July 2015, the average APC for a hybrid journal was GBP 1,882, while for a full open access journal it was GBP 1,354. Moreover, hybrid journals made up 80% of APC expenditure in 2014-2015.[108] Earlier studies also confirmed that open access publishing via hybrid journals is significantly more expensive than via fully open access journals.[109] The available evidence also shows that where funds are made available to cover APCs, it is more likely that a more expensive, hybrid solution will be chosen instead of a purely open access journal.[110] The beneficiaries interviewed for this study largely agreed with the plans to cease support for hybrid journal APCs in the Horizon Europe programme.[111]

Some beneficiaries expressed **a need to fund the APCs/BPCs for post-project publications that resulted from the grant activities**. In many cases, a

[108] Shamash, K. (2016). Article processing charges (APCs) and subscriptions: Monitoring open access costs. https://www.jisc.ac.uk/reports/apcs-and-subscriptions
[109] Björk, B.-C., & Solomon, D. (2014). Developing an effective market for open access Article processing charges. Final report. https://doi.org/10.6084/m9.figshare.4873532.v1
[110] van der Graaf, M. (2017). Paying for open access - The Author's Perspective. *Zenodo*. https://doi.org/10.5281/zenodo.438037
[111] https://www.openaire.eu/horizon-europe

publication based on Horizon 2020 activities is actually published after the formal end of the project. This is particularly common in the Social Sciences and Humanities, where among the main outputs (in addition to scientific articles) are books and book chapters, which usually take much longer to publish. In some cases, books or book chapters prepared on the basis of Horizon 2020 project activities might be published years after the end of the project. In such cases, the BPCs/APCs for open access publishing cannot be covered out of the project budget anymore. A number of beneficiaries indicated this as a common barrier to ensuring open access to all Horizon 2020 project outputs – in some cases, it might even prevent the publication of outputs prepared after the project has ended.

### *6.3 Effectiveness of the Horizon 2020 open access policy*

The effectiveness of the Horizon 2020 open access policy refers mainly to the extent to which the policy has succeeded in achieving its goals, and the progress of the policy. In addition, in this section we also summarise the key lessons learned on the basis of the evidence stemming from interviews and desk research, with regard to the impacts of the Horizon 2020 open access policy on researchers, other stakeholders and on society as a whole. Although a much wider exercise would be required in order to comprehensively assess the effects and impacts of the Horizon 2020 open access policy at individual, organisational and system level, the following sub-sections present a 'stock-taking' of evidence on the effects of Horizon 2020, collected during the previous stages of the present study.

**Effectiveness of Horizon 2020 open access policy – progress over time, variation by programme and discipline**

Our analysis of open access success rates among publications over time shows that on average, the open access rates among Horizon 2020 publications have increased steadily over the programme's duration, from just over 65% of peer-reviewed publications in 2014 to 86% in 2019 (for more details, see Section 2.1.2 above). The average open access rate for the whole period of 2014-2020 was around 83% of all peer-reviewed publications that resulted from Horizon 2020 grants.

Analysis of the evidence provided by the exercises carried out under Tasks 1 and 2 of the present study also shows that the effectiveness of Horizon 2020 **differs somewhat between different Horizon 2020 programmes**. In some Horizon 2020 programmes, for instance, a higher share of publications is open access than in others. Referring to Section 3.1.2, and examining the open access rates of those programmes with the *highest number of publications,* we find that European Research Council (ERC) grants have an open access share of 88%, well above the European Commission's average of around 83%, whereas Euratom, Twinning of Research Institutions, and Leadership in Enabling and Industrial Technologies (LEIT) grants have much lower rates (65%, 66% and 79%, respectively). More extensive analysis and in-depth, qualitative evidence would be necessary to identify the precise reasons behind this variation in open access rates between different Horizon 2020 programmes.

Further analysis of open access publications under Horizon 2020 also confirms that the percentage share of open access publications **varies by scientific field and specific discipline**. The highest percentage of open access publications can be found in medical and health sciences (88%) and natural sciences (82.8%), while the share was lower within agricultural and veterinary sciences (74.2%), engineering and

technology (77.9%), social sciences (78%) and humanities and arts (78.2%) (see Table 5).

More detailed analysis also shows that in some cases, the **variation in percentage share of open access publications also exists at the level of specific disciplines within particular scientific fields in Horizon 2020**. For example, in the field of natural sciences, the share of open access varies from 78% in chemical sciences to 86% in biological sciences. In most fields, however, variation between disciplines is much less pronounced, ranging between 2 and 5 percentage points (see Figure 11).

## Impacts and benefits of Horizon 2020 open access policy - Lessons learned

Qualitative evidence collected during our interviews with stakeholders (mainly project officers and beneficiaries) also shows that the current Horizon 2020 open access monitoring system is effective at ensuring compliance with open access obligations, and in guiding researchers towards fulfilling these obligations during the project lifecycle. Although the evidence shows that these obligations are often not fully met by the mid-point of the project, **towards their end, projects usually become increasingly compliant with the programme's open access requirements**. The key catalyst in this transformation is the support and feedback received by beneficiaries from project officers, who identify the most problematic areas of a project in terms of open access compliance, and provide guidance on how to correct existing gaps in compliance.

Interviews with beneficiaries also confirmed that open access to Horizon 2020 project outputs resulted first of all in the **wider outreach and dissemination of the research work across different fields and to the general public**. This, in turn, significantly increases the visibility and potential impact of the research. Anecdotal evidence shows that open access to Horizon 2020 outputs increased dissemination from several hundred books sold to over a million downloads of the same book.

> *"In my case, open access was very successful <…> in my field we normally would have sold around 600 copies of books, but in open access we have over a million of downloads. The success of my project is almost entirely based on open access."' (From an interview with a beneficiary)*

Qualitative evidence also shows that Horizon 2020 requirements and the subsequent need to report on open access to research outputs has a learning effect on beneficiaries. A number of stakeholders interviewed for this study indicated that the experience of fulfilling the Horizon 2020 open access obligations led to **increased awareness and knowledge among beneficiaries with regard to Open Science** concepts and principles, and the processes and measures required to ensure open access to research outputs. Anecdotal evidence shows that Open Science learning effects stemming from the Horizon 2020 open access policy were particularly significant among beneficiaries from countries in which open access is not very widespread or developed, whereas these effects were less distinct among the researchers from countries where the Open Science movement began earlier and is well advanced (e.g., in the UK).

Finally, as we mentioned in our discussion of the Horizon 2020 open access policy's intervention logic, the open access principles at EU level implemented via the Horizon 2020 programme often **encouraged other European research funders and**

**institutions to adopt similar open access policies.**[112] One of the most significant recent initiatives is Plan S, which was prepared by a group of national research funding organisations, with the support of the European Commission. According to Plan S, "with effect from 2021, all scholarly publications on the results from research funded by public or private grants provided by national, regional and international research councils and funding bodies, must be published in open access journals, on open access platforms, or made immediately available through open access repositories without embargo."[113]

[112] Guedj, D. & Ramjoué, C., European Commission Policy on Open Access to Scientific Publications and Research Data in Horizon 2020 Biomed *Data J.* 2015; 1(1): 11-14, https://doi.org/10.11610/bmdj.01102.
[113] https://www.coalition-s.org/about/

# 7 Annex

## *7.1 Article 29.2*[114]

**BOX 1.**

**29.2 Open access to scientific publications**

Each beneficiary must ensure open access (free of charge, online access for any user) to all peer-reviewed scientific publications relating to its results.

In particular, it must:

(a) as soon as possible and at the latest on publication, deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications;

Moreover, the beneficiary must aim to deposit at the same time the research data needed to validate the results presented in the deposited scientific publications.

(b) ensure open access to the deposited publication — via the repository — at the latest:

(i) on publication, if an electronic version is available for free via the publisher, or

(ii) within six months of publication (twelve months for publications in the social sciences and humanities) in any other case.

(c) ensure open access — via the repository — to the bibliographic metadata that identify the deposited publication.

The bibliographic metadata must be in a standard format and must include all of the following:

- the terms ["European Union (EU)" and "Horizon 2020"]["Euratom" and Euratom research and training programme 2014-2018'];
- the name of the action, acronym and grant number;
- the publication date, and length of embargo period if applicable, and
- a persistent identifier.

### 7.1.1 Article 29.2 – ERC specificities

The following specificities relating to the ERC are taken into account when considering the compliance of ERC grants with Article 29.2.[115]

- *"Moreover, for publications after the end of the action, if beneficiaries cannot provide open access within the time limits set by Article 29.2 without incurring additional costs for 'gold' open access, they may choose 'green' open access with an extended embargo period which goes beyond six/twelve months."*
- *"Finally, the ERC MGA foresees lighter requirements for bibliographic metadata, focusing only on a persistent identifier (i.e. a stable address/marker to identify the publication, such as a digital object identifier (DOI) or other systems)."*

[114] https://ec.europa.eu/research/participants/data/ref/Horizon 2020/grants_manual/amga/Horizon 2020-amga_en.pdf

[115] The ERC specific annotations in the AGA also include some guidelines for good practice.

### *7.2 Article 29.3 (relevant excerpt)*

The box below presents the excerpt of Article 29.3 that is **relevant to this study**. In other words, **it does *not* concern:**

1. Information that can be found only in data management plans (e.g. "the data management plan must contain the reasons for not providing access"). For a discussion of data management plans, see Section 7.4.1.1 of the Annex;

2. Direct communication with the European Commission (e.g. "data which is relevant for addressing a public health emergency, if specifically requested by the [Commission][Agency], and within the deadline specified in the request"); or lastly

3. "Tools and instruments" necessary for validating results. We were not able to collect such data, as they are neither requested by project beneficiaries in SyGMa, nor provided as metadata in repositories.

**BOX 2.**

**29.3 Open access to research data**

*[OPTION 1a for actions participating in the open Research Data Pilot: Regarding the digital research data generated in the action ('**data**'), the beneficiaries must:*

*(a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:*

*(i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;*

#### 7.2.1 Article 29.3 – ERC specificities

The following specificities relating to the ERC should be taken into consideration with regard to ERC frontier research actions participating in the Open Research Data pilot.

- *"The beneficiaries may opt out of the pilot at any stage – both before signing the GA and afterwards (through an amendment; see Article 55 Horizon 2020 General MGA). No reasons have to be provided for opting out. By opting out, they free themselves retroactively from the obligations associated with taking part in the pilot."*

### *7.3 Methodology for Horizon 2020 publications*

In this section, we describe the methodology followed in order to arrive at an authoritative list of Horizon 2020 publications, to cross-validate and enrich their metadata, and to build a *robust* and *reproducible* database of Horizon 2020 publications, their metadata and their indicators, as presented in Table 12.[116]

[116] In the MOAP Horizon 2020 DB, indicators are conveniently saved as Boolean (0,1) variables for each publication (e.g., open_access → is the publication open access, erc_only → is the publication linked to an ERC project only and so on) to facilitate aggregation into different levels of interest (per programme, per year, etc.).

The work conducted and the methodological steps taken are broken down into the following sub-sections:

- Compiling the list of Horizon 2020 publications
- Collecting, creating and triangulating metadata
- Article/book processing charges (A/BPCs)

### 7.3.1 Compiling the list of Horizon 2020 publications

The first task we undertook was to compile an authoritative list of Horizon 2020 peer-reviewed scientific publications – the set relevant for examining compliance with Article 29.2. Our approach was to **cross-validate Horizon 2020 publications** stemming from different sources, and to ascertain as accurately as possible, which of them were **peer-reviewed**. We begin by summarizing the methodology and findings, and in the next sub-section these are presented in detail, including the issues we encountered with the data.

The *main* data sources for this task were:

1. Data shared by the EC, as reported in the participant portal SyGMa[117] (henceforth, 'EC-Shared'); and
2. The OpenAIRE[118] Research Graph[119] (henceforth, 'ORG').

We also used the following *supplementary* sources:

3. Web of Science (WoS)[120]; and
4. Scopus.[121]

EC-Shared and ORG data were updated three times during this study in order to procure an up-to-date set. Considerable effort was put into cleaning the metadata entries in EC-Shared to achieve the highest quality matching with the other data sources and to *minimise* the number of *duplicates.*

Moreover, to construct *reliable* indicators, and based on our triangulation approach, any EC-Shared publications that *could not be found* in any other data source were *discarded* from the analysis; the next sub-section presents and discusses their characteristics.

On the other hand, as OpenAIRE ingests, merges and de-duplicates records across a host of content providers, publications found in ORG are likely to be cross-validated prior to this exercise.

After merging records from all data sources, we isolated those that have been **peer-reviewed** and are the relevant set for evaluating compliance to Article 29.2 by examining

[117] The European Commission's System for Grant Management
[118] https://www.openaire.eu/
[119] https://graph.openaire.eu/
[120] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[121] https://www.scopus.com/home.uri

1. the publication type (excluding grey literature such as pre-prints, reviews, reports, etc.) and

2. the venue of publication (identifying peer-reviewed venues).

To conclude, Table 30 presents some key characteristics of the MOAP Horizon 2020 publications database.

**Table 30. MOAP Horizon 2020 publications DB**

| **MOAP Horizon 2020** ***publications DB*** | **Total number of publications** |
|---|---|
| **peer-reviewed Horizon 2020 publications** | **154,185** |
| reported in SyGMa | 111,343 (72.2%) |
| found in Scopus (Jan 2021 version) | 121,571 (78.9%) |
| found in WoS (Jan 2021 version) | 115,518 (74.9%) |
| found in the OpenAIRE Research Graph (Mar 2021 version) | 152,211 (98.7%) |

#### *7.3.1.1 Compiling the list of Horizon 2020 publications: detailed methodology*

This section presents in detail the methodology to compile the MOAP Horizon 2020 publications database.

**Steps taken:**

1. EC-Shared data was cleaned.

2. Publications linked to Horizon 2020 projects were fetched from OpenAIRE.

3. ORG and EC-Shared were merged using clean DOIs and Titles.

4. This merged list was triangulated with WoS and Scopus, using DOIs and PMIDs.

5. EC-Shared publications that were not found in any other data source were discarded.

6. Peer-reviewed publications in the merged list were identified.

These steps resulted in **154,185 distinct** publications being included in the **MOAP Horizon 2020 publication DB**.

**Iterations:** during the study, we re-compiled and cleaned the database at various stages, as depicted in Table 31 below.

**Table 31. Data iterations in MOAP**

| Period of the study | EC-Shared Data | ORG instance | WoS and Scopus |
|---|---|---|---|
| Inception stage | August 2021 | October 2021 | - |
| Interim stage | January 2021 | January 2021 | January 2021 |
| Final stage of the study | March 2021 | March 2021 | - |

### *1. Cleaning the EC-Shared data*

The client shared with the research team data reported in SyGMa by beneficiaries. When a project beneficiary reports a publication as a project output, the system may offer suggestions from OpenAIRE. For each suggestion, the beneficiary then has three options:

**Option 1:** The beneficiary can confirm the OpenAIRE suggestion as a project output.[122] In this case, a pop-up shows the publication's metadata, in which:

- The *non-editable* fields are the repository link, title and authors of the publication.
- The *editable fields* are the DOI, type of publication, venue, open access route ('green' or 'gold') and embargo period, peer-review status, and a private/public partnership option.

**Option 2:** reject the OpenAIRE suggestions, i.e. state that this publication has been wrongly linked to the project.

**Option 3:** The beneficiary can manually enter a publication, together with the required metadata fields.

According to information supplied by the client, and to the best of our knowledge, the data we received up to now (EC-Shared) have not been consistently validated *ex-post*.

**Data issue:** *SyGMa*

Manual entry and editing of metadata fields in SyGMa is likely to produce a lot of 'dirty' data.

There is a lack of *ex-post* validation (e.g. by fetching the metadata from the repository URL provided).

DOIs are the only PIDs that can be reported in SyGMa.[123]

[122] When fetching the record from OpenAIRE the system also fetches (using a DOI) the same record from Crossref in order to fill any metadata elements missing from the OpenAIRE record.
[123] For example, ISBNs would be necessary for books.

**Data issue:** *cleaning the EC-Shared data*

To clean the data received from the client, we carried out the following steps:

- Due to the data having been badly encoded, we converted the it to UTF-8 characters, perhaps losing some letters here and there, mainly from titles.
- We removed line breaks in titles and journals names that would have been imported wrongly into the MOAP Horizon 2020 database.
- Visual inspection revealed several invalid DOIs; wherever possible, these were converted into the form 10.[…]/[…].

Other fields were found to be invalid, but it was not possible to clean them consistently. These fields included:

- Embargo period and processing charges.

Lastly, we created a **distinct identifier per publication** (as this was not supplied in the shared data),[124] by using a hash (MD5) of the title of the publication, concatenated with the DOI; when a DOI was not available, we used the project code.

### *2. Fetching publications linked to Horizon 2020 projects from OpenAIRE*

OpenAIRE ingests data from a large set of content providers[125] and de-duplicates records[126] so that the final merged record still includes all instances of the metadata element, while avoiding having duplicates in the Graph. OpenAIRE links research outcomes (publications, datasets, software, other research products) to projects in the following ways:

1. Data shared from the European Commission to OpenAIRE (as reported in SyGMa),
2. As metadata elements from data sources ingested into OpenAIRE,
3. Via text mining of acknowledgements, abstracts and full texts; and
4. Manually added by OpenAIRE portal users via the Link functionality.[127,128]

We began by collecting all OpenAIRE publications linked to Horizon 2020 projects.

[124] ORG assigns unique OpenAIRE IDs to each publication.
[125] OpenAIRE ingests data from the following types of data sources: repositories, open access journals and publishers, aggregators, entity registries, journal aggregators, and CRIS (Current Research Information Systems); https://explore.openaire.eu/search/find/dataproviders .
[126] https://www.openaire.eu/blogs/on-deduplication-in-the-openaire-infrastructure-1
[127] https://www.openaire.eu/claim-publication
[128] Originally, we had planned to run the Athena RC funding extraction algorithm on the Funding Text metadata field in WoS, thus potentially extracting additional Horizon 2020 publications available in WoS. However, their API does not allow for bulk retrieval of records. In other words, there is no easy way to get every single publication in WoS to check each Funding Text for links to Horizon 2020.

### *3. Merging ORG and EC-Shared using clean DOIs and titles*

Next, we merged the two datasets. Our goal was to **maximise the number of matches** between the two, so as to avoid duplicates in the final merged list. We gave priority to DOI matching over Title matching.

The **merged MOAP Horizon 2020 publication database** is a table of pairs of the form (OpenAIRE ID, EC-Shared ID), where both fields are non-empty for the *matched* publications (i.e., a record is found in both ORG and EC-Shared), and one field is empty for *unmatched* publications (e.g. ORG_ID_1234, 0), where an ORG record is not matched to an EC-Shared record.

Moreover, it is also possible for one OpenAIRE ID to be mapped to more than one EC-Shared ID, and vice versa. One such case is when two distinct EC-Shared IDs, linked to two distinct DOIs, are mapped to one OpenAIRE ID, because in ORG the two DOIs have been merged into the same record (i.e. they are different versions of the same document). Therefore, where we describe the process of merging the two datasets in Table 32 below in each case we present separately the number of distinct OpenAIRE and EC-Shared IDs.

**Table 32. Merging ORG and EC-Shared publications**

| | **MOAP HORIZON 2020 PUBLICATIONS DB** | |
|---|---|---|
| | *Unique OpenAIRE IDs* | *Unique EC-Shared IDs* |
| **Step 0:** Start with **206,445 ORG** and **179,517 EC-Shared** publications. | - | - |
| **Step 1:** Match publications based on clean **DOIs** *(i.e. in the form 10.[…]/[…]).* | 122,427 | 122,101 |
| **Step 2:** Match the remaining publications based on **title** *(titles normalised to lowercase, only A to Z characters, w/ length more than five characters and more than two words).* | 9,601 | 9,929 |
| **Step 3:** Search the OpenAIRE publications not linked to Horizon 2020 projects for additional matches to the remaining EC-Shared **DOIs**. | 5,418 | 5,227 |
| **Step 4:** Search the OpenAIRE publications not linked to Horizon 2020 projects for additional matches to the remaining EC-Shared **Titles** (as matched in Step 2) *(while applying the same pre-processing as previously).* | 4,681 | 4,494 |
| **Step 5:** Add the remaining/unmatched records from both databases. | 74,417 | 37,766 |
| **TOTAL** | 216,544 | 179,517 |

### 4. ***Triangulating the merged list with WoS and Scopus, using DOIs and PMIDs***

To triangulate the metadata elements in the MOAP publication list, we matched publications to records from WoS and Scopus using PMIDs and DOIs.

**Three types** of publications are included in the **merged MOAP list**: matched from ORG and EC-Shared, found in ORG only, found in EC-Shared only. Table 33 below presents the results of the triangulation exercise.

**Table 33. MOAP triangulation with WoS and Scopus**

| **MOAP POSTGRES** ***publications*** | **in WoS only** | **in Scopus only** | **in both** | **in neither** |
|---|---|---|---|---|
| **ORG and EC-Shared matches** *(results in number of OpenAIRE IDs)* | 2,739 | 7,242 | 79,914 | 52,245 |
| **in ORG, unmatched to EC-Shared** *(results in number of OpenAIRE IDs)* | 1,090 | 2,367 | 30,444 | 40,516 |
| **in EC-Shared, unmatched to ORG** *(results in number of European Commission IDs)* | 58 | 414 | 1,542 | 35,752 |

**Note:** publications in ORG are already likely to be *triangulated,* as OpenAIRE merges records from several content providers. The publications in the blue box in the table above are found *only* in the EC-Shared data, and are thus the only ones **not triangulated** by any source. Given the issues with metadata entries in SyGMa (see discussion above) and in order to maintain the quality of data in this study, **we have excluded these from the final MOAP Horizon 2020 publication database**.

### 5. ***Discarding EC-Shared publications not found in any other data source***

As mentioned above, we discarded 35,752 EC-Shared publications that could not be validated by any other source. Of those, **32,719 did not have a PID** in the reported data. Below, we present their main characteristics, as reported in SyGMa.

**Table 34. Characteristics of unmatched European Commission publications**

| Publication types (by number of unmatched European Commission publications) | | Top (sub-) programmes (by number of unmatched European Commission publications) | |
|---|---|---|---|
| conference_proceeding | 17,427 | EU.1.1. European Research Council (ERC) | 7,268 |
| other | 9,932 | EU.1.3. Marie Skłodowska-Curie Actions | 6,562 |
| peer_reviewed_article | 3,551 | EU.2.1.1. Information and Communication Technologies | 4,975 |
| book_chapter | 1,971 | EU.3.4. Smart, 'green' and integrated transport | 2,256 |
| thesis_dissertation | 1,572 | EU.3.2. Food security, sustainable agriculture and forestry, marine and maritime and inland water research and the bioeconomy | 2,145 |
| article | 784 | EU.3.3. Secure, clean and efficient energy | 2,111 |
| monographic_book | 663 | EU.1.4 Research Infrastructures | 1,755 |

**Data issue:** the majority of reported publications in EC-Shared that could not be found in any other data source *are missing a DOI.* For the rest, the largest share of unmatched reported publications come from ERC or the Marie Skłodowska-Curie grants, and mostly from conference proceedings.

## *6. Identifying peer-reviewed publications in the merged list*

Table 35 below presents the criteria used to identify the *peer-reviewed status* of a publications, so as to exclude non-peer-reviewed publications from the final list.

**Table 35. Identifying the peer-review status of Horizon 2020 publications**

| NON-PEER-REVIEW CRITERIA[129] | PEER-REVIEW CRITERIA |
|---|---|
| Publications that can only be found as grey literature (pre-prints, reports, etc.)[130] or only in repositories | Publications in WoS or Scopus[131] that are *not* grey literature (document type provided in both indices) |
| Articles without a DOI from Crossref[132] | Articles and conference proceedings in peer-reviewed venues[133] |
| Publications with "non-peer-reviewed" in the `refereed` ORG metadata field[134] | Publications with "peer-reviewed" in `refereed` ORG metadata field[135] |

Applying these criteria led to a total of 64,373 **non-peer-reviewed publications** being excluded from the final database.

**In conclusion, the MOAP Horizon 2020 publications database includes *154,185* publications, of which**:

- *152211* are uniquely identified by OpenAIRE IDs, and
- *1974* identified by EC-Shared IDs (matched only to WoS/Scopus, not to

### 7.3.2 Collecting, creating and triangulating the metadata

In this Section, we discuss the methodological steps taken to finalise the metadata records for the creation of indicators, including an assessment of their quality.

***Triangulation with WoS and Scopus***

We matched publications to WoS and/or Scopus as additional sources for triangulating metadata elements (beyond the triangulation already in the ORG data due to the multiple content providers ingested into OpenAIRE). The main benefit of this exercise was to validate the information regarding the *venues* of publications (journal/conference, publisher, ISSN/ISSBN, publication year), as both databases (WoS and Scopus) are built from publishers that are in principle authoritative sources for these metadata fields. As neither database includes repository information, we were unable to use them for the compliance indicators relating to repository metadata.

[129] We considered the possibility of excluding certain publications on the basis of having too few references in their bibliography (indicating a non-peer-reviewed article). However, upon closer examination, too much variation exists in the number of references even within scientific fields (FOS level 2). In fact, the standard deviation on the number of references within an FOS level 2 is usually as large as the mean number of references within the class. Thus, we removed this criterion as a basis for identifying non-peer-reviewed publications.
[130] We retained articles, conference objects/proceedings, books and book chapters and theses.
[131] Both databases include mostly peer-reviewed journals, and provide a document type that can be used to identify cases of grey literature.
[132] To the best of our knowledge, DOIs from Crossref are the most authoritative source of DOIs for peer-reviewed articles.
[133] The methodology uses Science-Metrix venue labels (https://www.science-metrix.com).
[134] In all their instances in the OpenAIRE Graph.
[135] In at least one of their instances in the OpenAIRE Graph.

***Open access route classification***

There are two ways to produce open access route indicators for publications: constructing them step by step from metadata elements (repository, journal, access rights, etc.), or using the open access route metadata field directly.

***Open access routes from Unpaywall and Scopus***. MOAP Postgres contains open access routes as metadata fields harvested from Unpaywall and Scopus. Both of these sources use the same open access route classification algorithm described in Figure 30 below.

*'Green' open access:* we note that, unlike the definition agreed upon for this study (Table 1), in Unpaywall's algorithm a publication cannot be both 'green' and 'gold'. In fact, 'gold' is given priority over 'green' (Unpaywall chooses a 'best open access location' for publications, with the publisher having priority over repository). We have therefore not used the 'green' open access information harvested from Scopus and Unpaywall.

*'Gold' open access:* since 'gold' is given priority over 'green' by Unpaywall, we were able to use those entries in the creation of the indicators. However, the routes *'gold'*, *'hybrid' and 'bronze' all* belong to the *'gold'* definition used in this study (Table 1), and are thus all aggregated into 'gold'.

So, without further ado, here's a flowchart-style description of how we determine the oa_status for each article:

- Ok, let's get started. **Is the article open access?**
  - no (`is_oa = false`): Ok, oa_status is **`closed`**. Done.
  - yes (`is_oa = true`). Let's learn more. **Where is the best copy of the article hosted?**
    - In a repository (`best_oa_location.host_type = "repository"`): Ok, oa_status is **`green`**. Done.
    - On the publisher website (`best_oa_location.host_type = "publisher"`). Let's learn more. **Is the article published in a fully-OA journal?**
      - yes (`journal_is_oa = true`): Ok, oa_status is **`gold`**. Done.
      - no (`journal_is_oa = false`). Let's learn more. **Is the article published under an open license?**
        - yes (`best_oa_location.license is not null`): Ok, oa_status is **`hybrid`**. Done.
        - no (`best_oa_location.license is null`): Ok, oa_status is **`bronze`**. Done.

Figure 30. Unpaywall's open access route classification algorithm

***Open access routes constructed in MOAP***

*'Green' open access:* we assigned 'green' open access to publications according to the definitions set out in Table 1: .

*'Gold' open access:*

1. We used the Directory of Open Access Journals (DOAJ, updated monthly)[136] and the ISSN-GOLD-OA 3.0[137] database (updated every three months) to identify all 'gold' open access journal titles and assign 'gold' open access to corresponding publications.

[136] https://doaj.org/
[137] https://pub.uni-bielefeld.de/record/2934907

2. To identify the remaining 'gold' open access publications (where the publisher provides open access, but it is not a 'gold' open access venue), we used ORG data to examine the licences from the original data sources that hosted the publications. When the original data source for a publication is a journal or publisher, and the licence is open (see the indicators and the discussion on licences in Section 3.1.2), we assigned the corresponding publications as having 'gold' open access.

***Research areas and FOS classification***

Discipline/field classification provides additional facets, as well as valuable insights and trends to explore, with regard to Horizon 2020 open access and open research data uptake. Moreover, it enables and facilitates advanced bibliometric/citation analysis; for example, the field-weighted citation indicators can be calculated on Horizon 2020/Horizon Europe publications at a granular level:

- Engineering and tech > civil engineering > transportation > railroad engineering
- Nanotechnology > nano-materials > nanostructures > graphene
- Medical and health > basic medicine > neurology > Parkinson's

Our classification system is based on two resources: (a) the OECD disciplines/fields of research and development (FORD) classification scheme, developed within the framework of the Frascati Manual and used to classify R&D units and resources in broad (first level [**L1**], one-digit) and narrower (second level [**L2**], two-digit) knowledge domains, based primarily on R&D subject matter; and (b) the EuroSciVoc, a multilingual taxonomy that represents all of the major **fields of science**[138] in five additional levels **(L3-L7),** connected to the above OECD levels (L1-L2). The discipline/FOS classification system is a publication-based classifier employing publication metadata (i.e. venue, references, citations, title, abstract) as they become available. It automatically assigns one or more FORD codes at both levels (L1/L2) to each publication. In addition, it automatically expands this two-level classification using the relevant EuroSciVoc fields of science, based on a content analysis of the publication abstract. We also note that datasets are classified into disciplines/fields of science using their links to other publications.

Below, we summarise the steps we followed in classifying the Horizon 2020 publication list for the MOAP study:

1. For each OpenAIRE ID, we fetched the metadata set (including venue, references, citations, title and abstract);
2. We then merge and de-duplicated citation metadata coming from different sources; our analysis is currently based on Crossref (version: Nov 2020), OpenCitations (version: Sept. 2020) and Microsoft Academic Graph (version: Nov 2020). All sources were integrated and made available through OpenAIRE;
3. OpenAIRE IDs were classified, along with their L1-L2 disciplines [FORD codes];
4. Classifications were expanded to include L3-L7 fields of science [EuroSciVoc codes].

[138] EuroSciVoc, managed by the Publications Office of the EU, is developed as a reference vocabulary for the Open Science community, and is currently used by the CORDIS website. The current version, used in this study, is version 1.2.

A total of 131,972 OpenAIRE IDs were automatically classified. Of these, 128,943 were assigned one or more FOS/FORD codes. The remainder (3,029) were classified as 'general science', and mostly consisted of articles published in multidisciplinary journals or mega-journals such as *PLoS One*, *Nature* and *Science*. Approximately 13,500 OpenAIRE IDs were excluded or not classified due to a lack of venue information and other metadata (e.g. missing references, lack of citations, no abstract). Excluded publications fell into one of the following subtypes: 'unknown'; 'part of book or chapter of book'; 'review'; 'report'; 'contribution to newspaper or weekly magazine'; 'book'; or 'external research report'.

### ***Embargoed publications***

Article 29.2 of the MGA requires that open access to a publication must be provided within a specific period. Specifically, if there is an embargo on a publication, it must expire within 6/12 months (depending on the scientific domain) of the date of publication.

We were able to estimate compliance with this policy for Horizon 2020 publications for which an **embargo end date** was available in their metadata (see Table 12). Nevertheless, we cannot accurately assess the quality of this indicator because it is impossible to know the **full set of originally embargoed publications,** as no data are available on the original access rights of publications. In other words, it is impossible to identify which publications were originally embargoed but are now open access, in order to examine their compliance.[139]

**Data issue/lesson learned:** because historical data, such as the *original* access rights and *the first date* a publication became open access, are not metadata elements typically exposed by repositories, *compliance* in terms of date of open access/embargo date cannot be accurately estimated.

### ***Metadata standards***

Open access compliance requires that the repository metadata follow certain *metadata standards.* We can examine the latter using the OpenAIRE Validator service.[140] The OpenAIRE Validator service has been developed to help repository managers assess the compatibility of their metadata records against the OpenAIRE guidelines.[141] Each version of the guidelines contains a number of validation rules, which may be either *mandatory* or *recommended*. Validation of a metadata record against the guidelines means making sure that the record complies with as many of the rules as possible. Every mandatory rule carries a weight, and the sum of the weights of all successful rules (normalised to 100) is the final score of the validated metadata record.

Since the target audience for the Validator are repository managers, *the expected input format of the service is the OAI-PMH protocol*, which is practically the only protocol used to harvest metadata records from repositories. This means that for the purposes of this study, *only 30,211 of the records were validated*, since the rest of the records come from other sources (Unpaywall, Microsoft Academic Graph, etc.) which do not offer up their metadata using the OAI-PMH protocol. In total, we fetched

[139] Embargo periods are included in EC-Shared, but as also confirmed by the client, this field is dirty and likely to be unreliable.
[140] Available at https://provide.openaire.eu.
[141] For the purposes of this study, we used the OpenAIRE guidelines for literature repositories v3.0, the most recent guidelines in use by OpenAIRE, available at https://guidelines.openaire.eu/en/latest/literature/index_guidelines-lit_v3.html .

103,717 distinct original metadata records from 506 data sources that match to 30,211 unique (de-duplicated) OpenAIRE records.

***Accessibility and interoperability: publications***

**Accessibility:** the DocUrlsRetriever[142] program was developed by Athena Research Center to check the accessibility of the full text of a publication. After inputting all URLs available for a publication (as metadata elements), the program connects with the corresponding web pages and uses various smart techniques to retrieve the **full-text URLs.** These techniques include, but are not limited to: checking the metadata records within the landing page's response body; applying text-mining to the internal links and the data that accompanies them; applying offline redirects; URL transformations and API calls for specific domains. The output of the program comprises retrieved full-text links. These are then plugged into another program that verifies if the links provide an accessible full text.

The final output includes the following information per URL link:

- whether it is valid;
- whether the file is accessible via the link; and
- whether the URL links directly to the file (i.e. if it is a direct link to the full text – if not, the linked site is crawled for the full text link)

For the purposes of the study, we identified 2.31 million potential URLs to publications and datasets, using both ORG and EC-Shared data. Processing this amount of information requires a considerable amount of computational power. Our solution was to share those URLs between eight virtual machines to distribute the load and reduce the total execution time. Even using this solution, it took a total of five days to obtain the final results.

**Interoperability:** to examine whether the accessible texts are also interoperable (i.e. in a machine-readable file format), their file format must also be considered. However, the algorithm described above looks specifically for PDF files, the most popular format for publications. PDFs can be machine-readable once converted into text; however, this process is prone to errors, due to various obstacles encountered in the conversion. Further post-processing/text normalisation is required before applying text mining. In other words, the publications that are accessible are also interoperable, conditional on the machine-readability of PDFs.

***Citation band (for breakdown of indicators)***

The MOAP Horizon 2020 database provides data on *the number of references and the number of citations of a* publication. Data on the number of publications citing a publication were merged from Microsoft Academic Graph (version: Aug 2020), OpenCitations (version: Sep 2020) and Crossref[143] (version: May 2020); all sources integrated are made available through OpenAIRE.

[142] https://github.com/LSmyrnaios/DocUrlsRetriever.
[143] Citation data are currently in OpenAIRE beta awaiting validation before they go into production. There are currently about 62M citation pairs (publication, citing publication).

***Metadata coverage for indicators***

Table 36 presents the final coverage of the metadata fields required to construct the MOAP Horizon 2020 publication indicators. Several iterations of cleaning the metadata elements were carried out, so that the numbers below refer, as far as possible, to *valid* entries.

**Table 36. Metadata gap analysis for publication indicators**

| METADATA ELEMENT | Coverage | Notes and data issues |
|---|---|---|
| **Number of funders,**<br>**Number of projects** | 98.7%<br>98.7% | |
| **Number of authors** | 98.7% | |
| **ORCID**[144] | 27.9% | Although not required by Article 29.2, and thus *not* essential for this study, good coverage of author IDs provides a great deal of assistance in linking research outputs and examining collaboration networks. |
| **Open access route classification**[145] | 95.2% | |
| **Data source** *(name, type)* | 100% | Including repositories and journals. |
| **Access rights**[146]<br>*in merged record:*<br>*in repository metadata:* | <br>100%<br>99.1% *of 'green' open access publications* | |
| **Date published**<br>*in merged record:*<br>*in repository metadata:* | <br>98.2%<br>98% *of 'green' open access publications* | Year published (as opposed to date: YYYY-MM-DD) was available for 99.3% of publications (in the merged record). |
| **Date of deposition in repository** | ~0% *of 'green' open access publications* | **Issue:** repositories do not normally expose this information. |
| **Embargo end date** | 4,593 *publications* | Impossible to assess coverage (see the Section 'Embargoed publications', above) |
| **Version deposited in repository** | 88.8% *of 'green' open access publications* | **Issue:** Unpaywall provides good coverage of versions in repositories, but this is not a common metadata element for other content providers. |

[144] Share of publications with at least one ORCID of an author of the publication provided in the metadata.
[145] As found in a publication's metadata.
[146] A publication's access rights being available in at least one of the data sources in which the record is found. Types of access rights available: open, embargo, restricted, closed.

| | | |
|---|---|---|
| **PID**<br>*in merged record:*<br>*in repository metadata:* | <br>99.8%<br>94.8% *of 'green' open access publications* | DOIs have been cleaned, but other PID types may include 'dirty' values. |
| **Valid URL**<br>*in merged record:*<br>*in repository metadata:* | <br>96.9%<br>88.6% *of 'green' open access publications* | These numbers include the URL links reported in SyGMa. The MOAP Postgres includes the validity and accessibility per link for each |
| **Licences** | 81.6% | |

### 7.3.3 APCs and BPCs

For this task, we identified and extrapolated the publishing costs for Horizon 2020 publications. One potential source of data that we **did not use** is the **processing charges available in European Commission data,** since close inspection shows that it is likely to be invalid (numbers not comparable with other sources).

The source of APCs in ORG is OpenAPC. This is the largest database for APCs paid by **academic institutions** and **funders**. Decentralised APCs paid by faculties or *individual authors* are *not covered in the database*, thus OpenAPC is not comprehensive.

First, we present some summary statistics on the MOAP Horizon 2020 publications database and its overlap with the OpenAPC data.

**Table 37. Original APCs/BPCs in MOAP**

| SUMMARY STATISTICS ON MOAP HORIZON 2020 PUBLICATIONS AND OPENAPC DATA | |
|---|---|
| Number of 'gold' publications | 86,767 |
| Number of 'gold' non-book publications (i.e. excluding books and book chapters) | 85,971 |
| Number of 'gold' non-book publications with APCs provided by OpenAPC | 4,423 (5.1% of 85,971) |
| Number of 'gold' books or book chapters | 935 |
| Number of 'gold' books with BPCs provided by OpenAPC | 13 |
| (min, max) of BPCs for these 9 books or book chapters | (527.12 EUR, 18,000 EUR) |

***BPCs***

As shown in Table 37, the BPCs available for Horizon 2020 publications are insufficient for extrapolating the values for the rest of the books and book chapters. Instead, we considered using the entire OpenAPC BPC database for the extrapolation.

Table 38 below presents the publishers in the MOAP Horizon 2020 publications database according to the number of 'gold' open access books or books chapters available, and the corresponding number of BPCs available in the OpenAPC BPC database.

Due to the overlap between the two databases being insufficient for an extrapolation exercise, and because Springer Nature is by far the most prominent publisher of Horizon 2020 'gold' books and book chapters, the possibility of obtaining the BPCs via direct communication and exchange of data with the publisher should be considered.

**Table 38. Top publishers of Horizon 2020 'gold' books/book chapters**

| **Top publishers** *(by number of books and book chapters in MOAP Horizon 2020)* | **NUMBER OF 'gold' books and book chapters in MOAP Postgres** | **NUMBER OF BPCs available in the OpenAPC BPC database** |
|---|---|---|
| **Springer Nature** | 438 | 15 |
| **Association for Computing Machinery** | 70 | 3 |
| **Elsevier** | 53 | 0 |
| **Wiley** | 29 | 0 |
| **Routledge** | 25 | 24 |
| **Society for Industrial and Applied Mathematics** | 15 | 0 |
| **Taylor and Francis** | 14 | 1 |

## ***Methodology used for extrapolating APCs***

As summarised in Table 37, APCs were available for only 5.1% of 'gold' Horizon 2020 publications. Therefore, in the interest of achieving accurate estimates, rather than using these APCs, we chose to extrapolate the values for the remaining publications from the entire OpenAPC APC database, which includes APCs for 123,337 publications.

We based the extrapolation of APCs on grouping publications from the MOAP Horizon 2020 publications database and the OpenAPC APC database according to three factors:

1. quantile of the Source Normalised Impact per Paper (SNIP) score in the CTWS database;[147]

[147] CTWS (2018). CWTS Journal indicators. Version 2018-05-01. Leiden University's Centre for Science and Technology Studies. http://www.journalindicators.com/Content/CWTS%20Journal%20Indicators%20May%202018.xlsx

2. whether the publication is pure 'gold' open access or 'hybrid' (Table 1);

3. the year of publication.

We decided on this grouping by following the findings of Schönfelder (2020)[148,149] which quantitatively identified the most significant predictors of APCs.

After publications were grouped, we took the average APC for the OpenAPC database and assigned it to all of the publications in the same group within the MOAP Horizon 2020 'gold' publications. The APCs for the 5.1% of publications from OpenAPC that originally had an APC were not altered.

This process resulted in APCs (either original or extrapolated) for:

- **66,306** publications (71.1% of 'gold' open access publications)

These were then used in the analysis presented in Section 3.2.

## *7.4 Methodology for Horizon 2020 datasets*

In this section, we present the methodology followed to converge the data into an authoritative list of Horizon 2020 datasets, to cross-validate and enrich their metadata, and to build a *robust* and *reproducible* database of Horizon 2020 datasets. Their metadata and indicators are presented in Table 23.[150]

### 7.4.1 Compiling the list of Horizon 2020 datasets

Our first task was to converge the data into an authoritative list of datasets that were **created/produced** by Horizon 2020 projects. Of these, the sets relevant for examining *compliance* with Article 29.3 are those that participated in the ORDP and did not opt out.

We first **cross-validated Horizon 2020 datasets** stemming from data reported to the European Commission, and the OpenAIRE Research Graph (which ingests a host of content providers). Second, we ascertained which of these were actually produced by the projects themselves. In other words, we wanted to identify those datasets that had not existed previously and were reused by the Horizon 2020 projects. The latter, reused datasets, were removed from the MOAP Horizon 2020 datasets DB. Below, we summarise our approach and findings. These are presented in greater detail in the next sub-section (including a discussion on data management plans as sources of datasets).

When this task was concluded, we had converted to a set of **6,231 distinct Horizon 2020 datasets**. The *main* data sources for this task were:

1. The data shared by client (EC-Shared), as reported in SyGMa; and

[148] Schönfelder, Nina. "Article processing charges: Mirroring the citation impact or legacy of the subscription-based model?" *Quantitative Science Studies* 1.1 (2020): 6-27, https://doi.org/10.1162/qss_a_00015
[149] We were unable to use the scientific domain of publications for the extrapolation as these data are not available for the publications in OpenAPC. However, domain was not found to be a significant determinant of APCs by Schönfelder (2020).
[150] In the MOAP Horizon 2020 database, indicators as conveniently saved as Boolean (0,1) variables for each dataset (e.g. open_access → is the dataset open access, in_repo → if the dataset was found in a repository harvested by OpenAIRE) to facilitate aggregation into different aspects of interest (per programme, per year, etc.).

**2.** Data from the OpenAIRE Research Graph (ORG)

ORG data was updated three times during this study in order to ensure the list of Horizon 2020 datasets was up to date.

As with publications, considerable effort was put into cleaning the metadata entries in EC-Shared to achieve the highest quality matching with ORG and to *minimise* the number of *duplicates.*[151]

To generate *reliable* indicators, and on the basis of our triangulation approach, any EC-Shared datasets that *could not be found* in any other data source were *discarded* from the analysis.

On the other hand, datasets in found OpenAIRE were likely to have been triangulated already given the host of content providers harvested by OpenAIRE. Moreover, beneficiaries in projects *not* participating in the ORDP are not able to report the datasets they produce in SyGMa; thus, as expected, there was a large number of datasets in ORG that were not found in EC-Shared.

Our final step was to isolate the datasets that were **created** by Horizon 2020 projects (as opposed to those that were reused by the projects), as these constitute the relevant set for evaluating compliance with Article 29.3. In order to do so, we identified as *newly produced* datasets those:

**1.** reported in SyGMa by project beneficiaries, and

**2.** those with a project reference in their metadata.

Table 39 presents the final MOAP Horizon 2020 datasets DB with some figures of interest.

**Table 39. MOAP database of Horizon 2020 datasets**

| MOAP Horizon 2020 datasets DB | Total number of datasets |
|---|---|
| **MOAP HORIZON 2020** | **6,231** |
| Datasets reported in SyGMa | 2,815 |
| Datasets produced in projects that participated in the ORDP and did not opt out | 5,244 |

#### *7.4.1.1 Compiling the list of Horizon 2020 datasets: detailed methodology*

This sub-section presents in detail the methodology we followed to compile the MOAP database of Horizon 2020 datasets.

[151] Unmatched datasets that appear that appear as distinct datasets, but are in fact the same dataset.

**Steps:**

1. EC-Shared data were cleaned.
2. Fetched datasets linked to Horizon 2020 projects from OpenAIRE.
3. ORG and EC-Shared were merged using clean DOIs and Titles.
4. EC-Shared datasets not found in OpenAIRE (including providing

After performing these steps, we arrived at **6,231 distinct** datasets in the **MOAP Horizon 2020 dataset** list, **5,244 of which were from ORDP participant projects** that did not opt out of the pilot. In the rest of this Section, we outline in detail the steps above.

### *1. Cleaning the EC-Shared data*

The client shared with the research team data reported in SyGMa by beneficiaries. At this time, it was noted that projects that do not participate in the ORDP are *not* able to report datasets as project outputs. A list of projects participating in the ORDP and those that opted out (along with their reasons of opting out) was also provided.

In a similar manner to reporting a publication, when a project beneficiary reaches the screen in SyGMa where they can report a dataset as a project output, they are presented with dataset suggestions from OpenAIRE. For each suggestion, they are offered the following options:

**Option 1:** Confirm the OpenAIRE dataset suggestion as a project output. In this case, a pop-up shows the dataset's metadata, where:

- The *non-editable* fields are the DOI, the repository link, the non-repository link, whether the dataset is accessible, and *the DOI of the linked publication* (if one exists);
- The only *editable field* is whether the dataset is reusable.

Note the difference with publication reporting, where most of the metadata fields are editable. Moreover, when a dataset if fetched from OpenAIRE, SyGMa also fetches the corresponding record (using a DOI) from *Crossref* in order to fill out metadata elements missing from OpenAIRE. Nevertheless, the majority of datasets can be found in DataCite, thus the process is unlikely to improve the metadata record of the dataset.

**Data issues:**

- When fetching a dataset from OpenAIRE, a user is *not* able to enter the DOI for the **linked publication**, thus significantly limiting the potential benefits of the reporting tool.

- Augmenting missing OpenAIRE metadata elements by looking for a dataset record in Crossref is *unlikely* to produce any results.

**Option 2:** Reject one the OpenAIRE suggestion, i.e., state that this dataset has been wrongly linked to the project.

**Option 3:** Manually enter a dataset, and its metadata fields.

Due to the issues described above, and since, to the best of our knowledge, SyGMa data shared with the contractor has *not* been validated *ex-post*, in this study we use only those European Commission datasets that can be validated by other sources. This yields a triangulated set of data and more reliable indicators.

**Data issue: *cleaning EC-Shared data.***

To clean the data received from the client, we carried out the following steps:

Due to data having been badly encoded, we converted it to UTF-8 characters, perhaps losing some letters here and there, mainly from titles (in an Excel file).

We removed line breaks in titles and journals names that would have been imported wrongly into the MOAP Postgres.

Visual inspection revealed several invalid DOIs; wherever possible, these were converted into the form 10.[…]/[…].

Lastly, we created a distinct identifier per dataset (as this is not supplied in the shared data), by using a hash (MD5) of the title of the dataset, concatenated with the DOI; where a DOI was not available, we used the project code.

***Fetching datasets linked to Horizon 2020 projects from OpenAIRE.***

We began by collecting all OpenAIRE datasets linked to Horizon 2020 projects. There are three sources of project-dataset links in the ORG:

1. As metadata elements from data sources ingested into OpenAIRE,
2. via text extraction (machine/deep learning) of acknowledgements, abstracts and full texts of publications; and
3. manually added by OpenAIRE portal users via the Link functionality.[152,153]

### *2. Merging ORG and EC-Shared using clean DOIs and Titles*

Next, we proceeded to merge the two datasets. The goal when merging **ORG** and **EC-Shared** Horizon 2020 datasets was to *maximise the number of matches* between the two, so as to avoid duplicates in the final list. We gave priority to DOI matching over title matching.

The **merged Horizon 2020 dataset** list is a table of pairs of the form (OpenAIRE ID, EC-Shared ID), where both fields are non-empty for the matched datasets (i.e. record found in both ORG and EC-Shared), and one field is empty for unmatched datasets (e.g. ORG_ID_1234, 0), when an ORG record was not matched to an EC-Shared record.

Table 40 presents the steps involved in merging the two sets.

[152] https://www.openaire.eu/claim-publication

[153] SyGMa data shared by the European Commission to OpenAIRE is currently in OpenAIRE Beta (not production), as an ongoing problem exists of project beneficiaries reporting publications as datasets.

**Table 40. Merging ORG and EC-Shared datasets**

| | MOAP DATABASE OF HORIZON 2020 DATASETS | |
|---|---|---|
| | ***Unique OpenAIRE IDs*** | ***Unique EC-Shared IDs*** |
| **Step 0:** We began with **6,958** ORG and **4,890** EC-Shared datasets. | | |
| **Step 1:** Datasets were matched on the basis of clean **DOIs** *(i.e. of the form 10.[…]/[…]).* | 1,830 | 2,180 |
| **Step 2:** The remaining datasets were matched on the basis of **title**<br>*(titles were normalised to lowercase, only A to Z characters, w/ length more than 5 characters and more than 2 words).* | 133 | 135 |
| **Step 3:** We searched those ORG datasets not linked to Horizon 2020 projects for additional matches to the remaining EC-Shared *DOIs.* | 448 | 454 |
| **Step 4:** We searched those ORG datasets not linked to Horizon 2020 projects for additional matches to the remaining EC-Shared *Titles* (as matched as in Step 2) | 404 | 349 |
| **Step 5:** We added the remaining/unmatched records from both databases. | 4,995 | 1,772 |
| **TOTAL** | **7,810** | **4,890** |

We note that datasets in the ORG were already likely to be cross validated, as OpenAIRE merges records from several content providers. However, those that can be found *only* in EC-Shared data (1,772 unmatched) **cannot be triangulated** with any source. Moreover, given the concerns expressed regarding the reporting of datasets on SyGMa, as well as the lack of metadata field coverage available in EC-Shared data, **we decided to exclude the unmatched EC-Shared datasets from the final MOAP Horizon 2020 Dataset list, as well as from the ORDP uptake and compliance indicators.**

### *3. Discard EC-Shared datasets not found in any other data source*

We discarded 1,772 EC-Shared datasets that could not be validated by any other source; of these, 704 did not have a PID in the reported data (only DOIs allowed in reporting tool). In Table 41, we present some key characteristics of the unmatched set.

**Table 41. Characteristics of unmatched European Commission datasets**

| TOP BENEFICIARY COUNTRY (by number of unmatched European Commission datasets) | | TOP (SUB-)PROGRAMMES (by number of unmatched European Commission datasets) | |
|---|---|---|---|
| United Kingdom | 1,315 | EU.2.1.1. Information and Communication Technologies | 418 |
| Germany | 1,303 | EU.1.2. Future and Emerging Technologies (FET) | 411 |
| Italy | 1,150 | EU.1.3. Marie Skłodowska-Curie Actions | 217 |
| France | 1,096 | EU.3.6. Europe in a changing world - inclusive, innovative, and reflective societies | 148 |
| Spain | 1,072 | EU.1.4. Research Infrastructures | 133 |
| Netherlands | 887 | EU.3.5. Climate action, environment, resource efficiency and raw materials | 117 |

**Data issue:** a large proportion of reported datasets that could not be found in the ORG *are missing a DOI in EC-Shared.* The highest numbers of unmatched reported datasets come from *Information and Communication Technologies* and *Future and Emerging Technologies* grants.

## BOX 3. SUPPLEMENTARY SOURCE OF DATASETS

In an effort to increase coverage, our analysis highlighted the need for an additional path: to datasets directly from the body text of the MOAP Horizon 2020 publications. This methodology entails a considerable number of steps: (a) Compile a collection of body texts from the MOAP Horizon 2020 publications; (b) Transform this into a machine-readable format while retaining its structural information; (c) Identify candidate datasets mentioned in text; (d) Isolate those ascribed to MOAP Horizon 2020 researchers as the publication's authors; (e) Extract additional metadata enriching the dataset description; and (f) Build and validate links between datasets and publications, against findings in the database.

Although our preliminary experiments have been promising, more resources are needed to produce satisfactory results.

Another possible source of datasets is the text mining of Data management plans (DMPs). To examine this possibility, we obtained from OpenAIRE's RDM Task Force[154] the full texts of 841 Horizon 2020 *open access* DMPs. These are deposited in the University of Vienna's PHAIDRA Repository.[155] We examined in detail the possibility of extracting key information from those full texts, namely: ORDP participation; datasets; their access rights and other metadata; the version of the DMP (final deliverable where datasets are already produced vs. inception phase deliverable where everything is only planned).

**Data issue:** Data management plans do not follow a fixed/streamlined format or vocabulary; this renders it difficult to extract information via text mining.

**Lessons learned: why a DMP standard is important, and why separate per-dataset documentation is needed**

DMPs have so far been produced and published in the form of plain text documents that contain narratives about how the data have been created, handled and managed by researchers and data managers throughout the project lifecycle, including provisions for their long-term curation, preservation and reuse after the project ends. To enrich traditional DMP documents and enable their automatic information exchange between systems, the Research Data Alliance (RDA) has developed an application profile for machine-actionable DMPs, named the **RDA DMP Common Standard**.[156] Major DMP tool providers apply this standard to enable the production of concrete DMPs whose information can be further integrated and validated. This has helped greatly, both in terms of interoperability and with providing a structure for the information contained in DMPs. On the other hand, compliance involves a set of *extra actions* to solve issues that delays adoption and uptake, such as:

- Extra modifications are needed on some DMP service providers' data model to maximise the standard's integration advantages (e.g., DMPOnline[157]).
- The standard does not necessarily highlight and/or solve the problem of DMPs recording a pool of information, *collectively about all project's datasets*, thus posing obstacles in the evaluation and exploitation of individual datasets.
- DMPs can be harvested in many ways, depending on how they have been published. They can be found *classified with diverse labels*, such as articles, reports etc. This means that repository providers need to specify the resource types of DMPs[158] in their systems, and promote them widely so that researchers are aware of and use them.

## 4. ***Identifying the merged list those datasets that have been produced (as opposed to simply used).***

Lastly, to identify datasets that are relevant for this study, we isolated those that were *produced* in the Horizon 2020 projects concerned, from those that existed previously and were simply reused. Our assumptions are presented in Table 42.

[154] https://www.openaire.eu/task-forces-in-openaire-advance
[155] https://phaidra.univie.ac.at/detail/o:1140797
[156] https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard
[157] https://openworking.wordpress.com/2021/02/22/towards-better-efficiency-integrating-data-management-plans-with-institutional-systems/
[158] http://vocabularies.coar-repositories.org/documentation/resource_types/

**Table 42. Identifying datasets produced by Horizon 2020 projects**

| ASSUMPTIONS |
|---|
| **A1.** All datasets **reported in SyGMa** are produced by the project. |
| **A2.** All datasets that **reference the project in their metadata** (as harvested from OpenAIRE) are produced by the project. |
| **A3.** Datasets linked to projects via OpenAIRE's **inference system** (text-mined) are not necessarily produced by the project. |

The first two assumptions are straightforward, in the sense that in both cases there is no incentive to report a dataset-project link unless the former was a project output. With respect to assumption A3, because OpenAIRE aims to ***link*** projects to research outputs, the inference system is currently *agnostic* towards the semantic relationship between a project and the linked dataset.

Thus, **for those Horizon 2020 datasets not found in EC-Shared or in the harvested OpenAIRE data,** it is not possible to verify that they were created by the projects. For the purposes of this study, we have therefore discarded **1,579 ORG** datasets that are linked to Horizon 2020 only via text mining.

As with the discarded set of unmatched European Commission datasets, we present here the statistical characteristics of the discarded ORG datasets. We do not observe bias in any direction, as the entities with the highest numbers of discarded datasets are also those that produce the most output overall.

**Table 43. Characteristics of discarded ORG datasets**

| TOP COUNTRY (by number of discarded ORG datasets) | | TOP (SUB-)PROGRAMMES (by number of discarded ORG datasets) | |
|---|---|---|---|
| United Kingdom | 859 | EU.1.1.<br>European Research Council (ERC) | 697 |
| Germany | 714 | H2020-EU.1.3.2.<br>Marie Skłodowska-Curie Actions Mobility | 217 |
| France | 516 | H2020-EU.1.3.1.<br>Marie Skłodowska-Curie Actions Initial training | 154 |
| Netherlands | 502 | EU.4.b.<br>Twinning of research institutions | 83 |
| Italy | 446 | H2020-EU.3.1.1.<br>Understanding health, wellbeing and disease | 67 |
| Spain | 434 | H2020-EU.3.1.2.<br>Preventing disease | 45 |

After these steps were carried out, the final **MOAP Horizon 2020** dataset list for this study included:

- ***6,231*** datasets, of which:
- ***2,815*** *were* reported in SyGMa (EC-Shared); and
- ***5,244*** were produced in projects that participated in the ORDP and did not opt out.

**Lesson learned:** it appears to be the case that beneficiaries do not report a large proportion of datasets, even when participating in the Pilot.

### 7.4.2 Collecting, creating and triangulating metadata

In this section, we discuss the work carried out to prepare indicators, including the assessment of their quality.

***Research areas and FOS classification***

Datasets were classified into research areas by inheriting the classification of the publications that were linked to them. Using this process, we were able to classify the 16.2% of datasets to which a publication is linked. For the methodology used to classify publications, we refer the reader to Section 7.3.2.

***Metadata standards***

The validation process followed was the same as that used for publication metadata (see Metadata Standards in Section 7.3.2), the only difference being that we used the OpenAIRE guidelines for data archives.[159]

***Accessibility and interoperability***

To establish the accessibility and interoperability of datasets, we used the same software, created by Athena RC, that was applied to publications (see Section 7.3.2), with the difference that instead of PDFs, the data file formats for which the software searched were as follows:

- xls, xlsx, csv, tsv, tab, json, geojson, xml, ddi, ods, rdf, zip, gzip, rar, tar,
- 7z, tgz, gz, gz3, bz, bz3, xz, sas, smi, por, ascii, dta, sav, dat, txt,
- tif, tiff, tfw, dwg, svg, sas7bdat, spss, sql, mysql, postgresql, sqlite, bigquery,
- shp, dbf, mdb, accdb, mat, pcd, bt, n3, ns3, nc, h4, h5, hdf, hdf4, hdf5, trs, opj, fcs,
- fas, fasta, keys, values.

We did not include data files in the form of images, as this would have led to a lot of false positives (any image on a webpage would have been tagged). The software is

[159] https://guidelines.openaire.eu/en/latest/data/index.html

currently being developed further, and will eventually be able to correctly tag sound and image data files.

***Metadata coverage for indicators***

To conclude the methodology section,

Table 44 presents the final coverage of the metadata fields required to construct the MOAP dataset indicators. We went through several iterations of cleaning the metadata elements, so that the numbers below, as far as possible, refer to *valid* entries.

**Table 44. Metadata gap analysis for dataset indicators**

| METADATA ELEMENT | Coverage | Issues |
|---|---|---|
| **Number of funders**<br>**Number of projects** | 100%<br>100% | |
| **Number of authors** | 100% | |
| **Author identifiers**<br>*Dataset metadata:* [160]<br>*linked publication metadata*[161]*:* | <br>0%<br>14.8% | Although not required by Article 29.3, and thus *not* essential for this study, good coverage of author IDs provides a great deal of assistance in linking research outputs and examining collaboration networks. |
| **Access rights**[162] | 92.4% | |
| **Linked publications** | 16.2% | As we cannot estimate the full set of linked publications, we cannot assess this coverage. |
| **Date of deposition in repository** | 0% | We were unable to find the 'date deposited' metadata in the original metadata records from repositories. |
| **PID**<br>*in record*<br>*in repository metadata (coverage of datasets in repository):* | <br>99.8%<br>86.7% | DOIs have been cleaned, but other PID types may include dirty values. |
| **Valid URL**<br>*in repository metadata (coverage of datasets in repository):* | <br>37.1% | |
| **Licence**<br>*in repository metadata*<br>*(coverage of datasets in repository):* | <br><br>66.9% | |

[160] Share of datasets with at least one ORCID for an author of the dataset provided in the metadata.
[161] Share of datasets with at least one ORCID for an author of the linked publication, provided in the publication's metadata.
[162] Access rights for the dataset being available in at least one of the data sources in which the record is found.

## Getting in touch with the EU

IN PERSON
All over the European Union there are hundreds of Europe Direct information centres.
You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

ON THE PHONE OR BY EMAIL
Europe Direct is a service that answers your questions about the European Union.
You can contact this service:
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696, or
– by email via: https://europa.eu/european-union/contact_en

## Finding information about the EU

ONLINE
Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU PUBLICATIONS
You can download or order free and priced EU publications from:
https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

EU LAW AND RELATED DOCUMENTS
For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

OPEN DATA FROM THE EU
The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

The report examines, monitors and quantifies compliance with the open access requirements of Horizon 2020, for both publications and research data. With a steadily increase over the years and an average success rate of 83% open access to scientific publications, key findings indicate that the European Commission's leadership in the Open Science policy has paid off. The study concludes with specific recommendations to improve the monitoring of compliance with the policy under Horizon Europe – which has a more stringent and comprehensive set of rights and obligations for Open Science. The data management plan and the datasets of the study are also available on data.europa.eu, the official portal for European data.

*Research and Innovation policy*

Publications Office
of the European Union