European Commission

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Written by PwC EU Services

March – 2018

Research and Innovation

Cost-Benefit analysis for FAIR research data - Cost of not having FAIR research data

European Commission
Directorate-General for Research and Innovation
Directorate A — Policy Development and Coordination
Unit A.2 — Open Data Policy and Science Cloud
Contact Athanasios Karalopoulos
E-mail Athanasios.Karalopoulos@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu
European Commission
B-1049 Brussels

Manuscript completed in March 2018.

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein.

More information on the European Union is available on the internet (http://europa.eu).

Luxembourg: Publications Office of the European Union, 2018

PDF ISBN978-92-79-98886-8 doi: 10.2777/02999 KI-02-19-023-EN-N

EUROPEAN COMMISSION

Cost of not having FAIR research data

Cost-Benefit analysis for FAIR research data

Written by PwC EU Services

Directorate-General for Research and Innovation
Open Science

Contents

List of figures

List of tables

EXECUTIVE SUMMARY

Technological advancements have made research and science more data intensive and interconnected, with researchers producing and sharing increasing volumes of data. In their effort to produce high quality data, researchers have to follow good data management and data stewardship practices such as the FAIR principles.

However, there has been no thorough analysis to determine the value of not having FAIR research data, within and across scientific disciplines, both in economic and non-economic terms, and to contrast it against the current situation where a majority of research data is not adhering to the FAIR principles.This report aims to fill these gaps by estimaing the cost of not having FAIR research data for the EU data market and EU data economy.

Our analysis relied on available studies that have focused on the quantitative value of research data that is findable, accessible, interoperable and reusable.

By looking at the impact of FAIR on research activities, collaboration and innovation, indicators were identified, defined and then quantified. Seven indicators were defined to estimate the cost of not having FAIR research data: Time spent, cost of storage, licence costs, research retraction, double funding, interdisciplinarity and potential economic growth.

To estimate the first five indicators, we first assessed the inefficiencies arising in research activities due to the absence of FAIR data. From the different levels of inefficiency, we computed the time wasted due to no having FAIR and the associated costs. Secondly, we estimated the cost of extra licences that researchers have to pay to access data that would otherwise be open with the FAIR principles. Thirdly, we looked at the additional storage costs linked to the absence of FAIR data. Unaccessible data leads to the creation of additional copies of the data (e.g. by journals or partner universities) which would otherwise not be required if the FAIR principles were in place. With insufficient data to estimate the last two indicators, we provided mostly qualitative considerations and findings instead.

Following this approach, we found that the annual cost of not having FAIR research data costs the European economy at least €10.2bn every year. In addition, we also listed a number of consequences from not having FAIR which could not be reliably estimated, such as an impact on research quality, economic turnover, or machine readability of research data. By drawing a rough parallel with the European open data economy, we concluded that these unquantified elements could account for another €16bn annually on top of what we estimated. These results relied on a combination of desk research, interviews with the subject matter experts and our most conservative assumptions.

Moreover, while building on top of other available studies and being heavily reliant on existing material, we have come to realise ourselves how important is to have FAIR research data. Not only the time invested in this study could have been reduced by a significant amount, but the content could have been enhanced if more material had been accessible and reusable.

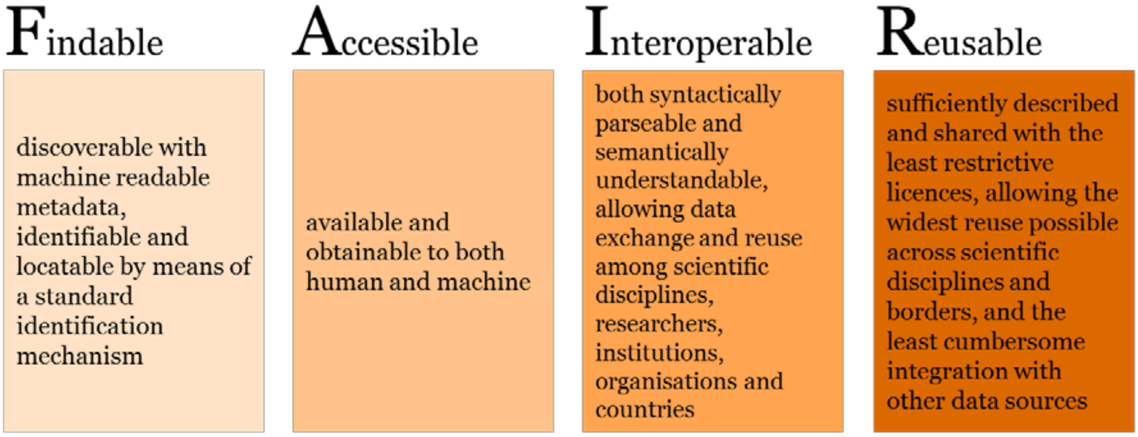
Finally, by estimating the qualitative and quantitative costs of not having FAIR data, this report will enable decision makers to make evidence based decisions about efficient ways to support the real-life implementation of the FAIR data principles. Researchers and research institutions will now be able to weight the cost of not having FAIR versus the cost of implementing the FAIR principles.

1. INTRODUCTION

1.1. Context

Technological advancements have made research and science more data intensive and interconnected, with researchers producing and sharing increasing volumes of data. In their effort to produce high quality data, researchers have to follow good data management and data stewardship practices. Beyond proper collection, annotation and archival, good data management and stewardship include the notion of long-term care of valuable digital assets, either alone or in combination with newly generated data. Good data management and stewardship is not a goal in itself but rather is a key conduit leading to easier and simpler data and knowledge discovery and evaluation, and to subsequent data and knowledge integration and reuse in downstream studies.

To maximise the value of science, research data should have four foundational characteristics[1]; they should be:



Findable

discoverable with machine readable metadata, identifiable and locatable by means of a standard identification mechanism

Accessible

available and obtainable to both human and machine

Interoperable

both syntactically parseable and semantically understandable, allowing data exchange and reuse among scientific disciplines, researchers, institutions, organisations and countries

Reusable

sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible across scientific disciplines and borders, and the least cumbersome integration with other data sources

Figure 1: Four foundational characteristics of FAIR

Findability, Accessibility, Interoperability and Reusability – the FAIR principles – intend to define a minimal set of related but independent and separable guiding principles and practices, which enable both machines and humans to find, access, interoperate and re-use research data and metadata.

1.1.1. The value of FAIR

The implementation of the FAIR principles can bring direct and indirect benefits to research stakeholders, from funders to researchers, and can have a positive impact on the quality and the ROI of research itself. Studies supporting the implementation of the FAIR principles talk about their positive impact on:

- Reducing duplication in research, in terms of time, effort and funding;
- Rigorous management and stewardship of digital resource helping researchers adhere to the expectations and requirements of their funding agencies[2];
- Scaling up research findings based on integrated and analysed existing data from multiple disciplines and regions[3];
- Enabling research to focus more on adding value activities such as interpreting the data rather than on searching, collecting or re-creating existing data[4]; and

[1] FAIR Principles described by GO-FAIR, https://www.go-fair.org/fair-principles/; FAIR Principles described by Force 11, https://www.force11.org/group/fairgroup/fairprinciples; and (Wilkison, Dumontier, & Mons, 2016).

[2] (Wilkison, Dumontier, & Mons, 2016)

[3] (Bonino da Silva Santos, et al., 2016)

[4] http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

- Enhance the science infrastructure to support knowledge discovery and innovation.

1.1.2. *On-going FAIR initiatives*

Multiple initiatives are currently working towards the implementation of the FAIR principles. We are listing the main ones below:

- The GO FAIR movement[5] which federates existing national initiatives that committed to the implementation of the FAIR principles. The GO FAIR movement makes collective choices for the implementation in terms of standards, good practices and protocols and proposes a technical infrastructure, change management and training for data stewards.
- Having recognised the importance of a FAIR-enabled data ecosystem for the implementation of the European Open Science Cloud[6], the European Commission - together with stakeholders and assisted by the FAIR Data Expert Group[7] - is working towards an action plan which aims to make research data FAIR and allow researchers to be able to do cross-disciplinary scientific data sharing and reuse. This action plan (aka the FAIR data action plan) covers all types of digital research objects and it is a collaborative instrument guiding the integration of the FAIR principles at European level, across borders and/or disciplines. In addition, the FAIR data action plan creates the conditions for the development of coherent national and/or discipline specific plans for the implementation of the FAIR principles at national and/or discipline level.
- The study on the possible implementation of the FAIR principles in Denmark, which was commissioned by the Danish Agency for Science and Higher Education. This study offers a cost-benefit analysis for a future implementation of the FAIR principles for public-funded research in Denmark.
- Numerous universities and research institutes, which have openly embraced the FAIR movement and are promoting the FAIR principles. Examples include the Dutch Techcentre for Life Science[8], Leiden University[9], Utrecht University[10], and Maastricht University[11].
- ELIXIR which is a distributed infrastructure for life science information. ELIXIR[12] is committed to enable the availability of FAIR research data within the framework of EOSC[13]. EXILIR Nodes together with EMBL-EBI[14], coordinate and integrate Bioinformatics resources across Member States (i.e. by providing databases, analysis tools, interoperability services, etc.) with the end goal of making information freely available to the science community.

1.1.3. *Challenges for the implementation of the FAIR principles*

The fact that the FAIR principles are not common practice yet is due to numerous reasons. Some concern the lack of awareness in the research community[15] about how to share data, in which format, what information or metadata should be provided etc. Others touch upon existing cultures and behaviours in conducting research, from research funders forbidding researchers to share their data to researchers not even considering that the data they produce can be valuable for others, the lack of attention

[5] https://www.go-fair.org/

[6] European Cloud Initiative - Building a competitive data and knowledge economy in Europe (COM(2016) 178 final)

[7] The Commission has established the Expert Group on FAIR data to support the Research and Innovation policy development on Open Science.

[8] https://www.dtls.nl/fair-data/fair-data/

[9] https://www.universiteitleiden.nl/en/research-dossiers/data-science/leiden-silicon-valley-of-fair-data

[10] https://www.uu.nl/en/research/research-data-management/guides/costs-of-data-management

[11] https://www.maastrichtuniversity.nl/research/data-science-um/research

[12] ELIXIR Is a distributed infrastructure comprising 180 leading universities and centres of excellence

[13] https://www.elixir-europe.org/system/files/elixir_statement_on_fair_data_management.pdf

[14] European Molecular Biology Laboratory (EMBL) – European Bioinformatics Institute (EBI)

[15] Interview with Barend Mons, 2018-01-17

given to the preparation of a data management plan, missing metadata[16], various competing standards for research data and metadata, and the lack of persistent identifiers for data, datasets and metadata[17].

Moreover, many researchers and organisations are still reluctant to apply the FAIR principles and share their datasets because of real or perceived costs, including time investment and money.

In order to drive the implementation of the FAIR principles in Europe, the European Commission together with a number of pioneering European research stakeholders is taking activities and measures that aim at raising awareness about costs and benefits of FAIR data, and is encouraging funding bodies to set guidelines or support the development of an infrastructure for publishing FAIR data.

This current study is positioned exactly in this context. Although the discussion is very often around the costs and benefits of the FAIR principles' implementation both in economic and non-economic terms (and even there a global view is still to be developed), few think about the current costs and opportunity losses from not having the FAIR principles implemented (or having very small partial implementations in specific disciplines and countries). Assessing the cost of not having FAIR data will provide quantified facts to back-up the utility of the FAIR principles and will help convincing research stakeholders to invest in their implementation.

This study presents and applies a quantitative methodology for estimating the cost of not having FAIR research data in Europe as well as cost estimations. The methodology is well documented, easily applicable in different countries and disciplines and hence easily repeatable in the future.

The scope of this study comprises direct and indirect costs resulting from not implementing with the FAIR principles in Europe. This includes research carried out by public, private and non-governmental organisations. For this reason, the report did not look into the cost to implement FAIR research data in Europe nor the potential loss of revenue incurred by the researchers as a result of making data free and open.

1.2. Intended audience

This report is intended primarily for:

(1) Research funders: to raise awareness on the cost of not having FAIR research data, e.g. funding redundant work because earlier research was not FAIR. With this report, funders can better appraise the importance of FAIR and include where relevant compliance to the FAIR principles as a condition for researchers to obtain funding.

(2) Research (data) infrastructures: to raise awareness about the impact of not having FAIR research data on e.g. the cost of storage, and other activities of research infrastructures.

(3) Research performing organisations: to raise awareness about the cost of not having FAIR research data for researchers, e.g. in terms of research quality, additional time spent on research, and citations. By identifying the cost of not having FAIR research data, this report helps researchers to justify investing into compliance with the FAIR principles.

[16] (Zahedi, Haustein, & Bowman, 2014), and (Parsons, Grimshaw, & Williamson, 2013)

[17] (Johnson, Parsons, Chiarelli, & Kaye, JISC Research Data Assessment Support - Findings of the 2016 data assessment framework (DAF) surveys, 2016), Stehouwer & Wittenburg, 2014 and Tenopir C. , et al., 2011

1.3. FAIR versus open research data

This study adopts the H2020 Programme[18] definition of research data, i.e. *information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion, or calculation*. The types of data covered by this definition are first, the underlying data (the data needed to validate the results presented in scientific publications), including the associated metadata (i.e. metadata describing the research data deposited) and second any other data (for instance curated data not directly attributable to a publication, or raw data) including also the associated metadata.

FAIR data and open data are two distinct concepts which are however coming closer and closer. (Mons, et al., 2017) distinguish the concept of FAIR from the concept of open, saying: "In the envisioned Internet of FAIR Data and Services, the degree to which any piece of data is available, or even advertised as being available (via its metadata) is entirely at the discretion of the data owner." The reasoning behind is that "FAIR" only speaks of the need to:

- Describe a machine-readable or human process for accessing discovered data;
- Openly and richly describe the context within which those data were generated, to enable evaluation of its utility;
- Explicitly define the conditions under which they may be reused; and
- Provide clear instructions on how they should be cited when reused.[19]

However, research, data is not truly reusable unless it is open, i.e. available under an open licence and at marginal costs (in most cases at zero cost), and openness comes often hand in hand with the implementation of the FAIR principles. As part of the Open Science Agenda, the European Commission defined the ambition to make FAIR <u>and</u> open access the default for data sharing in scientific research. Moreover, the Council of the European Union concluded[20] that "as open as possible, as closed as necessary" is the underlying principle for optimal reuse of research data.

In this study, we therefore consider the cost of not having FAIR <u>and free open access</u> research data whenever discussing the FAIR principles.

1.4. Structure of the report

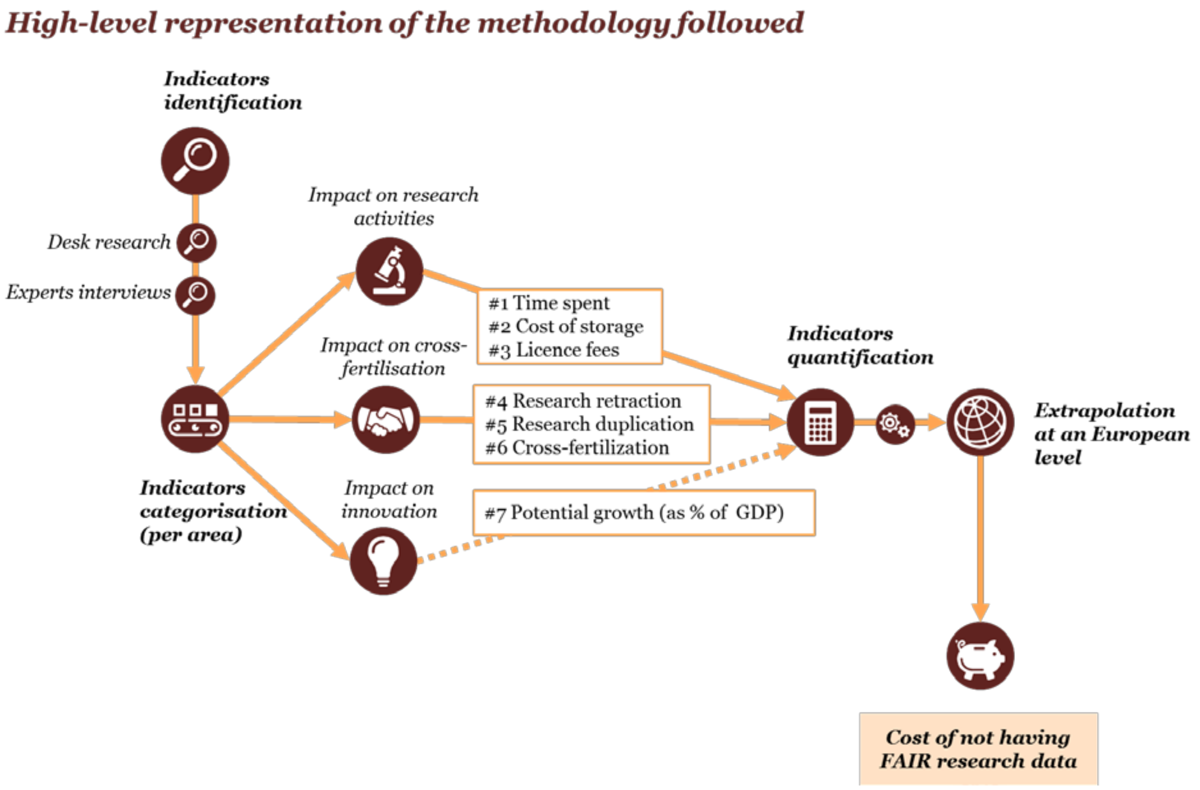
The remainder of this study is structured as follows:

- Chapter 2 "Approach" lays out our approach for identifying indicators to measure the cost of not having research data, describes each indicator and the methodology for quantifying them.
- Chapter 3 "Cost calculation" presents the results of the quantification exercise for the cost of not having FAIR data.
- Chapter 4 "Conclusion" discusses the main findings and observations of the study.

[18] (Directorate-General for Research & Innovation (European Commission), 2017)
[19] (Data Citation Synthesis Group, 2014)
[20] (Council of the European Union, 2016)

2. APPROACH

2.1. Overview



High-level representation of the methodology followed
Indicators identification
Desk research
Experts interviews
Impact on research activities
#1 Time spent
#2 Cost of storage
#3 Licence fees
Impact on cross-fertilisation
#4 Research retraction
#5 Research duplication
#6 Cross-fertilization
Impact on innovation
#7 Potential growth (as % of GDP)
Indicators categorisation (per area)
Indicators quantification
Extrapolation at an European level
Cost of not having FAIR research data

Figure 2: High-level representation of the methodology followed

In this study, we started by identifying and selecting economic indicators which drive the cost of not having FAIR research data. These indicators - linked to the causes of the cost occurring due to not having FAIR research data - were categorised and quantified to the extent possible. Furthermore, this approach is based solely on secondary data[21].

In order to quantify the indicators, assumptions and assertions were made to allow us to estimate the cost of not having FAIR research data. The method for estimating the indicators and the necessary assumptions are described.

Different criteria were considered for identifying and quantifying the indicators:

- **Preciseness**: the indicator is related to specific activities, which are defined in clear terms;
- **Reliability**: the indicator is based on facts and consistently measurable over time, which ensure the sustainability of this estimation;
- **Measurability**: the indicator can be quantified using existing tools and methods; and
- **No overlaps**: each indicator can be measured independently from the other indicators.

2.2. Indicators identification

In order to measure the cost of not complying with the FAIR principles, we first needed to assess their impact on research data. Therefore, we detailed in Annex I the various

[21] Secondary data refers to information that already exist, in our case information that we collected rather than created.

facets of each principle, and reviewed the metrics proposed by recent related works[22] for estimating the compliance of research data to the FAIR principles. These facets and metrics are used for identifying the economic indicators linked to non-FAIR research data. Economic indicators not related to the FAIR principles were discarded.

We projected the impact of the FAIR principles on research data to researchers, repositories and publishers, and innovation. The identification of potential economic indicators for the costs of not having FAIR research data also benefited from the desk research and interviewing experts.

The following three areas regroup the different economic indicators identified as part of the cost of not having FAIR research data. These areas are applicable to all sectors (i.e. academic, private, public, non-profit).

- ***Impact on research activities:*** This category of indicators relates to activities taking place during research. It concerns, but is not limited to, research development, planning of the research project, creation of the data, collection of the data, pre-processing and cleansing of the data, integration and transformation of the data, etc. In these activities, time wasted due to research data not being FAIR is identified and quantified. This category also includes the duplication of storage for research data that is not FAIR (i.e. research data being stored unnecessary on different repositories). This duplication is quantified in terms of storage costs. In addition, we also include in this category the costs faced by researchers for finding, accessing and analysing research data that is not open. For instance, time spent to obtain the data (e.g. registration process), associated fees and licensing costs, etc.
- ***Impact on collaboration:*** here is an opportunity cost for the research community but also individuals and organisations using research data. This cost manifests itself in the form of missed opportunities for collaboration or cross-fertilization between research groups within and across disciplines. It is important to note that the estimation of this cost is very sensitive to the assumptions taken.
- ***Impact on innovation:*** This category relates to the impact of research data not being FAIR on innovation and consequently on the European economy. Ample empirical evidence demonstrates that research is a “key driver of productivity and economic growth”[23]. Likewise, non-FAIR research data can have an impact on the number of patents filed, missed opportunities in terms of new business or products, lower job creation, etc.

In the remainder of this study, we differentiate between academic and non-academic researchers, the latter encompassing the private, public and non-profit sectors. For academic staff, the term ‘researcher’ includes all people for whom research is one of the core activities such as professors, PhD students, post-doctorates, assistant professors etc. For non-academic staff, the term ‘researcher’ includes all people who are not academics but qualified as ‘research intensive’ (i.e. whose core activity is research).

2.2.1. *Impact on research activities*

When analysing existing studies on the cost related to research activities, we identified a set of activities to which the cost indicators can be linked. These activities have been compiled from existing literature about data life cycle[24] and research activities[25].

Most of the costs faced by research stakeholders can be linked directly to one or multiple of the following activities (see also Figure 2):

[22] (Dunning, de Smaele, & Böhmer, 2017)
https://via.hypothes.is/https://www.force11.org/group/fairgroup/fairprinciples
http://datafairport.org/fair-principles-living-document-menu
(Wilkison, Dumontier, & Mons, 2016)
https://www.force11.org/group/fairgroup/fairprinciples
https://via.hypothes.is/https:/content.iospress.com/articles/information-services-and-use/isu824#ref019

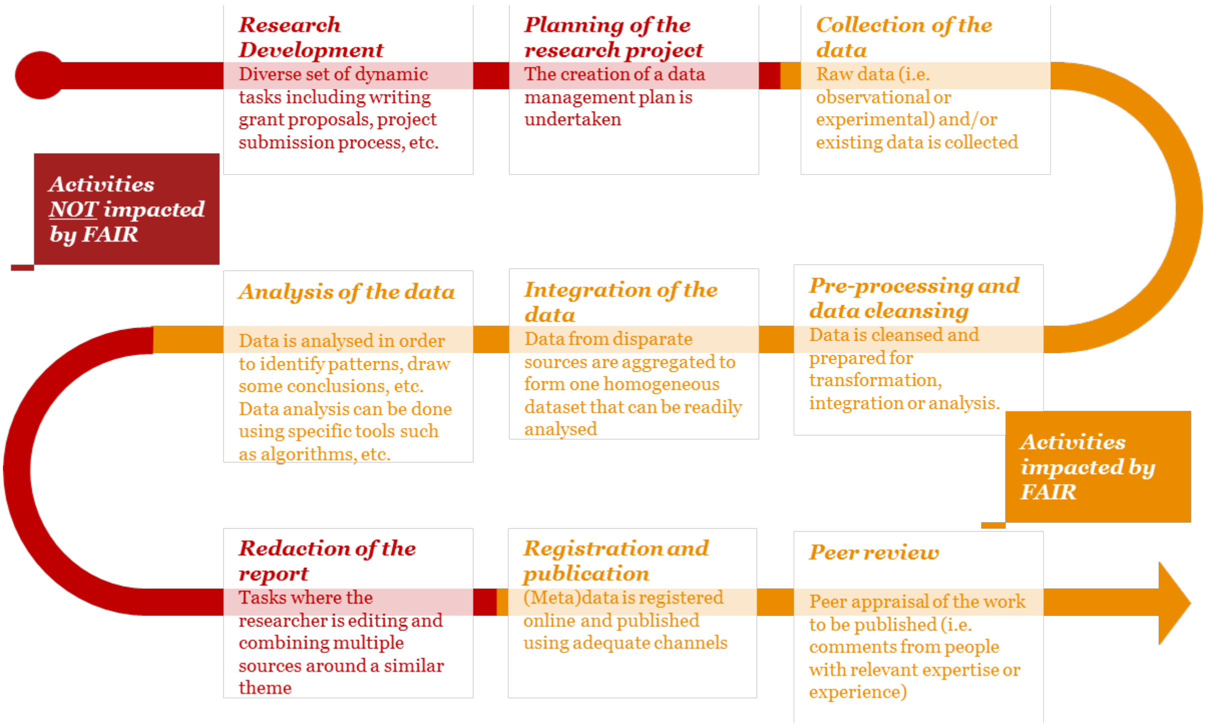
[23] (Directorate-General for Research and Innovation (European Commission), 2017)

[24] (Data Life Cycle | DataONE, n.d.)

[25] (Ziker, 2013)

- Research development
- Planning of a research project
- Creation and collection of the data
- Pre-processing and data cleansing
- Integration of the data
- Analysis of the data
- Redaction of the report
- Registration and publication
- Peer review

The FAIR principles do not have direct impact on all those activities, which is why *research development*, *planning of the research project* and *redaction of the report* are left out in our analysis.



Research Development
Diverse set of dynamic tasks including writing grant proposals, project submission process, etc.

Planning of the research project
The creation of a data management plan is undertaken

Collection of the data
Raw data (i.e. observational or experimental) and/or existing data is collected

Pre-processing and data cleansing
Data is cleansed and prepared for transformation, integration or analysis.

Integration of the data
Data from disparate sources are aggregated to form one homogeneous dataset that can be readily analysed

Analysis of the data
Data is analysed in order to identify patterns, draw some conclusions, etc. Data analysis can be done using specific tools such as algorithms, etc.

Redaction of the report
Tasks where the researcher is editing and combining multiple sources around a similar theme

Registration and publication
(Meta)data is registered online and published using adequate channels

Peer review
Peer appraisal of the work to be published (i.e. comments from people with relevant expertise or experience)

Figure 3: Research activities

2.2.1.1. *Creation and collection of the data*

Creation and collection of the data is the step in the research life cycle where data is created (i.e. through observations, experiments or simulations), and potentially useful and existing data is found and obtained. This step also includes the performance of veracity checks on the data.

In this activity of the research life cycle, not having FAIR has an impact on the time, storage costs and licence costs required to find, access, manipulate and reuse (meta)data underpinning research papers and studies which has not been made (properly) available, for different reasons:

- Time lost due to unfindable, not human understandable or with unstructured or incomplete metadata. Many researchers do not publish the minimum set of metadata for their studies, required for other researchers to find, access and reuse it. Metadata of poor quality hampers findability. Furthermore, as described in Annex I, the metadata should be rich enough for external readers to directly understand the necessary information about the study and its supportive data.

- The cost of time from the human expertise needed to read and understand the data. Having the data well documented can result in a much lower overhead in terms of capturing the essential information inside the dataset and work with it (i.e. sufficient understanding of the field research).
- The lack of a single point of access[26] is increasing the time required by researchers to authenticate to different publication channels, to search into each channel etc. The FAIR movement is advocating a single point of access, where you could query for metadata which would instantly point to the layer where the data are located, thus reducing considerably time to access (meta)data. In the event of unavailable existing data underpinning published findings, time is required to re-create a dataset.
- Data related to a specific research or a whole research discipline is often stored in isolation, in information silos. It often occurs that data is partially available from different points of access[27], causing the researcher to spend more time to retrieve all the data he/she needs. This researcher may also need to transfer data from one repository to another which would increase overall storage costs.
- Data is duplicated across one or more repository, which causes researchers to spend additional time selecting the right datasets.
- Making the research data compliant to the FAIR principles will have a positive impact on the open access of scientific publications and will consequently reduce the access fees (i.e. licence costs) faced by researchers which will enhance the citation ratio (i.e. reuse) of their publications[28].
- Finally, unstructured metadata makes it hardly machine-readable which is one of the many perspectives advocated by the FAIR principles. Non-machine-readable and non-interoperable metadata, which cannot be indexed by search engines hampers greatly the findability of the associated data[29].

  Machine analysing data for pattern instead of human capital could reduce the time spent by researchers and all the costs associated (e.g. licence costs to access specific software, external expertise, data transfer due to remote access impossible) to it down to almost zero. The current state of play forces researchers to read and understand the metadata manually, and hampers automated data search and discovery.

Moreover, difficulties in assessing data quality and integrity because of un-adequate (meta)data makes it impossible for researchers to share and analyse research data in a trusted environment across technologies, disciplines and borders.

2.2.1.2. Pre-processing and data cleansing

Pre-processing and describing the data is the step in the research life cycle where inspection and checks are performed to improve and ensure data quality. Data modelling and data cleansing also take place in this step.

Not having FAIR has an impact on the time necessary to create, curate, cleanse, reprocess and keep the data and metadata up to date. Maintenance of low quality metadata incurs significant costs and can be a source of errors. This cost increases with time for various potential reasons: human factors, standards changed etc.

For instance, time is needed to run quality checks (e.g. run scripts/queries) on both existing and raw data collected. Similarly, time is required to improve the quality of (meta)data and to transform data when necessary (e.g. to identify and fix problems).

[26] E.g. EOSC ambitions to be “a one-stop-shop to find, access, and use research data and services from multiple disciplines and platforms.”

[27] A single point of access should not be confused with a single repository. The first would only harvest and store the metadata while the second refers to all types of data.

[28] (Van Noorden, 2013)

[29] (Stehouwer & Wittenburg, 2014)

FAIR research data should be machine readable, thus reducing the time required for checking quality. The underlying reasons are:

- Curation methodologies are often invented in a moment of need and consequently cannot be replicated for other metadata, which results in inefficiencies in the future; and
- Lack of documentation for the wide heterogeneity of pre-processing operations being carried out, depending on the data that has been generated.

2.2.1.3. Integration of the data

Integration of the data is the activity in the research life cycle where data from disparate sources is aggregated to form one homogeneous dataset that can be analysed. As research data volume is surging[30], data integration becomes necessary to be able to aggregate data from multiple sources (and from different formats).

Not having FAIR has an impact on the time required to correlate/transform two or more datasets in a machine-readable and harmonised format (from proprietary formats to standardized ones), or time required to understand that two different datasets cannot be integrated due to non-interoperable vocabularies (i.e. metadata) used to define datasets. As a rule of thumb, in a data analysis project, data cleansing of poor quality data can take up to 80% of the total effort.

It should be noted that bridging the gap between semantic classifications and annotations is a challenge due to the dynamic nature of research, the dynamics of classifications and unhandy tools[31].

In the private sector, integration costs are particularly high in the case of joint research ventures, as time to integrate heterogeneous data is considerable[32]. This may sometimes lead to the non-utilization of datasets.

Moreover, experts in some communities developed their own guidelines on metadata. While guidelines provide a high degree of flexibility, and thus representational capacity, it first requires experts to make use of them[33] and secondly, the guidelines are often not aligned with recognized standards, hindering human and machine readability.

2.2.1.4. Analysis of the data

Analysis of the data is the activity in the research life cycle where the data is processed with the ultimate purpose to extract useful information, to elicit insights and eventually to formulate observations and conclusions. Analysis of the data or data modelling is often performed using specific software's, tools or methodologies.

In this step of the research life cycle, not having FAIR has an impact on the time required to analyse data that is not properly documented or is not complete and at the right level of quality. According to a recent paper, "there are growing concerns about replicability and reproducibility of research results: a recent survey indicated that more than 70% of researchers have tried and failed to reproduce another scientist's results, and more than half agree that there is a significant crisis of reproducibility" (Baker, 2016). It also pertains to the time required to verify the findings of other studies in order to decide whether the methodology and the results are correct and could be the object of a publication.

2.2.1.5. Registration and publication

Registration and publication is the step in the research life cycle where data is accurately and adequately described using the relevant metadata. The (meta)data is then registered

[30] Interview with Barend Mons, 2018-01-17
[31] Peter Wittenburg
[32] Interview with Barend Mons, 2018-01-17
[33] Second year report on RDA Europe analysis programme

(i.e. stored) online e.g. on web servers and published using the adequate channels. Not having FAIR has an impact on both time and storage costs, namely:

- Time to register (meta)data and to maintain a research repository with (meta)data not FAIR, or to keep a specific database up-to-date with a certain level of interoperability.
- Storage costs related to the volume of data duplicated in different repositories, independently from the preservation mechanism used. For example, data inaccessible or wrongly described could lead to duplications within and between repositories.

The publication and indexation processes and requirements vary between repositories, increasing the time spent on those activities.

Without accessible data and publications, journals, universities and funders often each hold a separate copy of the underlying research data on a repository. This unnecessary redundancy drives the overall costs of storage up.

2.2.1.6. Peer review

Peer review is the step in the research life cycle where research work is evaluated by other researchers with relevant expertise or experience before the work can be endorsed, published or presented.

In this step of the research life cycle, not having FAIR has an impact on the time spent because of redundant reviews. Reviews of research data and publications are rarely shared, and even when they are, the lack of industry-wide standard means that each entity solicits its own reviews before making a decision[34].

Moreover, each peer review can be time consuming due to the lack of data supporting the conclusions of the article reviewed.

According to the trade association for academic and professional publishers[35], peer review accounts for 15 million hours of time a year. The process takes on average up to two or three persons and the average acceptance rate is about 50%. Peer review is often seen as a quality insurance tool for improving the quality of research studies. However, the process is often redundant across the scientific community and therefore sometimes unnecessary.

2.2.2. Impact on collaboration

Collaboration applied to science is all about combining research data and findings from different organisations within and across disciplines to produce new and better outputs. Hence, it is much impacted due to research data not being FAIR. A survey[36] conducted at the University of Sheffield showcased that 31% of the academic staff agreed with the following statement: "Lack of access to data generated by other researchers or institutions has restricted my ability to answer research questions". Another study[37] refers to 50.1% of the respondents having restricted ability to answer scientific questions due to a lack of access to data generated by other researchers.

Furthermore, a recent study[38] put forth that 73% of academics surveyed said that having access to published research data would benefit their own research. In correlation with that figure, survey respondents indicated that sharing research data is important for doing research in their field. Having FAIR research data could thus enable the reuse of research data and hence collaboration. A better collaboration could indirectly and positively impact cross-fertilization[39] and thus innovation.

[34] (American Journal Experts, n.d.)
[35] (Ware & Mabe, 2012)
[36] (Cox & Williamson, 2014)
[37] (Tenopir, et al., 2011)
[38] https://www.elsevier.com/about/open-science/research-data/open-data-report
[39] Cross-fertilization refers to the mixing of data from different disciplines to produce a better result.

The probability of finding reports likely depends on their availability in indices such as Google Scholar and Scopus or in particular research journals and portals. By increasing the usage of publicly available dissemination channels, FAIR increases the probability for reports to be found and reused.

In addition, a lot of papers remain uncited[40] because of poor collaboration. Data reuse can also have a huge impact on citation and thus collaboration and cross-fertilization as shown by a recent study[41]. It has been proven that studies that made data available in a public repository received more citations than similar studies for which the data was not made available.

Implementing the FAIR principles could partially increase the collaboration between scientific communities. For instance, the survey conducted at the University of Sheffield found that 64% of the respondents would use the data of others, if it was easily accessible. Between 50 and 60% of the researchers surveyed would be willing to share data with others via a central data repository without restrictions. However, quantifying the direct impact of FAIR on collaboration is challenging as one would have to assess the value of the research that would have been done with FAIR against that of the research that has been done without.

Because of this, we look at the impact of not having FAIR on collaboration by using the following proxies:

- Research retraction: research which is retracted as a consequence of not having FAIR cannot contribute to the progress of science. This includes research that is retracted because of errors, non-reproducibility, fraud, plagiarism, etc[42]. We postulate that the FAIR principles would decrease fraud and increase research quality which would be observed in a decrease of the number of articles retracted.
- Research (funding) duplication: redundant research does not contribute to science.
- Cross-fertilization: research made possible thanks to the FAIR principles which would not have been possible otherwise (e.g. because the data on which it is based would not have been reusable).

2.2.3. Impact on innovation

The impact on innovation represents an opportunity cost of not having FAIR research data. In a survey published in ‘Data Sharing by Scientists: Practices and Perceptions’[43], 64% of respondents answered that the lack of access to data generated by other researchers or institutions is a major impediment to progress in science. Another study[44] supports this idea with a figure up to 43%, with the sole difference that they mentioned a *major* impediment to progress in their *subject discipline*. From these observations, we conclude that having FAIR research data would have a positive impact on innovation as the opposite is claimed to slow down progress. Beyond research and science, it can be reasoned that businesses are facing pitfalls too as data availability, use and combination are crucial for developing new and differentiating existing services and products[45].

Not implementing the FAIR principles impacts innovation in different ways, here is a non-exhaustive list:

- Lack of access to high value data prevents the development of innovative services and the creation of new business models.

[40] (Davis, 2012)
[41] (Piwowar & Vision, 2013)
[42] (Fang, Steen, & Casadevall, 2012) and (Wager & Williams, 2011)
[43] (Tenopir, et al., 2011)
[44] (Cox & Williamson, 2014)
[45] (Wittenburg, Costs of FAIR Compliance and not being FAIR compliant, 2017)

- Non-FAIR research data will also prevent the use of Machine Science[46] and what it entails. Data not being machine readable, it will not be possible for a machine to process the growing quantity of data and extract insights at a level the human is not capable of doing.
- Innovation and progress are made possible by building upon the results of previous work. If research data is not available, new research will always add to the same baseline, thereby hampering innovation.
- Inability to easily identify possible research collaborators, partners and experts due to research data not being FAIR.
- Loss of data stemming from poor data management or the absence of a clear retention policy.
- Lack of clarity about licences and data use conditions prevents the use of non-FAIR data in value-added projects, due to litigation and licence violation risks.

Overall, the impact of FAIR on innovation translates itself in terms of unrealised economic growth and job creation.

2.3. Selected indicators

Based on the above three areas and our analysis, the indicators we identified are as follows:

Areas	Indicators
Impact on research activities	1. Time spent 2. Cost of storage 3. Licence costs
Impact on opportunities for further research	4. Research retraction 5. Double funding 6. Cross-fertilization
Impact on innovation	7. Potential economic growth (as % of GDP)

Table 1: Selected indicators per area

Indicator #1: Time spent

In order to cost the time wasted during research activities, we estimate the total time spent by research staff to perform the research-related activities listed above. From that total time we derived the time wasted due to non-FAIR data and transformed it into a financial value using the average salaries for researchers per country, factoring in differences for academic and non-academic research.

Indicator #1 measures a direct cost faced today by researchers and funders.

Indicator #2: Cost of storage

The cost of storage relates to the electronic storage costs for additional redundant copies of the data that would otherwise not be needed if the data was FAIR. The total cost of storage is estimated based on the number of copies, the size of datasets and the storage duration.

Indicator #2 measures a direct cost faced by research organisations today.

[46] Machine Science is defined by the increasing use of computational resources in research-related activities. In our case, Machine Science specifically refers to machine-readability and reusability of data.

Indicator #3: Licence costs

This indicator looks at the licence costs to use data which could have been made available as open and FAIR data.

Indicator #3 measures a direct cost paid today by research organisations and funders.

Indicator #4: Research retraction

Research retraction measures the cost of research which would not have been retracted if the FAIR principles had been respected.

Indicator #4 measures an indirect cost faced by researchers today.

Indicator #5: Double funding

While measuring what could have been, i.e. the benefits of more cross-fertilisation, is not directly possible. We postulate that non-FAIR research leads to duplication of research projects and the costs linked to redundant research mirrors at least in part the value of the research that could have been funded instead. This indicator does not include the duplication of some research activities between different research projects. These are included in indicator #1: time spent.

Indicator #5 measures a direct cost faced by funders today.

Indicator #6: Interdisciplinarity

Interdisciplinarity refers to the added value of new research combining several academic disciplines which would be possible thanks to the FAIR principles, compared to the value of research that would be done otherwise without FAIR.

Indicator #6 estimates an opportunity cost due to the unrealised benefits of FAIR.

Indicator #7: Potential economic growth

Potential economic growth refers to the GDP growth and the number of jobs that would be created if the FAIR principles were applied widely. By unlocking the value of research and facilitating the progress of science, FAIR has a positive impact on innovation which translates itself into job creation and higher GDP.

Indicator #7 estimates an opportunity cost due to the unrealised benefits of FAIR.

2.4. Indicators quantification

For quantifying the selected indicators, we started by collecting data at the microeconomic scale, namely data applicable to individual research projects and activities. Once we quantified the costs of not having FAIR data at the micro level, we extrapolated those results to the level of the European research economy.

Because of the extrapolation from micro to macro level, error margins and estimation intervals at the micro level may cause the global estimate to be inaccurate. Therefore, we took the most conservative estimates with regard to the costs of not having FAIR data and the frequency at which these costs occur. Indicators which are not clearly quantifiable have been excluded. Consequently, the actual true cost of not having FAIR data may be significantly higher, due to e.g. unquantifiable spill-over effects or qualitative externalities that remain unaccounted for.

When collecting the data for the microeconomic indicators, we applied the following criteria to maximize the representativeness of the data:

- Discipline specificity: data related to specific research discipline will be preferred over aggregated data for e.g. a research programme. This will allow us to extrapolate information more accurately, as different discipline have different weights.

- Geographical coverage: our datasets cover multiple countries (e.g. Denmark, Finland, Netherlands, UK), mostly European. Knowing the research expenditures and research areas of each country allows us to extrapolate information more accurately. But due to a shortage of quality resources, results from non-European regions (e.g. Australia) were included in our report when the research infrastructure was similar.
- Reusability: data pertaining to cases which are not likely to reoccur have been avoided to the extent possible.

2.4.1. Indicator #1: Time spent

In order to estimate the time[47] wasted on specific research activities, it is essential to have an approximated idea of the time spent by researchers on each research activities.

The cost of time spent was identified by multiplying the time wasted by not having FAIR data with the average wage of academic and non-academic researchers respectively. This calculation will take into account the number of researchers and the average wages for each of the 28 Member States.

$$Cost\ of\ time\ spent\ = Time_{Wasted} \times Wages$$

The time wasted by researchers is identified by factoring the time spent by researchers on research activities[48] and the inefficiencies due to non-FAIR data.

$$Time_{Wasted}\ = Time_{Activities} \times Inefficiency\ rate$$

The share of time spent by researchers on research activities has been studied by (Ziker, 2013) and (Tenopir, et al., 2011) among others. For our calculations, we used an average compiled from seven studies, the two referred above and (Nur Farah Naadia Mohd Fauzi, Abd Rashid, Ahmad Sharkawi, Farihah Hasan, & Aripin, 2016), (Court, 2012), (Cheol Shin, Arimoto, Cummings, & Teichler, 2014), (Houghton & Gruen, 2014). Moreover, we also dissociate between scientific papers (reports) and supporting data in our calculations wherever relevant.

Overall, the time inefficiencies due to the lack of FAIR data have been estimated to 3.12% for academic researchers according to the share of time spent on each of the following research activities and the associated ineffeciences. Knowing which share of time is spent on research by academic and non-academics researchers[49], we extrapolated this figure to 4.47% of time inefficiencies for non-academic researchers.

2.4.1.1. Creation and collection of the data

We estimated that 31.52% of time spent for finding data could be saved if the FAIR principles were applied. This figure is based on aggregated statistics[50] about the use and quality of the metadata (i.e. type of standards, if existing) and assumptions on the time lost depending on the availability and quality of the metadata.

The time for finding secondary data is directly linked to the quality and richness of the available metadata. Consequently, we estimate the time in accordance with four levels of metadata availability:

- No metadata available;
- Metadata not following any standards;

[47] For the remaining of this section, the term ‘Time’ will be used in equations as a percentage of time but for convenience, we will keep the notation ‘Time’.

[48] The research activities are defined in section 2.2.1.

[49] (Tenopir, et al., 2011)

[50] (Tenopir, et al., 2011)
(Royal Veterinary College, University of London, n.d.)
(Parsons, Grimshaw, & Williamson, 2013)

- Metadata following local/proprietary specifications; and
- Metadata following international recognised standards such as DataCite[51], DCAT-AP[52] Dublin Core[53], DDI[54], or SDMX[55].

To the time for finding the right secondary data, we add the time lost in searching for data in multiple access points. Each researcher loses time for accessing relevant data from journals where an authentication is required. This time lost consists off the time required to authenticate to the journal as well as the time for managing an account (e.g. creating, authenticating, changing password). We assume this time to be rather low. However, implementing the FAIR principles would decrease the number of times a researcher will be required to authenticate himself or herself.

2.4.1.2. Pre-processing and data cleansing

While the interoperability and reusability principles of FAIR can facilitate the (pre-) processing of existing data, the time saved in direct correlation to the FAIR principles was mainly due to a reduction in the effort necessary for the initial cleansing of the data. However, the time spent in this step is often hard to dissociate from the time for transformation and integration activities. We therefore estimated the cost of time for this step together with the step integration of the data.

Furthermore, we understood from interviews with researchers and experts that most of the time saved by FAIR pertains to other research activities and the impact on the production of data by researchers is marginal.

2.4.1.3. Integration of the data

The probability of being allowed to reuse an article and the time required to identify this authorisation depends on the licence provided and the easiness with which a researcher can find this licence. From there, we estimated the average time required to identify and understand a licence for secondary data and estimated that FAIR could help reduce this time by 1.46%.

While we found that the time needed to identify the licence is low, it should be noted that every time a dataset or document cannot be reused, the time spent finding it has essentially been wasted.

Furthermore, the probability of being able to integrate multiple sources of secondary data together directly depends on the formats and quality of the data to be integrated and the objectives pursued by the research. For estimating the feasibility of the integration, we look at the percentage of data replicable, if the format of the data is machine readable or not, and if data quality issues can be resolved or not.

- Replicable data:
    - Machine readable; or
    - Non-machine readable, for which we differentiate pdf from other formats.
- Non-replicable data:
    - Resolvable; or
    - Non-resolvable.

For each of the data quality levels above, we estimated the time lost for being able to reuse the data and how FAIR would impact this time. For example for resolvable data

[51] http://schema.datacite.org/
[52] https://joinup.ec.europa.eu/page/dcat-ap
[53] http://dublincore.org/specifications/
[54] http://www.ddialliance.org/Specification/
[55] https://sdmx.org/

with quality issues, we based ourselves on a survey from (Biology, 2015) with the time required to fix bad quality data.

The impact of FAIR on the time and probability of interoperating data is twofold. Firstly, FAIR pushes researchers to use data formats which are “formal, accessible, shared, and broadly applicable for knowledge representation”[56] for human and machine readability. As many reports are only accessible in pdf format, the machine readability is limited. Secondly, the reusability FAIR principle stipulates that data must be “richly described with a plurality of accurate and relevant attributes”[57] enabling a human or a machine to “decide if the data is actually useful in a particular context”[58].

Overall, we estimated that FAIR could help reduce time spent for integrating the data by 28.66%.

2.4.1.4. *Analysis of the data*

The impact of not having FAIR on the time spent analysing the data is difficult to dissociate from the time spent due to interoperability and reusability issues faced when integrating the data.

Due to the lack of specific data regarding the impact of FAIR on this activity, we estimated the additional time needed specifically for analysis to be insignificant.

2.4.1.5. *Registration and publication*

When it comes to registration and publication, we found that the main impact of FAIR is not on the time spent registering and publishing data, but on the number of redundant copies being made due to inaccessible data. These aspects are quantified by indicator #2: Cost of storage.

2.4.1.6. *Peer review*

We assumed that 10% of time spent on peer reviews could be saved if the FAIR principles were applied. FAIR impacts peer review time by increasing the quality of research and results reproducibility. While we consider that the need for peer reviews would remain mostly, FAIR will help reduce the time needed to perform the reviews.

We then factor this assumption with the amount of peer reviews and the amount of time spent on peer reviews from existing literature to determine the cost of not having FAIR.

The FAIR principles would have a second impact on the time spent by researchers on peer reviews. Every time a published report is retracted from a journal for reasons such as errors in the methodology or misconduct of the authors, the time external peer reviewers had spent on this article is wasted. As the FAIR principles would have an impact on the amount of reports retractable as described in indicator #3: Research retraction, it would also contribute to reduce the time wasted by peer reviewers.

2.4.2. *Indicator #2: Cost of storage*

For publishers and data repositories, the FAIR principles would reduce the cost of storing data by reducing the need for redundant copies. In order to quantify this cost, we use the following information:

- Research data volumes per researcher in Europe per year[59].
- The proportion of datasets that are stored in multiples places[60].
- Cloud and University storage costs

[56] https://www.go-fair.org/fair-principles/i1-metadata-use-formal-accessible-shared-broadly-applicable-language-knowledge-representation/
[57] https://www.go-fair.org/fair-principles/r1-metadata-richly-described-plurality-accurate-relevant-attributes/
[58] https://www.go-fair.org/fair-principles/r1-metadata-richly-described-plurality-accurate-relevant-attributes/
[59] Based on sources such as (Addis, 2015) and (Van Tuyl & Michalek, 2015)
[60] (Parsons, Grimshaw, & Williamson, 2013)

Based on different pricing models[61], we noted that the cost for storing data permanently has a linear relationship to the volume of data stored.

The following assumptions[62] were used to quantify the added cost of storage due to non-fair data:

- Empiric evidence[63] shows that on average data is stored in 4.63 repository. In a perfect open and collaborative world, the data would not need to be stored in more than one repository. However, FAIR is not a panacea for unnecessary redundancy. In order to remove unnecessary redundancy, new internal rules and cultural changes are also required in research (funding) institutions. While FAIR can contribute to making this change happen, we take the conservative assumption that implementing the FAIR principles will directly reduce the number of redundant copies of the data by 20%. The graph below shows the distributions of the number of copies (i.e. repositories used) in blue and the average sizes of datasets per number of copies.

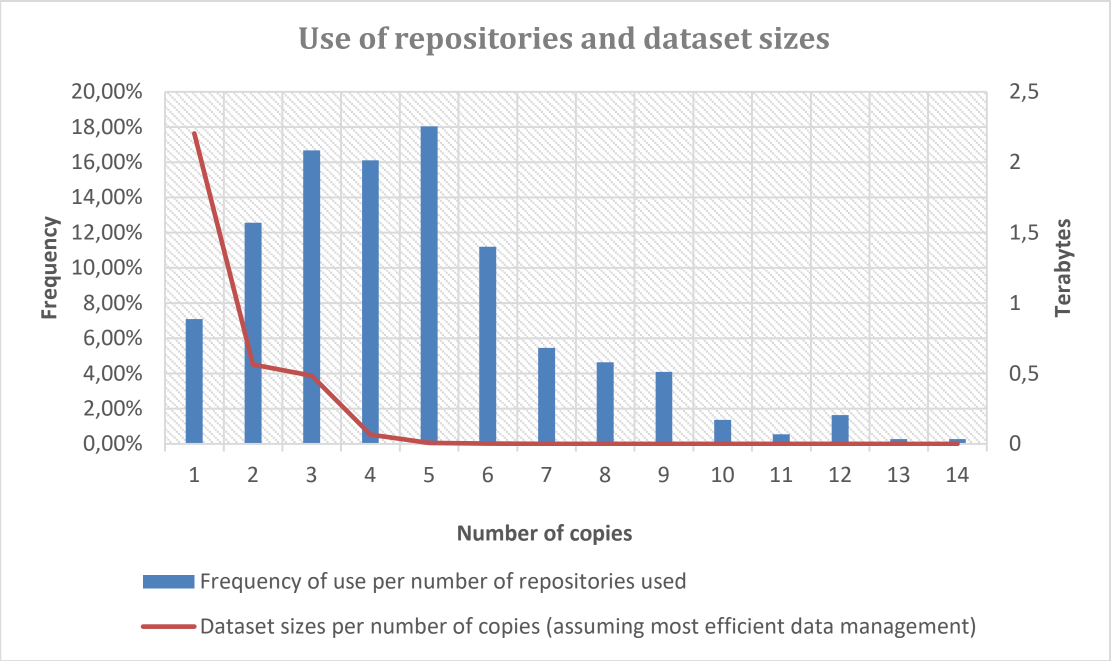


Figure 4: Distribution of volume per number of repositories

If we assume that the largest datasets are better managed (because the storage costs are higher) and map the distributions like presented above[64], then having a single copy of the data would reduce the total stored volume by 25.67%. This implies that a 20% reduction across the board is both feasible and conservative as in reality, not all of the largest datasets would be kept on a single repository.

- Backups which are made as part of the operation of an infrastructure or a repository are not regarded as separate copies of the data. For our calculation, we consider backups as a technical measure to prevent the loss of data on one copy of the data.
- As we rely heavily on surveys and data from the UK (according to the literature, UK being one of the most advanced countries in terms of data management) we assume that the data on which we based our assumptions and calculations can be extrapolated to the EU-28;
- We understand permanent storage as a minimum of 20 years;

[61] https://aws.amazon.com/s3/pricing/ and (Royal Veterinary College, University of London, n.d.)
[62] Arkivum, Estimating Research Data Volumes in UK HEI, 2015.
[63] (Parsons, Grimshaw, & Williamson, 2013)
[64] This assumption also gives the most conservative figures.

- The total volume of data is estimated based on existing studies which use buckets for asking researchers about the data volumes used. These buckets do not have a linear distribution, e.g. 1-50 MB, 50-100MB, 100-250MB etc.[65] As a consequence, the extrapolations made from such surveys are kept within each of the buckets used;
- Next to the non-linear distribution, the last proposition of the buckets is most of the time open-ended, e.g. >1TB. This may create of bias in our calculation for the storage of the largest files, since one answer in this category could be 2TB or 10TB, with a significant influence on the average.

The cost pricings per volume of data fluctuate with time and vary according to the pricing model of the repository. In this study, we use the pricing model offered by an UK college to its own researchers[66], and the data storage pricing models of AWS[67], Azure[68] and Google[69]. Finally, due to the limited availability of figures on volumes of data per researchers and redundancy, the scope of our calculation is limited to academic research. Extending our results to the whole research sector in EU would assume that all researchers in Europe produce the same quantity of data per year.

2.4.3. Indicator #3: Licence costs

First, we calculated the percentage of reports which specify a reusable licence in their metadata and we looked at the places in which the licence was actually indicated. Different uses of licences were evaluated:

- If there is no licence, the researcher either:
  - Do not reuse the report and its data; or
  - Reuse the report and its data without knowing if he or she can.
- If there is a licence in the metadata of the report or of the data, it can be:
  - A standardised machine readable licence; or
  - A local human readable licence; and
  - An open licence; or
  - A closed licence which imposes a fee for reusing the data.

In order to quantify the licence costs associated to not having FAIR and open access, we gathered statistics on the use repartition and we looked at which percentage of data could be made open. By aggregating studies[70] on open access to data, we consider that 71.5% of the data could be made open. This leaves 28.5% of the academic research data which must remain closed due to privacy and security reasons.

Secondly, we factored the percentage of data which is already open today. According to the European Commission[71], in 2015, 48% of the data is available in open access (i.e. gold and green open access). Similarly a recent study estimated, also for 2015, that 45% of the data is available in open access.

Thirdly, we looked at the subscriptions and access costs for research institutions. We found data for public research organisations in United Kingdom[72] and Finland[73] which we used to extrapolate the costs of licences for academic researchers in the EU28 based on the number of researchers per country. This indicator only looks at licence costs incurred

[65] (Van Tuyl & Michalek, 2015) and (Addis, 2015)
[66] (Royal Veterinary College, University of London, n.d.)
[67] https://aws.amazon.com/pricing/
[68] https://azure.microsoft.com/en-us/pricing/calculator/#storage1
[69] https://cloud.google.com/products/calculator/
[70] (Tenopir, et al., 2011) and (Johnson, Parsons, Chiarelli, & Kaye, JISC Research Data Assessment Support - Findings of the 2016 data assessment framework (DAF) surveys, 2016)
[71] https://ec.europa.eu/research/openscience/index.cfm?pg=access&section=monitor#viz1489066430689
[72] (Lawson, Meghreblian, & Brook, 2015)
[73] (Lahti, 2016)

today. It does not include the hypothetical licence costs for data that would have been reused if it had been available in open access.

2.4.4. Indicator #4: Research retraction

In order to quantify the cost of articles retraction, we first identified the volumes and reasons for retracting research. Various studies[74] estimate the reasons (e.g. non-reproducibility, errors, fraud, plagiarism, etc.) and the amount of retracted articles (0.04% of all articles). Based on discussions with experts and calculations, we assume that FAIR would help reduce the amount of retracted articles by 51.44%. This percentage is based on a study[75] looking at the causes for research retraction and our assumptions for the impact of FAIR on each if these causes (i.e. Errors, non-reproducibility, fraud (suspicion), multiple submission, plagiarism, other misconduct). We then factored the time spent by researchers on the retracted research to evaluate its cost.

Besides the decreasing percentage of articles retractable, the FAIR principles would also impact the average time required before an article is retracted. However, as the cost related to this average retraction time is difficult to estimate, we did not quantify it.

2.4.5. Indicator #5: Double funding

Our estimations of the costs of double funding focus on public research programmes and research grants. In order to identify to which extent double funding occurs, we rely on the work of Harold R. Garner, Lauren J. McIver and Michael B. Waitzkin[76] which estimates double funding by using full-text algorithms to identify overlaps in grant applications. From their work, we used the percentage of grants application with suspicious overlaps and the ratio depicting the average size of the first award compared to the potentially overlapping one.

Their conclusions for US research programmes are extrapolated to European research programmes using the average EU contribution for H2020 funded projects together with the total number of public research grants in EU.

The prevention of double funding relies primarily on anti-plagiarism tools which compare automatically research with available existing articles and data. This method is only effective to the extent that the plagiarised research is available in a machine-readable format, which would be ensured by the application of the FAIR principles. Therefore, we consider that at least 80% of double funding could be avoided with FAIR, as a few will likely find new creative methods for plagiarism.

Finally, while some private research-funding organisation also use grants, there is no data which could be used to support the calculations. As a consequence, our estimates cover double funding in public funded research.

2.4.6. Indicator #6: Interdisciplinarity

In order to quantify interdisciplinarity, we need to identify the volume of research made possible through FAIR-enabled cross-fertilisation. However, there are few studies that attempt to quantify the impact of interdisciplinarity and those that do rely on case studies[77]. For this reason, the impact of FAIR on interdisciplinarity cannot not be estimated with the data currently available.

However, while there is no data to estimate the value of cross-fertilization, it should also be noted that a new research project made possible with FAIR would always be replacing another research project which would have taken place otherwise. The impact of FAIR on cross-fertilization does not in fact increase the total volume of research performed.

[74] (Horbach & Halffman, 2017), (Masic, 2012), (Fang, Steen, & Casadevall, 2012), and (Wager & Williams, 2011)
[75] (Wager & Williams, 2011)
[76] (Garner, McIver, & Waitzkin, 2013)
[77] E.g. (Björkdahl, 2009)

Therefore, the impact of FAIR on interdisciplinarity is limited to the additional value of new research compared to that of the research which would have been done without FAIR.

2.4.7. Indicator #7: Potential economic growth

In order to quantify the potential economic growth, several approaches were followed, based on the work of Beagrie & Houghton (2014, 2016) and also the work of the Tauri Group[78]. Therefore, we used the following perspectives to look at the impact of FAIR research data on innovation, which later on can be transformed into an opportunity cost:

- Spinoff: relying on NASA's contractor work, we estimated the impact on innovation through the European spinoff network, as spinoffs stem from the academic world and are heavily reliant on research data. We assumed that due to research data not being FAIR a certain share of potential spinoffs does not exist. However, the multiplication effect of FAIR could not be quantified.
- Efficiency impact: we estimated the efficiency impact achieved due to FAIR research data. Based on calculations in section 2.4.1, the time lost due to inefficiencies stemming from non-FAIR data can be regarded as time freed up and reinvested to do other research-related activities. In other words, due to FAIR research data (i.e. access to new resources) a researcher could produce a same level output at a lower cost, which is considered to be an efficiency gain. However, we could not estimate to which extent the time saved by FAIR would be reinvested in innovative new research and the value thereof.
- Contingent valuation: contingent valuation consists of estimating the value of non-market goods and services based on preference theory[79]. In this case, individuals are asked what they would pay for a good or a service in a hypothetical market situation (i.e. pay for something that is actually free). Following this logic, the ideal process to quantify the FAIR research data impact on innovation would be to administer a questionnaire with a hypothetical scenario. The questionnaire would be created in such a way to arrive at a coefficient depicting the probability to which FAIR research data would contribute to a researcher's work. Later on, the coefficient would be transformed in a willingness to pay to benefit from FAIR research data[80]. However, performing an extensive survey to collect this information was not possible in the timeframe and context of our study which relies primarily on existing material.
- Loss of data: we tried to identify volumes of data lost due to not having FAIR research data and estimate the residual value of the lost data. However, without information on when and how much data is lost in relation to not having FAIR, we could not estimate the cost of data lost. Another issue is that not all data retains value equally over time. For instance, traffic information and weather forecasts lose most of their value very rapidly, while archaeological data retains value long after discovery.
- Return on public investment: finally, we also tried to measure the return on public investment on FAIR research, compared to regular research. According to Houghton, the social return on public investment of R&D should probably be somewhere 20-60%. However various problems arise, first research is both publicly and privately funded, and the distribution is unknown. And secondly, what is the proportion of the return on public investment that FAIR is contributing to.

The link between innovation and growth has been thoroughly studied by economists and researchers alike. However, such links are always illustrated with specific case studies, from which it is impossible to extrapolate. For instance, a case study at an institutional level would have to be extrapolated at discipline, national and European level. That is not

[78] https://www.nasa.gov/sites/default/files/files/SEINSI.pdf
[79] Reference theory states that a good or service which contributes to human welfare has economic value.
[80] To be also included, the calculation of the cost of (re)creating the data.

including the fact that a case study is generally discipline specific, and therefore cannot be a true reflection of another discipline.

Furthermore, the impact of not having FAIR on economic growth can only be quantified if the impact of FAIR on innovation (e.g. new science, research projects or patents) can be quantified accurately and with no overlaps. At the time of writing, there is insufficient data to support the assumptions needed to quantify the economic benefits of FAIR.

Instead, we will present mostly qualitative findings with regard to the economic benefits of FAIR.

3. COST CALCULATION

We estimate the annual cost of not having FAIR data to a minimum of **€10.2bn per year**. The actual cost is likely to be much higher due to unquantifiable elements such as the value of improved research quality and other indirect positive spill-over effects of FAIR research data.

The following picture shows the repartition of the cost per indicator compared to the total likely cost of not having FAIR research data, which includes a rough estimate based on figures for open data. On the left side of the graph, the impact on innovation, would account for over 60% of the likely cost of not having FAIR research data, while the minimum true cost of not having FAIR research data, encompassing indicators #1 to #5 accounts for the remaining 40%.

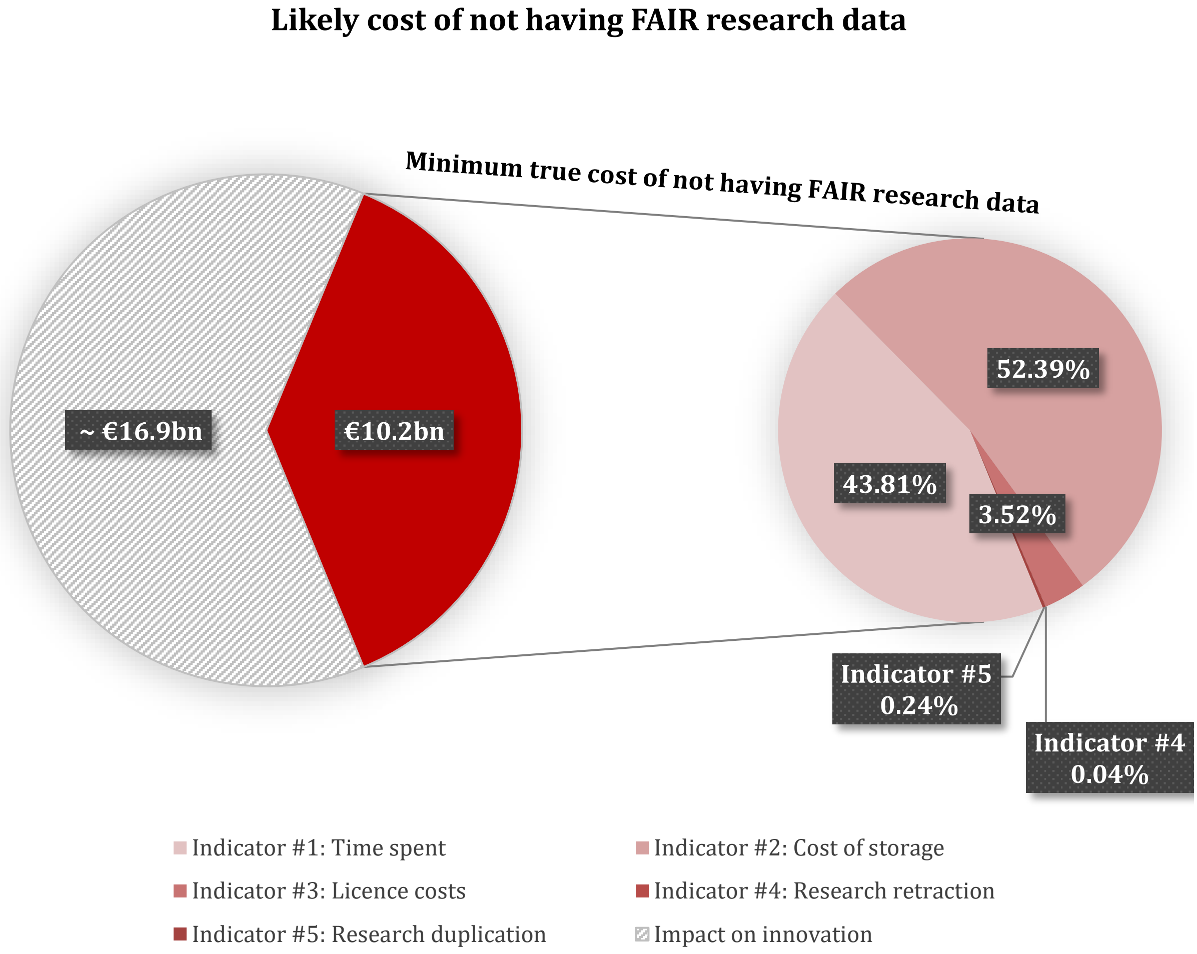


Figure 5: Cost breakdown

The right part of the graph shows that for the minimum true cost of not having FAIR research data, the indicator #1: time spent and the indicator #2: cost of storage account for most of the minimum true cost of not having FAIR research data.

The following sections present the main input data and assumptions used in the calculations for each indicator. The indicators and the approach to quantifying them are described in chapter 2. Our calculations, which rely exclusively on secondary data, are openly available[81].

3.1. Indicator #1: Time spent

[81] https://ec.europa.eu/info/files/cost-not-having-fair-research-data-data-and-calculations_en

The cost of time spent because of non-FAIR research is **€4.5bn per year**. It should be noted that having data solely for the public sector, the impact of not having FAIR on Non-Academics is extrapolated from the impact of not having FAIR on Academics. The table below presents the data that was used to estimate this indicator.

Cost of time lost by Non-Academics	***Unit***	***2017***
Avg. salary for a researcher in the private sector / *Non-Academics*	EUR (€)	€61,864
Avg. salary for a researcher in government / *Non-Academics*	EUR (€)	€52,853
Time dedicated to research by Non-Academics	Time	50.00%
Impact of not having FAIR on Non-Academics	Time	4.47%
Number of researchers in the private sector / *Non-Academics*	#	923,997
Number of researchers in government / *Non-Academics*	#	269,963
Total cost of time lost by Non-Academics		**€3,221,109,809**
Cost of time lost by Academics		
Avg. salary for a researcher in higher education / *Academics*	EUR (€)	€54,484.1
Time dedicated to research by Academics	Time	34.96%
Impact of not having FAIR on Academics	Time	3%
Number of researchers in higher education / *Academics*	#	727,348
Total cost of time lost by Academics		**€1,238,244,725**
Total cost of time lost due to not having FAIR research data		**€4,459,354,534**

Table 2: Indicator #1 calculation

3.2. Indicator #2: Cost of storage

The cost of redundant storage because of non-FAIR research is **€5.3bn per year**. The table below presents the data that was used to estimate the cost of storage that would not be needed of the FAIR principles were applied.

Inputs on data usage	***Unit***	***Value***
Average volume of data created per researcher/year	TB	2.45
Average cost of storing data per TB/year	EUR (€)	€122
Average number of repositories where data is stored	#	4.63
Average data retention period	Years	10
Reduction in the number of repositories needed thanks to FAIR	%	20%
Average number of repositories where data is stored after the FAIR implementation	#	3.70
Result: Cost of storage per researcher	EUR (€)	€2,776
Input on the number of researchers		***2017***
Total number of researchers in EU28	#	1,921,308
Total cost of unnecessary storage due to non-FAIR research data		**€5,333,228,576**

Table 3: Indicator #2 calculation

It should be noted that the average volume of data created per researcher/per year encompasses all types of data (e.g. audio and video recording, pictures, documents, spreadsheet, markup/code, etc.).

3.3. Indicator #3: Licence costs

The cost of licence costs due to the absence of open access is **€360m per year**. The table below presents the data that was used to estimate the value of the cost incurred for not having full open access to research data.

Inputs on open access	***Unit***	***2017***
(Research) data currently in open access	%	56.09%
(Research) data currently not available in open access	%	43.91%
Total proportion of (research) data that could be made open	%	71.47%
Closed (research) data which is not open for acceptable reasons (e.g. privacy)	%	28.53%
Additional data which would be made open by applying FAIR	%	31.38%
Licence costs for the academic research organisations in EU28	EUR(€)	€1,141,161,883
Total licence costs due to the absence of open access	***EUR (€)***	***€358,095,416***

Table 4: Indicator #3 calculation

Due to a lack of data, we could only estimate the licence costs faced by researchers in the public sector. The actual licence costs due to the absence of FAIR and open research data is thus likely even higher.

3.4. Indicator #4: Research retraction

The cost of research retraction because of non-FAIR research is **€4.4m per year**. The table below presents the data that was used to estimate the value of time lost due to research retraction which could have been avoider with FAIR.

Cost of time lost by Non-Academics	***Unit***	***2017***
Avg. salary for a researcher in the private sector / *Non-Academics*	EUR (€)	€61,864
Avg. salary for a researcher in government / *Non-Academics*	EUR (€)	€52,853
Time dedicated to research by Non-Academics	Time	50.00%
Time lost due to research retraction	Time	0.0085%
Reduction of retracted articles with FAIR	%	51.44%
Impact of not having FAIR on Non-Academics	Time	0.0044%
Number of researchers in the private sector / *Non-Academics*	#	923,997
Number of researchers in government / *Non-Academics*	#	269,963
Total cost of time lost by Non-Academics	**EUR (€)**	**€3,144,844**
Cost of time lost by Academics		
Avg. salary for a researcher in higher education / *Academics*	EUR (€)	€54,484.1
Time dedicated to research by Academics	Time	34.96%
Time lost due to research retraction	Time	0.0059%
Reduction of retracted articles with FAIR	%	51.44%
Impact of not having FAIR on Academics	Time	0.0031%
Number of researchers in higher education / *Academics*	#	727,348
Total cost of time lost by Academics	**EUR (€)**	**€1,208,927**
Total cost of time lost due to retraction		***€4,353,772***

Table 5: Indicator #4 calculation

It should be noted that FAIR also benefits the quality of articles that are not retracted. These benefits could not be estimated but it can be reasoned that retracted articles are only a small portion of the research which could benefit from FAIR. This was confirmed during interviews with experts who indicated that while FAIR discourages fraud, it also improves research quality across the board by improving reproducibility.

3.5. Indicator #5: Double funding

The cost of double funding because of non-FAIR research is **€25m per year**. The table below presents the data that was used to estimate this indicator.

Input about research duplication	***Unit***	***2017***
Total percentage of pairs with suspicious overlaps	Percent	0.03%
Average EU contribution in H2020 funded projects	EUR (€)	€1,783,788
Estimated number of public research grants in EU28	#	189,014
Total number of suspicious overlaps within the hypotheses for grants in EU28	#	50
Total funds allocated to the suspicious pairs	EUR (€)	€89,185,346
Average size of the first award of the pair compared to the second	#	1.9
Total value of the funds for duplicate grants	**EUR (€)**	**€30,753,568**
Reduction of duplication due to FAIR	Percent	80.00%
Avoidable cost of double funding	***EUR (€)***	***€24,602,854***

Table 6: Indicator #5 calculation

3.6. Indicator #6: Interdisciplinarity

The potential impact on interdisciplinarity that is missed due to not having FAIR research data could not be estimated reliably. It is acknowledged that the impact of FAIR on interdisciplinarity does not in fact increase the total volume of research performed, but rather enhance specifics aspects of the research itself, such as better output quality, greater collaboration and reusability, etc. On this basis, we identified several elements that led us to believe the tangible impact of FAIR on interdisciplinarity, but also the contribution of cross-fertilization to the cost of not having FAIR research data:

- Interdisciplinarity relies in part on reproducibilicy and requires transparency about tools, methods and data used. This leads to an increased reliability of the findings underlying scientific publications. Thanks to FAIR research data, the overall quality of research would be improved.
- For a vast majority of researchers surveyed[82], lack of access to data and poor quality of data is restricting interdisciplinarity and preventing good research quality. Introducing FAIR research data could drive to a greater cross-fertilization.
- FAIR would allow researchers to access disparate data from other disciplines, thereby giving them the opportunity to gain new insights and thus facilitating knowledge sharing.

To sump up, more interdisciplinarity through the FAIR principles could bring sciences disciplines closer and increase the current rate of data reuse[83], allowing to explore new

[82] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3126798/table/pone-0021101-t008/
https://www.elsevier.com/about/open-science/research-data/open-data-report
(Cox & Williamson, 2014)
[83] (Womack, 2015)

possibilities oustide the traditionnal way of conducting research, which would benefit the science community and other areas connected to it.

3.7. Indicator #7: Potential economic growth

The potential economic growth that is missed due to not having FAIR research data could not be estimated reliably just as well. However, we identified a number of elements which lead us to believe the missed economic spill-over effects of FAIR may constitute the largest component in the cost of not having FAIR research data:

- FAIR and openness have a direct positive impact on the number of citations. All things equal, FAIR research will be reused more often and thus have more value.
- Resources in open access constitute the second most used source of information after collections from one's own institution, according to survey[84]. By increasing the accessibility and reusability of research data, FAIR would also increase the societal value of research as open research is reused more often.
- Of the minority of researchers who make their data available electronically for others, the majority does so because they are required to do so, according to a study[85]. Applying the FAIR principles at institutions or funders' level would thus greatly improve the availability of research data, thereby creating value.
- Moreover, empirical evidence[86] demonstrates that the share of publications in open access has been steadily increasing since the 90s. In the absence of constraining policy or incentives, this strongly suggests that there are benefits for making research more accessible.

Finally, as the European data economy is estimated to be roughly the same size as the European research expenditures (both ~ €300bn[87]), a parallel can be drawn with open data. The likely economic benefits for open data were estimated[88] between €11.7bn and €22.1bn per year in Europe by 2020 and it could therefore be expected for the economic benefits of FAIR to fall in the same range.

[84] http://www.sr.ithaka.org/wp-content/uploads/2016/06/SR_Report_UK_Survey_Academics_2015_06152016.pdf
[85] http://www.ijdc.net/index.php/ijdc/article/view/10.1.210/393
[86] https://ec.europa.eu/research/openscience/index.cfm?pg=access&section=monitor#viz1489066430689
[87] (European Commission, 2017) and (Eurostat, 2017)
[88] (European Commission, 2017)

4. CONCLUSION

Interpreting the overall cost of not having FAIR research data as a single value overlooks many non-quantifiable benefits of FAIR. Nonetheless, at €10.2bn per year in Europe, the measurable cost of not having FAIR research data makes an overwhelming case in favour of the implementation of the FAIR principles.

To put this into perspective, research expenditures in Europe amounted to €302.9bn in 2016. While the minimum true cost of not having FAIR can be seen as only 3% of all research expenditures, €10.2bn per year is 78% of the Horizon 2020 budget per year and ~ 400%, of what the European Research Council and European research infrastructures receive combined.

To top this, figures for the open data economy suggest that the impact on innovation of FAIR could add another €16bn to the minimum cost we estimated.

While this study does not account for the cost of implementing FAIR, if we assume that the additional costs allocated for data management are up to 2.5% of all research expenditures, this would leave a positive balance of ~ €2.6bn per year from the implementation of the FAIR principles. Moreover, not all the costs for implementing the FAIR principles would be recurrent. Once the proper infrastructure in place, one could expect the net benefits from the FAIR principles to increase.

Our study presents results for the EU research economy as a whole, however the cost of not having FAIR varies strongly from discipline to discipline. In some data intensive disciplines such as genomics or crystallography, (some of) the FAIR principles have already been implemented already without the need for a quantified cost-benefit analysis.

As a result, some deductions can be drawn from the work accomplished:

- FAIR is part of a broader movement which is changing the way science is done, with the emergence of data stewardship and a growing momentum in favour of openness. Accordingly, it is essential that the necessary infrastructures and policies are implemented in order to fully benefit from the FAIR principles and maximise the value of research data.
- With regards to the cost of not having FAIR research, time spent and cost of storage are the most significant measurable cost drivers. In other words, FAIR would have a considerable impact on the time we spent manipulating data and the way we store data.
- Nevertheless, we are confident that the FAIR principles would greatly contribute to the science ecosystem and innovation in Europe, but the lack of data linking FAIR and innovation impedes the quantification of the impact of FAIR on innovation.

The cost of not having FAIR research was computed from five quantifiable indicators: time spent, cost of storage, licence costs, research retraction and double funding. These indicators cannot possibly cover all the benefits of FAIR i.e. on innovation and economic growth. Moreover, very conservative assumptions were used and some limitations apply to these indicators, for instance:

- With time spent, we did not factor that the time wasted due to not having FAIR could be reinvested in research which would lead to a certain return on investment.
- With cost of storage, our calculations cover only academic research, because of a lack of data for the private sector.
- With licence costs, our calculations cover only the cost borne by public research organisation, because of a lack of data for the private sector.
- With research retraction, we leave out an unknown quantity of research that is not retracted but also suffers of poor quality and non-reproducibility to some degree because of non-FAIR data.

We also could not fully consider the impact of not having FAIR on machine readability and usability. Data volumes produced are growing and in a few years from now researchers will not be able to deal and process those volumes manually. If the data is FAIR, machines could assist researchers in finding and analysing research data which would positively impact the science ecosystem, in terms of time freed up, preciseness, volumes analysed, and velocity but also in terms of new insights drawn.

Therefore, we are confident that the true cost of not having FAIR research data is much higher than the estimated €10.2bn per year. To estimate it fully, additional data would need to be collected, in particular for research in private organisations.

As the first world producer in volume of research data, Europe has placed research at the core of its development strategy. By unlocking additional value from research data, adopting the FAIR principles is not only a question of efficiency. Not doing so also comports a clear cost which is diverting resources from the next scientific breakthrough.

BIBLIOGRAPHY

Addis, M. (2015). Estimation of research data volumes per researcher in UK HEI. doi:10.6084/m9.figshare.1577539

American Journal Experts. (n.d.). *Peer Review: How We Found 15 Million Hours of Lost Time*. Retrieved from American Journal Experts: https://www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time/

Baker, M. (2016). *1500 scientists lift the lid on reproducibility.* Retrieved from https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Beagrie, N., & Houghton, J. (2014). *The Value and Impact of Data Sharing and Curation.* Retrieved from http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

Beagrie, N., & Houghton, J. (2016). *The Value and Impact of the European Bioinformatics Institute.* Retrieved from https://beagrie.com/static/resource/EBI-impact-report.pdf

Biology, A. I. (2015). ASCB Member Survey on Reproducibility. Retrieved from http://www.ascb.org/wp-content/uploads/2015/11/final-survey-results-without-Q11.pdf

Björkdahl, J. (2009). Technology cross-fertilization and the business model: The case of integrating ICTs in mechanical engineering products. *Research Policy*, 1468-1477. Retrieved from https://www.researchgate.net/publication/46489026_Technology_cross-fertilization_and_the_business_model_The_case_of_integrating_ICTs_in_mechanical_engineering_products

Bonino da Silva Santos, L. O., Wilkinson, M., Kuzniar, A., Kaliyaperumal, R., Thompson, M., Dumontier, M., & Burger, K. (2016). *AIR Data Points Supporting Big Data Interoperability.* Retrieved from https://www.researchgate.net/publication/309468587_FAIR_Data_Points_Supporting_Big_Data_Interoperability

Charles Beagrie Limited. (2011). *User Guide for Keeping Research Data Safe - Assessing Costs/Benefits of Research Data Management, Preservation and Re-use.* Retrieved from https://beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf

Cheol Shin, J., Arimoto, A., Cummings, W. K., & Teichler, U. (2014). *Teaching and Research in Contemporary Higher Education.* Springer. doi:10.1007/978-007-6830-7

Council of the European Union. (2016, May 27). *OUTCOME OF PROCEEDINGS: The transition towards an Open Science system - Council conclusions.* Retrieved from http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

Court, S. (2012, October). An analysis of student:staff ratios and academics' use of time, and potential links with student satisfaction. Retrieved from https://www.ucu.org.uk/media/5566/An-analysis-of-studentstaff-ratios-and-academics-use-of-time-and-potential-links-with-student-satisfaction-Dec-12/pdf/ucu_ssranalysis_dec12.pdf

Cox, A., & Williamson, L. (2014). *The 2014 DAF Survey at the University of Sheffield.* Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/10.1.210/393

Data Citation Synthesis Group. (2014, February). *Joint Declaration of Data Citation Principles*. (M. M., Editor, & F. 11, Producer) doi:10.25490/a97f-egyk

Data Life Cycle | DataONE. (n.d.). Retrieved January 2018, from DataOne | Data Observation Network for Earth: https://www.dataone.org/data-life-cycle

Davis, P. (2012). *How Much of the Literature Goes Uncited?* Retrieved from The Scholarly kitchen: https://scholarlykitchen.sspnet.org/2012/12/20/how-much-of-the-literature-goes-uncited/

Directorate-General for Research & Innovation (European Commission). (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020.* Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Directorate-General for Research and Innovation (European Commission). (2017). *The economic rationale for public R&I funding and its impact.* Publications Office of the European Commission. Retrieved from https://publications.europa.eu/en/publication-detail/-/publication/0635b07f-07bb-11e7-8a35-01aa75ed71a1/language-en

Dunning, A., de Smaele, M., & Böhmer, J. (2017). *Are the FAIR Data Principles fair?* 4TU.ResearchData. Retrieved from https://zenodo.org/record/321423#.Wku301WnE3F

European Commission. (2017). *Final results of the European Data Market study measuring the size and trends of the EU data economy.* Retrieved from https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

European Commission. (2017). *Review of the Directive on the re-use of public sector information (Directive 2013/37/EU).* Retrieved from https://ec.europa.eu/info/law/better-regulation/initiatives/ares-2017-4540429_en

Eurostat. (2017). *R & D expenditure*. Retrieved from Eurostat: http://ec.europa.eu/eurostat/statistics-explained/index.php/R_%26_D_expenditure

Fang, F. C., Steen, G. R., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 17028-17033. doi:https://doi.org/10.1073/pnas.1212247109

Garner, H., McIver, L., & Waitzkin, M. (2013). Same work, twice the money? *Nature, 493*, 599-601. Retrieved from https://www.healthra.org/download-resource/?resource-url=/wp-content/uploads/2013/11/Garner_Nature_1_2013.pdf

Horbach, S., & Halffman, W. (2017). The extent and causes of academic text recycling or 'self-plagiarism'. *Elsevier*, 11. doi:https://doi.org/10.1016/j.respol.2017.09.004

Houghton, J., & Gruen, N. (2014). *Open Research Data - Report to the Australian National Data Service (ANDS).* Retrieved from http://apo.org.au/system/files/53613/apo-nid53613-72236.pdf

Johnson, R., Parsons, T., Chiarelli, A., & Kaye, J. (2016). *Jisc Research Data Assessment Support - Finding of the 2016 data assessment framework (DAF) surveys.* doi:10.5281/zenodo.177856

Johnson, R., Parsons, T., Chiarelli, A., & Kaye, J. (2016). *JISC Research Data Assessment Support - Findings of the 2016 data assessment framework (DAF) surveys.* doi:10.5281/zenodo.177856

Kejser, U. B., Nielsen, A. B., & Thirifays, A. (2011). *Cost Model for Digital Preservation: Cost of Digital Migration.* The International Journal of Digital Curation. Retrieved from http://www.ijdc.net/index.php/ijdc/article/viewFile/177/246

Lahti, L. (2016). *Scientific journal subscription costs in Finland 2010-2015: a preliminary analysis.* rOpenGov. Retrieved from http://ropengov.github.io/r/2016/06/10/FOI/

Lawson, S., Meghreblian, B., & Brook, M. (2015). *Journal subscription costs - FOIs to UK universities.* Retrieved from figshare: https://figshare.com/articles/Journal_subscription_costs_FOIs_to_UK_universities/1186832

Masic, I. (2012). Plagiarism in scientific publishing. *Acta Informatica Medica*, 208-213. doi:10.5455/aim.2012.20.208-213

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L., & Wilkinson, M. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use, vol. 37, no. 1*, pp. 49-56.

Nur Farah Naadia Mohd Fauzi, P., Abd Rashid, K., Ahmad Sharkawi, A., Farihah Hasan, S., & Aripin, S. (2016). A survey on required time allocated vs. actual time spent by academic in pursuit of key performance indicators (KPIs) at departme,t of quantity surveying, KAED, IIUM. 42-48. Retrieved from http://jsrad.org/wp-content/2016/Issue%204,%202016/7jj.pdf

Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., & Sesink, L. B. (2012). *An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS.* International Journal on Digital Libraries. doi:10.1007/s00799-012-0092-1

Parsons, T., Grimshaw, S., & Williamson, L. (2013, 02 06). Research Data Management Survey. Retrieved from http://eprints.nottingham.ac.uk/1893/1/ADMIRe_Survey_Results_and_Analysis_2013.pdf

Peter, Larry, Raphael, Gary, & Beth. (2015). *FAIR and RDA DFT: sharing the same key messages.* Retrieved from https://b2share.eudat.eu/api/files/e2d84fea-b4e2-41d1-8d4e-6ce2d315b2b9/comparison-fair-dft-v2.pdf

Piwowar, H. A., & Vision, T. J. (2013). *Data reuse and the open data citation advantage.* Retrieved from https://peerj.com/articles/175/?utm_content=buffer9059d&utm_source=buffer&utm_medium=twitter &utm_campaign=Buffer

Royal Veterinary College, University of London. (n.d.). *Costing Research Data Management*. Retrieved from https://www.rvc.ac.uk/research/about/research-data-management/before-a-project/costing-research-data-management : https://www.rvc.ac.uk/research/about/research-data-management/before-a-project/costing-research-data-management

Stehouwer, H., & Wittenburg, P. (2014). *RDA Europe: Data Practices Analysis.* Research Data Alliance Europe. Retrieved from https://b2share.eudat.eu/api/files/0312c522-968c-43e2-8127-9772d68ba084/iCORDI-D2%205-final-submit.pdf

Tenopir, C., Allard, S., Douglass, K., Umur Aydinoglu, A., Wu, L., Read, E., . . . Frame, M. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*. doi:https://doi.org/10.1371/journal.pone.0021101

Van Noorden, R. (2013). Open access: The true cost of science publishing. *Nature*. Retrieved from https://www.nature.com/news/open-access-the-true-cost-of-science-publishing-1.12676

Van Tuyl, S., & Michalek, G. (2015). Assessing Research Data Management Practices of Faculty at Carnegie Mellon University. doi:10.7710/2162-3309.1258

Wager, E., & Williams, P. (2011). Why and how do journals retract articles? An analysis of Medline retractions 1988-2008. *Journal of medical ethics*, 9. doi:10.1136/jme.2010.040964

Ware, M., & Mabe, M. (2012). *The stm report - An overview of scientific and scholarly journal publishing.* Retrieved from http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf

Wilkison, M. D., Dumontier, M., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. doi:10.1038/sdata.2016.18

Wittenburg, P. (2017). *Costs of FAIR Compliance and not being FAIR compliant.* doi:10.23728/b2share.e184bd1ff12d45269de80c3f3e443eb7

Wittenburg, P., Ritz, R., & Berg-Cross, G. (2015). *RDA Data Foundation and Terminology - DFT: Results RFC.* Retrieved from https://b2share.eudat.eu/api/files/266e2ea2-6200-4b9d-9d3b-ed6d03ff93a5/DFT%20Core%20Terms-and%20model-v1-6.pdf

Womack, P. R. (2015). *Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics.* doi:10.1371/journal.pone.0143460

Zahedi, Z., Haustein, S., & Bowman, T. D. (2014). *Exploring data quality and retrieval strategies for Mendeley reader counts.* Retrieved from http://www.asis.org/SIG/SIGMET/data/uploads/sigmet2014/zahedi.pdf

Ziker, J. N. (2013, October 13). Time Allocation Workload Knowledge Study, Phase 1 Report. ResearchGate. Retrieved from https://www.researchgate.net/publication/308761975_Time_Allocation_Workload_Knowledge_Study_Phase_1_Report

Annex I FAIR PRINCIPLES

The table below describes different criteria required to adhere to the FAIR principles. The left column lists the 15 facets corresponding to FAIR, while the right column presents metrics to assess the FAIR compliance of a digital resource[89]. The wording '(meta)data' is used in cases where the Principles should be applied to both metadata and data.

Table 7: FAIR Principles and facets against a (meta)dataset

The FAIR Guiding Principles	FAIR Proposed Metrics[90]
To be Findable:	
F1. (meta)data are assigned a globally unique and eternally persistent identifier	*Identifier Uniqueness* Presence of an URL linking to a registered identifier scheme, which uniquely identify the digital resource. Some non-exhaustive examples are RN, IRI, DOI, Handle, trustyURI, LSID, etc. *Identifier Persistence* Presence of an URL linking to a policy that would manage changes in the identifier scheme.
F2. data have rich metadata description (defined by R1 below)	*Machine-readability of metadata* Attributes to optimize their discovery. Structured and rich metadata attributes such as title, creator, publication date, keyword(s), temporal and spatial coverage, etc. Ideally an URL should link to a document that contains machine-readable metadata for the digital resource.
F3. (meta)data are registered or indexed in a searchable resource.	*Resource Identifier in Metadata* Metadata must explicitly contain the identifier for the digital resource it describes thus an URL of the metadata and the IRI of the digital resource it describes must be provided.
F4. metadata specify the data identifier.	*Indexed in a searchable resource* The persistent identifier of the resource and one or more URLs that give search results of different search engines must be provided.
To be Accessible:	
A1. (meta)data are retrievable by their identifier using a standardized communications protocol **A1.1** the protocol is open, free, and universally	*Access Protocol* An URL describing the protocol whether it is an open and free protocol, closed protocol, protocol with royalties must be provided. *Access authorization* In case of restrictions, the protocol by which the content can be accessed must be fully specified (i.e. whether authorization is needed and a description of the process to obtain access to restricted

[89] https://www.go-fair.org/fair-principles/
https://zenodo.org/record/321423#.WhV7oFWnE3E
https://www.nature.com/articles/sdata201618
https://www.force11.org/group/fairgroup/fairprinciples
https://via.hypothes.is/https:/content.iospress.com/articles/information-services-and-use/isu824#ref019

[90] https://github.com/FAIRMetrics/Metrics/blob/master/ALL.pdf

implementable **A1.2** the protocol allows for an authentication and authorization procedure, where necessary	content).
A2. (Meta)data are accessible, even when the data are no longer available	*Metadata longevity* An URL to a formal metadata longevity plan must be provided, as metadata must remain discoverable, even in the absence of the data.
To be Interoperable:	
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.	*Use a Knowledge Representation Language* Necessary use of languages that are capable of representing concepts in a machine-readable manner. An URL linking to the specification of such a language must be provided.
I2. (meta)data use vocabularies that follow FAIR principles	*Use FAIR Vocabularies* The metadata values and qualified relations should themselves be FAIR, for example, terms from open, community-accepted vocabularies published in an appropriate knowledge exchange format. An IRIs representing the vocabularies used for (meta)data must be then provided.
I3. (meta)data include qualified references to other (meta)data	*Use Qualified References* Relationship within (meta)data, and between local and third-party data, have explicit and 'useful' semantic meaning.
To be Reusable:	
R1. meta(data) have a plurality of accurate and relevant attributes.	Aspects of metadata that help one to evaluate how reusable a dataset is. Authors should not attempt to define the possible downstream usages, but rather should provide as many attributes as possible, beyond those required for their anticipated downstream use.
R1.1. (meta)data are released with a clear and accessible data usage licence	*Accessible Usage Licence* (meta)data are released with a clear and accessible data usage licence. An IRI of the licence (e.g. its URL) for the data licence and for the metadata licence must be provided. Achieving machine-readability is a plus and is possible by referring to one of the licences at http://purl.org/NET/rdflicense
R1.2. (meta)data are associated with detailed provenance	*Detailed provenance* Provenance information such as who/what/when produced the data and why/how was the data produced should be associated with the data. Two URLs must be provided. One pointing to the vocabularies used to describe citational provenance, the second one pointing to one of the vocabularies (likely domain-specific) used to describe contextual provenance.
R1.3. (meta)data meet domain-relevant community standards	*Meets community standards* Provide a certification, from a recognized body, saying that the resource is compliant with the community standards.

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: http://europa.eu/contact

ON THE PHONE OR BY E-MAIL
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 electronic mail via: http://europa.eu/contact

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: http://europa.eu

EU PUBLICATIONS
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

EU LAW AND RELATED DOCUMENTS
For access to legal information from the EU, including all EU law since 1951 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/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

FAIR research data encompasses the way to create, store and publish research data in a way that they are findable, accessible, interoperable and reusable. In order to be FAIR, research data published should meet certain criteria described by the FAIR principles. FAIR originated from the current patchy data management practices in EU, which are not optimal yet. Several local initiatives, as well as global ones, are making the move towards an infrastructure supporting the FAIR principles in order to get the most of research data. This report aims to estimate the cost of not having FAIR research data for the EU economy based on a series of measurable indicators, which were defined based on existing studies and interviews with subject matter experts.

Studies and reports

Publications Office