Handling Open Research Data within the Max Planck Society – Looking Closer at the Year 2020

Martin Boosen https://orcid.org/0009-0009-3989-5087,
Michael Franke https://orcid.org/0000-0002-2661-8242,
Yves Vincent Grossmann https://orcid.org/0000-0002-2880-8947,
Sy Dat Ho https://orcid.org/0000-0002-6218-4146,
Larissa Leiminger https://orcid.org/0000-0002-6491-3197, and
Jan Matthiesen https://orcid.org/0000-0001-6548-3654
Max Planck Digital Library https://ror.org/0061msm67
rdm@mpdl.mpg.de

February 2024

Abstract

This paper analyses the practice of publishing research data within the Max Planck Society in the year 2020. The central finding of the study is that up to 40% of the empirical text publications had research data available. The aggregation of the available data is predominantly analysed. There are differences between the sections of the Max Planck Society but they are not as great as one might expect. In the case of the journals, it is also apparent that a data policy can increase the availability of data related to textual publications. Finally, we found that the statement on data availability "upon (reasonable) request" does not work.

Contents

1 Introduction

Data sharing is becoming a new standard in research. Data accessibility is increasingly being discussed as an important aspect of scientific reproducibility. Funding agencies and public authorities are increasingly demanding that their money be used to make research data as openly accessible as possible. Many efforts are being made to ensure that the conscious handling of research data becomes an established practice in everyday scientific life.
However, there is sometimes a gap between aspiration and reality. The reasons for this are to be found on the side of the scientists, the publishers, and on the side of the infrastructure providers. Although there is often some kind of agreement – voluntary or involuntary – to share data. But it is still remarkably rare. To change this, the aim would be to take the pain out of data sharing.[1] This and other debates around data sharing are discussed with reference to the Max Planck Society in the following chapters.

As already mentioned data sharing is perceived as an important element of intra-scientific exchange. At the same time, the understanding of data sharing is hampered by different interpretations of the pair "data" and "sharing". Focusing on research data from the perspective of the Max Planck Society as well as on its practices this paper acknowledges two ways of understanding this term. On the one hand, research data is the basis for scientific findings and their publication.[2] On the other hand, published research data is increasingly emerging as a genre of scholarly output in its own right. The two come together again when we ask about the accessibility of the data underlying published texts.

The two sides of the research data coin – wanting to commit to data sharing, but failing to do so – are the focus of the first section. A general spectrum of data sharing is briefly outlined. The main emphasis is on developments since 2010. Here, the positions and initiatives involved in bringing data accessibility more into focus are traced. The second section examines the perspective of the Max Planck Society by presenting the open data and sharing practices of the decade of the 2010s. The main aim is to show which institutional frameworks have been established and which intentional vagueness (in the sense of a desired added value) has been maintained. The third section takes a closer look at the Max Planck Society and shows what its publication behaviour looks like in 2020 in terms of research data. This evaluation will take up the most space, as it formulates the data collection and the descriptive analysis of the results. The paper concludes with an outlook on the decade of the 2020s and possible developments in data sharing for the Max Planck Society.

---

[1] See for example the recently published overview Hutson 2022.

[2] See the old "Rules of Good Scientific Practice" of the Max Planck Society from 2000 resp. 2009, Max Planck Society 2009, p. 4.

2　Observations on Sharing Research Data

Data sharing is not a phenomenon of the 21st century. Sharing knowledge has long been an ingrained cultural practice of mankind. At the same time, data sharing seems to be on the rise since the beginning of the third millennium, not least because of the possibilities digitalization brings. This development is new because it has given rise to a new genre of publication called "data" or "dataset". This data sharing is accompanied by several phenomena.

First and foremost, the open exchange of data is becoming increasingly important in this context. Open means barrier-free access to the data. This is usually done by attaching the data to the textual publication or by publishing the data as a separate dataset in a repository. Especially in the latter case, there are already signs that data publishing is becoming a genre in its own right. This goes hand in hand with increased visibility of research data. They can gain visibility through open or restricted access options. Nevertheless, they are accessible, or at least there is knowledge of their existence. Therefore, especially for research data with restricted access, it is important that the metadata, i.e. the descriptive information about the data, is freely available. This makes it much easier and faster for other researchers to learn of the existence of the research data and then request access. This is adding to the recognition as an entity in their own right.

Regardless of this, data can also be reused in ways that were not foreseen by the original authors. The more conscious use of data licences facilitates such reuse and reinterpretation. Particularly in a scientific context, future questions cannot always be guessed at from current perspectives. That is why it makes sense to present research data in such a way that new questions can be discussed in the future. Parallel to software, licensing is increasingly becoming a familiar process. This creates legal certainty, which makes access easier – in a legal sense. In many cases, this increase in visibility is accompanied by an increase in citations by other scientific results. It is therefore not surprising that the bibliometric study by Colavizza and colleagues found that there is "*a citation advantage, of up to 25.36% (± 1.07%), with articles that have a [..] link to a repository via a URL or other permanent identifier*".[3]

For the phenomena of sharing data openly, we see – in view of Germany – a strong acceleration of discussion contributions in the 2010s, but an increasing acceptance of the forms of data publication. For this reason we present a chronology for the recent past. An important milestone and turning point are the FAIR Data Principles[4], which were published in 2016. Not only do they cover the decade of the 2010s, but they also provide explicit guidance on the publication of research data. At the same time, for the first time, the FAIR Principles provide a simple and concise way of articulating – FAIR – what has

---

[3]Colavizza et al. 2020, p. 14.
[4]Wilkinson et al. 2016.

often been expressed in more descriptive terms when dealing with research data.[5] Such a German frame of reference is naturally integrated into European and worldwide developments. Research is mostly international. Nevertheless, in the area of normative discussions, the Max Planck Society mainly operates in Germany, so this is the primary focus of this text.

To reflect these developments, the following chapter discusses the evolution prior to the FAIR Data Principles 2016. This is followed in chapter 2.1 by a description of the FAIR acceleration phase from 2016 to the present. These two parts serve as a derivation to the concrete study year 2020. We will then focus in chapter 2.2 on specific aspects. They become particularly relevant again in the following chapter, i.e. the empirical analysis of publication behaviour. First, the text focuses on the particular aspect of data availability statements to illustrate how the approach to data accessibility has changed. However, this would be difficult to explain without the institutional framework. Therefore, the following two subsections discuss developments in institutional data policies in 2.4 and funders 2.5 to conclude the general perspective on research data sharing. Finally, this chapter also serves as a general introduction and transition in terms of content and timing for the next parts of the text.

2.1 Discussion on Sharing Research Data before 2016

Looking back, it is often difficult to pinpoint the exact start of a discussion. Therefore, temporal categorisations often have something artificial or arbitrary about them. The same is true here. Nevertheless, the dilemma remains that start and end points must be specified. One of the first publications with specific recommendations worldwide on research data was a 53-page publication by the OECD in 2007.[6] In particular, the aim was to develop a set of guidelines, based on commonly agreed principles, to facilitate cost-effective access to publicly funded digital research data.[7] The sequencing of human DNA – successfully completed in 2003 – was used to study many biological processes, so open sharing of research data was already implicit in the project.

In the German context, the first significant contribution to the discourse can be seen in 2009. One of the DFG's first decentralised recommendations for action on research data was the ”Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimardaten” of January 2009.[8] This

---

[5]Due to the abundance of introductions, handouts and presentations, an explanation of the FAIR principles will not be given here. For the German context, however, the following is a suitable overview see Linne et al. 2021.

[6]OECD 2007. Particularly worth reading are Principles A to M, which attempt to define concrete guidelines for the design of data access.

[7]OECD 2007, p. 3.

[8]Ausschuss für wissenschaftliche Bibliotheken und Informationssysteme and Unterausschuss für Informationsmanagement 2009. An English translation of this would be ”Recommendations for the Secure Preservation and Provision of Digital Research Primary Data”.

was shortly followed by the Alliance of German Science Organisations' "Principles for the Handling of Research Data" in 2010.[9] Both recommendations were intended to set initial standards for the concrete handling of research data.

The "Kommission zur Zukunft der Informationsinfrastruktur"[10], the first joint institution with the focus on research data in the Federal Republic of Germany was established in 2009 on the recommendation of the Gemeinsame Wissenschaftskonferenz (GWK). This commission, with the participation of many German scientific organisations and institutions, also had its own working group on research data.[11] This was followed in 2012 by recommendations from the German Science and Humanities Council (WR), which called for the professional communities or actors in interdisciplinary research fields to develop quality criteria for the generation of research data and guidelines for appropriate data management, where these do not already exist.[12] As an interim result, it was noted in 2015 that the field of research data is a very dynamic and diverse field.[13]

In addition to these recommendations for action by institutional stakeholders, research and education institutions have been issuing their own data policies since the beginning of the 2010s. Based on the recommendations mentioned above, it can be observed for the German context that the number of data policies and universities and research institutions has increased significantly since 2014.[14] Already in 2011, the German debate on data policy is characterised by opposite poles of recommendation – as recommendations in the hope that they will be followed out of conviction – and obligation – so real commitments.[15] This also fits in well with the fact that in Horizon 2020 (running from 2014 to 2020) the European Commission has for the first time called for a conscious handling of data, for example through data management plans (with the possibility to opt-out).[16] In retrospect, the long-term development towards obligatory requirements regarding research data at Horizon Europe is already indicated here.

With these developments towards active management of research data, the

---

[9] Allianz der deutschen Wissenschaftsorganisationen 2010.

[10] A translation of this may be "Commission on the Future of the Information Infrastructure.

[11] Kommission Zukunft der Informationsinfrastruktur 2011, line 490-506.

[12] Wissenschaftsrat 2012, p. 56.

[13] Franke et al. 2015, p. 2.

[14] The best overview of research data policies in the German-speaking world is provided by www.forschungsdaten.org/index.php/data_policies. It is worth mentioning the recommendations of the German Rectors' Conference (Hochschulrektorenkonferenz) of 2014, Hochschulrektorenkonferenz 2014. The two-year period between the WR recommendation and the first policies can be reasonably explained by an internal development and establishment process. The introduction of a data policy in an institution simply requires different resources and time.

[15] For the German context, see in particular Pampel and Bertelmann 2011.

[16] European Research Council 2017.

positive aspects of data sharing and citation are being observed. Publishers have also begun to adopt data policies and request research data associated with submitted articles.[17] In this context, we are beginning to see how a change in publishing behaviour – towards data sharing – is affecting the behaviour of scientists. For example, for some supporters, the willingness to share research data is linked to the strength of the evidence and the quality of reporting of statistical results.[18]

Some evaluations have shown that the behaviour of scientists is moving towards open sharing of research data. *"However, there is increased perceived risk associated with data sharing, and specific barriers to data sharing persist."*[19] It is interesting to note that the scientific communities have given themselves their own guidelines or recommendations on how best to deal with research data in their environment.[20] For some aspects, this development can be repeated as well at a general level, towards universal principles for the handling of (research) data.

2.2 Understanding Data Sharing Practices after Announcing the FAIR Data Principles in 2016

The publication of the FAIR Data Principles[21] in 2016 can, in retrospect, be seen as an important point on the way to an evolving awareness of data sharing. Two things in particular were central to this. First, there was broad acceptance of the principles. And secondly, it was now possible to give a name to both the process and the outcome, namely FAIR. This naming, as an explicit accumulation of knowledge, should not be underestimated in the success of the FAIR Principles.[22]

The FAIR Data Principles were adopted quite rapidly by the European Commission in 2016 and documented in the "Guidelines on FAIR Data Management in Horizon 2020"; [23] However, it did not stop at the introduction of guidelines. Already in 2017, the Open Research Data Pilot covered all thematic areas of Horizon 2020, which *"aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects"*.[24] The aim

---

[17]An example would be PLOS Biology, with the intention of increasing data availability and transparency, Bloom, Ganley, and Winker 2014.

[18]Wicherts, Bakker, and Molenaar 2011.

[19]Tenopir et al. 2015, p. 1.

[20]See for example Goodman et al. 2014.

[21]Wilkinson et al. 2016.

[22]A general overview of the introduction of the FAIR Data Principles in the first years can be found in Thompson et al. 2019. At the same time, this already contains an existing criticism of the FAIR Data Principles, namely that they are often mentioned performatively, but the concrete application sometimes leaves much to be desired.

[23]European Commission 2016. See also later the report on turning FAIR into reality from 2018, European Commission 2018b.

[24]European Commission 2018a.

was to create incentives to convince scientists to organise their data according to the FAIR Data Principles.

Two years later, the European Commission produced an analysis of the cost of FAIR research data and the lack of application of the FAIR data principles. According to this document, "*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.*"[25] From this perspective, the opportunity cost of not having the FAIR Data Principles was high. Beyond the EU as a major initiator, there have been many other developments towards the concrete application of the FAIR Data Principles.

According to a survey conducted by the European Commission in 2022, recognition of research data management and sharing is already partly present, but could be significantly higher, according to the participants.[26] There are different approaches to encourage different actors to structure their data according to the FAIR Data Principles and, where possible, to make them publicly accessible. There are explicit commitments, e.g. through research data policies of institutions, deliverables from funders or demanded data availability statements by journal publisher. At the same time there are incentives to promote the conscious use of data, e.g. a data index[27] as a bibliometric measure, to help the independent genre of 'data publication' gain more recognition.[28] There are many initiatives, ideas and new ways of doing things in the scientific communities that try to facilitate and promote the application of the FAIR Data Principles. However, it is difficult to generalise such developments in a meaningful way by subject or other criteria. Because "*data sharing perceptions and practices are highly variable among academic disciplines.*" [29]

Nevertheless, the genre of analysis of research data sharing behaviour has become common.[30] There are two broad categories into which this can be grouped. The first category is working on data sharing behaviour within a research field. Such publications are particularly concerned with understanding how data is handled within a discipline.[31] For example, to what extent is open sharing of research data a quasi-standard there? Or are there good reasons for restricting access to data? Are the FAIR data principles widely applied? How is the reproducibility of research data valued within the

---

[25]European Commission 2019.

[26]See in particular Neuroth and Oevel 2021, pp. 552-553 with a geographical focus on Germany and similar discussions is particularly interesting in this context.

[27]Hood and Sutherland 2021.

[28]The development of data journals are an example of this.

[29]Pujol Priego, Wareham, and Romasanta 2022, p. 238. In this whole discussion about the implementation of the FAIR Data Principles, however, it is striking that there is comparatively little explicit criticism of them.

[30]See also, for example, the meta-analysis of various studies Donner 2022.

[31]See for example J. A. Borghi and Gulick 2021, Crüwell et al. 2022, Houtkoop et al. 2018, Jeng and He 2022, Leonelli 2017, Mandeville et al. 2021, Rousi 2022 and more.

scientific community? And where is research data published? The overall aim of these analyses is to understand what patterns of behaviour currently exist and evolve. These questions are often compared to a desired state. At the same time, there are also comparisons between different disciplines, as well as cross-disciplinary analyses.[32]

The second category of analyses is primarily interested in data publication from an institutional perspective. This perspective is mostly taken by infrastructure providers such as libraries and data centres. It focuses on the question of where data should be published by researchers affiliated to their own institution.[33] In addition to the collection of bibliometric statistics, this is a motivation for the continuous improvement of institutional research data services. In the German context, the Charité Dashboard on Responsible Research[34] is particularly worthy of mention. There, the institution's open data and code publications can be viewed in quasi-live mode. Most of these studies of data sharing behaviour focus on the issue of restricting access without justification. It is interesting to emphasise here that in many cases after 2016, the main focus was on the concrete implementation of the FAIR Data Principles.

2.3 Data Availability Statements: Make Data "Accessible" on Reasonable Request

"*All that glisters is not gold.*"[35] The same sometimes applies for research data. In theory, reproducibility is part of good research. But in reality, reproducibility of data-based findings is sometimes difficult or almost impossible. The stumbling block is often a data availability statement, which publishers now routinely require for scientific publications. The aim is to document where the data, on which the hypotheses of the publication are based, is available. The proportion of data availability statements varies. In some cases, the coverage of articles is quite high, depending on the discipline and on the journals' data policies. For example, in 2018, 93.7% of 21,793 PLOS articles and 88.2% of 31,956 BMC articles had data availability statements.[36]

Unfortunately, the existence of an availability statement says nothing about its content. Often, it is stated the data is "available upon reasonable request". In one reasoned study with N=1792 statements about "93% authors either did not respond or declined to share their data" after a request.[37] As a comparison of preprints with their published versions has shown, that "*data availability*

---

[32]See for example Enwald et al. 2022, Feger et al. 2020, Gabelica, Bojčić, and Puljak 2022, Tedersoo et al. 2021, Thoegersen and Borlund 2022 and more. This list could be extended. However, it should be clear that such analyses are being carried out and are available.

[33]Examples of such studies are J. Borghi 2021, Quigley, Chan, and Clift 2022, Read et al. 2021, Gend and Zuiderwijk 2022 and more.

[34]https://quest-dashboard.charite.de. See also Iarkaeva et al. 2022.

[35]William Shakespeare (1564–1616), Merchant of Venice, act II, scene 7.

[36]Colavizza et al. 2020, p. 7.

[37]Gabelica, Bojčić, and Puljak 2022.

statements [...] are a good first step, but are insufficient to ensure data availability".[38] To be clear, there is a lot of frustration by getting research data with these statements of data availability that end up coming to nothing.

As already indicated, data sharing practices and the availability of data on request vary between scientific disciplines.[39] On the one hand, it is quite a challenge to work out the different disciplinary cultures in order to compare these different frames of reference. On the other hand, it is also unsatisfactory to postulate that everything is complicated without showing a higher level of detail. Ultimately, however, there are many different reasons why data sharing does not take place.[40] With a view to the Max Planck Society and the year 2020, we will take a closer look at some of these reasons in chapter 4.8.

2.4 Open Research Data Policies in 2020

Open sharing of research data is not the norm and, unlike the application of the FAIR principles, it is not always appropriate in all areas. At the same time, many European scientific institutions have policies that provide legal certainty for the free sharing of research data.[41] This issue is also increasingly being taken up by publishers, so that more and more data policies are being established.[42] Research organisations, such as the German Helmholtz Association and the French Institut Pasteur, have issued guidelines on the management of research data.[43] What is generally striking about these processes towards a data policy is that they require a lot of time and energy from all stakeholders. Internal coordination and the development of a common understanding are resource-intensive. Funders are well aware of this. The Swiss National Science Foundation, for example, evaluated the introduction of its Open Data Policy in 2020 and calculated the associated funding expenditure.[44]

For the Max Planck Society, it should be noted that there is no general research data policy for 2020. Individual institutes and departments have dealt with this issue in different ways. Based on the Harnack Principle, however, quite different solutions have emerged locally in the institutes and departments.[45] One example is the internal guideline on the handling of research data at the Max Planck Institute for the Study of Collective Goods.

---

[38]McGuinness and Sheppard 2021, p. 2.

[39]See a meta-study such as Tedersoo et al. 2021.

[40]See generally Gomes et al. 2022.

[41]But it should be clear that this is not a purely European phenomenon. Similar issues are being debated in other regions and countries. See for example the review of open research data policies in China Zhang et al. 2021.

[42]See for example the overview Hrynaszkiewicz et al. 2020. For a general overview of research data policies in journals in the 2010s and their evolution, see in particular Dearborn, Marks, and Trimble 2018.

[43]Helmholtz Open Science Office 2019 and Institut Pasteur 2021.

[44]Milzow et al. 2020, p. 4-5.

[45]Laitko 2015.

It was made mandatory for all employees in 2018. [46] It defines research data, regulates ownership, documents local data management and agrees on the implementation of the regulations. Another example of a local policy is the Biomaterials Department at the Max Planck Institute of Colloids and Interfaces.[47] This defines the procedures for handling samples in a departmental data policy.

At the same time, it should not be hidden that in many cases there are no explicit procedures, recommendations, etc. at the Max Planck Institutes. This can be seen as a sign that the existing workflows with research data are already carried out at a high level. At the same time, it should be noted that there is often an implicit procedure for research data. This does not always lead to a common approach to research data management and sometimes causes local problems.

2.5 Perspective of Third-Party Funders on Open Research Data

In addition to this intrinsic motivation, there is also the perspective of funders on data sharing. There is a clear tendency to make open access to research data a condition of grant approval. This applies not only to European, but to national funders. For the Max Planck Society, both the European Commission and the German Research Foundation were particularly important in 2020, each accounting for nearly a third of approved external funding.[48]
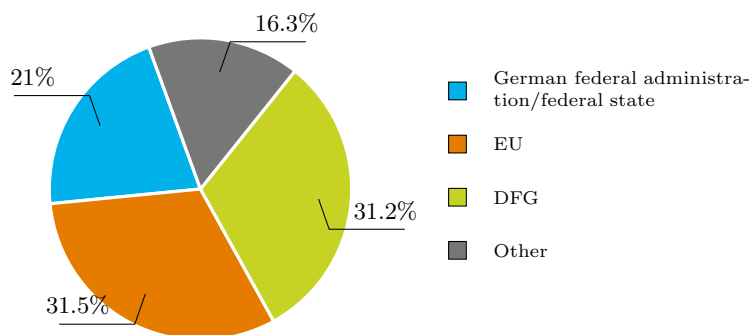

Figure 1: Distribution of third-party funds of the Max Planck Society in 2020

Research data has become increasingly important to the European Commission since the 2010s. Horizon 2020 explicitly includes research data and, in particular, data management plans.[49] At that time, a data

---

[46] Max Planck Institute for Research on Collective Goods 2018.

[47] Simon 2020.

[48] Max Planck Society 2021, p. 45

[49] For a good overview of these developments with static evaluations, see especially European Commission 2021, p. 49-63.

management plan (DMP) was not yet mandatory. At the same time, the availability of data management plans led to increased writing of such. In a highly competitive environment, it was no longer always possible to opt-out this element in relation to competitors. The already mentioned Open Research Data Pilot provided an opportunity for the Commission to promote the openness of research data.[50] In addition, the Commission has repeatedly tried to point to new developments. Only two years after the publication of the FAIR Data Principles, an attempt was made to put them into practice by releasing concrete financial resources, not just statements of intent.[51]

With the European Open Science Cloud (EOSC), the Commission tried early on to establish both a funded infrastructure for research data.[52] Open research data played a special role. On the part of the Max Planck Society, the MPCDF was particularly involved, especially at EOSCpilot with its own science demonstrators.

Compared to the European Commission, the DFG's developments in open research data up to 2020 are rather modest. This is primarily due to the self-governing nature of German science. Structurally, this is preceded by discussions within the discipline until, for example, a recommendation from a DFG commission is adopted. An early example of this is the 2015 Guidelines in the context of biodiversity research.[53] In retrospect, the emergence of the German National Research Data Infrastructure (NFDI) at the national level is probably the point at which research data management gained importance in Germany, and in the DFG in particular. The first call for NFDI consortia was launched in 2019. In 2020, the first consortia were named and the second call was published.[54] For discussions within the scientific community, the developments around the NFDI on Open Research Data can hardly be overestimated.[55]

---

[50]See for example European Research Council 2017.

[51]European Commission 2018b.

[52]For a good and concise summary of the history of the EOSC, see in particular Rat für Informationsinfrastrukturen 2023, p.13. For a perspective on the EOSC for Max Planck scientists see also Grossmann 2021c.

[53]Deutsche Forschungsgemeinschaft 2015.

[54]For statistical evaluations of these two rounds, see in particular Deutsche Forschungsgemeinschaft 2020a and Deutsche Forschungsgemeinschaft 2020b.

[55]For an evaluation of the first two NFDI calls with a focus on the participation of the Max Planck Society, see Grossmann 2021a.

3 Open Research Data in the Max Planck Society

Open research data has been of declared importance within the Max Planck Society since the beginning of the millennium. The Berlin Declaration of 2003 also refers specifically to data in its definition of Open Access contributions.[56] In 2020, however, the Max Planck Society did not have a general data policy or other recommendations for dealing with research data. At the same time, this does not mean that research data, and in particular open access to them, are not important to Max Planck scientists. An analysis by the Max Planck PhDnet Open Science Group, for example, highlights the potential of open research data for early career researchers.[57] Such discussions within the Max Planck Society are always embedded in the normative framework provided by the rules of good scientific practice.

3.1 Good Scientific Practices within the Max Planck Society

In 2000, the Senate of the Max Planck Society adopted new rules to ensure good scientific practice.[58] These were amended in 2009 and replaced by new rules in 2022 without widespread communication.[59] It is therefore important to bear in mind that the 2000 and 2009 regulations still applied in the year 2020. But it was already clear in 2020 that the regulations would change. This was initiated by the DFG and the adoption of the Guidelines for Safeguarding Good Research Practice, which became binding for applicant institutions such as the Max Planck Society.[60] The situation in 2020 was therefore characterised by an impending change in the normative framework for research data within the Max Planck Society.[61]

In the year 2020 Max Planck Rules for Safeguarding Good Scientific Practice, research data (in German "Forschungsdaten") were not yet referred to by this term. Rather, it was usually referred to as data or primary data. Nevertheless, the data had to be processed according to discipline-specific rules. Primary data had to be kept for ten years. It was also necessary to ensure that clear and comprehensible documentation was given, for example in laboratory notebooks. Access to the data had to be guaranteed for authorised interested parties.[62] However, there was already open research data. Various services for

---

[56]"*Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.*" Max Planck Society 2003.

[57]Toribio-Flórez et al. 2021, p. 3.

[58]Max Planck Society 2009.

[59]Max Planck Society 2022.

[60]Deutsche Forschungsgemeinschaft 2019.

[61]For an Max Planck internal perspective on this situation see especially Franke 2020.

[62]Max Planck Society 2009, p. 2 and p. 4.

open research data are already existing in 2020.

3.2   Max Planck Services for Open Research Data

The central infrastructure and service units of the Max Planck Society have been offering open research data services for some time. Some services for open research data were therefore already available within the Society in 2020. Three services are briefly presented here as examples to document an impression at this point in time.

In summer 2020, the MPCDF launched a CKAN instance for Pandora. This is a data repository for open archaeological data for the Max Planck Institute for the Science of Human History[63] and its collaborators. The repository has since supported the Pandora initiative to make historical and archaeological data more accessible and discoverable.

Since 2014, the Max Planck Digital Library offers Edmond, a repository for open research data. Here, Max Planck scientists and their collaborators could freely publish their data. A total of 84 datasets were published through this service in 2020.[64]

With GRO.data, the GWDG offers a service comparable to the Göttingen eResearch Alliance. Here, open research data can be made available via a repository, too. This service, as well as advice on open research data, was one of the services provided by the GWDG to Max Planck researchers in 2020.

With the World Data Climate Center, the German Climate Computing Centre (DKRZ) offers a community-specific data repository. Research data can also be published there as Open Research Data. The Max Planck Institute for Chemistry, for example, made use of this service in 2020 quite far.[65]

3.3   Max Planck Lighthouse Projects for Open Research Data

In 2020, the Max Planck Society had a number of Open Research Data projects that have developed quite successfully.

An illustrative example is the FACES[66] platform of the Max Planck Institute for Human Development. The website, developed in 2009 in collaboration with the Max Planck Digital Library (MPDL), offers a collection of high quality

---

[63]The former Max Planck Institute for the Science of Human History was renamed the Max Planck Institute for Geoanthropology in 2022.

[64]https://s.gwdg.de/G14Bgw.

[65]See also Wittenburg et al. 2019, p. 259-260.

[66]See https://faces.mpdl.mpg.de. For more information, see the Ebner, Riediger, and Lindenberger 2010.

images of human faces, grouped by age group, gender and a set of six different facial expressions (emotions).

Strictly speaking, this data is not entirely freely available: use of most of the data is restricted to scientific purposes and requires registration/login. However, this unique set of data has proven to be extremely fruitful, with a steady stream of diverse scientific publications over more than a decade.[67]

Another very successful project that has been running for a number of years is Movebank[68], which is run by the Max Planck Institute for Animal Behaviour together with various partners around the world. This platform collects and processes current and historical animal movement data from various sources around the world. As most of the data is freely available, it can be used in a variety of ways.

With a project start in 2007, Movebank has been around for a long time - with continuous growth and development of the platform. The number of annual publications that can be traced back to Movebank data has also remained high: many hundreds of publications in total.[69]

Max Planck's Fritz Haber Institute is heavily involved in the NOMAD Repository & Archive[70]. Maintained by the Novel Materials Discovery (NOMAD) Laboratory, it is the world's largest repository of input and output files from all major computational materials science programs. The repository contains data in raw format, while the archive provides normalised data in a common, machine-processable format.

After starting with the repository in 2014, the project has expanded and developed significantly in the following years as part of the Horizon 2020 European Center of Excellence, NOMAD CoE.[71]

---

[67]Scopus metrics show a steadily growing annual citation rate for FACES since 2010 until today, with a total of more than 700 citations.

[68]https://www.movebank.org.

[69]https://www.movebank.org/cms/movebank-content/literature.

[70]https://nomad-lab.eu/services/repo-arch.

[71]Wittenburg et al. 2019, p. 260-261.

4 Publications by the Max Planck Society in the Year 2020

The previous two chapters show that research data and open access were already widely discussed and in some places already implemented within the Max Planck Society in 2020. At this point in time, there were no general obligations to make research data publicly available. Therefore, this year can be used as a baseline from where to observe future developments. In order to put this understanding on a statistically sound basis, we present in detail our sample of publications by Max Planck scientists in 2020. The focus of this analysis is the handling of research data. It will demonstrate how and where the Max Planck scientists published their data in 2020.

4.1 Method

We manually analysed publications with at least one author from a Max Planck Institute. First, we assessed whether the publication was either empirical[72] or non-empirical, i.e. theoretical. Theoretical publications, which are often found in legal research, but also in mathematics or engineering, are not usually expected to contain research data. We then looked for a statement on data availability or similar. If data was stated to be available, we tried to access the data and categorised the result (all data, some data, no data) and the type of data obtained (raw data, analysed data, summarised data).[73] If the data availability statement said that the data were "available on request", we contacted the corresponding author (not necessarily from a Max Planck Institute) and asked for the data. We also checked whether the use of software was mentioned and whether this software was available.

4.2 Approach

The publication repository of the Max Planck Society MPG.PuRe contained 15,850 publication references from 2020 (the population, measured on 29/11/2022) that had a publication type for which research data could be expected (journal article, monograph, book, book chapter, edited volume, contribution to edited volume, conference paper, dissertation). Since December 2021, we took a pseudo-randomised[74] sample of 1,040 publications. We then tried to access the publications, preferably on the publisher's platform. Where this was not possible, we tried to obtain a post-print version from the repository, the author's institute, other libraries or through interlending (n=985). Finally, we looked for a preprint (n=12) or manuscript

---

[72]We defined an empirical paper as one that draws conclusions based on information from the real world. After a long discussion we decided not to consider literature as such information. For example, literature reviews are treated as non-empirical works

[73]i.e. only tables for figures or single values

[74]we created the MD5 hash of the publication identifier and sorted the list by this hash. Later updates were merged into the list

(n=20). There were 21 publications that we were not able to access despite thorough searches. We excluded these from the sample, which left us with a total N_{total}=1,019.

4.3 Descriptive Analysis

The vast majority of publications were journal articles (85,3%), other publication types followed far behind (see table 1).

Table 1: Publication Types.

	Frequency	Percent	in population
Journal article	869	85.3	83.3
Thesis	52	5.1	5.2
Book chapter	35	3.4	4.1
Conference paper	31	3.0	4.2
Book	13	1.3	0.9
Contribution to collected edition	13	1.3	1.5
Monograph	5	0.5	0.6
Collected edition	1	0.1	0.3
Total	1019	100.0	100.0

The Max Planck Society is divided into three sections: the Chemical, Physical and Technical Section (CPTS), the Biological and Medical Section (BMS), and the Humanities and Social Sciences Section (GSHS). Publications are distributed among the sections as shown in the table 2. Combinations of sections reflect collaborations between institutes from different sections.

Table 2: MPG sections

	Frequency	Percent	in population
CPTS	555	54.7	55.6
GSHS	245	23.9	23.4
BMS	214	20.0	19.4
CPTS,BMS	2	1.1	1.4
Central	2	0.2	0.1
GSHS,BMS	1	0.1	0.0
Total	1019	100.0	100.0

In 2020, Max Planck scientists published their research results in many different places. The absolute numbers of all publications by Max Planck scientists certainly show similarities to the national figures provided by the ESAC Market Watch for Germany.[75] For the publications selected in this study, this means that the top three publishers were Springer Nature, Elsevier and Wiley. As figure 2 shows, in 2020, one fifth of the selected empirical

---

[75] For more details and the code see mainly Jahn and Hobert 2019.

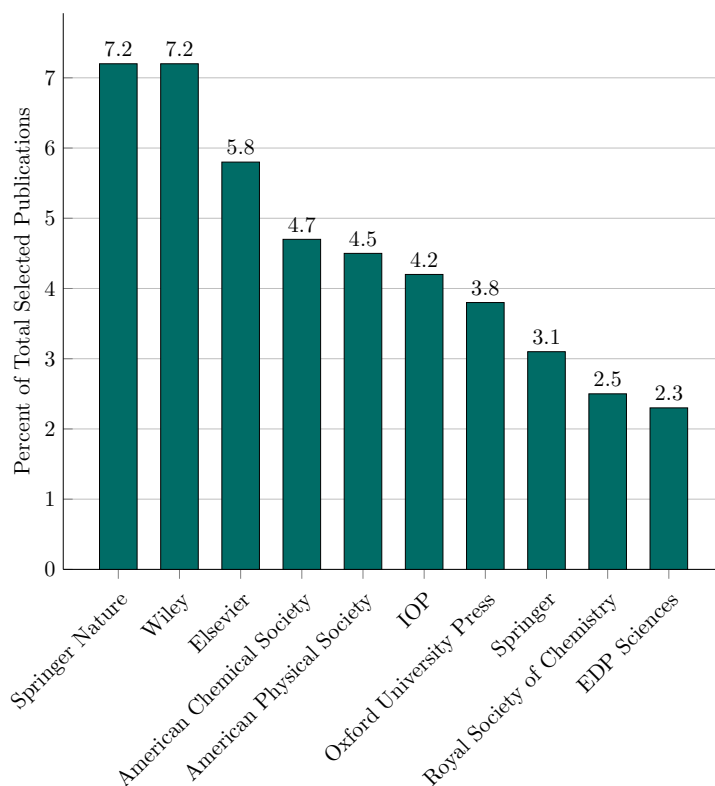publications were published at the top three.

Figure 2: Ten Most Frequent Publisher of the Empirical Selected Publications

4.4 Selected Publications

We analysed a total of 1,019 data publications in more detail. The CPTS accounted for slightly more than half of this sample, see figure 3. This result is to be expected due to the relatively large number of institutes and staff in this section. In addition, about a quarter and a fifth of the publications came from the BMS and GSHS, respectively. Collaborations involving Max Planck scientists from several sections account for only 1.4% of the total number of publications.
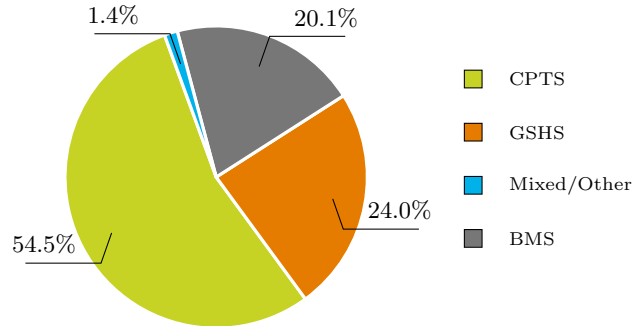
Figure 3: Distribution of Publications by Max Planck Sections

An explicit look at the GSH section with its 244 publications is particularly remarkable in this distribution. Here, comparatively different disciplinary cultures are linked. The more detailed distribution in figure 4 shows the distribution of research fields. The majority of GSHS publications in 2020 came from the human sciences. This is followed at a distance by publications from the fields of law and the humanities.[76] Publications in the social sciences have the smallest share in the sample.
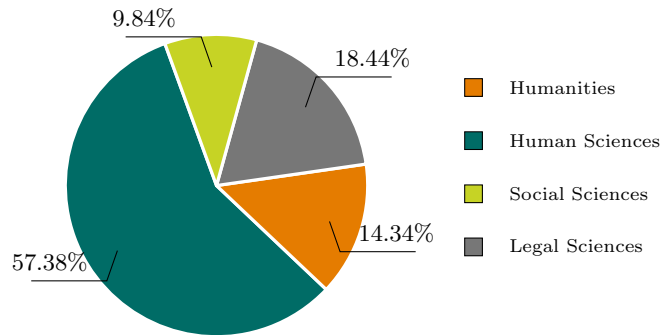

Figure 4: Distribution of Publications within the GSH Section

The core of a bibliometric analysis of research data is mentioned in textual publications. However, it is important to remember that theoretical work does not necessarily have to include or be based on research data. Behaviours and characteristics can be postulated on the basis of theoretical assumptions and deductions. These would only have to be supported or refuted by empirical data, for example, in a second step. For this reason, to gain this insight, we

---

[76]Our internal breakdown of the GSH section can be found in the corresponding data publication, see 8.1.

inquired to what extent each publication could be considered theoretical. This classification is important in the following, as we only analyse the handling of research data in 2020 in the case of non-theoretical publications. This is because we limit ourselves to dealing with the management of research data. The inclusion of theoretical works, which are not supposed to have research data, would put the results of the analysis into an additional negative context.

According to this classification shown in figure 5, about 30% of the randomly selected papers are non-empirical. In contrast $N_{empirical} = 708$ are empirical publications.
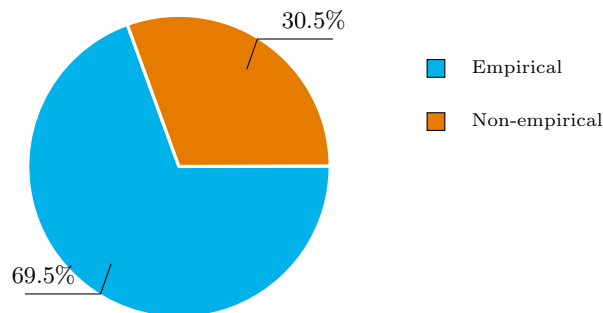

Figure 5: Classification According to Empirical and Non-Empirical

In relation to the scientific domains described above, the highest ratio of empirical publications are in biological/medical research (82%) whereas in humanities, no empirical works appear in this sample (see figure 6). We excluded theoretical work from the sample.
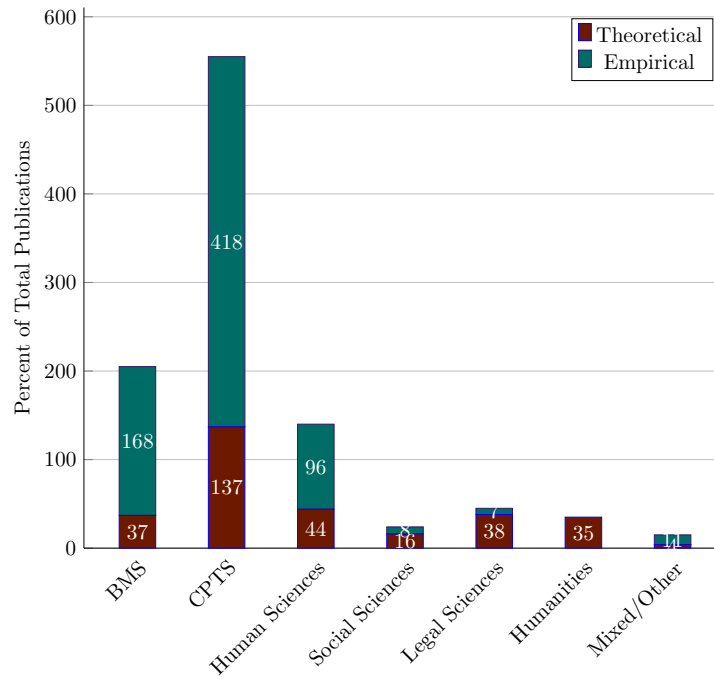
Figure 6: Empirical and theoretical publications in different domains

4.5 Results

Figure 7 shows our sample, where 40% of the empirical publications provide research data. Conversely, this means that research data is not available in 60% of the publications with empirical focus.
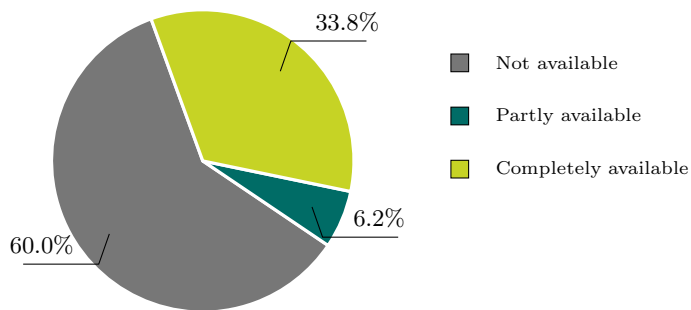

Figure 7: Research Data Availability within Selected Empirical Publications

However, the claimed 40% of available research data $N_{empirical}$ is even more limited. As figure 7 visualises, open research data is actually completely available for only 137 publications. This is partially the case for a further 43 (6,2%) publications. For a total of 84 publications it is stated that the research data will be made available upon "reasonable request". In theory, these publications should be accessible, but in reality this is not always the case. This phenomenon will be discussed in more detail in the following chapter 4.8. To anticipate this, in our case only a fraction of the research data was actually made available; as shown in figure 8.
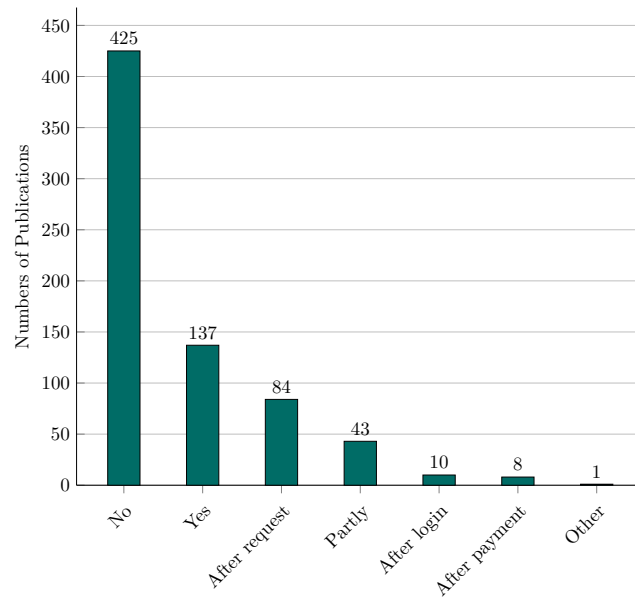

Figure 8: Types of Availability.

The handling of research data in theses, especially doctoral theses, is striking: Only 18.2% of the 44 theses have (some) research data, whereas 81.8% have none.

Table 3: Available data per publication type with $n \geq 5$

	partly(%)	complete(%)	sum(%)	n
Journal article	6,2	36,0	42,3	641
Conference paper	0,0	23,5	23,5	17
Thesis	6,8	11,4	18,2	44
Total	6,1	33,9	40,0	702

At the same time, a closer look at the aggregation level of the data in

empirical publications shows that 31.5% of the research data is in its raw state, see figure 9. In contrast, only 11.7% of research data consists of a series of individual values. Most of the publications (56.8%) offer research data in an analysed form.
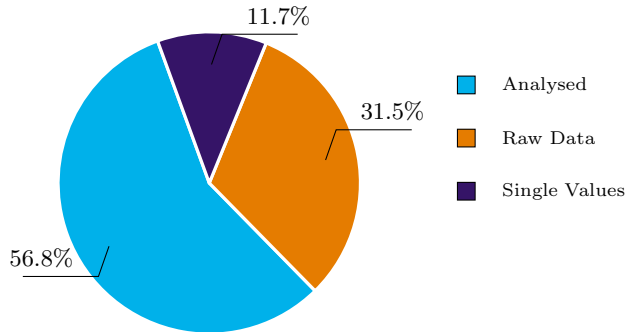

Figure 9: Data aggregation level

The availability of data also varies widely among the publishers that released the works, shown in table 4. Data is mostly available in high-impact journals like Cell and Nature and in the open-access journals of Frontiers and Copernicus. At the lower end of the scale are the learned societies with the exception of the American Geophysical Union where data is available for every publication. However, data policies can be an important point for clarifying these differences. Publishers have very different policies on data availability and reproducibility. As shown by Canadian colleagues, data policies have become increasingly common in journals since the mid-2010s.[77] It seems quite plausible that there is a causal relationship between data availability and the existence of a publisher's data policy.

---

[77]Dearborn, Marks, and Trimble 2018, pp. 381-382.

Table 4: Available data per publisher with $n \geq 5$

	partly(%)	complete(%)	sum(%)	n
American Geophysical Union	0.0	100.0	100.0	8
Cell Press	20.0	73.3	93.3	15
Frontiers	0.0	90.9	90.9	11
Copernicus Publications	16.7	66.7	83.3	12
Nature	9.8	68.6	78.4	51
Academic Press	25.0	50.0	75.0	8
BioMed Central	0.0	71.4	71.4	7
Oxford University Press	7.4	51.9	59.3	27
AAAS	0.0	50.0	50.0	10
American Institute of Physics	20.0	20.0	40.0	5
MDPI	20.0	20.0	40.0	10
Wiley	7.8	29.4	37.3	51
Springer	4.5	31.8	36.4	22
Cambridge University Press	0.0	33.3	33.3	6
EDP Sciences	6.3	25.0	31.3	16
Pergamon	0.0	28.6	28.6	7
Royal Society of Chemistry	0.0	22.2	22.2	18
Elsevier	0.0	22.0	22.0	41
American Chemical Society	9.1	6.1	15.2	33
IOP	3.3	10.0	13.3	30
American Physical Society	0.0	3.1	3.1	32
Total	6.4	36.2	42.6	420

Similar to the different publishers, there are differences in data availability between the sections of the Max Planck Society. In terms of data availability within the three sections, an average of 33.8% ($N_{\text{Empirical with data}}$=239) of empirical publications are linked to accessible data. An average of 6.2% ($N_{\text{Empirical with partly data}}$=44) of the publications in the sections have partial data. The figure 10 shows the statistical variation between the sections. The data availability within the CPTS is slightly below average, while the BMS and the GSHS are slightly above average. Combining this with the values for partial data availability, we see that within the BM section almost every second publication contains research data. In the CPT section, however, this is the case for every third publication.

At the same time, it is important to have a look at the similarities in these figures. The difference between the BMS and the CPTS is only 15%. As a result it can be argued that data availability within the Max Planck Society differs only slightly between the individual sections. No section outperforms another by a multiple in terms of percentage data availability. A long-term perspective is particularly interesting for these figures. Here it might be possible to observe whether different developments and internal discussions regarding the availability of research data in publications take place in different subject groups or times.
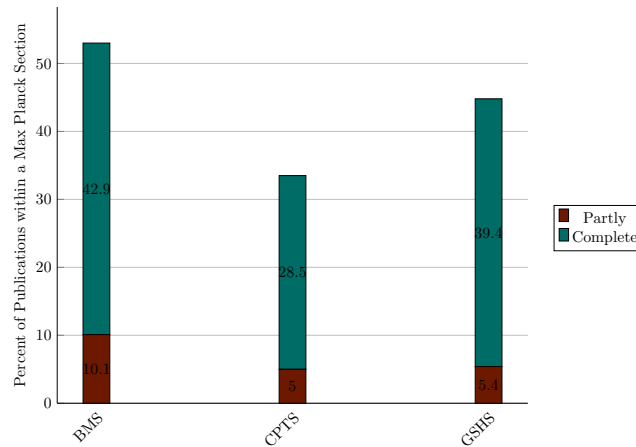
Figure 10: Data Availability within the Max Planck Sections of the Empirical Selected Publications

4.6 Data Licenses

The FAIR principles mentioned in chapter 2.2 have increasingly become a quasi-standard. The key idea behind these principles is to make the descriptive aspects of data explicit. This concerns, for example, a statement about the possibility of using the data, which is made explicit by convention through a licence. It is therefore an interesting aspect to ask about the use of licences within the here presented Max Planck sample. This provides a better perspective on the concrete application of the FAIR principles, independent of the often communicated relevance of the principles.

If we ask this question of the 239 available publications whose research data we were able to access, the result is somewhat surprising. Three quarters of the data do not have their own licence. It is therefore unclear how and under what conditions the data could be reused. In contrast, only a quarter of the data had an explicit licence. Of these, the CC BY 4.0 licence was clearly the most widely used, at 15%. CC0 is another licence worth mentioning, with two percent. The remaining, many licences are lost in a kind of background noise at 6.8%.

4.7 Data Repositories

As with licences, looking at the used data repositories in our sample reveals a behaviour in relation to data availability. Here, it is particularly interesting to see which category of repository is represented. Figure 11 shows the distribution of publications by repository: the first value is striking. However, it should be noted that the use of no repository rather indicates that the data was for example attached directly to the publication. Data papers as a genre

are also possible. So there are other ways of making data available than using data repositories.

The largest number of Max Planck researchers (4.8%) of the publications sample have made their data available via the National Center for Biotechnology Information. This US data repository specialises in the storage of molecular biology data. It has become the quasi-standard repository for data in this field of research. Despite its institutional reference, it can be classified as a subject specific repository due to its focus on specific topics. In addition, ENA, PRIDE and Allele have been used as other subject-specific repositories for the availability of research data. They all have a focus on bioinformatics. In this research environment it can be observed that a research culture for the use of repositories already exists. Some of the repositories are even internally differentiated, such as proteomics data in PRIDE. In addition, SIMBAD stands out as a subject-specific repository for astronomical research data. After all, almost 2.0% of all research data in our example was published there.

In addition to subject-specific repositories, the use of generic data repositories by Max Planck scientists can be observed in 2020. Figure 11 shows that they were used frequently in our sample. It is remarkable that `https://github.com`, a primary software repository, was used as a data repository in 4.1% of cases. The institutional data repository of the Max Planck Society, Edmond, had been used (1.7%).
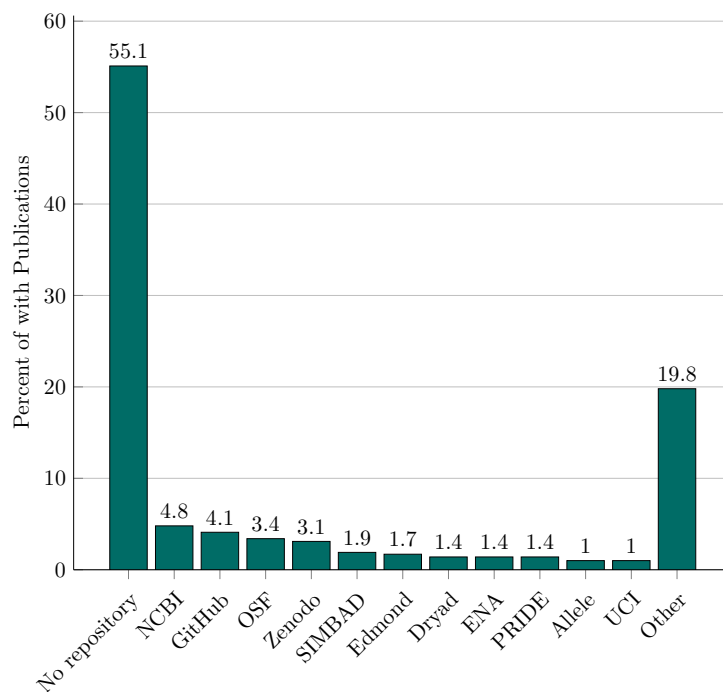
Figure 11: Data Repositories at the Available Research Data

4.8 Data "Available" Upon Request

Data is not always readily available. As discussed in chapter 2.3, this phenomenon is associated with some problems. In our sample, this is exemplified by the 84 publications for which data are described as available upon (reasonable) request. It was possible to obtain data on request from about 20% of the publications. Specifically, about 8% of the corresponding authors could not be contacted because their 2020 email address no longer worked and about half of the authors did not answer at all. At the same time it was also observed that, despite justified requests, the data was not available. Figure 12 visualises this.

In fact, it is alarming that more than three-quarters of the research data claimed to be available are in fact unavailable. With 84 publications related to this, the data base is comparatively small. Nevertheless, it can be said that the enquiries and correspondence leading up to the final result were lengthy, time-consuming and often frustrating. Similar to other studies described above, our experience shows that obtaining data usually involves a long email correspondence with the authors.[78]
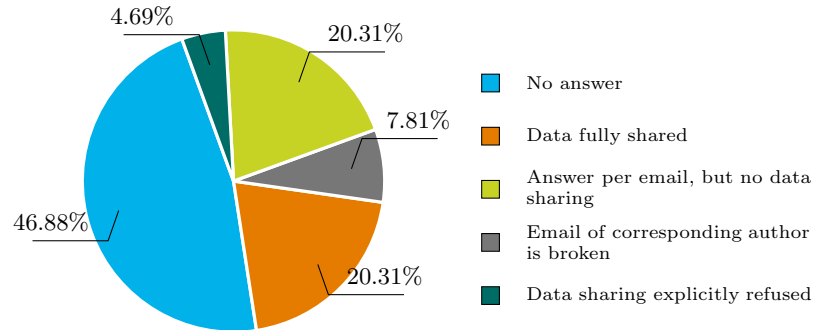
---

[78]Tedersoo et al. 2021, p. 8.

Figure 12: Distribution of responses to a data request

4.9 Research Software in Publications

Software is often needed to reproduce the analysis of research data and to be able to follow the path of scientific knowledge. Research software was mentioned in some form in just over 41% of the selected publications (see figure 13). Research software was only marginally the focus of this study. Nevertheless, it is remarkable how clear the treatment of the accessibility in comparison to research data is.
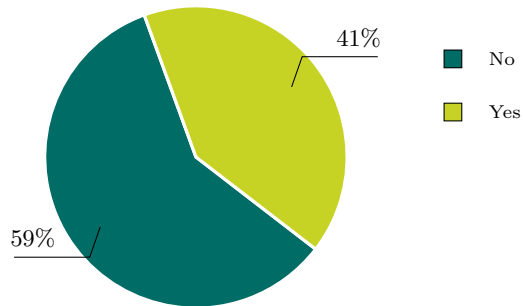

Figure 13: Mentioned Software in the Selected Publications

Of particular note is the more explicit use of research software in publications compared to research data. Research software is mentioned in 288 of the N_{total} = 708 selected empirical publications. And of these 288 software packages, 226 – i.e. 78% – are openly accessible. This is a significant difference compared to the figures from openly available research data.[79] This is in line with other

---

[79] However, it should not be concealed that even with research software not all elements are

research on the open availability of research software. For example, in 2021, research software on `https://github.com` was already mentioned in over 20% of all publications on `https://arxiv.org/`.[80]

5 Discussion

"*There is nothing either good or bad, but thinking makes it so*."[81] To cite Shakespeare, how we value open research data depends on our own point of view and context. The selection of publications from 2020 presented here can therefore also be discussed from different perspectives on the part of the Max Planck Society. We have identified five key aspects from the analysis of publications in 2020 and its relation to data availability.

1. Expectations and actual results differ significantly by **type of data availability**. Particularly in empirical work one would expect the research data to be available in some way, whether as open research data or with restricted access. Since in terms of good scientific practice, it should be possible to reproduce the results. Considering this statement, only 40% of research data available or partially available for empirical publications is not much. Theoretically, one could expect 100% availability. In all previous studies, however, the found numbers were at a very low level.[82] Open or restricted access – for which there are good reasons – would be irrelevant for the time being. Reproducibility would mainly be guaranteed in both cases. However, almost 60% of empirical work is without data. For an excellent research organisation like the Max Planck Society, which sees itself as one of the leading scientific organisations operating in the field of basic research, there is still potential for a greater data availability.

2. The **aggregation** of the available data is predominantly analysed. But there is also raw data. There is a discussion about laboratory data and how far it can or should be published.[83] There is therefore no definite answer to the question of when data should be published. This is probably a case-by-case decision. For the Max Planck Society in its diversity, it does not seem sensible to formulate a general rule for the degree of aggregation.

3. A **data policy** can increase the availability of data associated with textual publications. Such normative requirements on the part of publishers lead and will increasingly lead to available data in the near future. It can be assumed that publishers, funders and scientific organisations will increasingly develop such normative frameworks for scientists. However, this is not the same as

---

always accessible and sharing with others is refused here. Overall, however, this is on a much lower level. See also the observations from Assel and Vickers 2018, p. 832

[80]Escamilla et al. 2022.

[81]William Shakespeare (1564–1616), Hamlet, act II, scene 2.

[82]See chapter 2.2.

[83]See for example Pinel, Prainsack, and McKevitt 2020, p. 192.

demanding open access to research data. There may be good reasons for publishing data with restricted access, for example in clinical trials or industrial contract work.

"*Nothing can come of nothing.*"[84] Greater visibility of their research results and the application of Open Science can, for example, be motivation for a data policy. Such a normative framework is indispensable if the Max Planck Society or individual institutes want to motivate their own scientists in this direction. Recent developments at the Helmholtz Association in particular show how such a path could be taken in German basic research.[85]

4. The concept of **data availability statements** like "Data available on reasonable request" do not work as expected. Response rates are low. The communication effort is usually high. Both our experience and other studies have shown that there is a mismatch between effort and return. Storing the research data in a data repository – with open or restricted access – would eliminate this problem.

For the Max Planck Society, this may lead to a kind of recommendation that research data should be published in suitable infrastructures. These could be research data repositories, data journals or similar solutions.

5. The data handling culture of the **sections of the Max Planck Society** do differ from each other. At the same time, the differences in the publication of research data are not as great as one might expect. This suggests that the Max Planck Society has already taken aspirations towards a data sharing culture. For example, the new rules on good scientific practice mention research data and the explicit handling of them quite often.[86] However, if we compare the Max Planck Society with, for example, the Helmholtz Association and its use of research software, or the Charité and its open research applications, there is still a lot of potential that can be employed within the Max Planck Society.[87]

---

[84]William Shakespeare (1564–1616), King Lear, act I, scene 1.

[85]See for example Helmholtz Open Science Office 2022 and Ferguson et al. 2021.

[86]A brief overview of the research data aspect of the new rules can be found in Grossmann 2021b.

[87]Helmholtz-Gemeinschaft 2019, Iarkaeva et al. 2022 and Nachev, Taubitz, and Riedel 2023.

6 Perspectives

Analysing the publications by Max Planck researchers in 2020 is the past. What has happened since then?

Open Science and Open Research Data are becoming increasingly relevant in the German research landscape. This is clearly indicated, for example, by the positioning of the DFG in autumn 2022.[88] There are many advantages to sharing data and code for a culture of science.[89] With a focus on Germany, the NFDI will play an increasing role in this. One of its first successes is that research data and its management have not only reached the scientific community but German decision-makers, e.g. in politics. It is therefore not a far-fetched thesis to assume that research data will become an increasingly important topic in the coming years. The debate has already lost some of its drama. It is now less a question of "why" and more a question of "how".

All this manifests itself in the fact that the topic of research data is increasingly being dealt with locally at the Max Planck Institutes. The new guidelines on good scientific practice from the end of 2022 also made a significant contribution to this.[90] There are local working groups, initiatives and committed colleagues looking for local solutions in their departments and institutes. At the same time, a Max Planck-wide network of RDM experts is evolving. Events such as the regular RDM workshops of the Max Planck Digital Library are evidence of this.[91] However, it remains to be seen whether the developments at the individual institutes will be merged. It also remains to be seen whether concepts such as "data stewardship" will become established in individual disciplines.

There is already a clear need for more knowledge about research data within one's own institution. Services such as the Charité Dashboard on Responsible Research mentioned above, the HMC Dashboard on Open and FAIR Data in Helmholtz or the French Open Science Monitor provide a first glimpse of what such bibliometric services might look like. The transition to evaluation methods and research assessment is, of course, imminent. Since the Max Planck Society has a real interest in such metrics, it would certainly be in everyone's interest to be able to offer such services (internally) in the long term. Such dashboards with a focus on research data can, for example, also open up longitudinal perspectives on how Max Planck scientists deal with research data. Nevertheless, it is clear that such analyses of data availability as we have presented here should be carried out more frequently. It will be interesting to see how publishing behaviour changes as a result of the developments mentioned. In the end,

---

[88] Deutsche Forschungsgemeinschaft 2022, pp. 4-5.
[89] See also the detailed discussion of the advantages and disadvantages in Gomes et al. 2022.
[90] Max Planck Society 2022.
[91] See `https://rdm.mpdl.mpg.de/mpdl-services/workshops/`.

however, it is said, "[b]*e great in act, as you have been in thought.*"[92]

---

[92]William Shakespeare (1564–1616), King John, act V, scene 1.

7 Abbreviations

- BMC = BioMed Central
- BMS = Biological and Medicine Section of the Max Planck Society
- CPTS = Chemistry, Physics and Technology Section of the Max Planck Society
- DFG = Deutsche Forschungsgemeinschaft
- DMP= Data Management Plan
- DOI = Digital Object Identifier
- ENA = European Nucleotide Archive
- EOSC = European Open Science Cloud
- ESAC = Efficiencies and Standards for Article Charges
- FAIR = Acronym of findability, accessibility, interoperability, and reusability regarding data principles, see Wilkinson et al. 2016
- GSHS = Human Sciences Section of the Max Planck Society
- GWK = Joint Science Conference
- HRK = Hochschulrektorenkonferenz (German Rectors' Conference)
- MPCDF = Max Planck Computing Data Facility
- MPDL = Max Planck Digital Library
- MPI = Max Planck Institute
- MPS = Max Planck Society
- NCBI = National Center for Biotechnology Information
- NFDI = National Research Data Infrastructure
- OECD = Organization for Economic Co-operation and Development
- ORDP = Open Research Data Pilot
- CC = Creative Commons
- PLOS = Public Library of Science
- PRIDE = Proteomics Identifications Database
- SIMBAD = Set of Identifications, Measurements and Bibliography for Astronomical Data
- URL = Uniform Resource Locator
- WR = German Science and Humanities Council (Wissenschaftsrat)

8 Statements and Comments

8.1 Data and Code Availability Statement

The data and code are freely available in Edmond via
`https://doi.org/10.17617/3.XI0LP5`. Via Edmond the scripts can also run
directly in Mybinder using
`https://mybinder.org/v2/dataverse/10.17617/3.XI0LP5`.

8.2 Author contributions

The study design was set up by Franke and Grossmann. The drawing of the
sample was programmed by Franke. Ho and Matthiesen were mainly
responsible for data collection and research. Ho and Matthiesen were also in
charge of the graphical representations of the results. Ho and Franke have
evaluated and analysed the data collected. The texts for the open data
examples were written by Boosen. Other text elements including the
bibliography were written by Grossmann. Corrections and linguistic
adjustments were done by Leiminger.

References

Allianz der deutschen Wissenschaftsorganisationen (2010). "Principles for the Handling of Research Data". In: DOI: 10.2312/ALLIANZOA.035.

Assel, Melissa and Andrew J. Vickers (2018). "Statistical code for clinical research papers in a high-impact specialist medical journal". In: *Annals of internal medicine* 168.11, pp. 832–833. DOI: 10.7326/M17-2863.

Ausschuss für wissenschaftliche Bibliotheken und Informationssysteme and Unterausschuss für Informationsmanagement, eds. (2009). *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten.* URL: https://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf (visited on 08/22/2023).

Bloom, Theodora, Emma Ganley, and Margaret Winker (2014). "Data Access for the Open Access Literature: PLOS's Data Policy". In: *PLOS Biology* 12.2, e1001797. DOI: 10.1371/journal.pbio.1001797.

Borghi, John (2021). *Identifying the who, what, and (sometimes) where of research data sharing at an academic institution.* DOI: 10.5281/zenodo.4737831.

Borghi, John A. and Ana E. van Gulick (2021). "Data management and sharing: Practices and perceptions of psychology researchers". In: *PLOS ONE* 16.5, e0252047. DOI: 10.1371/journal.pone.0252047.

Colavizza, Giovanni et al. (2020). "The citation advantage of linking publications to research data". In: *PLOS ONE* 15.4, e0230416. DOI: 10.1371/journal.pone.0230416.

Crüwell, Sophia et al. (2022). *Investigating the Effectiveness of the Open Data Badge Policy at Psychological Science Through Computational Reproducibility.* DOI: 10.31234/osf.io/729qt.

Dearborn, Dylanne, Steve Marks, and Leanne Trimble (2018). "The Changing Influence of Journal Data Sharing Policies on Local RDM Practices". In: *International Journal of Digital Curation* 12.2, pp. 376–389. DOI: 10.2218/ijdc.v12i2.583.

Deutsche Forschungsgemeinschaft (2015). *Richtlinien zum Umgang mit Forschungsdaten in der Biodiversitätsforschung.* URL: https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_fors (visited on 05/23/2023).

– (2019). *Code of Conduct.* URL: https://wissenschaftliche-integritaet.de/en/code-of-conduct/ (visited on 08/22/2023).

– (2020a). *Nationale Forschungsdateninfrastruktur Statistische Übersicht zu den Förderentscheidungen in der ersten Ausschreibungsrunde.* URL: https://www.dfg.de/download/pdf/foerderung/programme/nfdi/20200626_nfdi_foerderentscheidur (visited on 08/22/2023).

– (2020b). *Nationale Forschungsdateninfrastruktur Statistische Übersichten zum Antragseingang (Zweite Ausschreibungsrunde, September 2020).* URL:

https://www.dfg.de/download/pdf/foerderung/programme/nfdi/nfdi_auswertung_2020.pdf (visited on 08/22/2023).

Deutsche Forschungsgemeinschaft (2022). *Open Science as Part of Research Culture. Positioning of the German Research Foundation.* DOI: 10.5281/zenodo.7194537.

Donner, Eva Katharina (2022). "Research data management systems and the organization of universities and research institutes: A systematic literature review". In: *Journal of Librarianship and Information Science*, p. 096100062110702. DOI: 10.1177/09610006211070282.

Ebner, Natalie C., Michaela Riediger, and Ulman Lindenberger (2010). "FACES - A database of facial expressions in young, middle-aged, and older women and men: Development and validation". In: *Behavior Research Methods* 42.1, pp. 351–362. DOI: 10.3758/BRM.42.1.351.

Enwald, Heidi et al. (2022). "Data sharing practices in open access mode: a study of the willingness to share data in different disciplines". In: *Information Research: An international electronic journal* 27.2. DOI: 10.47989/irpaper932.

Escamilla, Emily et al. (2022). *The Rise of GitHub in Scholarly Publications.* DOI: 10.48550/ARXIV.2208.04895.

European Commission (2016). *H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020.* URL: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020- (visited on 08/22/2023).

– (2018a). *Open Research Data (ORD) - the uptake in Horizon 2020 - Data Europa EU.* URL: https://data.europa.eu/data/datasets/open-research-data-the-uptake-of-the-pilot-in-the-fi (visited on 08/22/2023).

– (2018b). *Turning FAIR into reality - Final Report and Action Plan from the European Commission Expert Group on FAIR Data.* DOI: 10.2777/1524.

– (2019). *Cost-benefit analysis for FAIR research data - Cost of not having FAIR research data.* DOI: doi/10.2777/02999.

– (2021). *Monitoring the open access policy of Horizon 2020: final report.* LU. DOI: 10.2777/268348.

European Research Council (2017). *Guidelines on: Implementation of Open Access to Scientific Publications and Research Data: In projects supported by the European Research Council under Horizon 2020.* Guideline Version 1.1. URL: https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-

Feger, Sebastian S. et al. (2020). "'Yes, I comply!': Motivations and Practices around Research Data Management and Reuse across Scientific Fields". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, 141:1–141:26. DOI: 10.1145/3415212.

Ferguson, Lea Maria et al. (2021). *Indikatoren für Open Science: Report des Helmholtz Open Science Forum.* DOI: 10.48440/os.helmholtz.024.

Franke, Michael (2020). *Forschungsdaten-Policies in der MPG*. URL:
https://info-bib.mpg.de/wp-content/uploads/2020/09/2020.Franke.pdf
(visited on 08/22/2023).

Franke, Michael et al. (2015). "Positionspapier "Research data at your
fingertips" der Arbeitsgruppe Forschungsdaten". In: DOI:
10.2312/allianzfd.001.

Gabelica, Mirko, Ružica Bojčić, and Livia Puljak (2022). "Many researchers
were not compliant with their published data sharing statement: a
mixed-methods study". In: *Journal of Clinical Epidemiology* 150, pp. 33–41.
DOI: 10.1016/j.jclinepi.2022.05.019.

Gend, Thijmen van and Anneke Zuiderwijk (2022). "Open research data: A
case study into institutional and infrastructural arrangements to stimulate
open research data sharing and reuse". In: *Journal of Librarianship and
Information Science*, p. 09610006221101200. DOI:
10.1177/09610006221101200.

Gomes, Dylan G. E. et al. (2022). "Why don't we share data and code?
Perceived barriers and benefits to public archiving practices". In:
Proceedings of the Royal Society B: Biological Sciences 289.1987,
p. 20221113. DOI: 10.1098/rspb.2022.1113.

Goodman, Alyssa et al. (2014). "Ten Simple Rules for the Care and Feeding of
Scientific Data". en. In: *PLOS Computational Biology* 10.4, e1003542. DOI:
10.1371/journal.pcbi.1003542.

Grossmann, Yves Vincent (2021a). "Participation by Max Planck Members at
the National Data Infrastructure". In: *RDM Information Platform for Max
Planck Researcher*. URL:
https://rdm.mpdl.mpg.de/2021/07/16/participation-by-max-planck-members-at-the-national-da

– (2021b). "Research Data within the New Rules of Conduct for Good
Scientific Practice by the Max Planck Society". In: *RDM Information
Platform for Max Planck Researcher*. URL:
https://rdm.mpdl.mpg.de/2022/12/12/rules-of-conduct-for-good-scientific-practice-by-the-m

– (2021c). "The European Open Science Cloud: A Very Short Summery". In:
RDM Information Platform for Max Planck Researcher. URL:
https://rdm.mpdl.mpg.de/2021/10/04/the-european-open-science-cloud-a-very-short-summery/.

Helmholtz Open Science Office (2019). *Empfehlungen für Richtlinien der
Helmholtz-Zentren zum Umgang mit Forschungsdaten*. DOI:
10.2312/os.helmholtz.002.

– (2022). *Helmholtz Open Science Policy*. URL:
https://os.helmholtz.de/open-science-in-helmholtz/open-science-policy/
(visited on 08/22/2023).

Helmholtz-Gemeinschaft (2019). *Empfehlungen zur Implementierung von Leit-
und Richtlinien zum Umgang mit Forschungssoftware an den
Helmholtz-Zentren. Positionspapier*. DOI: 10.2312/OS.HELMHOLTZ.008.

Hochschulrektorenkonferenz, ed. (2014). *Management von Forschungsdaten -
eine zentrale strategische Herausforderung für Hochschulleitungen*. URL:
https://www.hrk.de/positionen/beschluss/detail/management-von-forschungsdaten-eine-zentra
(visited on 08/22/2023).

Hood, Amelia S. C. and William J. Sutherland (2021). "The data-index: An author-level metric that values impactful data and incentivizes data sharing". In: *Ecology and Evolution* 11.21, pp. 14344–14350. DOI: `10.1002/ece3.8126`.

Houtkoop, Bobby Lee et al. (2018). "Data Sharing in Psychology: A Survey on Barriers and Preconditions:" in: *Advances in Methods and Practices in Psychological Science* 1.1, pp. 70–85. DOI: `10.1177/2515245917751886`.

Hrynaszkiewicz, Iain et al. (2020). "Developing a Research Data Policy Framework for All Journals and Publishers". In: *Data Science Journal* 19.1, p. 5. DOI: `10.5334/dsj-2020-005`.

Hutson, Matthew (2022). "Taking the pain out of data sharing". In: *Nature* 610.7930. DOI: `10.1038/d41586-022-03133-5`.

Iarkaeva, Anastasiia et al. (2022). *Semi-automated extraction of information on open datasets mentioned in articles*. preprint. DOI: `10.17504/protocols.io.q26g74p39gwz/v1`.

Institut Pasteur (2021). *Politique de gestion et partage des données de la recherche et codes logiciels*. DOI: `10.25490/A97F-EGYK`.

Jahn, Najko and Anne Hobert (2019). *subugoe/oa2020cadata: Publisher-level data for internal review*. DOI: `10.5281/zenodo.3519004`.

Jeng, Wei and Daqing He (2022). "Surveying research data-sharing practices in US social sciences: a knowledge infrastructure-inspired conceptual framework". In: *Online Information Review*. DOI: `10.1108/OIR-03-2020-0079`.

Kommission Zukunft der Informationsinfrastruktur (2011). "Gesamtkonzept für die Informationsinfrastruktur in Deutschland – Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder". In: p. 254. URL: `https://www.hof.uni-halle.de/web/dateien/KII_Gesamtkonzept_2011.pdf` (visited on 05/23/2023).

Laitko, Hubert (2015). "Das Harnack-Prinzip als institutionelles Markenzeichen: Faktisches und Symbolisches". In: *"Dem Anwenden muss das Erkennen vorausgehen": Auf dem Weg zu einer Geschichte der Kaiser-Wilhelm-/Max-Planck-Gesellschaft*. MPRL – Proceedings. DOI: `10.34663/9783945561010-04`.

Leonelli, Sabina (2017). "Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology". In: *Data Science Journal* 16.0, p. 32. DOI: `10.5334/dsj-2017-032`.

Linne, Monika et al. (2021). "GO FAIR und GO CHANGE: Chancen für das deutsche Wissenschaftssystem". In: *Praxishandbuch Forschungsdatenmanagement*. Ed. by Markus Putnings, Heike Neuroth, and Janna Neumann, pp. 215–238. DOI: `10.1515/9783110657807-013`.

Mandeville, Caitlin P et al. (2021). "Open Data Practices among Users of Primary Biodiversity Data". In: *BioScience* 71.11, pp. 1128–1147. DOI: `10.1093/biosci/biab072`.

Max Planck Institute for Research on Collective Goods (2018). *FDM-Policy*. The policy is only available internally at the Max Planck Institute for Research on Collective Goods. It can be requested from the local library.

Max Planck Society (2003). *Berliner Erklärung*. URL: https://openaccess.mpg.de/Berliner-Erklaerung (visited on 08/22/2023).

– (2009). *Regeln zur Sicherung guter wissenschaftlicher Praxis*. URL: https://web.archive.org/web/20220531145138/https://www.mpg.de/199493/regelnWissPraxis.pdf (visited on 08/22/2023).

– (2021). *Jahresbericht der Max Planck Gesellschaft 2020*. URL: https://www.mpg.de/17035587/jahresbericht-2020.pdf (visited on 08/22/2023).

– (2022). *Verhaltensregeln für gute wissenschaftliche Praxis – Umgang mit wissenschaftlichem Fehlverhalten*. URL: https://www.mpg.de/199493/regelnWissPraxis.pdf (visited on 08/22/2023).

McGuinness, Luke A. and Athena L. Sheppard (2021). "A descriptive analysis of the data availability statements accompanying medRxiv preprints and a comparison with their published counterparts". In: *PLOS ONE* 16.5, e0250887. DOI: 10.1371/journal.pone.0250887.

Milzow, Katrin et al. (2020). *Open Research Data: SNSF monitoring report 2017-2018*. DOI: 10.5281/zenodo.3618123.

Nachev, Vladislav, Jan Taubitz, and Nico Riedel (2023). *Charité Dashboard on Responsible Research*. URL: https://github.com/quest-bih/dashboard (visited on 08/22/2023).

Neuroth, Heike and Gudrun Oevel (2021). "Aktuelle Entwicklung und Herausforderungen im Forschungsdatenmanagement in Deutschland". In: *Aktuelle Entwicklung und Herausforderungen im*. De Gruyter Saur, pp. 537–556. DOI: 10.1515/9783110657807-029.

OECD (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Paris. DOI: 10.1787/9789264034020-en-fr.

Pampel, Heinz and Roland Bertelmann (2011). "'Data Policies' im Spannungsfeld zwischen Empfehlung und Verpflichtung". In: *Handbuch Forschungsdatenmanagement*. Bad Honnef, pp. 49–62. URL: https://opus4.kobv.de/opus4-fhpotsdam/frontdoor/index/index/docId/195.

Pinel, Clémence, Barbara Prainsack, and Christopher McKevitt (2020). "Caring for data: Value creation in a data-intensive research laboratory". In: *Social Studies of Science* 50.2, pp. 175–197. DOI: 10.1177/0306312720906567.

Pujol Priego, Laia, Jonathan Wareham, and Angelo Kenneth S. Romasanta (2022). "The puzzle of sharing scientific data". In: *Industry and Innovation* 29.2, pp. 219–250. DOI: 10.1080/13662716.2022.2033178.

Quigley, Niamh, Janice Chan, and Julie Clift (2022). *The role of Australian institutional repositories in sharing academic research: Research report*. DOI: 10.25917/S5A6-R623.

Rat für Informationsinfrastrukturen (2023). *Föderierte Dateninfrastrukturen für die wissenschaftliche Nutzung*. URL: https://nbn-resolving.org/urn:nbn:de:101:1-2023021709.

Read, Kevin B. et al. (2021). "Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis". In: *Canadian Medical Association Open Access Journal* 9.4, E980–E987. DOI: 10.9778/cmajo.20200303.

Rousi, Antti Mikael (May 2022). "Using current research information systems to investigate data acquisition and data sharing practices of computer scientists". In: *Journal of Librarianship and Information Science*, p. 096100062210930. DOI: 10.1177/09610006221093049.

Simon, Lena (2020). *Datenmanagement in der Abteilung Biomaterialien*. 43. MPG-Bibliothekstagung, 11. September 2020. URL: https://info-bib.mpg.de/wp-content/uploads/2020/11/BT43-Simon.pdf (visited on 08/22/2023).

Tedersoo, Leho et al. (2021). "Data sharing practices and data availability upon request differ across scientific disciplines". In: *Scientific Data* 8.1, p. 192. DOI: 10.1038/s41597-021-00981-0.

Tenopir, Carol et al. (2015). "Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide". In: *PLOS ONE* 10.8, e0134826. DOI: 10.1371/journal.pone.0134826.

Thoegersen, Jennifer L. and Pia Borlund (2022). "Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing". In: *Journal of Documentation*. DOI: 10.1108/JD-01-2021-0015.

Thompson, Mark et al. (2019). "Making FAIR Easy with FAIR Tools: From Creolization to Convergence". In: *Data Intelligence* 2.1-2, pp. 87–95. DOI: 10.1162/dint_a_00031.

Toribio-Flórez, Daniel et al. (2021). "Where Do Early Career Researchers Stand on Open Science Practices? A Survey Within the Max Planck Society". In: *Frontiers in Research Metrics and Analytics* 5, p. 586992. DOI: 10.3389/frma.2020.586992.

Wicherts, Jelte M., Marjan Bakker, and Dylan Molenaar (2011). "Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results". In: *PloS One* 6.11, e26828. DOI: 10.1371/journal.pone.0026828.

Wilkinson, Mark D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1, p. 160018. DOI: 10.1038/sdata.2016.18.

Wissenschaftsrat (2012). *Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020*. Berlin. URL: https://www.wissenschaftsrat.de/download/archiv/2359-12.pdf.

Wittenburg, Peter et al. (2019). "FAIR Practices in Europe". In: *Data Intelligence* 2.1-2, pp. 257–263. DOI: 10.1162/dint_a_00048.

Zhang, Lili et al. (2021). "A Review of Open Research Data Policies and Practices in China". In: *Data Science Journal* 20.1, p. 3. DOI: 10.5334/dsj-2021-003.