Wojciech Fenrich, Krzysztof Siewicz, Jakub Szprot

Towards Open Research Data in Poland

OPEN SCIENCE PLATFORM

UNIVERSITY OF WARSAW
Interdisciplinary Centre for Mathematical and Computational Modelling
www.icm.edu.pl

Wojciech Fenrich, Krzysztof Siewicz, Jakub Szprot

Towards Open Research Data in Poland

Wydawnictwa ICM
Warszawa 2016

ISBN 978-83-63490-14-0

Published by
Wydawnictwa ICM
ul. Pawińskiego 5a
02-106 Warszawa

UNIVERSITY OF WARSAW
Interdisciplinary Centre for Mathematical and Computational Modelling
www.icm.edu.pl

Proofreading
Marta Hoffman-Sommer

Editing
Michał Starczewski

Cover and title page design, graphics, typesetting and text makeup
Jakub Rakusa-Suszczewski

OPEN SCIENCE PLATFORM

Table of Contents

List of Figures

Introduction

In this report, we understand "research data" as all data that is collected or created during and for the purposes of scientific research. We understand "open data" in the sense of the Panton Principles for Open Data in Science: "By open data in science we mean that it is freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself."[1]

The benefits of opening research data are now widely recognized. First of all, it allows for the verification of claims made in research articles and books, thus being fundamental to enabling proper evaluation of science and preserving scientific rationalities. It helps to avoid the duplication of research efforts and maximize the benefits from the efforts already made. It opens up the possibilities of combining data from different sources across disciplinary and institutional boundaries. Moreover, researchers are not the only beneficiaries. Entrepreneurs can re-use open research data as the basis for their new, innovative products and services - or as a means to improve them. Government agencies and public institutions will find open research data useful in defining and implementing their evidence-based policies and improving the effectiveness and efficiency of their activities and procedures. Finally, the public can hope that opening of data will add to the transparency and public accountability of research, as well as its usefulness to the society.

Policies on research data are introduced by many institutions and primary research funders. Most of them are concerned with research data management in the first place. Data sharing is an important aspect of data management and as such has to be taken into account in any comprehensive data policy. From the Polish perspective, a particularly important example is the Open Research Data Pilot in Horizon 2020. The Pilot applies both to the data needed to validate the results presented in publications and to other data. For 2016 - 2017, it covers selected core areas of Horizon 2020, but individual projects outside these areas can participate in it voluntarily (opt-in). Projects can also partially (for selected datasets) or entirely (for all datasets) opt out at any stage, given legitimate reasons (EC provides a list of such reasons, which is not a closed one). Costs involved in the implementation of the Pilot are eligible as part of the grant. As for the specific requirements imposed on the projects participating in the Pilot, their Grant Agreements include Article 29.3, which reads as follows:
„Regarding the digital research data generated in the action ('data'), the beneficiaries must:

1 P. Murray-Rust, C. Neylon, R. Pollock, J. Wilbanks, Panton Principles. *Principles for open data in science*, 2010, http://pantonprinciples.org/.

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

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

(ii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';

(b) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves)."[2]

The Commission summarises the approach taken in the Pilot with the words "as open as possible, as closed as necessary".

In Poland, the Ministry of Science and Higher Education issued in October 2015 the document „Directions of the development of open access to research publications and research results in Poland". While the policy set out in this document focuses on publications, presenting a series of recommendations for research funders, scientific institutions and publishers, it also mentions research data. The Ministry states that open access to publicly funded or co-funded publications and data should be a general principle; exemptions from this principle should be based on objective, reasonable grounds. It is recommended to make research data openly available in a way that does not result in disclosure of secrets or violation of commercialisation interests. While making data available, institutions and researchers should take into account current best practices in the field. The Ministry itself will take steps aimed at analysis of the current state of affairs regarding research data in Poland, as well as identification of best practices, strategies and policies. It will also undertake consultations with key stakeholders.

The aim of this report is to provide a starting point for the debate about opening research data in Poland. It consists of two main parts: an analysis of the legal framework and the presentation of the results of an empirical study (a survey and a qualitative study) conducted amongst the Polish academic community. In the appendix, we provide a list of recommendations for the Polish academia.

2 Horizon 2020 Annotated Model Grant Agreements, Version 2.1, 30 October 2015, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf, p. 215.

Legal framework

A key aspect in the understanding of openness of research data are legal issues. There is a debate in the open access movement about the exact meaning of the word "open" with regard to publications. It has resulted in a distinction between "gratis" and "libre" open access. The former focuses on availability without payment, while the latter stresses that additional broad reuse rights are necessary. Such a distinction results in another dilemma, namely about the exact scope of reuse rights sufficient to qualify as "libre". While this distinction and the following dilemma is often discussed in the debate about open access to publications, it appears that "open" with regard to (research) data has always been understood as including broad reuse rights.[3] This kind of approach can be found in such documents as the OECD Declaration on Access to Research Data from Public Funding (2004)[4] or the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003).[5]

Therefore, open data should be understood as data not just publicly available without charge, but also without material legal restrictions of reuse. There is also a popular consensus that the level of acceptable restrictions is minimal, which follows from the standard licenses that are often used for data (CC0, CC-BY or ODbL, we describe them in more detail later). So, open data is data either not subject to exclusive rights at all, or subject to a free license granted by the owner of such rights.

The issue of exclusive rights to data in the context of open data is at least twofold. First, there is the question of the rights of the researcher or research institute that collected the data vis-a-vis third party rights. Second, there is the question of the multiplicity of rights and their applicability at different levels of a structured object such as a dataset.

3 For a thorough overview see: P. Heinz, S. Dallmeier-Tiessen, Open Research Data: From Vision to Practice, 2014, http://book.openingscience.org/vision/open_research_data.html.

4 "RECOGNISING that open access to, and unrestricted use of, data promotes scientific progress and facilitates the training of researchers", OECD, Declaration on Access to Research Data from Public Funding, http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157.

5 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, 2003, http://openaccess.mpg.de/Berlin-Declaration.

Third party rights

Exclusive rights to data constitute an obstacle towards reuse that is a bit different to those that result from such rights with regard to software or other copyrighted works such as scientific publications. This is because datasets are regularly subject not only to exclusive rights of the researcher or research institute that collected the data, but also to exclusive rights that apply to individual pieces of data (e.g., copyrights to photographs, publicity rights of data subjects, etc.). This is a case similar to the situation that sometimes occurs in the Free Software world, where an author of a program releases it under a free license, but such a license may not release the program from a software-related patent that may be granted to a third party but still cover a particular application of the said program. In the Free Software movement, existence of third party rights has not been considered as a defining factor of the notion "Free Software". Such rights are rather treated as restrictions of user freedoms to a particular piece of Free Software that is subject to them. The debate about open data has not yet reached the point where this issue would have been analysed; the definitions so far ambiguously refer to a lack of restrictions, but do not specify whether these "restrictions" include also third party rights.

Different rights to and in a dataset

There is an important legal distinction between a dataset as a whole and individual items included in the dataset. These can be separately subject to different rights that have to be separately analysed.

Generally speaking, the most probable exclusive right attaches at the level of the dataset as a whole. It is the so-called *sui generis* right that follows from Directive 96/9[6] and its implementations in EU member states. Strictly speaking, this exclusive right covers only datasets that meet the definition of a "database", i.e. "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means" (Directive 96/9 Art. 1.2) and which show "that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database" (id. Art. 7.1). There is no measurement of "substantiality" included in the law, but on the other hand it follows from the case law that the necessary investment does not include investment in the creation or verification of the database's individual elements.[7] Also the investment does not have to be financial, but could also be of other nature. It follows that currently the "investment" criterion is sufficiently broad that it may be expected that *sui generis* protection will often attach to datasets. Notably, *sui generis* protection is independent of any other exclusive rights that might apply to the dataset as a whole, for example copyrights.

Apart from the above, individual pieces of data included in the dataset may be subject to separate exclusive rights, such as copyrights, publicity or privacy rights, personal data protection regulations, public sector information reuse restrictions, etc.

6 Directive 96/9/EC of the European Parliament and of the Council, http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML.

7 See: Judgement of the Court of 9 November 2004, The British Horseracing Board Ltd and Others v William Hill Organization Ltd., C-203/02, ECR I-10415, http://curia.europa.eu/juris/liste.jsf?num=C-203/02.

Again, the debate about open data has not reached a point where the above issues would have been analysed in detail. Although there is a general recognition of the existence of *sui generis* rights that might apply irrespectively of copyright protection, we have not found any analyses that differentiate between legal protection at the level of the dataset as a whole and at the level of individual elements of the dataset.

Open data licences

Open data can be perceived as at least the third wave of "openness" in the digital environment. First, we have experienced the Free Software/Open Source movement. Second, Open Culture and in particular Open Access to scientific publications. Third, there is open data. So, open data does not need to invent anew tools that allow to provide for openness, rather it only needs to adjust the existing tools where necessary. From the legal point of view, such tools are open licenses.

The development of Free and Open Source Software has been coupled with the drafting of standard licenses adjusted for software (with a prominent exception of GNU FDL, designed for software documentation and later used e.g. by Wikipedia). With the dawn of Open Culture, the Creative Commons organisation delivered a set of licenses adjusted to other objects of copyright and neighbouring rights. But the early versions of Creative Commons licenses did not cover database rights, nor did they account for the peculiarities of databases sufficiently, at least from the point of view of some open data enthusiasts from the Open Knowledge Foundation. Hence, they have come up with a set of licenses designed especially for open (research) data - including the Open Database License (ODbL). Notably, this is the only set of licenses which explicitly recognizes the different levels of rights in a database (rights to the database as a whole and rights to individual elements of the database). In the meantime, Creative Commons have launched the 4.0 version of the Creative Commons licenses which now include database rights. Creative Commons also provides the CC0 public domain dedication and license.

Currently, a prospective open data contributor may choose at least from among the Creative Commons and Open Knowledge Foundation "offerings". Of these, CC0 is a tool that receives much attention, as it results in the least set of user obligations. CC0 waives all copyright, neighbouring rights, and database rights (or freely licenses them if the waiver is not effective under the applicable law), but does not even require attribution as is the case with CC-BY or ODbL. The preference for CC0 results from the perceived burden that would follow from "attribution stacking" in case data is extensively reused and included in derivative datasets. CC0 does not pose such a burden, as there is no such obligation at all. Many open research data repositories require CC0 or at least suggest it as a default license.

The above licenses are royalty-free, but one has to bear in mind that open data is not always gratis. This is especially the case with Big Data, which requires investing in considerable storage and transfer capacities. User payment for transfer is often required in such a case, even if the data itself is under a free license. This, however, is not a contradiction, because free licenses do not prevent charging for the delivery of data, and any such payment does not affect user rights under the license.

The practice of open (research) data

The state of the debate is reflected by the practice of open data science, and it is particularly visible in open data repositories. Different open data repositories differently approach the issue of third party rights as well as the issue of multiplicity of rights. Here, we provide a general overview of these practices, compiled using a non-representative choice of example repositories. We start our overview with Zenodo, which is a suggested open data repository for the Horizon 2020 Open Research Data Pilot mentioned above. This makes Zenodo an important repository, capable of setting standards for the future of Open Science, at least in the EU. Then we move on to Figshare, a popular repository open for a wide variety of data, which is important precisely because of its scope. Third, we choose Dryad, which is important as an example of a US-based repository, and also because it is focused on scientific (mainly biological) and medical data, which makes it an interesting sector-specific example. The two last examples are data repositories for the Netherlands and the United Kingdom, which makes them important because they are a *de facto* standard for Member States particularly advanced in practicing Open Science.

Zenodo

Zenodo (https://zenodo.org/) is a research data repository with an option to make the data openly available. Zenodo is being developed within the OpenAIRE project funded by the European Commission, and it is hosted by CERN.

From the point of view of an uploader, Zenodo allows to specify a "license for files", which is a single choice made when uploading a given set of files. There is no clear distinction whether the choice applies to the dataset as a whole, or (also) to each individual file in the dataset. However, there is certain recognition of the problem of third party rights, as it is clearly stated on the Zenodo web page that "You are responsible for respecting applicable copyright and license conditions for the files you upload" (although it's in the fine print). Also, the brief terms and conditions of Zenodo laconically require that "Users shall respect copyright and applicable license conditions. In particular, reconstructing the full texts of articles from snippets is not allowed. Download and use of information from Zenodo does not amount to a transfer of intellectual property."

Figshare

Figshare (https://figshare.com/) is a research data repository which requires all deposited data to be released under CC0. Figshare is run by a company controlled by Macmillan Publishers. Similarly to Zenodo, the information at the Figshare site does not go into detail whether the CC0 waiver applies only to the rights to the dataset as a whole, or also to its individual elements. Third party rights are also recognised and addressed in T&C, which requires that users do not infringe these rights.

Dryad

Dryad is a repository for "data underlying scientific and medical publications" run by a nonprofit membership organization funded partially by the US National Science Foundation (http://datadryad.org/). Similarly to Figshare, it requires that the rights to the deposited data be waived under CC0. Its terms and conditions require that the submitter holds sufficient rights to apply CC0, without resolving whether this applies to the dataset as a whole or also to its individual elements. However, Dryad reserves to itself the right to review the contents for statements

potentially incompatible with the CC0 waiver, which suggests that the waiver is intended to apply to all possible rights included in the dataset. This also indirectly follows from Dryad's FAQ.[8]

3TU.Datacentrum

3TU.Datacentrum is a Dutch repository for data (http://data.3tu.nl/repository/) run by the Delft University of Technology. There is no general information on the site or in metadata about rights and licenses applicable to datasets deposited in the repository, any such information has to be sought in the files of individual datasets. Downloading many, if not all, datasets requires registration.

UK Data Archive

The UK Data Archive is a UK repository for social science and humanities data run by the University of Essex (http://www.data-archive.ac.uk/). The data is made available under a specific End User License, but the repository allows to deposit open data as well, under either the UK Open Government License or Creative Commons BY 4.0. This is allowed only for data "which are not personal", but there is also no clear guidance whether the license applies to the dataset as a whole or additionally to the dataset's contents. The fact that personal data may not be made available openly in the UK Data Archive gives raise to the interpretation that the license is intended to cover both the dataset and its contents.

Legal framework and current practice

It follows from the above brief overview that the issue of third-party rights is recognized in practice to some extent. The fact that one dataset may be subject to many different rights is also recognized. This is made particularly explicit as far as personal rights are concerned, such as publicity, privacy or data protection - many repositories make it clear that such third party rights make it impossible to deposit data, unless anonymised.

However, there is hardly any recognition of the legal challenges that follow from the fact that the legal protection may apply not only to the dataset as a whole, but also to individual elements of a dataset. The reviewed repositories do not require clarification on this from data providers. So, information whether the license applies to all such rights or only to the rights to the dataset as a whole is only implicit unless the data providers themselves clearly state it in their datasets (e.g. in additional textfiles, etc.). Thus, there seems to be an implicit assumption of the repositories that the free license is a confirmation that all such rights have been cleared. However, it does not follow that this is the understanding of all data providers.

Thus, we may sum up that the legal solutions used in open research data include in particular releasing the data under free licenses such as CC-BY or dedicating them to the public domain (e.g. by using CC0), with an implicit assumption that the license or dedication applies to all rights controlled by the data submitter. There is also an understanding that the submitter should adequately clear third-party rights and especially respect the privacy of data subjects. But since any such rights constitute a burden on reuse, and their complete clearance by the submitter is not possible in some situations, the general conclusion is that users have to scrutinize these issues on their own.

8 Dryad, http://datadryad.org/pages/faq#deposit ("It is important to note that if you have data that, due to pre-existing agreements, cannot be released under the terms of CC0, please do not submit that data to Dryad").

Legal aspects in the Polish system

The Polish legal system concerning open data is based generally on the same principles as described above. Data can be subject to such exclusive rights as copyrights and *sui generis* rights, there may be many different rights attached to one dataset, these rights may belong to different persons, and the rights attach at different levels of the dataset. Apart from rights that may be transferred or licensed, there are also other types of rights such as personal rights, publicity rights, and data protection.[9] We describe them as they operate in Poland below. We start with rights attaching at the level of the whole dataset: *sui generis* and copyrights. Then, we move on to discuss rights applicable to individual elements of a dataset. In the last part of this chapter we describe the process of removing legal obstacles from the Polish legal perspective.

Rights to datasets as a whole

Sui generis

The most important right that attaches at the level of the whole dataset is the *sui generis* right to databases. This right follows from the EU Directive 96/9 (mentioned above), which has been implemented in Poland in the act of 27 July 2001 on the protection of databases. The protection applies to databases that satisfy the definition specified in art. 2.1.1, stating that the database is a collection of any materials or elements gathered systematically or methodologically, individually accessible in any way, that has required qualitatively and/or quantitatively a substantial investment in order to make, verify, or present the contents of the collection (the definition roughly matches the respective wording of directive 96/9). The Polish act clearly incorporates the criterion of bearing the investment risk to the definition of the maker (producer) of the database in art. 2.1.4 (directive 96/9 mentions this only in recitals).

This means that the "substantial investment" and "risk of the investment" are crucial factors in determining the applicability of *sui generis* protection to a particular database. Introduction of such criteria also means that in the usual context of institutional research, the rights will initially materialize with the research institute conducting the whole process of making, verifying or presenting the database, not with the individual scientists themselves (this is a material difference from how the Polish law treats copyrights to research). Only if the investment risk of the whole process can be attributed to individual scientists, the *sui generis* right will be held by them initially. Consequently, unless explicitly agreed upon, this issue may be quite unclear in projects performed by consortia. Namely, the rights may be shared within a consortium if the investment and the risk can be attributed to more than one participant.

9 Researchers often turn to data gathered or stored by public-sector bodies, such as statistical data, data from registries, spatial information, etc. There are specific legal issues related to public-sector information, which call for a separate analysis.
Here, we can only briefly mention that in Poland, there is a general act that regulates access and re-use of such information, but also a whole set of sector-specific legal acts that may apply to individual datasets (such as the provisions regulating access and re-use to public spatial information). The general re-use provisions are the implementation of the directive 2003/98 on the re-use of public sector information, but sector-specific acts not always conform to the directive.
It follows from these provisions that mere public availability of public-sector information does not automatically allow its reuse, even for scientific purposes only. Public-sector bodies may specify conditions on reuse, and conditions of different bodies do not have to be compatible. From the technical point of view, data is rarely made available through APIs and in conformity with uniform (open) technical standards, that would in particular allow to obtain updated information automatically and make it more easily interoperable with other data. Rather, the basic idea of the whole system is that data is made available in chunks, upon request (which must specify, in particular, how the data will be processed), without any obligation to make the data technically more usable for the reuser.

As a sidenote, explicit inclusion of the risk criterion as a condition of protection may make it harder for publicly-funded research institutes to invoke *sui generis* rights. The risk may not exist if the funding is provided without a direct link to the project's output or its quality. Also, there may be no risk if the database is produced in the fulfilment of a public task. But everything depends on the circumstances of a particular case. Here, it is important to note that art. 6.2 of the Polish database act introduces a presumption that the maker (producer) is an entity whose name is made publicly available in relation to the dissemination of the database. This means that if such a notice is present, the lack of investment risk has to be proven by anyone who contests that fact.

Sui generis protection results in the need of obtaining the maker's (producer's) consent for the extraction and re-utilization of data (all data or a substantial part of it). Consent is not necessary if the data is used as illustration, for educational or research purposes, provided it is non-commercial and the source is mentioned (art. 8.1.2). While the wording of the Polish act suggests that "illustration", "academic purposes", and "research purposes" are alternatives, the wording of Directive 96/9 is clear that only "illustration for teaching or scientific research" is allowed. It is the only situation that allows a researcher to utilize a substantial part of a database without consent, which means that *sui generis* protection effectively prevents data mining (clearly exceeding "illustration").

It follows that under Polish law we may refer as open data only to such data that is available under a free license explicitly covering database rights. Such licenses include, for example: CC (in 3.0 *sui generis* protection was waived, in 4.0 it is retained, but it is licensed in the same way as copyrights and neighbouring rights), CC0, and the Open Database License.

Copyright

Irrespectively of the *sui generis* protection described above, datasets as a whole may also benefit from the protection of copyright law. Art. 3 of the Polish Copyright Act explicitly states that "collections, anthologies, selections, creative databases are subject to copyright, even if they consist of unprotected materials, if the selection, composition, or combination is creative, without affecting the protection of works used in them."

It is thus possible that a given dataset benefits from double protection - as a copyrighted "collection, etc." and under the *sui generis* regime. However, one does not imply the other, since the conditions of protection are different. There is no requirement for investment risk in the copyright law, but creativity is required. For copyright protection of the dataset as a whole, there has to be creativity in the selection, composition or combination of individual elements of the dataset. Creativity in the elements themselves is irrelevant for the protection of the whole, but it may still hinder the reuse of the whole - since the whole has to be reused in compliance with the requirements applicable to every individual part.

Additionally, in case the whole is subject to copyright, the removal of parts may not be a plausible option, since the mutilated whole may be regarded as an infringing derivative, or a breach of the author's personal right of integrity.

Different conditions of protection mean in particular that *sui generis* rights and copyright may initially subsist in the hands of different persons. The former in the hands of the bearer of the investment risk, the latter in the hands of the creative contributor. Notably, although the Polish Copyright Act provides for a default transfer of works

made in the course of employment to the employer, it also specifies an exception from that rule in the case of scientific works. Rights to scientific works made in the course of employment at a scientific institution belong to individual authors (art. 14). It is not clear how this relates to scientific collective works, because collective works are subject to yet another specific rule - the rights to such works as a whole belong to the producer as a matter of law, while the rights to individual elements to their authors (art. 11). So, if there is no explicit agreement on this issue, it remains an open question who exactly owns the rights to datasets compiled by a researcher employed at a scientific institution.

Copyrights are subject to a different set of limitations than *sui generis* rights. Research freedom is currently quite narrow due to limitations specified in art. 27 of the Polish Copyright Act, which allows scientific institutions (not individual researchers) to illustrate their teaching materials with excerpts from third-party works and to use such works for research purposes. However, in the case of double protection (*sui generis* and copyright) the resulting scope of scientific freedom is limited to the least common denominator, which in that case is the narrow "illustration" exception in the database act without a clear authorization for any wider use for research purposes.

Moreover, as we have already mentioned, there may be other rights attached to individual elements of a dataset, which may constitute an additional restriction on reuse. Such elements may be copyrighted, or subject to other rights (privacy, data protection, ISP reuse, etc.).

Rights to individual elements of a dataset

Data gathered in the course of scientific research is as diverse as science itself. One thus cannot exclude that a particular piece of data is subject to copyright, as could be the case for many research projects in the social sciences and humanities, or arts: collections of articles, books, paintings, photographs, expressions, etc. Even if a piece of data is an uncopyrightable representation of a fact due to lack of creativity in its representation, it may still be subject to use and reuse restrictions that follow from other laws. Here, we focus on various laws that protect privacy, as this is probably the most often mentioned obstacle in the way to public availability of research data.

Personal data protection

Probably the most often mentioned restriction is privacy and data protection. Indeed, many research projects gather personal data. Under the Polish implementation of directive 95/46 on personal data protection, personal data is any information about an identified or identifiable natural person. Only such information that does not allow to identify a person without investing excessive costs, time, or activities, is not personal data. Processing personal data encompasses every operation on the data, so it includes not only internal processing in the course of scientific research, but also making the data (publicly) available.

Personal data processing must always conform with the act of 29 August 1997 on personal data protection. There are many obligations that follow from this act, in particular to exercise special care in order to protect the interests of data subjects. The data has to be processed in accordance with the act, collected for legitimate and

specific goals, processed to reach those goals, it must be correct and adequate to the goals, and anonymized if not necessary to reach the goals (Art. 26.1 of the act). For scientific purposes, it is possible to process data gathered for other reasons, but only as long as the processing does not breach the rights and freedoms of the data subject. There are also some other examples where obligations that follow from the act are liberalized when data is processed for scientific purposes. However, processing personal data for these purposes is still subject to the act in general. Also, these more liberal provisions hardly apply in a situation where the data is made publicly available, since most probably "scientific purposes" will be construed narrowly in case of a dispute in such a case.

Other privacy-related provisions

The personal data protection act is not the only act that regulates the use of personal data. First, there are some acts which introduce additional requirements in specific cases, but these apply only to specific research projects. Second and more importantly, apart from the above administrative law, privacy is protected in Poland by private law - Civil Code provisions on "personal goods" and Copyright Act provisions on the right of publicity. The private law protection adds an additional layer of conditions that have to be followed to legally use and reuse data. "Personal goods" protection from the Civil Code is an implication of the legal recognition of people's dignity - the data has to be processed in such a way that does not threaten this dignity. A breach of personal goods is decided by applying a sophisticated balancing test, invoking weighting of values (e.g., privacy and the people's right to know). The right to publicity allows to control the distribution of one's likeness (such as photographs), and features a list of exceptions from this rule which does not easily match the list of conditions that legitimize processing of personal data (while one's likeness usually constitutes one's personal data).

Removing legal obstacles

Opening data requires that all relevant legal restrictions of reuse are identified and removed to the maximum possible extent. The identification of restrictions is a repetitive activity that has to be included in the whole process of data gathering and processing (including making available).

Data gathering

The most important thing is that the identification cannot be performed only from the point of view of the researcher or researchers that gather the data, and only for their own purposes. While open data may be used by anyone, sometimes for much different purposes, so the process has to include others' needs as well. Certainly, this is possible to a limited extent only, since it is not possible to predict all users and uses of the data. But it does not follow that users should be left alone even if they have to perform the identification and removal of restrictions on their own. The initial data collector is definitely most capable of gathering important metadata that minimizes the burden necessary to legally reuse the data by subsequent users. This in particular implies that without a valid reason, researchers should not remove any information from data they gathered, even if it seems useless from their point of view. Valid reasons include for example privacy or data protection requirements.

Given the narrow scope of statutory scientific freedom, a necessary step is obtaining the relevant rights in order to release the data under an open license, if they do not belong to the researcher who intends to make

the data openly available already. But this may not be possible for various reasons, of which the following two are most important. First, some right owners may not be willing to allow to license them under an open license. Second, some rights are not waivable.

The first problem may be ultimately solved only through a change of law which would carve out a more broad scope of scientific freedom. Otherwise, the only solution is careful negotiation with the right owners. Obviously, it is much easier to negotiate before the data is gathered, and the person best suited for such negotiations is the one who gathers the data, not subsequent users of the data.

Second, as far as unwaivable rights are concerned, such as privacy, data protection or rights of publicity, it may still be possible to obtain the data subjects' consent, in a form sufficiently broad to allow for a certain degree of reuse. From the point of view of open science it is necessary that such consent clearly covers not only internal processing for the purposes of the given project, but also processing for the purpose of other projects that may reuse the data. There are certainly legal limits to the scope of such consents, and too broad a consent may simply be legally invalid. But this is yet another argument for maintaining in the data as much metadata as possible, including the texts of the data subjects' consent, and possibly some contact information for subsequent users who might need to obtain further consent on their own.

Processing

The process of removing legal obstacles is a continuous one, even if the data is not released outside of the project. Even in the case of internal processing, one has to constantly monitor whether the current use of data is in line with the licenses and consents obtained, and, if necessary, to amend them.

Making available

Given a narrow and vague scope of statutory scientific freedom, free licenses are an important legal tool for open data science not only with regard to *sui generis*, but also to copyright protection. Licenses such as CC, CC0 or the Open Database License clearly cover copyrights and neighbouring rights and allow for broad use of the licensed works, relieving users from doubts whether a given use is still within the narrow and vague statutory freedom. It is also important that there is clear information whether a license applies only to the dataset as a whole, or additionally to its individual elements, in order to clarify what restrictions remain. The legal tool usually used to remove legal restrictions on reuse are standard free licenses. In case of data, these usually are CC-BY, ODbL, or CC0 (some organizations require or simply advocate CC0 as the tool which leaves the least restrictions - we elaborate on various licenses above). The removal of restrictions by licensing, however, is possible only if the licensor holds the necessary rights (or an authorization to license them on behalf of the holder, a license to sublicense, etc.).

Results of an empirical study

The survey study

For this report a survey study was conducted among Polish researchers working in one of three types of academic institutions: universities, institutes of the Polish Academy of Sciences, and research institutes. The main goal of the study - as well as of the qualitative study presented in the next chapter of this report - was to explore the phenomenon of academic data sharing in Poland (experiences with and attitudes towards the issue, enablers and obstacles, level of knowledge about legal and technical aspects).

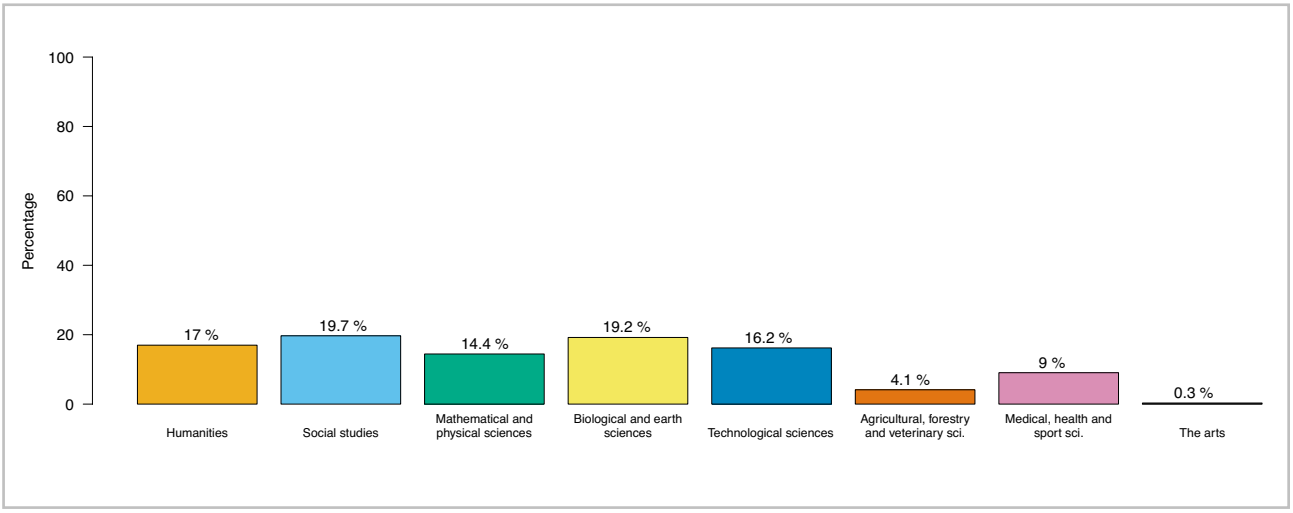
The questionnaire that we used was for the large part inspired by a study conducted by Benedikt Fecher, Sascha Friesike, Marcel Hebing, Stephanie Linek and Armin Sauermann[10], although some issues were omitted, reformulated or added to better fit the Polish academic and legal milieu. Of course, the sole responsibility for the study lies with the authors of this report.

The web-based questionnaire (created using Google Forms) was open from July 3rd until August 13th, 2015. To encourage researchers to participate in the survey, an invitation e-mail was sent to all Polish universities (both public and private), institutes of the Polish Academy of Sciences and research institutes. A necessary e-mail database was built upon information available from POL-on - a governmental database containing information on Polish scholarly units.

As a result, 630 researchers completed the questionnaire, of which 286 (45.4%) were women and 344 (54.6%) were men. Exactly half of the respondents held a doctoral degree; 17% were habilitated doctors and 11.9% - full professors. 21.1% of participants had - according to Polish nomenclature - a professional title (MA, MSc, BA, BSc or similar). 19.7% of the respondents represented social studies, 19.2% - biological and earth sciences, 17% - humanities, 16.2% - technological sciences and 14.4% - hard sciences (Fig. 1). Less than 10% of the participants represented medical, health and sport sciences (9%), agricultural, forestry and veterinary sciences (4.1%) and the arts (0.3%).

10 B. Fecher, S. Friesike, M. Hebing, S. Linek, A. Sauermannn, *A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing*, 2015. DIW Berlin Discussion Paper No. 1454. Available at SSRN: http://ssrn.com/abstract=2568693 or http://dx.doi.org/10.2139/ssrn.2568693. The questionnaire used in their study can be found on GitHub: https://github.com/data-sharing/persistent/blob/master/dsa-02/questionnaire-en.pdf.
See also: B. Fecher S. Friesike, M. Hebing, What Drives Academic Data Sharing? PLoS ONE 10(2): e0118053. doi: 10.1371/journal.pone.0118053.

Fig. 1. Research areas



Because of the way the respondents were approached and selected, the results of this study cannot and should not be extrapolated to the whole population of researchers working in the three abovementioned types of institutions. Thus, all results described below can be related only to the group of researchers that took part in the study.

General questions concerning sharing data

The majority of respondents had a generally positive attitude towards data sharing. More than 70% (37.2% strongly and 36.9% rather - Fig. 2) agreed that researchers should share their data, although it should be emphasized that we have not asked here about any specific way of sharing, such as, for instance, open data sharing. This joint support for sharing is not dependent on the researchers' professional titles or academic degrees, but stronger support (reflected in the answer "strongly agree") was most frequent among researchers with a vocational title only (MA's, BA's etc. 43,9% - Fig. 3) and least frequent among habilitated doctors (25.2%).

Almost 60% agreed (22.6% strongly and 36.4% rather - Fig. 4) that sharing data brings more benefits than losses to the researcher. The vast majority of respondents also agreed (almost 53% - strongly and 30.7% - rather - Fig. 5) that free access to research data produced by other researchers contributes to the advancement of science.

Despite this rather positive attitude towards sharing data, only two fifths of the researchers claimed that in their discipline sharing data is a common thing, while only slightly less - 36.6% - claimed that it is not (Fig. 6). Again, there were differences between researchers depending on the academic degrees held: only 27.3% of participants with a professional title picked "Strongly agree" or "Rather agree", while more than half (52.3% and 50.7%, respectively - Fig. 7) of the habilitated doctors and full professors chose these answers. Thus, early-career researchers may differ from their senior colleagues in their views of what is common in academia.

Fig. 2. Researchers should share their research data

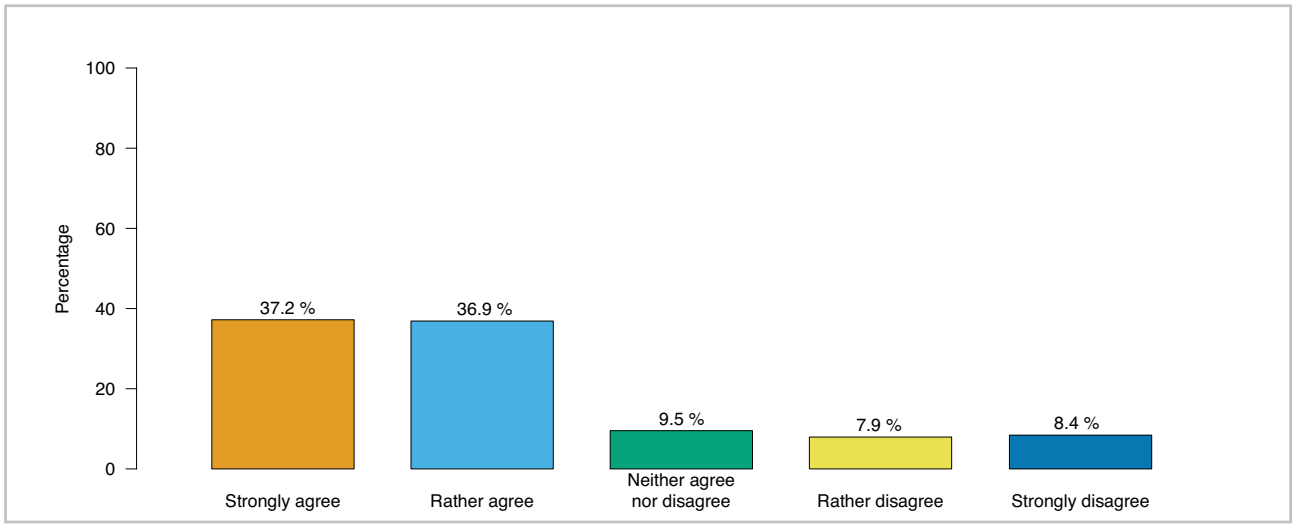


Fig. 3. „Researchers should share their research data" by "Scholarly degree or title"

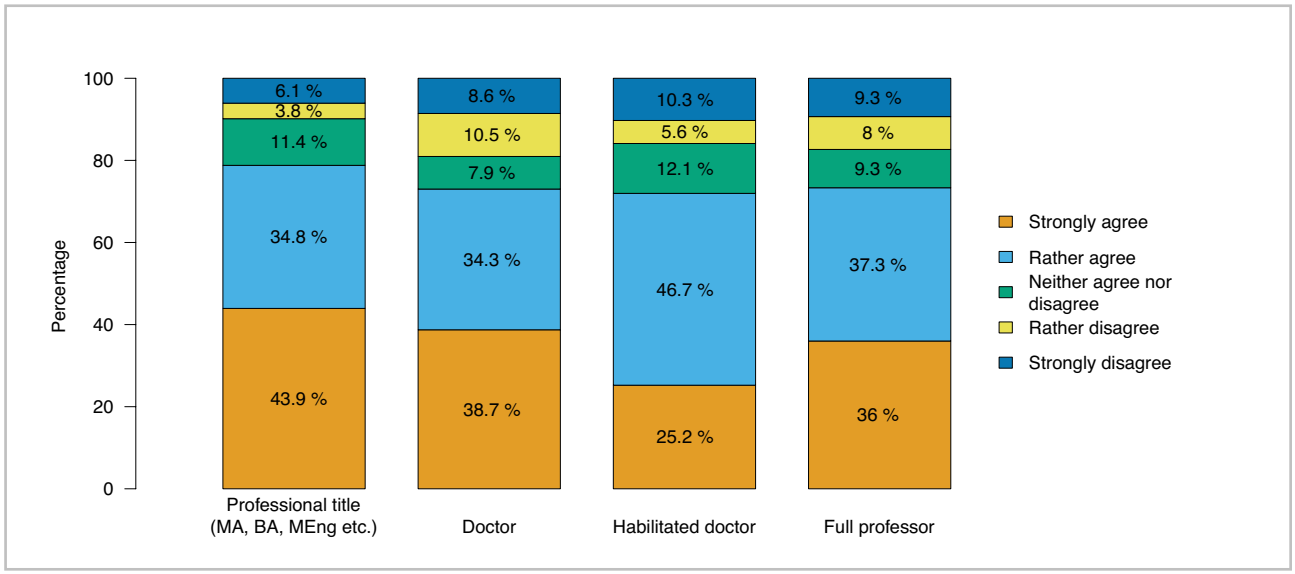


Fig. 4. Sharing data brings more benefits than losses to those who share it

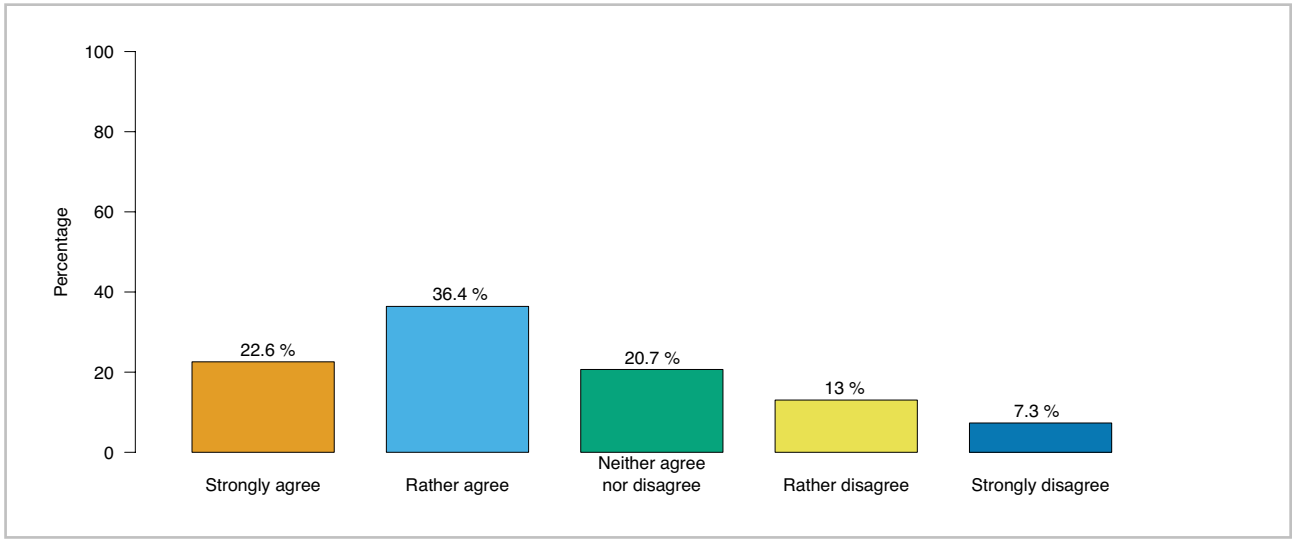


Fig. 5. Unfettered access to research data produced by other researchers contributes to the advancement of science

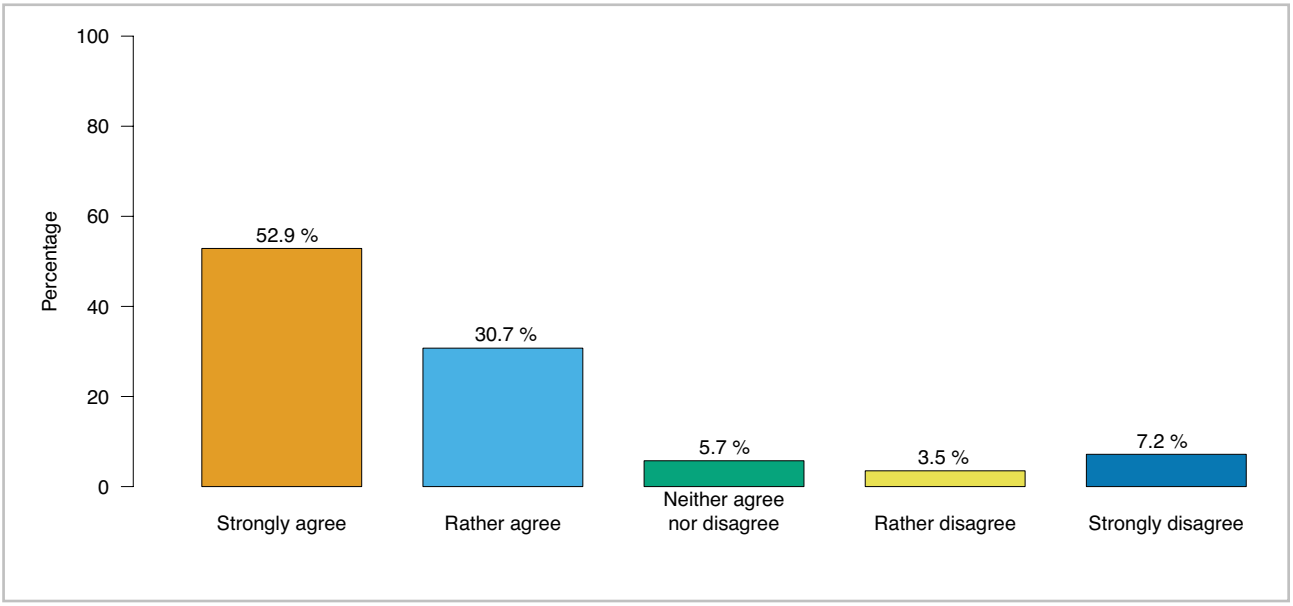


Fig. 6. In my discipline, sharing data is a common thing

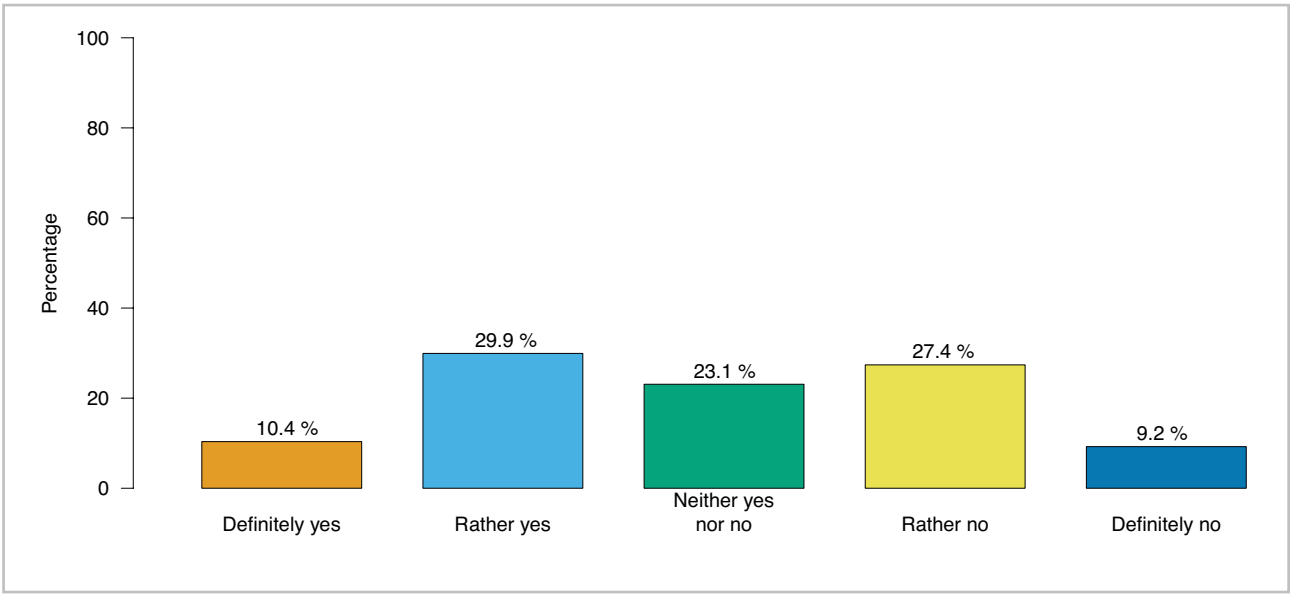
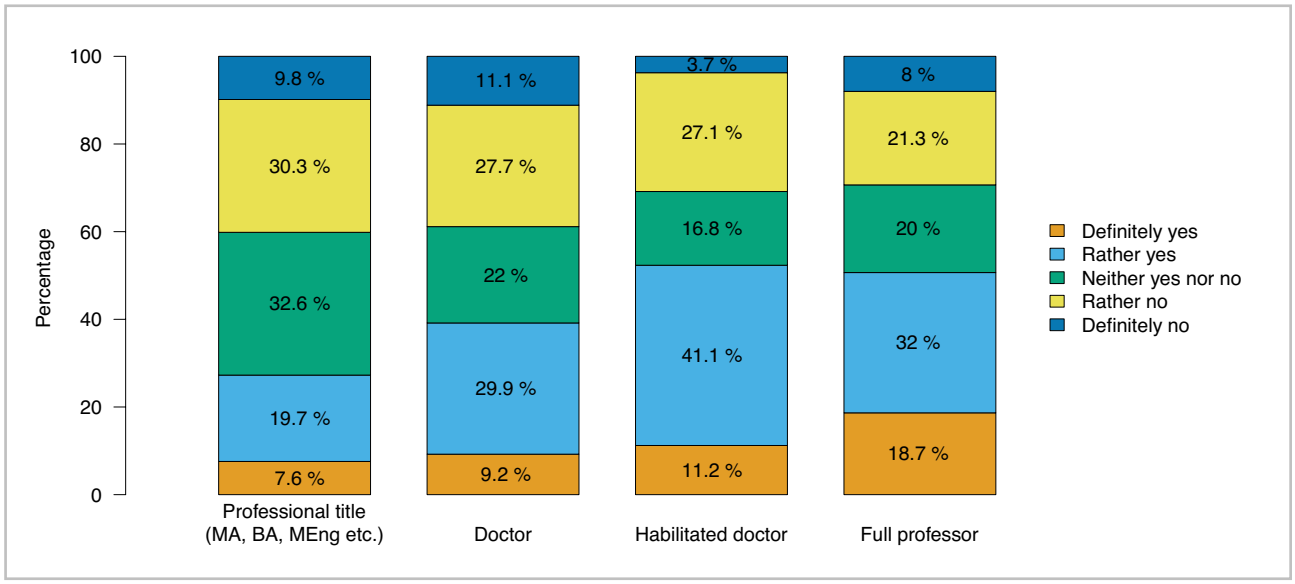


Fig. 7. „In my discipline, sharing data is a common thing" by „Scholarly degree or title"



Although the respondents had a generally positive attitude towards sharing data, their knowledge on this subject turned out to be limited. The majority of them agreed that they know where on the Internet they would be able to publicly share their research data. 19.6% of them were strongly convinced about it and more than 30% percent answered "Rather yes", but at the same time more than one third gave an opposite answer (25.3% "Rather no" and 11.5% - "Definitely no" - Fig. 8). Here, the higher the academic degree, the higher the joint frequency of answers "Definitely yes" and "Rather yes" (41.9% for researchers with MA up to 61.3% for full professors - Fig. 9).

When asked if the sentence "I know where on the Internet I could find publicly available research data shared by other researchers, that could help me in my work" adequately describes their situation, the respondents answered in a similar manner: 21.2% of participants picked "Definitely yes", 37.4% - "Rather yes" and almost one third - "Rather not" or "Definitely not" (Fig. 10). Here again the highest fraction of positive answers was from full professors (61.3%, although the fraction of habilitated doctors was nearly as high), and the lowest among MA's (40.9%).

Fig. 8. I know where on the Internet I could publicly share my own research data

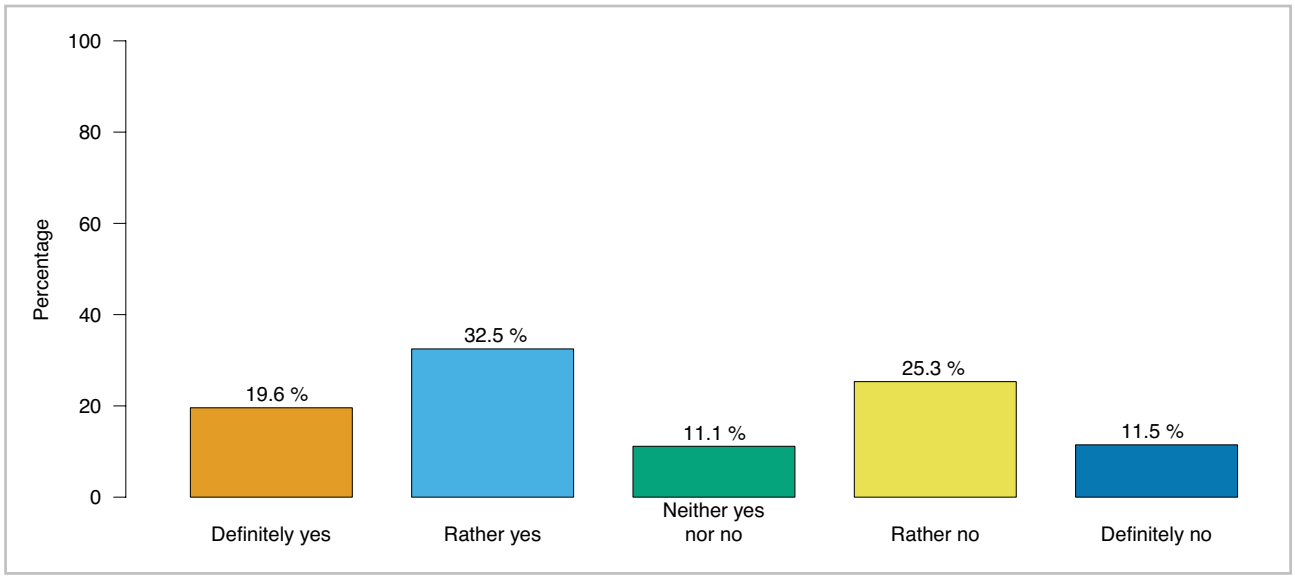


Fig. 9. „I know where on the Internet I could publicly share my own research data" by „Scholarly degree or title"

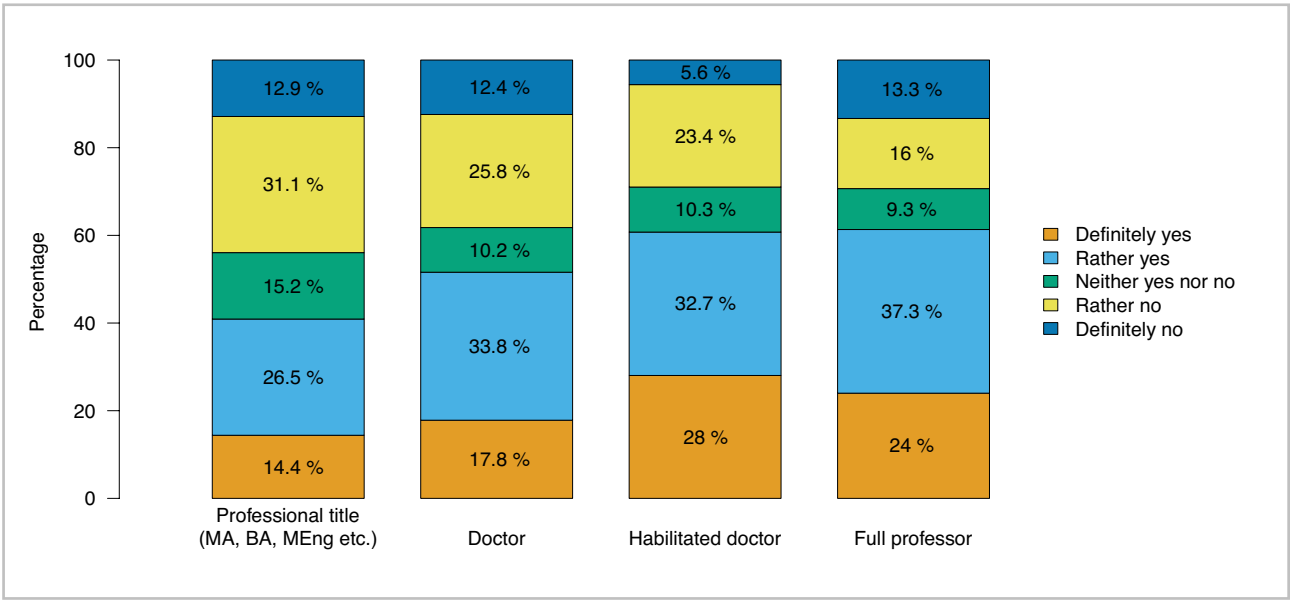
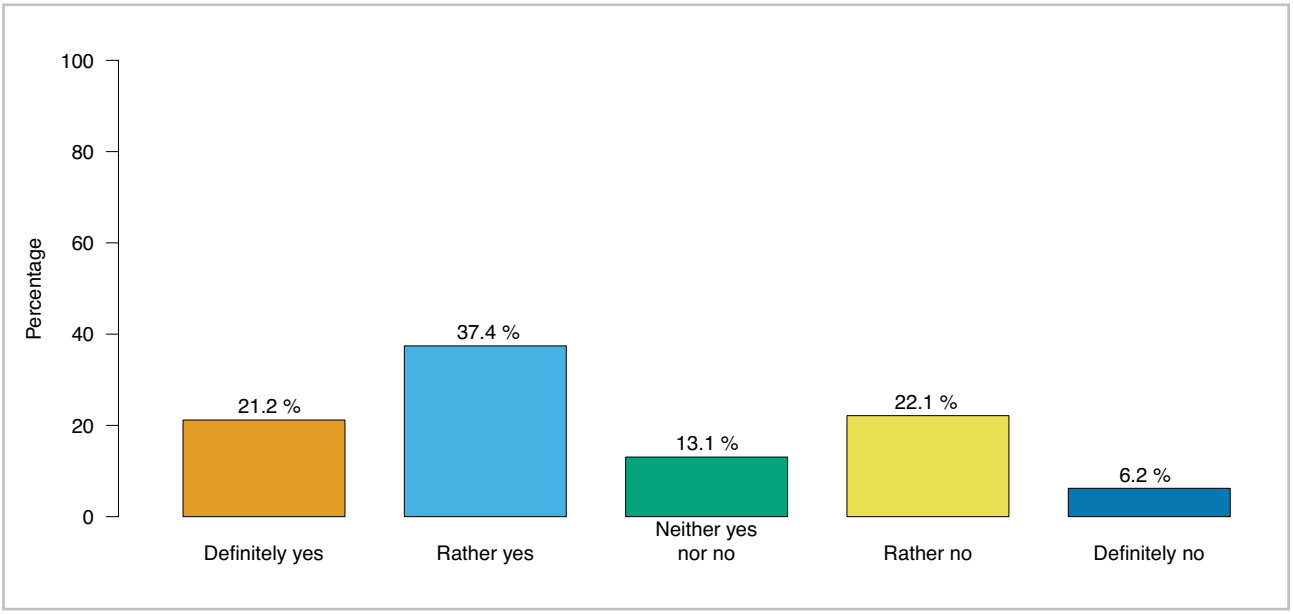


Fig. 10. I know where on the Internet I could find publicly available research data shared by other researchers, that could help me in my work



The participants were also asked if the statement "I would be able to cite in my publication a dataset that was made publicly available on the Internet by other researchers and make it in accordance with one of the existing citation styles" adequately describes their situation. Here again a vast majority declared that they know how to do it (35.6% - strongly and 41.3% - rather, Fig. 11), and again positive answers were most common among habilitated doctors and full professors (83.2% and 81.5%, respectively - Fig. 12).

The majority of respondents also declared that they are not discouraged by the fact that a journal requires them to deposit the data related to a paper. 57.5% disagreed (strongly or rather) that the sentence "If a scholarly journal requires that authors make data related to an article publicly available, I feel discouraged

from publishing in this journal" relevantly describes their situation. Slightly more than 10% agreed, but a significant fraction of respondents picked the neutral answer "Neither yes, nor no" (18.6%, Fig. 13) or claimed that they do not know or have never heard of such journals (13.5%).

Two of the questions in the survey aimed to test the respondents' knowledge about legal issues related to sharing data and using data shared by others. They were asked if the sentence "According to the law, for research purposes you can use every dataset publicly available on the Internet, provided that you indicate its source" is true or false (the option "I don't know" was also available). Almost 73% answered that the sentence is true, while 15.7% picked "False" and 11.4% "I don't know" (Fig. 14). With regard to the sentence "The decision whether to share research data or not always lies solely with the researchers who produced the data", 35% of respondents claimed that it is true, 48.6% that it is false and 16.4% picked the answer "I don't know" (Fig. 15). In both these cases, the wrong answers (that the sentences are true) were most popular among full professors (84.3% for the first question and 47.3% for the second - Fig. 16, Fig. 17).

In both cases a significant fraction of participants picked inappropriate answers, given the complexity of legal relations that may accompany even a simple dataset. For example, *sui generis* rights will rather belong to a research institute or consortium, not to individual researchers. Also, the collected data might include third party rights that have to be cleared before the data may be opened.

Fig. 11. I would be able to cite in my publication a dataset that was made publicly available on the Internet by other researchers and make it in accordance with one of the existing citation styles

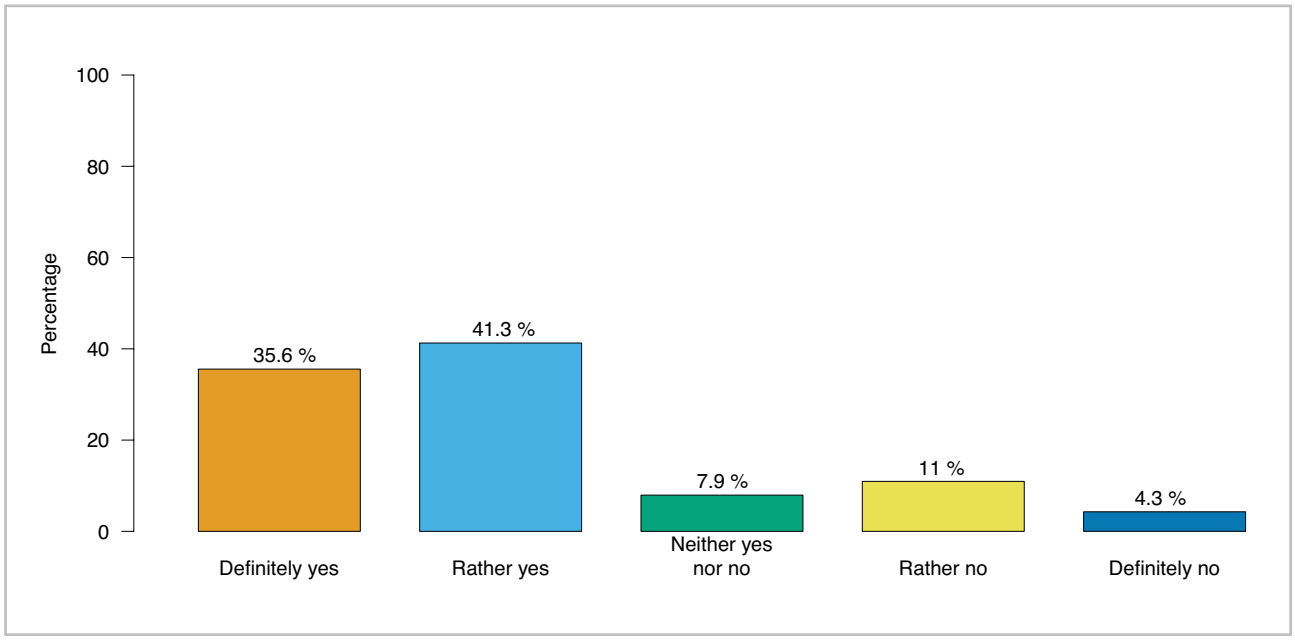


Fig. 12. „I would be able to cite in my publication a dataset that was made publicly available on the Internet by other researchers and make it in accordance with one of the existing citation styles" by „Scholarly degree or title"

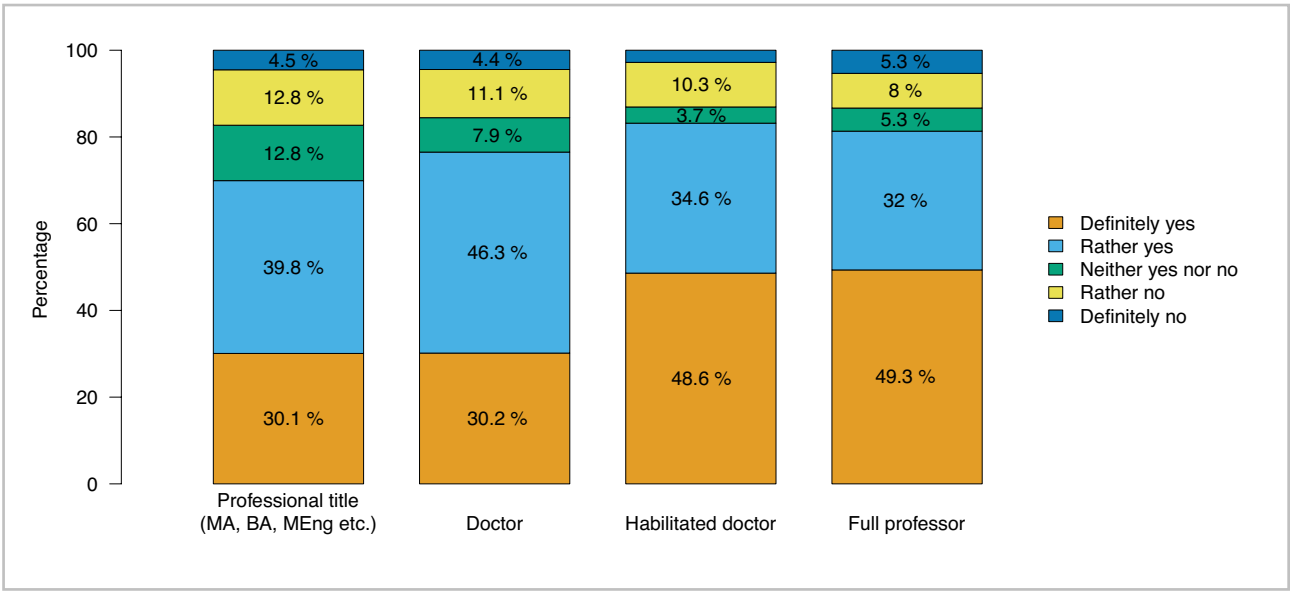


Fig. 13. If a scholarly journal requires that authors make data related to an article publicly available, I feel discouraged from publishing in this journal

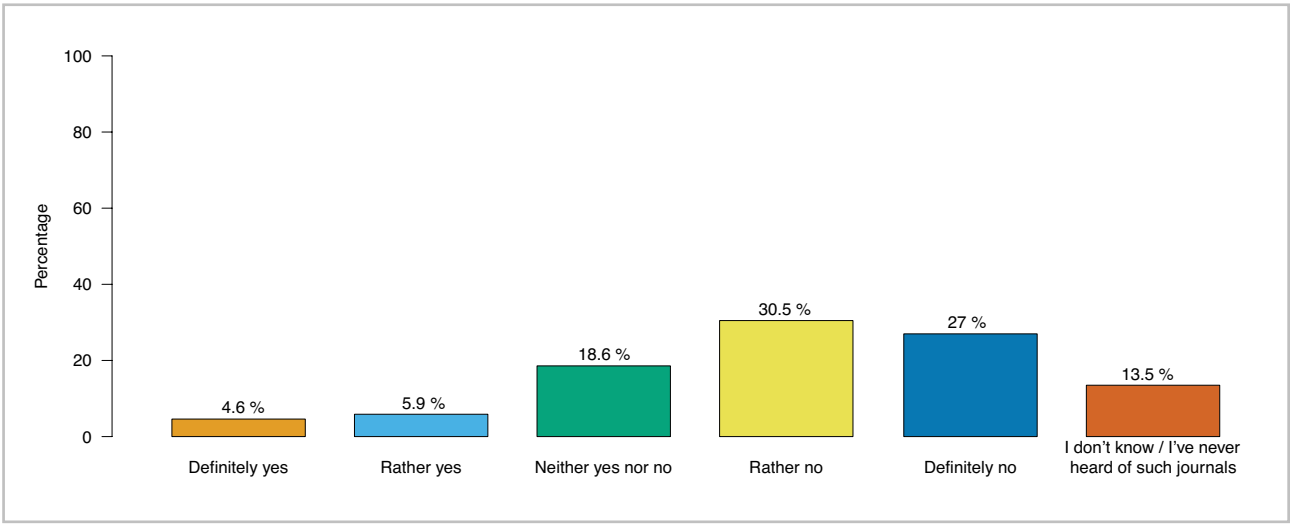


Fig. 14. According to the law, for research purposes you can use every dataset publicly available on the Internet, provided that you indicate its source

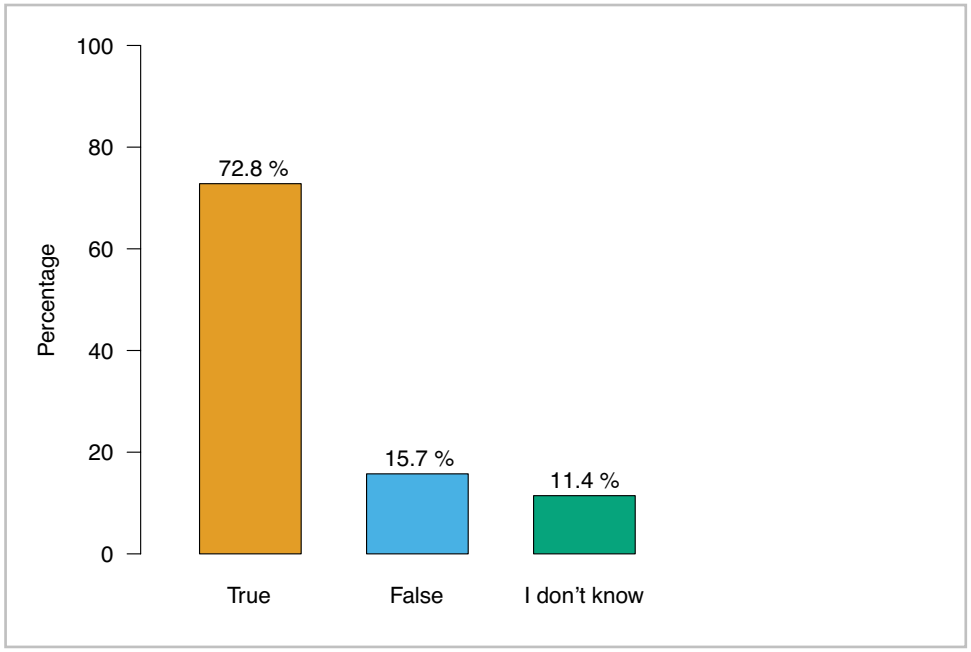


Fig. 15. The decision whether to share research data or not always lies solely with the researchers who produced the data

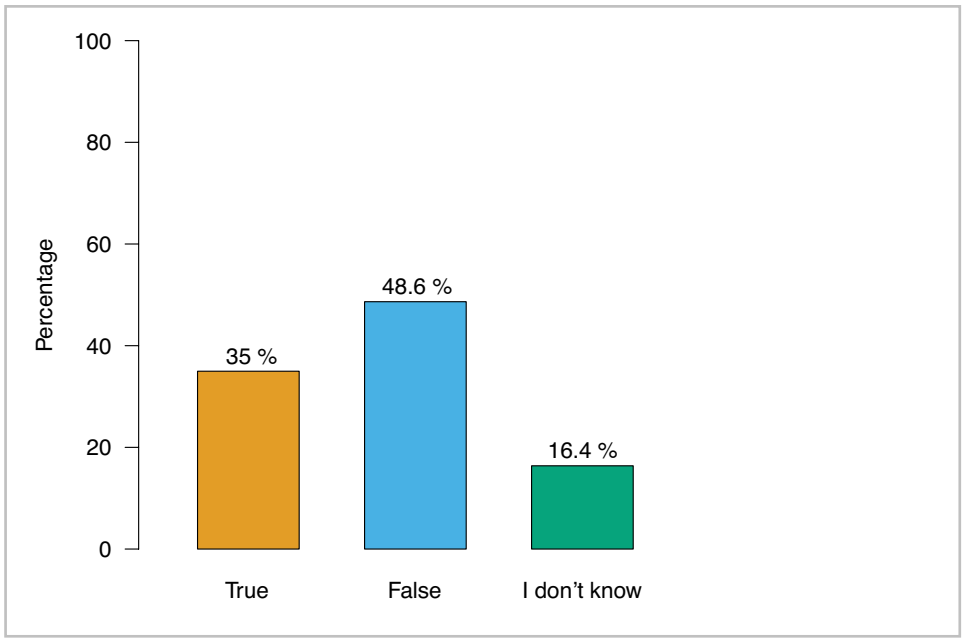


Fig. 16. „According to the law, for research purposes you can use every dataset publicly available on the Internet, provided that you indicate its source" by „Scholarly degree or title"

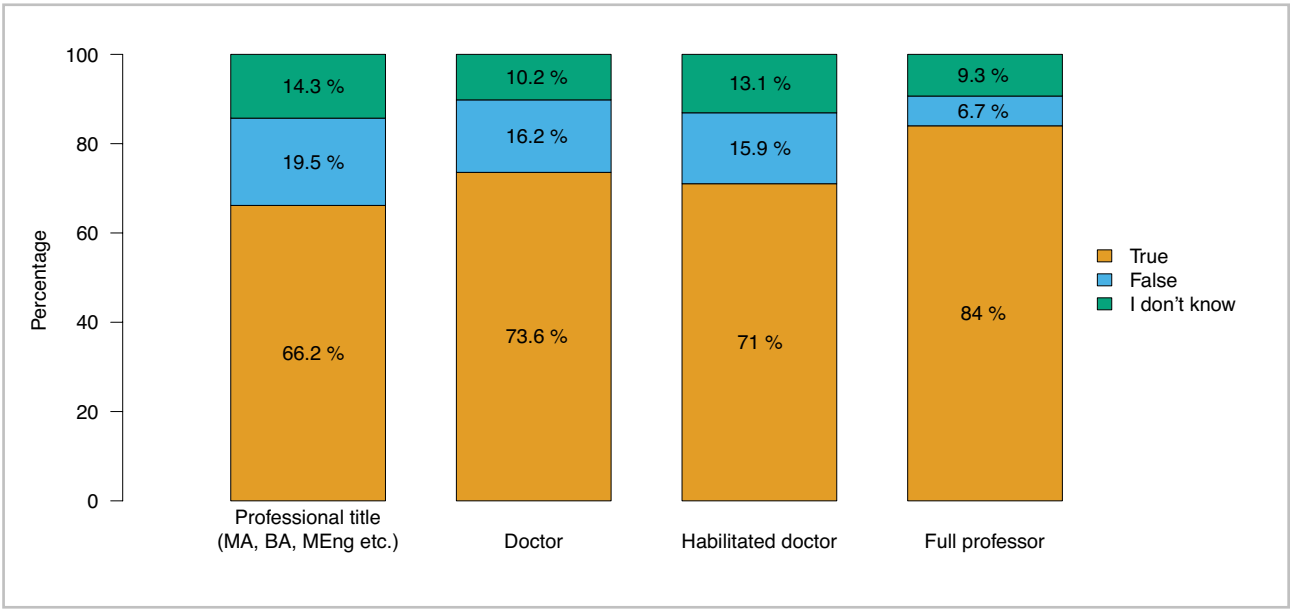
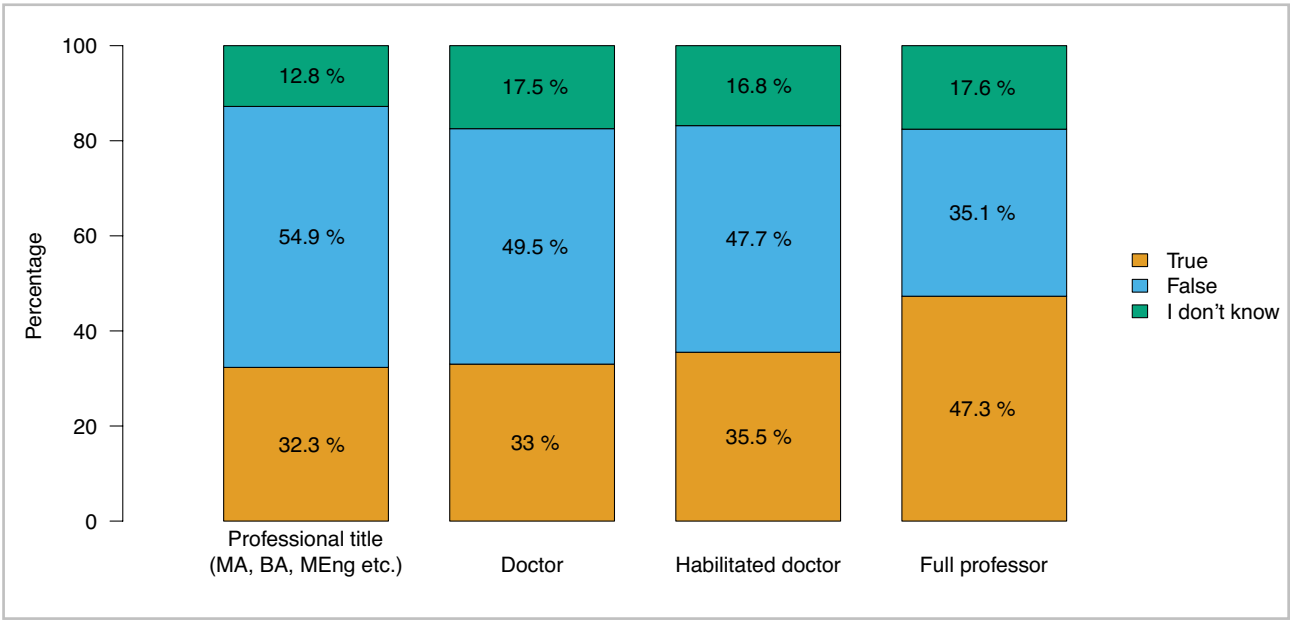


Fig. 17. „The decision whether to share research data or not always lies solely with the researchers who produced the data" by "Scholarly degree or title"



Enablers for sharing data

The questionnaire contained a set of questions concerning various factors that may act as enablers when making a decision about sharing or not sharing of research data. Here participants were asked to describe each factor as "Very important", "Rather important", "Neither important, nor not important", "Rather not important" or "Not at all important".

The enabler considered important by the largest fraction of respondents was that sharing data could bring a citation. The survey included two separate questions about citing: in one we asked if the citation of a dataset

is important, in the other if the citation of a publication describing it is important. Both questions resulted in similar answers (Fig. 18 and 19): in the first case 67.4% of the respondents assessed the factor as very important and 26.1% as rather important, while in the second case it was 71.7% and 23.8%, respectively. Having enough time to prepare all planned publications based on the data was also considered important by more than 90% of the participants (60.1% - very important, 30.1% - rather important - Fig. 20).

Another factor considered very important or rather important by the vast majority of participants (53.2% and 30.3%, respectively, Fig. 21), was whether sharing of the dataset would be taken into account during the evaluation of scholarly achievements. This factor was more important for early-career researchers (MAs) than for full professors (Fig. 22). The fact that sharing data will enable new scholarly contacts was considered very important by 48.3% of the respondents and as rather important by 38.6% (Fig. 23), while active technical support provided by an employer was important for 35.4% and 35.5% (Fig. 24), respectively. Setting the allowed and unallowed purposes of data reuse was very important to 39.7% and rather important to 34.8% of the respondents (Fig. 25).

Becoming a co-author of a publication resulting from research conducted by other researchers but based on the shared data would be very important to 33.8% and rather important to 34.8% of the respondents (Fig. 26), while establishing who will and who will not have access to the data was very important to 27.2% and rather important to 32.9% (Fig. 27).

Relatively the smallest fraction of respondents considered important receiving financial gratification for sharing data (19.7% - very important, 24.3% - rather important, Fig. 28). Here the largest fraction of participants chose the answer "neither important, nor not important", while almost 26% found this factor not important.

It is worth noting that factors considered important by the largest groups of survey participants are related to securing scholarly benefits which influence career development. These benefits may be connected either with sharing data (citations) or with not sharing (having enough time to publish all planned publications). Also, for the questions related to enablers of data sharing, in the majority of cases the joint frequency of answers indicating importance (or those indicating strong importance) was higher among women than among men (Fig. 29-36).

Fig. 18. If you were making a decision concerning whether or not to share your research data, how important would it be for you that thanks to sharing the data would be cited by other researchers?

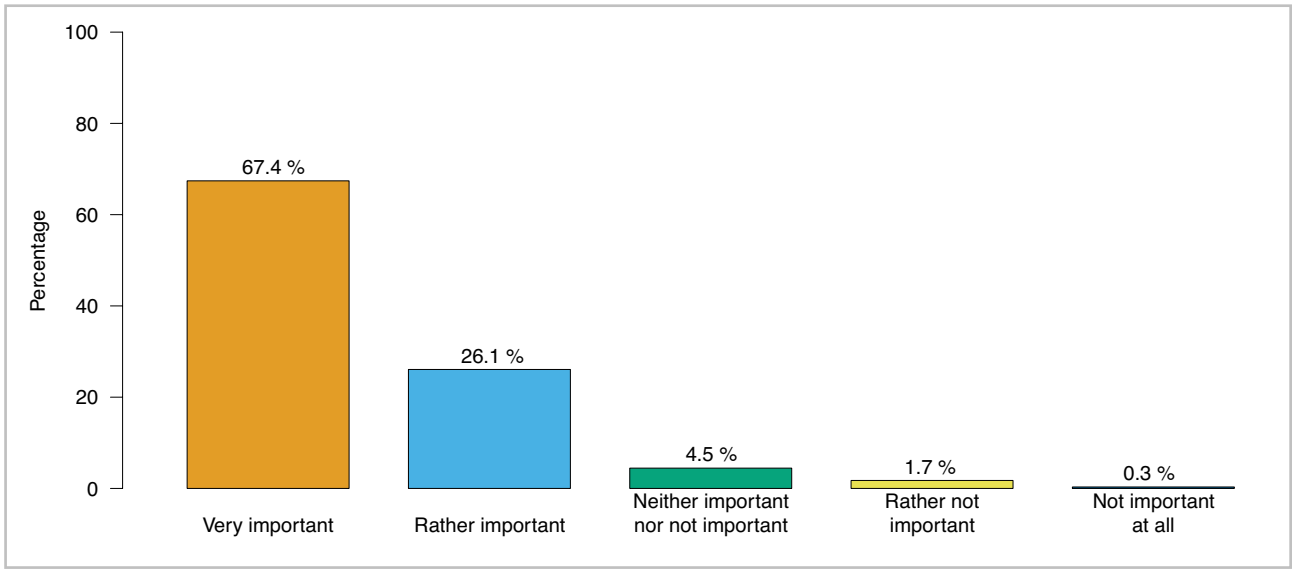


Fig. 19. If you were making a decision concerning whether or not to share your research data, how important would it be for you that thanks to sharing the data a publication describing it would be cited by other researchers?

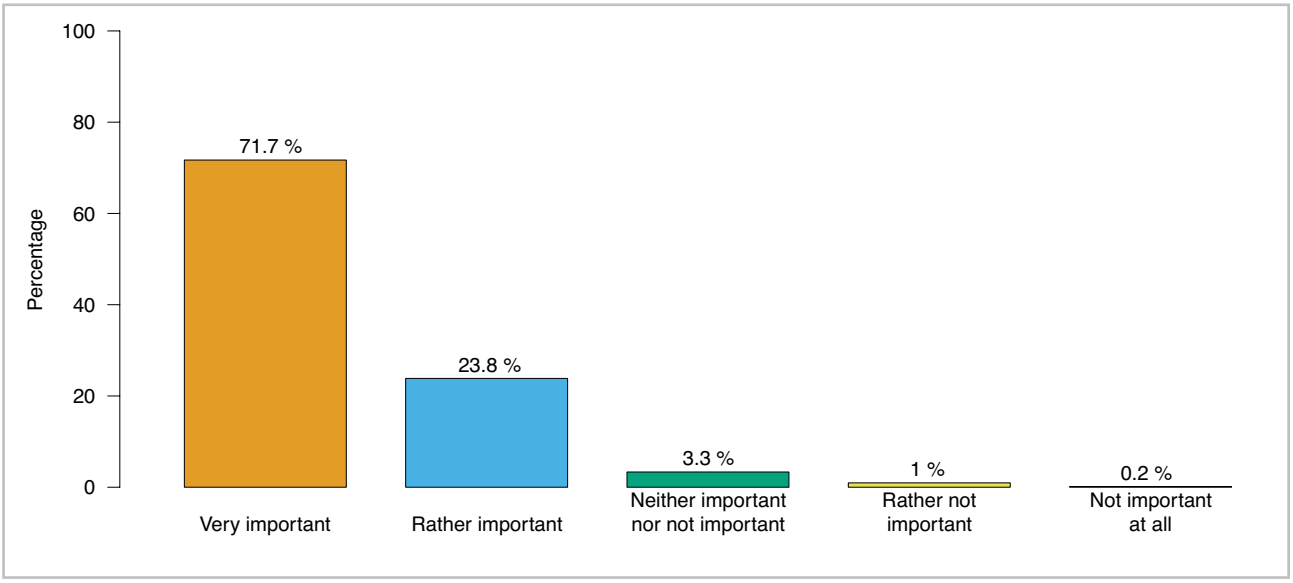


Fig. 20. If you were making a decision whether or not to share your research data, how important would it be for you that before you share the data you have enough time to prepare on its basis all planned publications

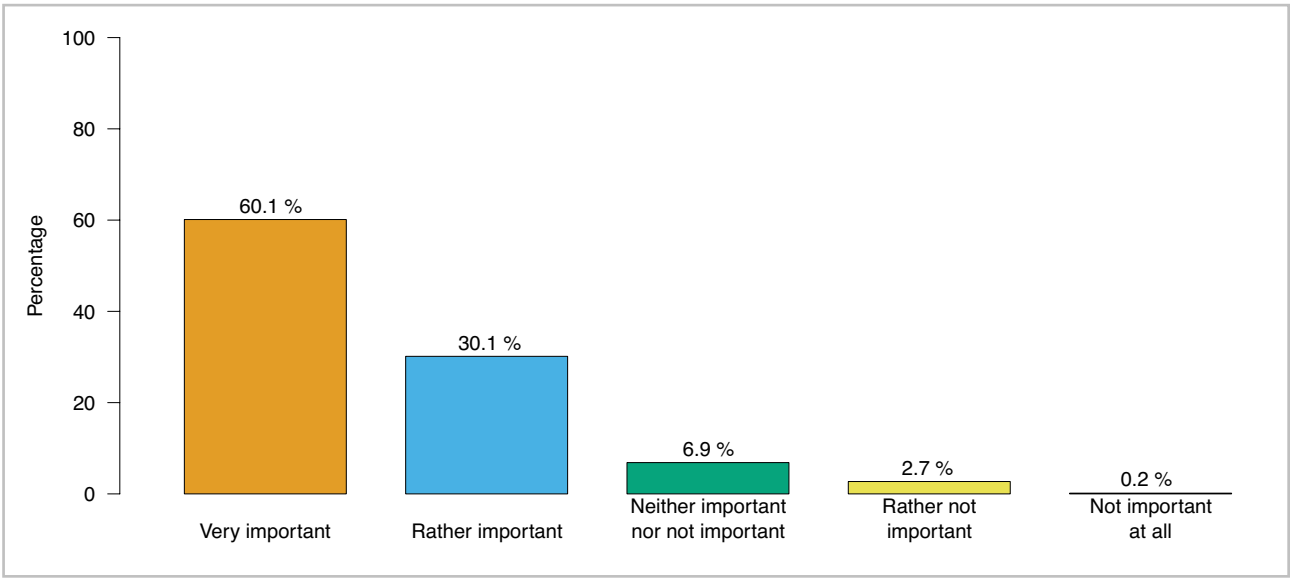


Fig. 21. If you were making a decision concerning whether or not to share your research data, how important would it be for you that the sharing is taken into account when evaluating your research achievements?

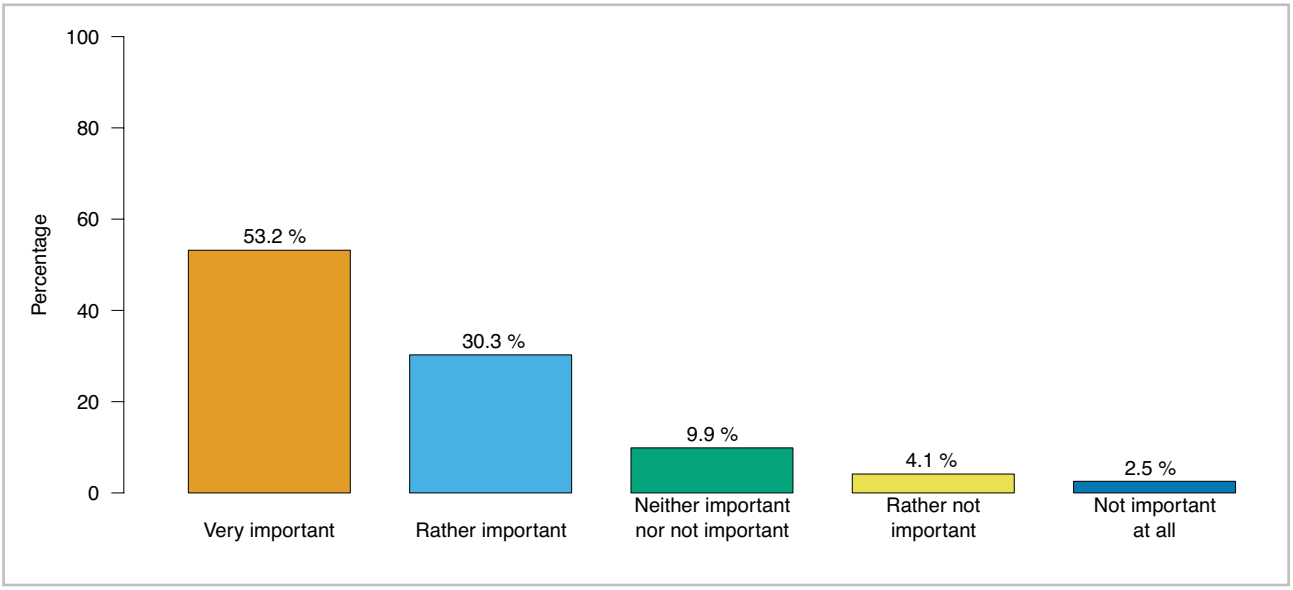


Fig. 22. „If you were making a decision concerning whether or not to share your research data, how important would it be for you that the sharing is taken into account when evaluating your research achievements?" by „Scholarly degree or title"

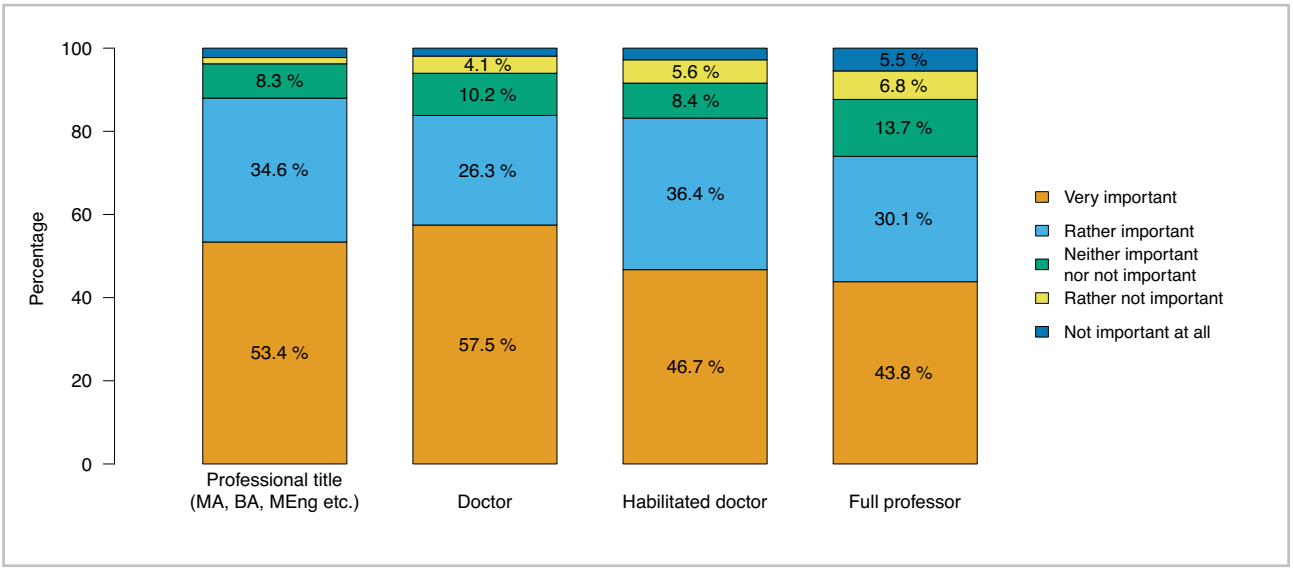


Fig. 23. If you were making a decision concerning whether or not to share your research data, how important would it be for you whether or not the sharing of this data would enable you to establish new contacts with other researchers?

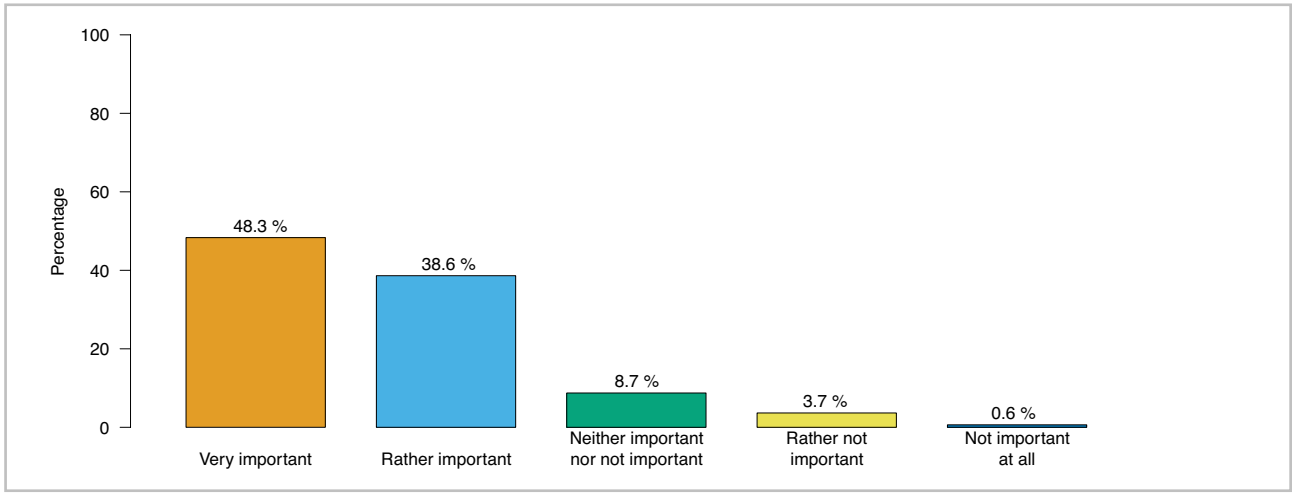


Fig. 24. If you were making a decision concerning whether or not to share your research data, how important would it be for you that you receive technical support from your employer?

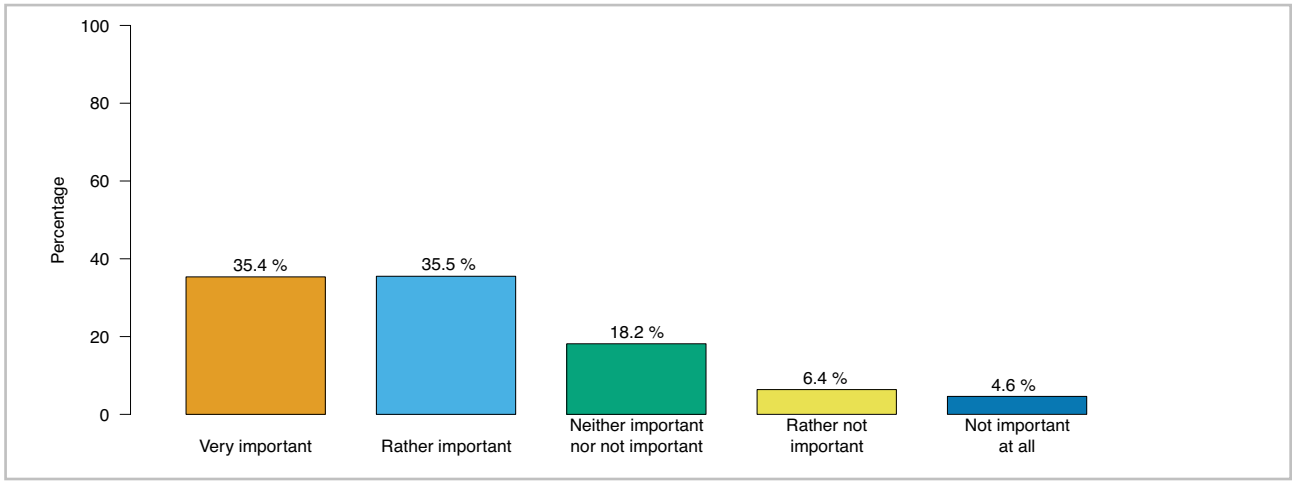


Fig. 25. If you were making a decision whether or not to share your research data, how important would it be for you to establish, for what purpose other users would be allowed to use the data, and for what purpose they would not?

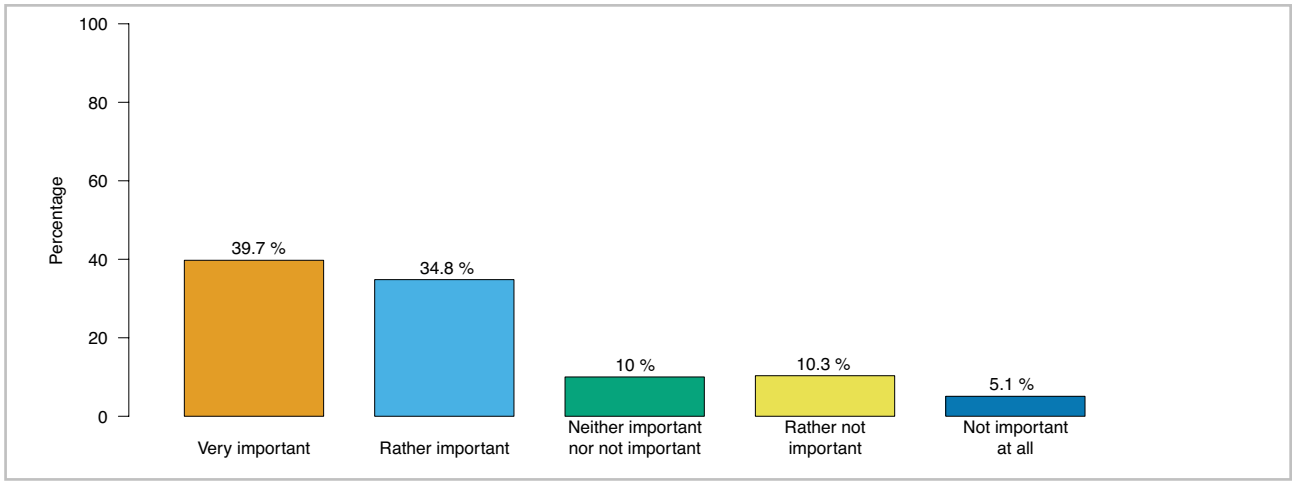


Fig. 26. If you were making a decision concerning whether or not to share your research data, how important would it be for you to become a coauthor of a publication resulting from research done by other researchers but based on the data you shared?

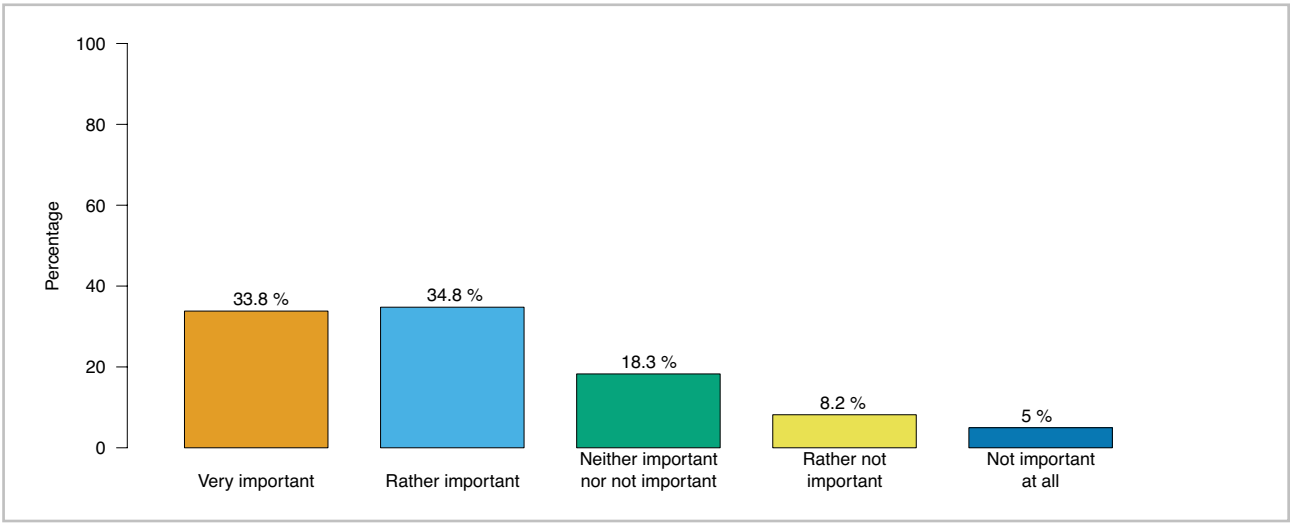


Fig. 27. If you were making a decision concerning whether or not to share your research data, how important would it be for you to establish who will have acces to the data and who will not?

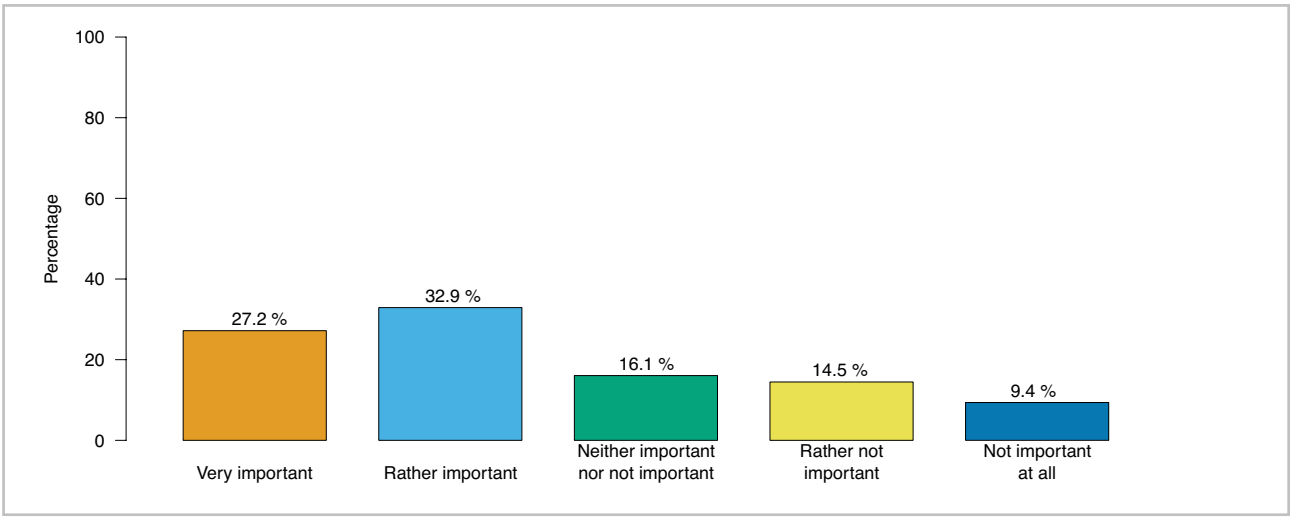


Fig. 28. If you were making a decision concerning whether or not to share your research data, how important would it be for you to receive a fee for sharing the data?

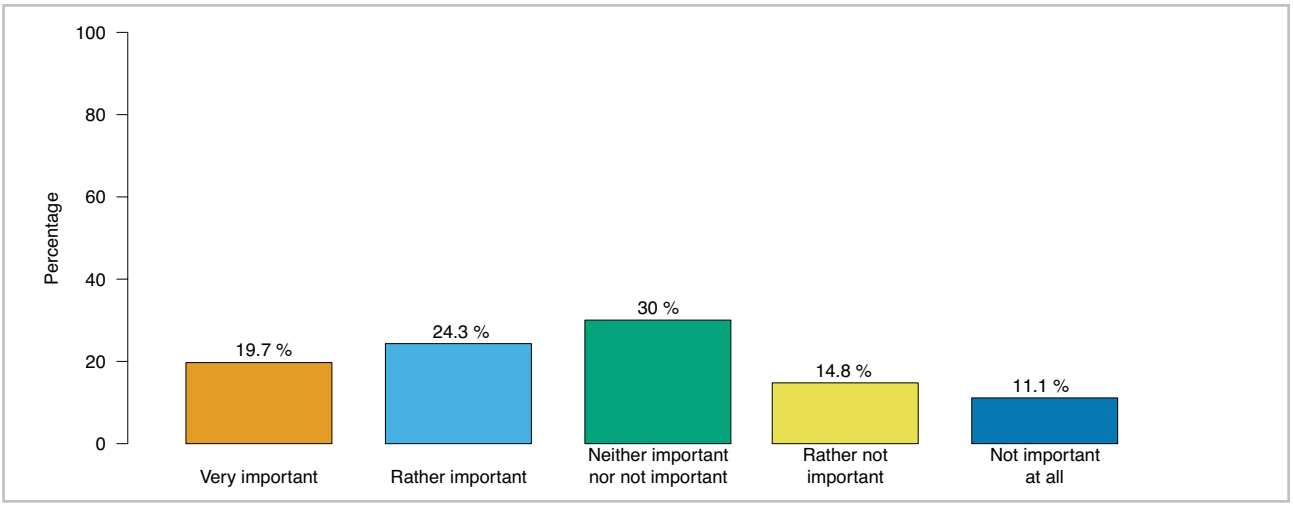


Fig. 29. „If you were making a decision whether or not to share your research data, how important would it be for you to establish, for what purpose other users would be allowed to use the data, and for what purpose they would not?" by „Gender"

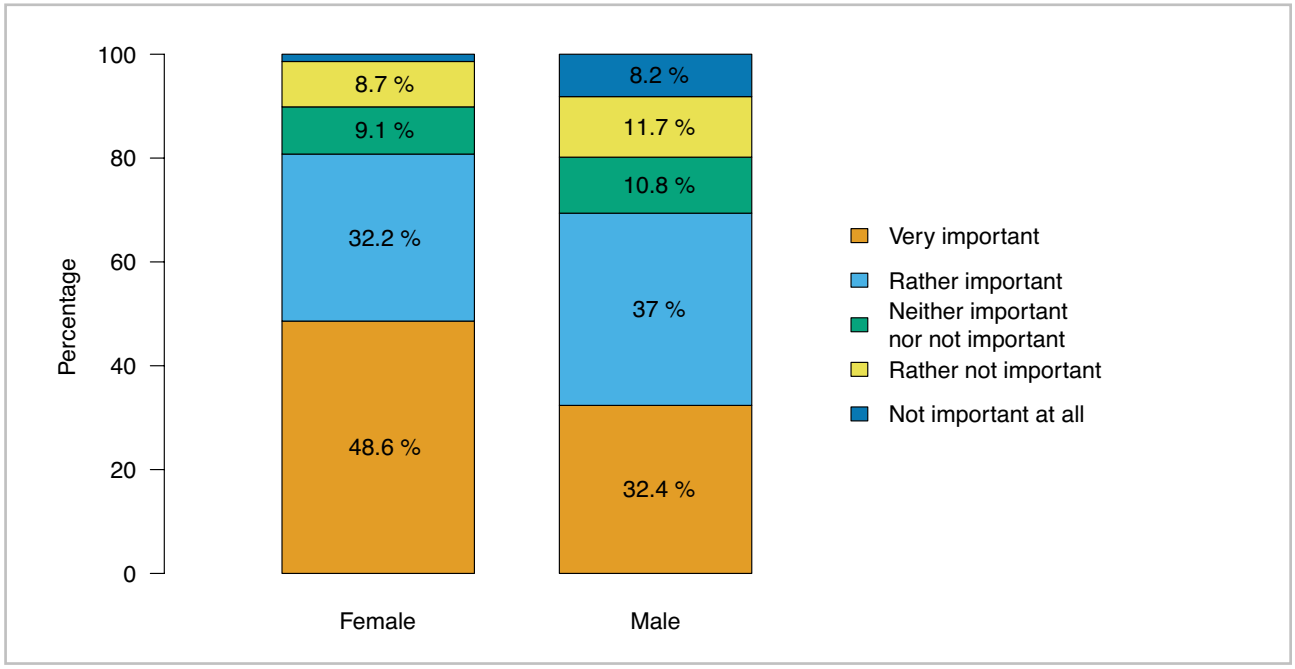


Fig. 30. „If you were making a decision concerning whether or not to share your research data, how important would it be for you to establish, who will have acces to the data and who will not?" by „Gender"

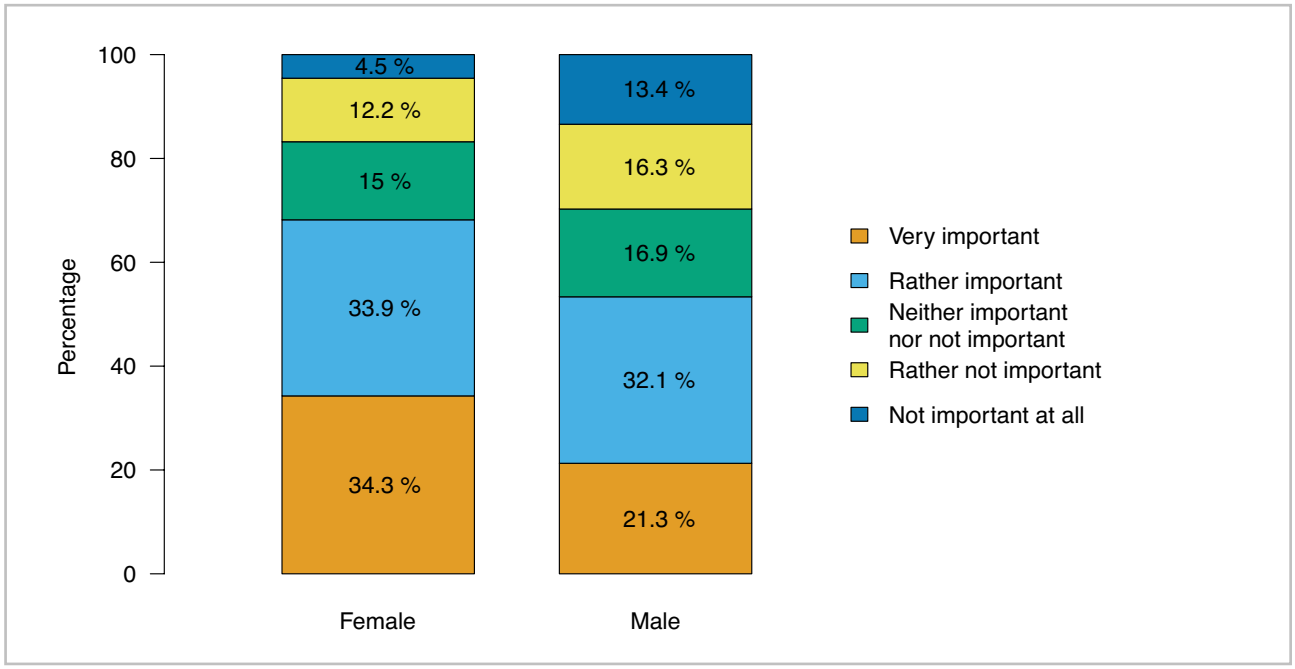


Fig. 31. „If you were making a decision concerning whether or not to share your research data, how important would it be for you whether or not the sharing of this data would enable you to establish new contacts with other researchers?" by „Gender"

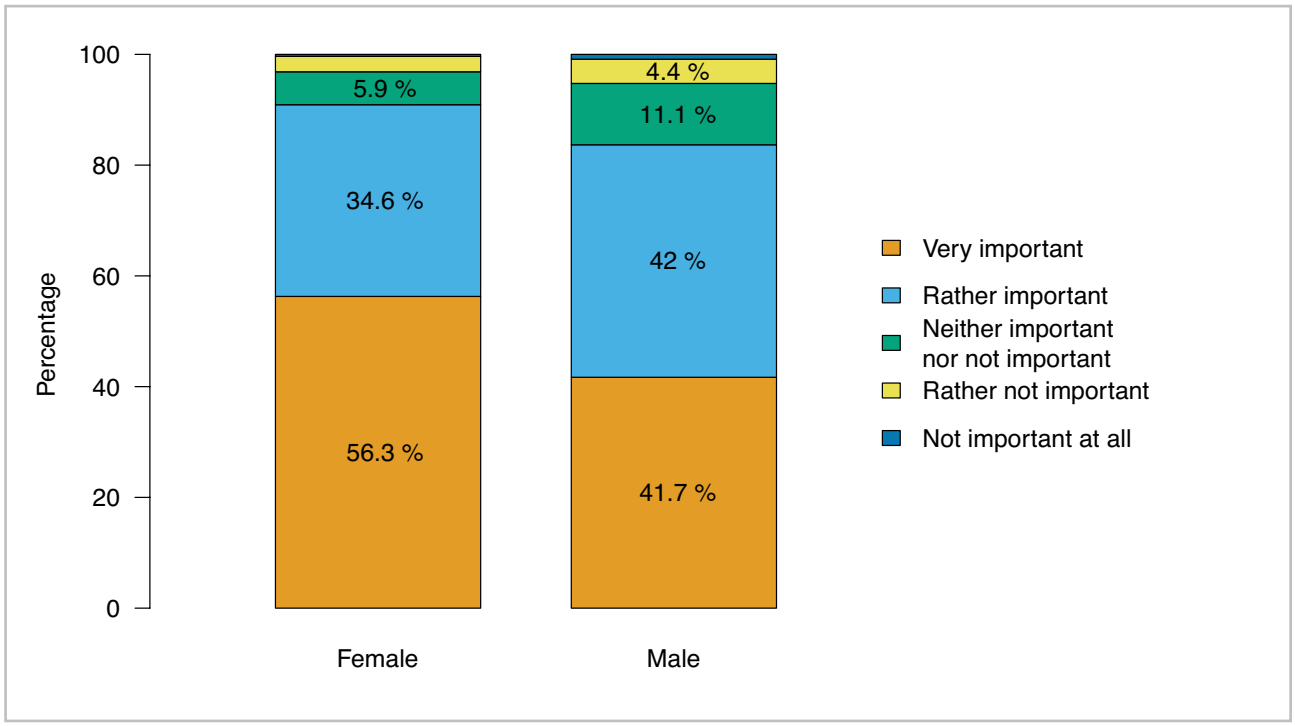


Fig. 32. „If you were making a decision concerning whether or not to share your research data, how important would it be for you that you receive technical support from your employer?" by „Gender"

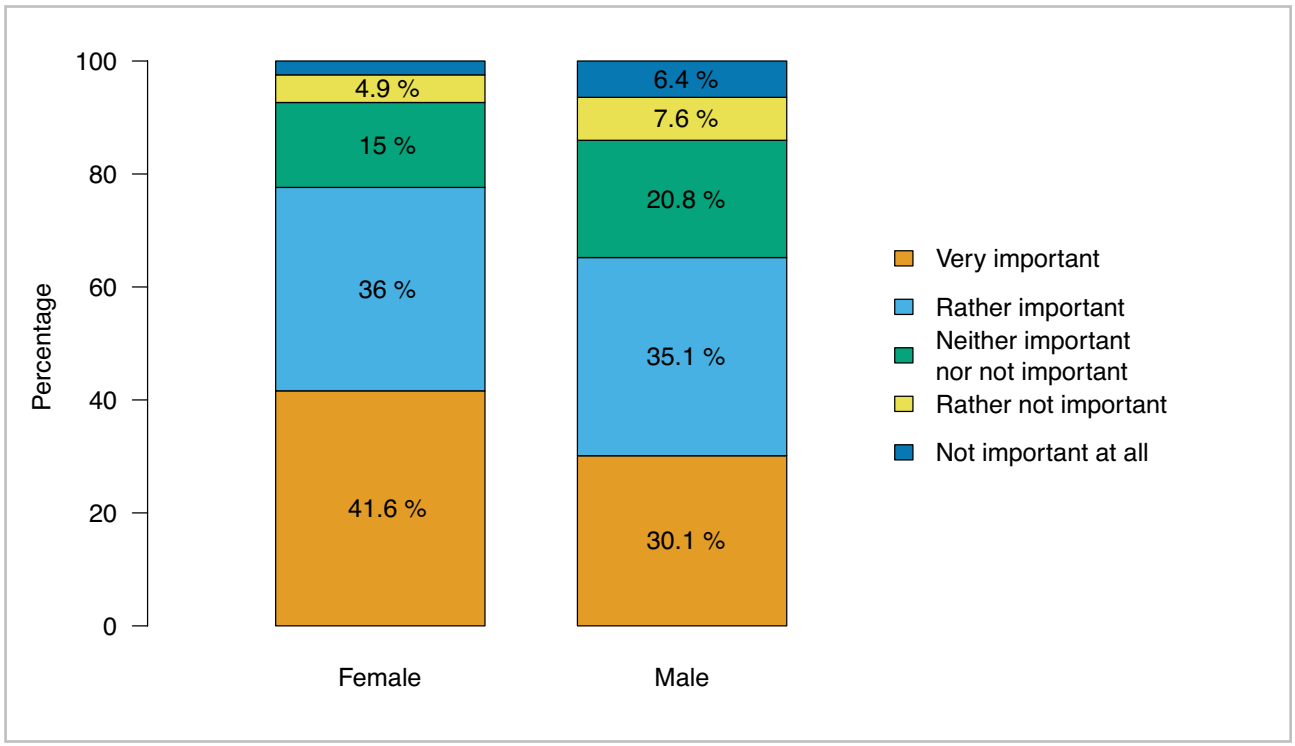


Fig. 33. „If you were making a decision concerning whether or not to share your research data, how important would it be for you that thanks to sharing the data would be cited by other researchers?" by "Gender"

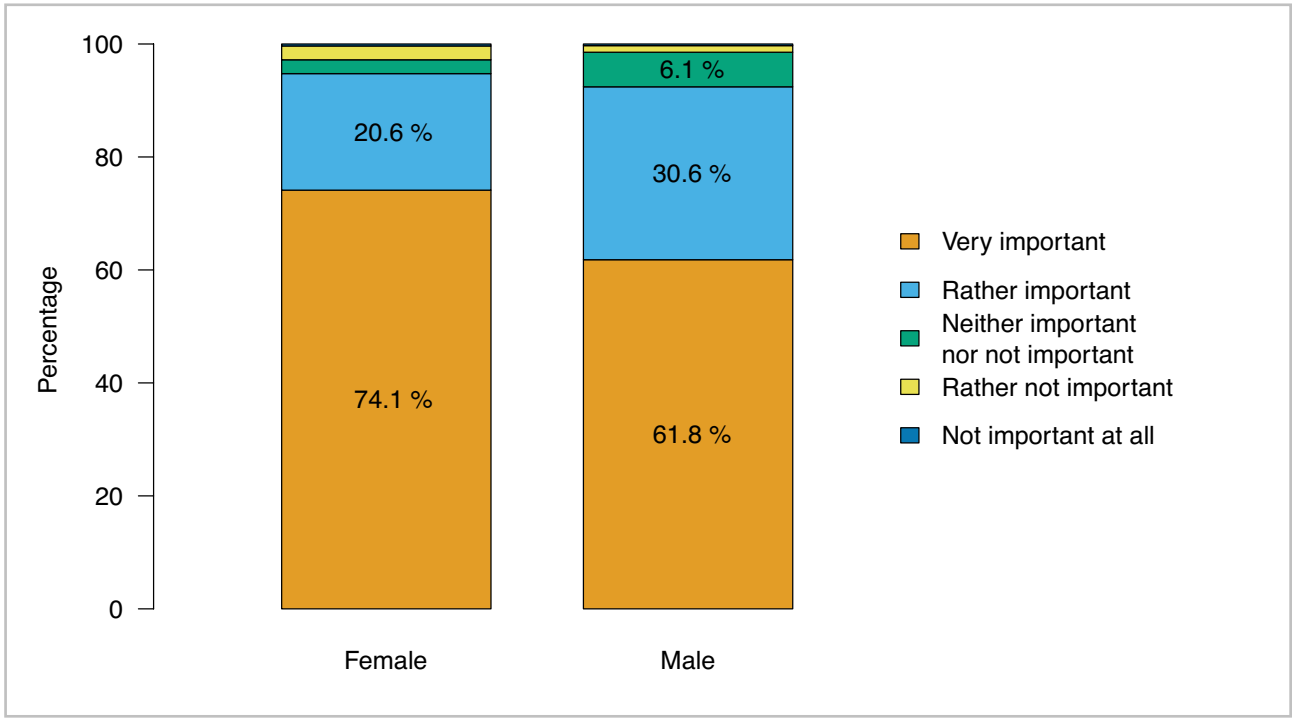


Fig. 34. „If you were making a decision concerning whether or not to share your research data, how important would it be for you that thanks to sharing the data a publication describing it would be cited by other researchers?" by "Gender"

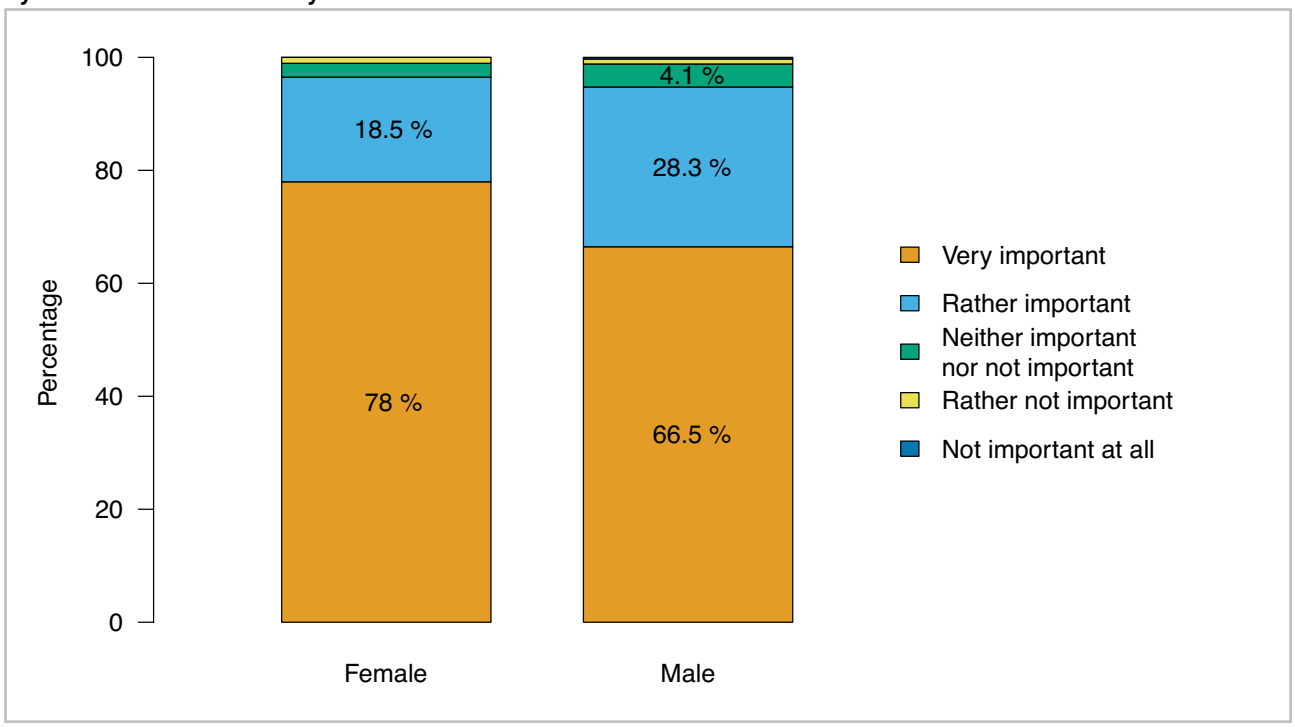


Fig. 35. „If you were making a decision whether or not to share your research data, how important would it be for you that before you share the data you have enough time to prepare on its basis all planned publications" by „Gender"

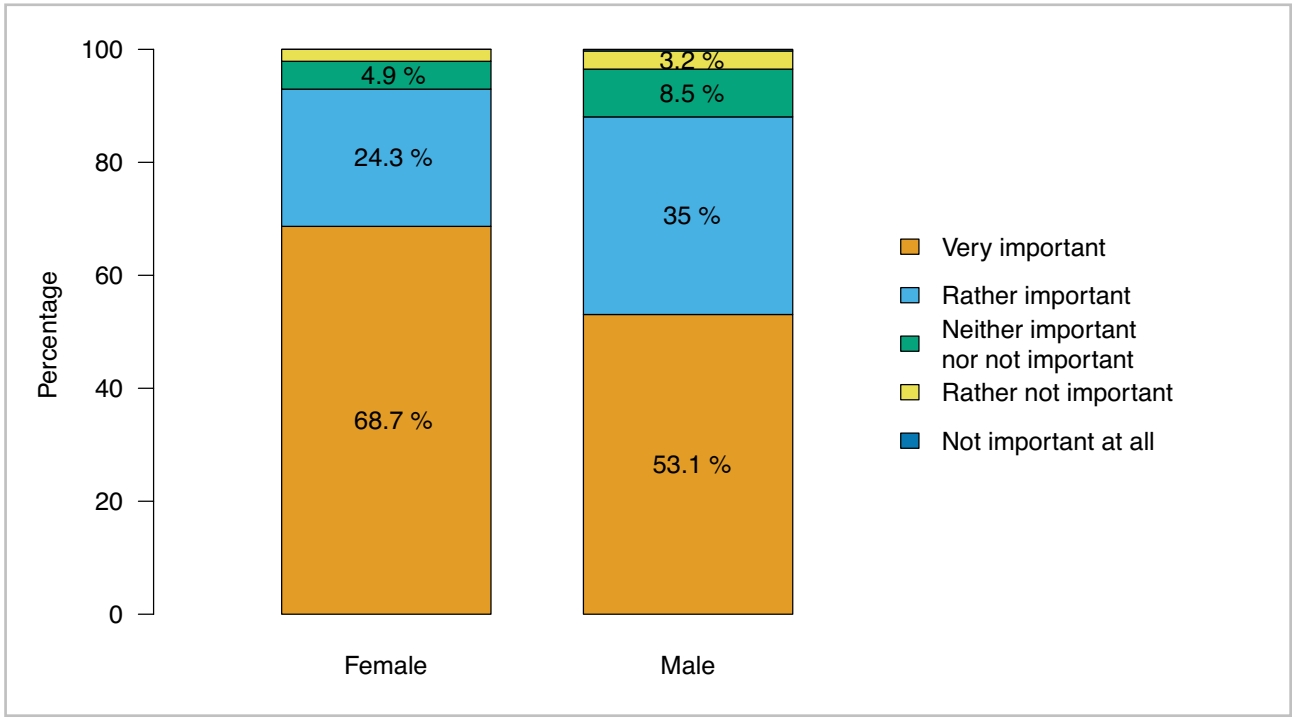
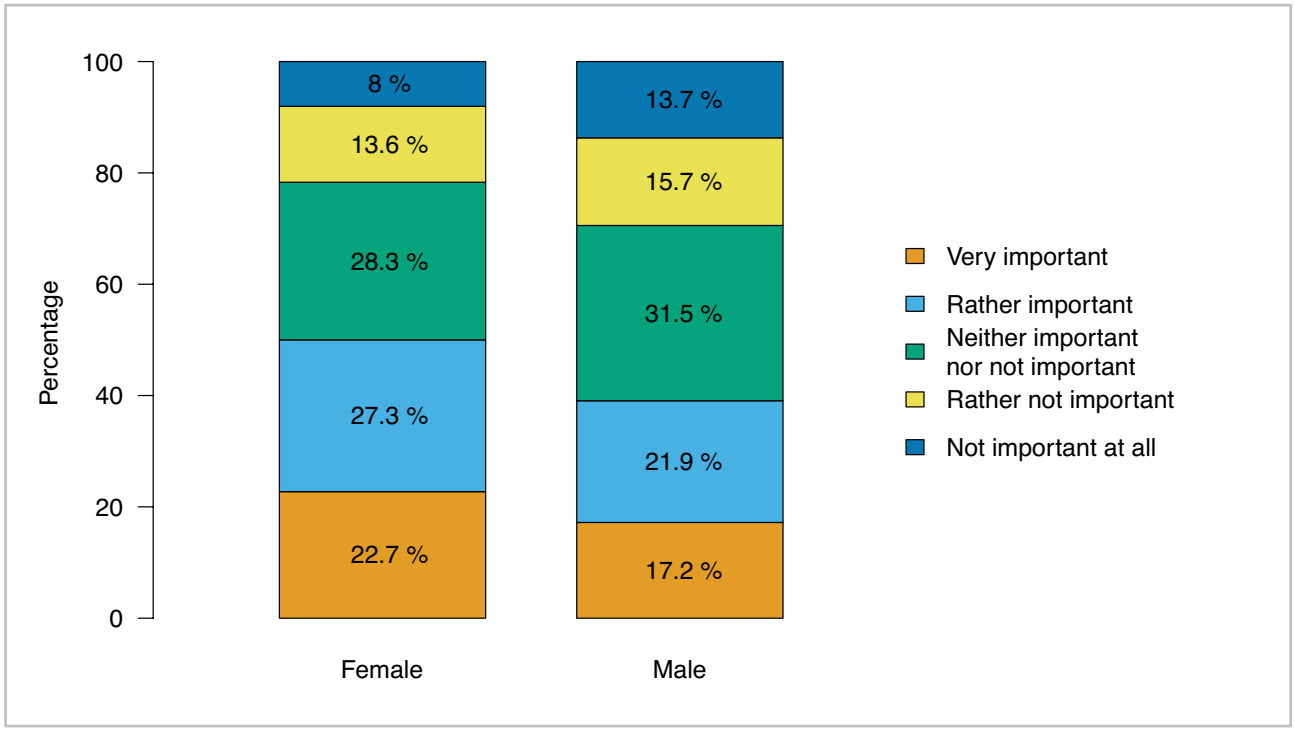


Fig. 36. „If you were making a decision concerning whether or not to share your research data, how important would it be for you to receive a fee for sharing the data?" by „Gender"



Factors hindering data sharing

The next section included questions about factors that hinder the sharing of data. Here, the respondents were asked if they agree or disagree with a series of statements saying that if it was up to them, they would share their data under given circumstances.

The observations from the previous section are corroborated by the answers to the first question in this section. The vast majority of participants disagreed (46.4% strongly and 33.1% rather - Fig. 37) with a sentence stating that they would share their data even if it would enable other researchers to publish before them on a topic they are planning to elaborate on. It was the only factor that was assessed by such a large fraction of respondents as something that could effectively stop them from sharing data. Also the possibility of incorrect interpretation of data transpired as a factor that could prevent a significant (though smaller) percentage of participants from sharing data: 19.9% and 16.6% of the respondents disagreed (strongly and rather, respectively) that they would share their data if there was a risk of misinterpretation (Fig. 38). On the other hand, 10.7% and 30.8% strongly agreed and rather agreed, and a significant fraction of respondents (22%) chose the neutral answer "neither agree, nor disagree".

In all other cases, the dominant fraction of respondents chose the answers "Strongly agree" or "Rather agree". Even under circumstances where there would be a possibility of criticizing or falsifying the research of a researcher who shares his/her own data, 26.4% strongly agreed and 37.3% rather agreed that they would share the data anyway (Fig. 39). A significant effort made to create the data also wouldn't prevent the majority of participants from sharing data (18.5% and 44.2% of the respondents chose the answers "Strongly agree" and "Rather agree", Fig. 40).

On the other hand, if sharing data required an effort, this could discourage 29.2% and 11.3% of the survey participants from sharing (answers "Rather disagree" and "Strongly disagree", respectively, while 4.6% and 31.2% of the respondents picked "Strongly agree" and "Rather agree", and 23.6% chose the neutral answer, Fig. 41). This may mean that for a significant fraction of respondents the easiness of sharing is important, as they do not want to put additional resources into activities necessary for sharing.

Fig. 37. If it were up to me, I would share my research data even if as a result someone else could use the data to publish an article or monograph on a subject that I was planning to work on myself

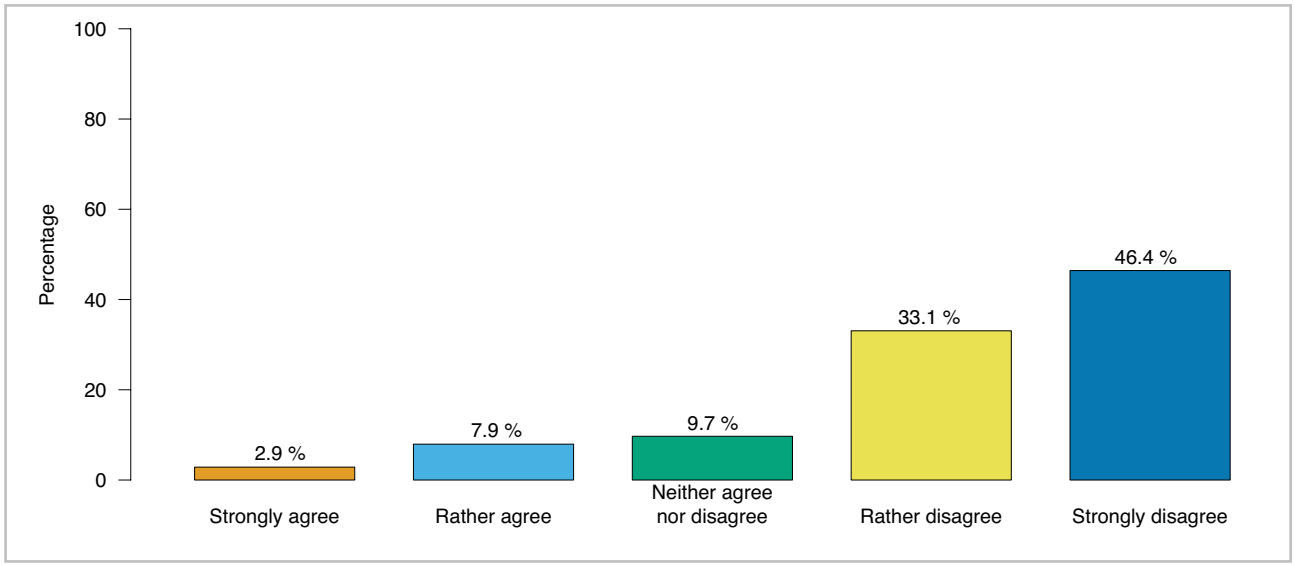


Fig. 38. If it were up to me, I would share my research data even if as a result it could be misinterpreted

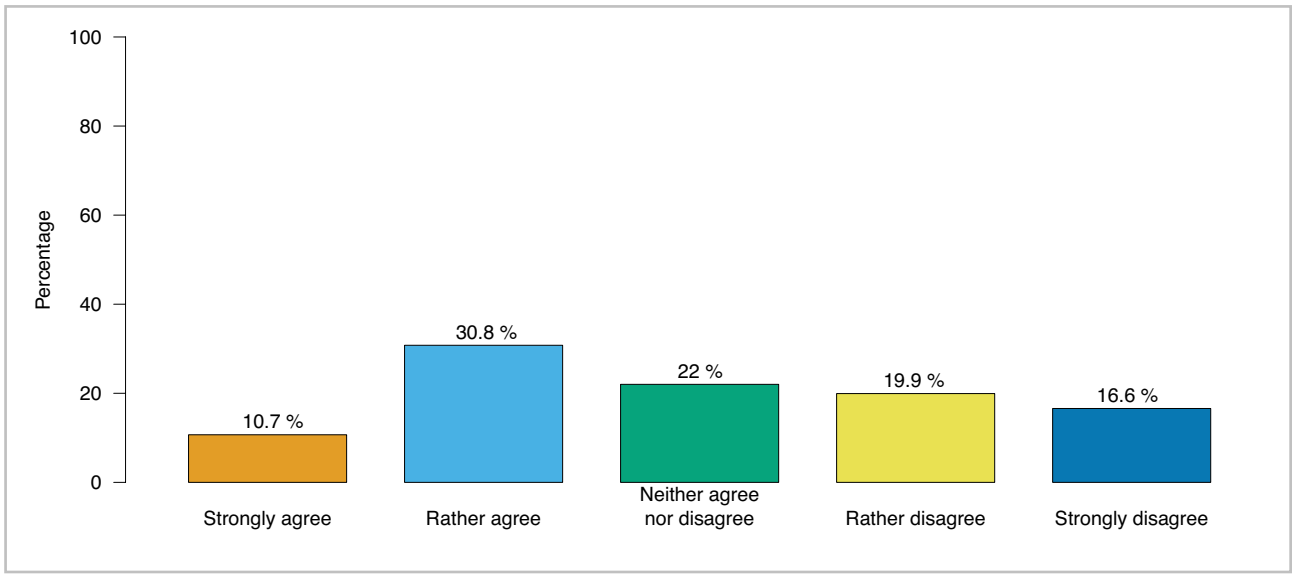


Fig. 39. If it were up to me, I would share my research data even if as a result others could criticize or falsify the results of my research

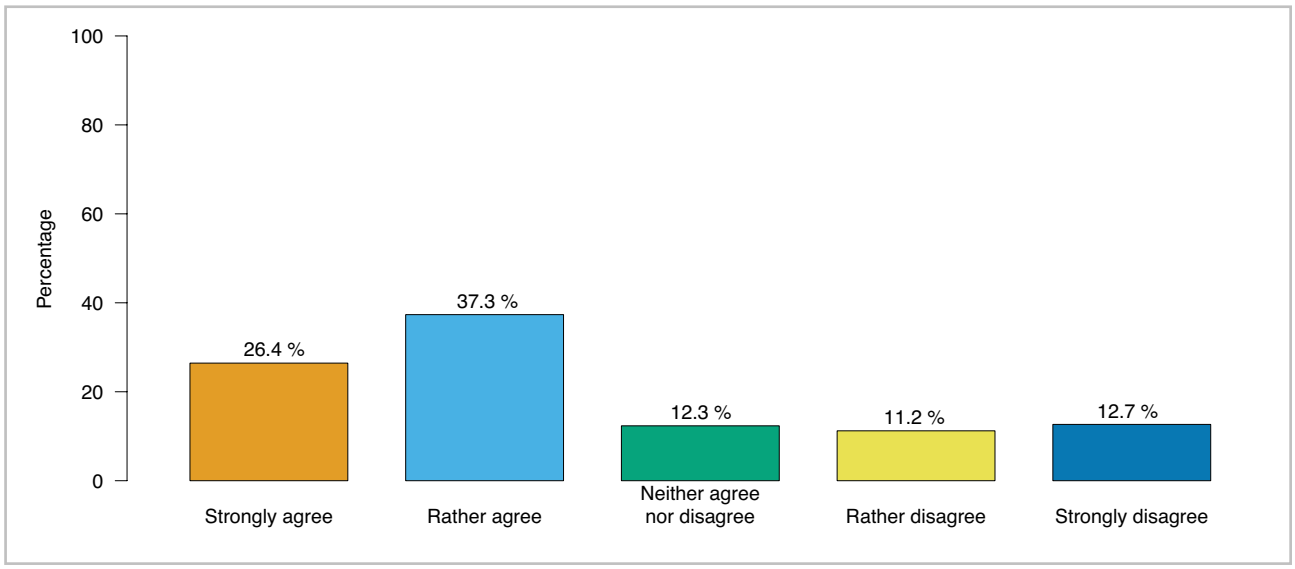


Fig. 40. If it were up to me, I would share my research data even if creating the dataset had required a significant effort

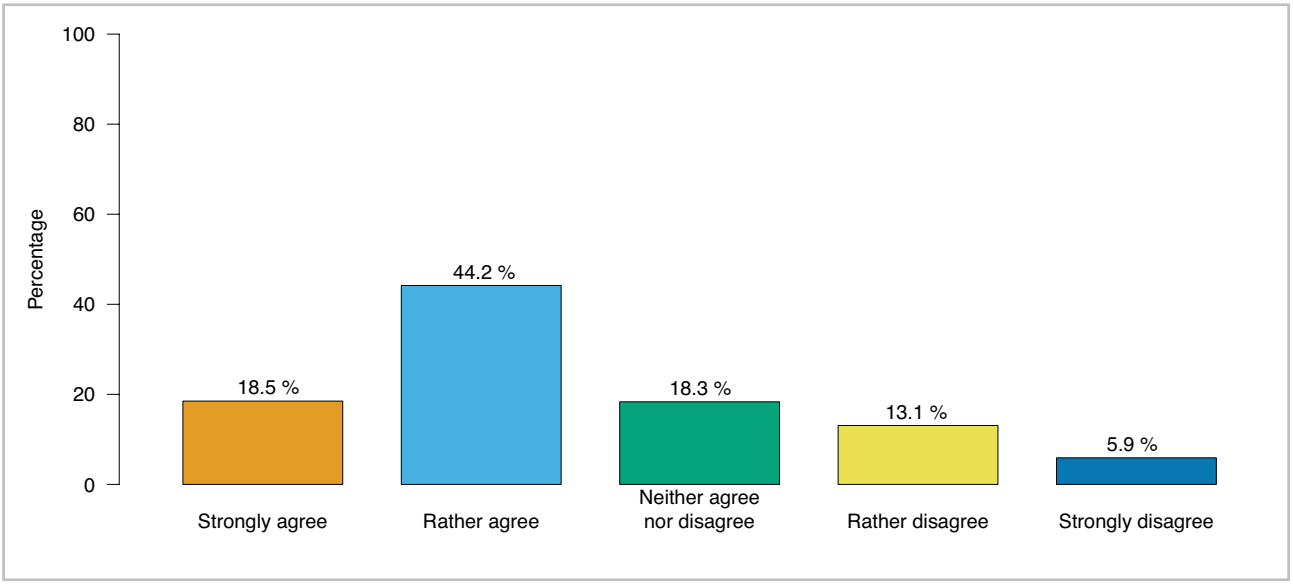
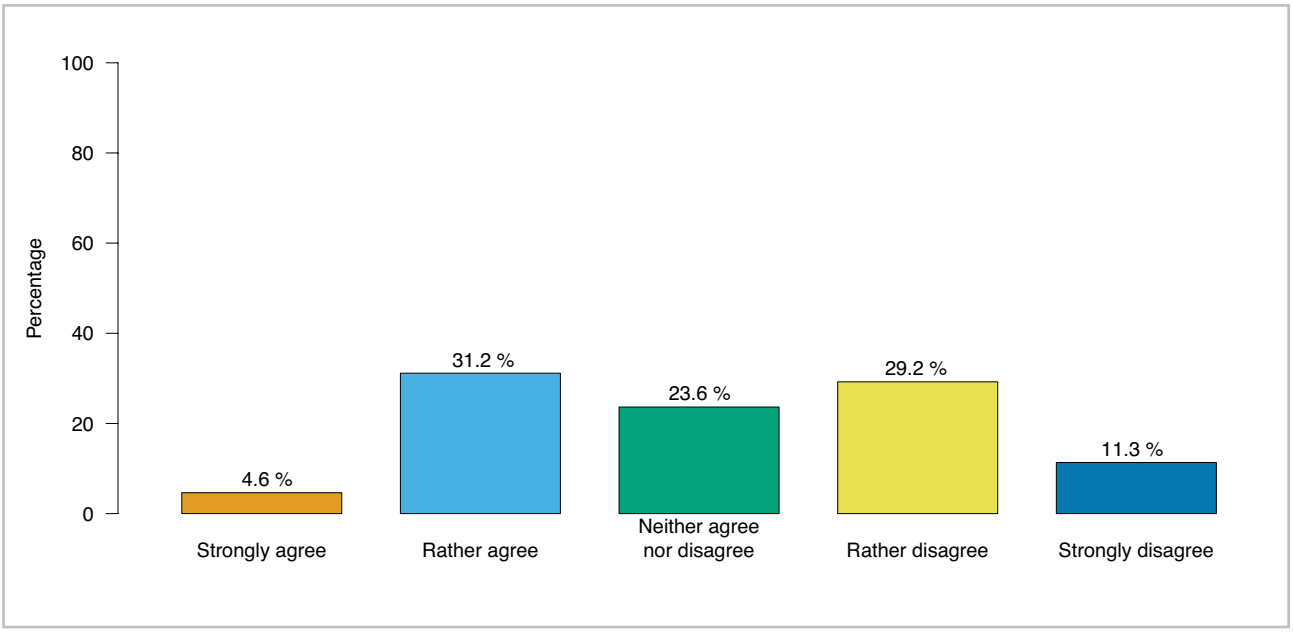


Fig. 41. If it were up to me, I would share my research data even if sharing it required a significant effort



Whom to share with?

The vast majority of respondents had no objections to sharing data with researchers known to them personally: almost 90% of them answered that if the decision concerning sharing or not sharing data was up to them, they would share it with such researchers (Fig. 42). Only slightly less respondents (47.8% strongly and 37.7% rather, Fig. 43) agreed that they would share it with researchers from the research unit they work in and with researchers working on similar topics (39.1% and 45.5%, Fig. 44). In this last case, the more cautious

answer ("rather") prevailed; this might be explained by the fact that a researcher working on similar topics may be perceived rather as a direct competitor than a colleague.

Importantly, questions about sharing data with researchers who conduct noncommercial research versus those conducting commercial research yielded significantly different distributions of answers. While more than 85% of the respondents agreed that they would share their data with researchers conducting noncommercial research (45.9% - definitely yes, 39.2% - rather yes, Fig. 45), the fraction of researchers willing to share with researchers conducting commercial research was significantly lower (13.1% - definitely yes; 30.4% - rather yes, Fig. 46). More than one third of them would not share data with such researchers (22% - rather not, 12.8% - definitely not, with the highest fraction of negative answers – 47.2% - among habilitated doctors, Fig. 47), while 21.7% picked the neutral answer "neither yes, nor no".

Almost half of the respondents declared that they would publicly share their data with everyone without any exceptions (14.8% of them answered "Definitely yes", while 33.3% "Rather yes", Fig. 48), while about one third chose "Rather not" (16.4%) or "Definitely not" (12.6%). The neutral answer "Neither yes nor no" was also picked by a relatively significant fraction of respondents (22.8%).

Fig. 42. If the decision concerning whether or not to share your research data lied with you, would you be ready to share it with researchers known to you personally?

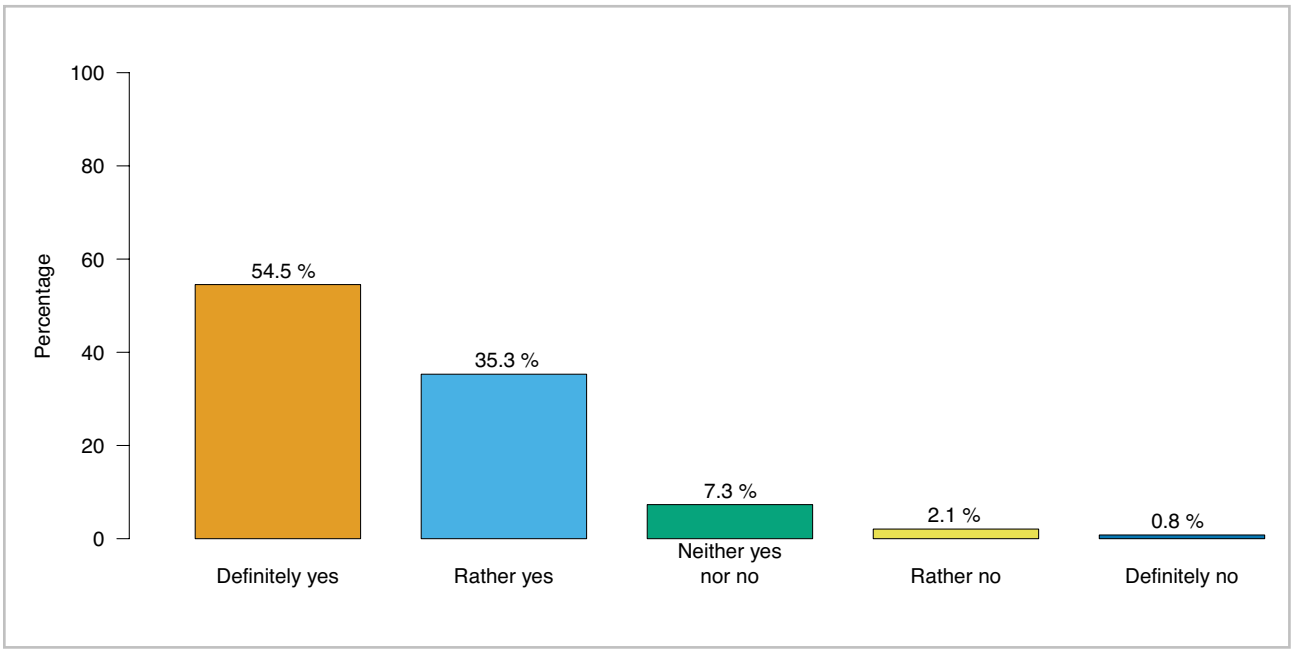


Fig. 43. If the decision concerning whether or not to share your research data lied with you, would you be ready to share it with researchers working at your research unit?

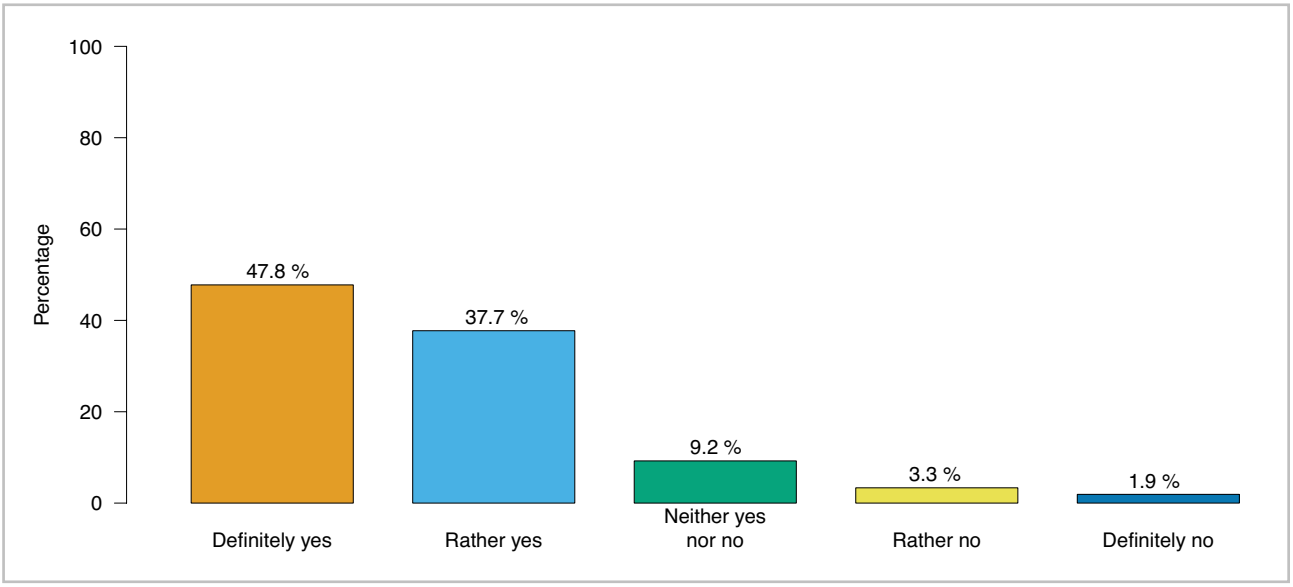


Fig. 44. If the decision concerning whether or not to share your research data lied with you, would you be ready to share it with researchers working on similar topics?

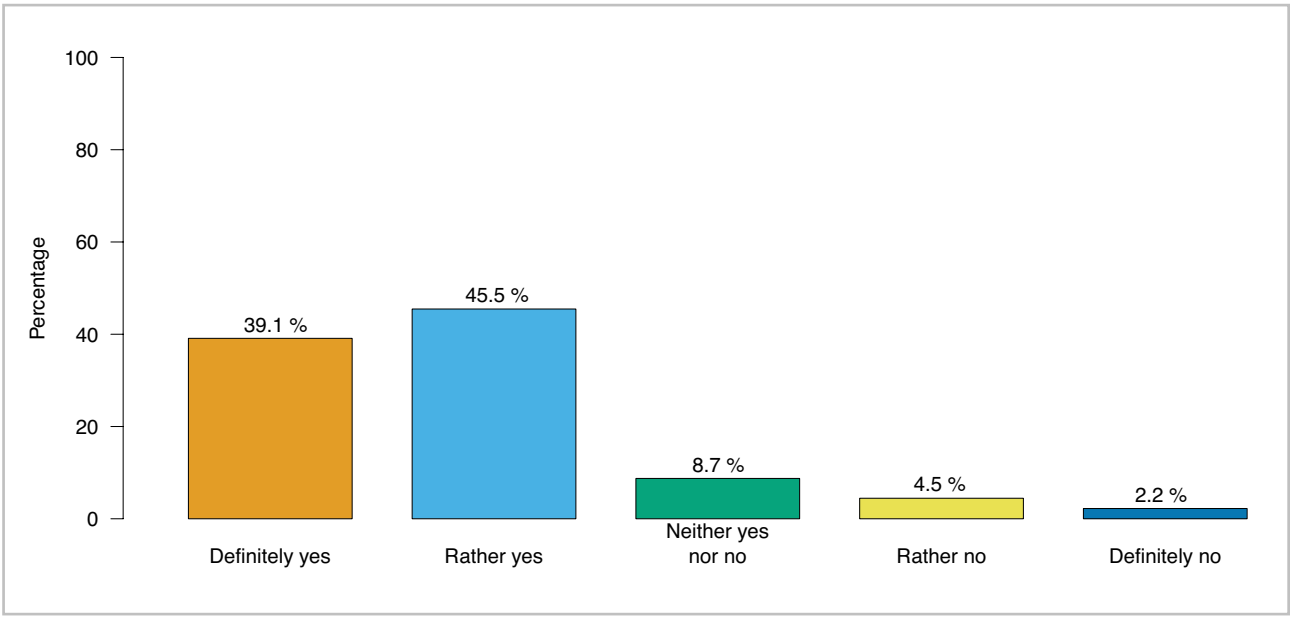


Fig. 45. If the decision concerning whether or not to share your research data lied with you, would you be ready to share it with researchers conducting noncommercial research?

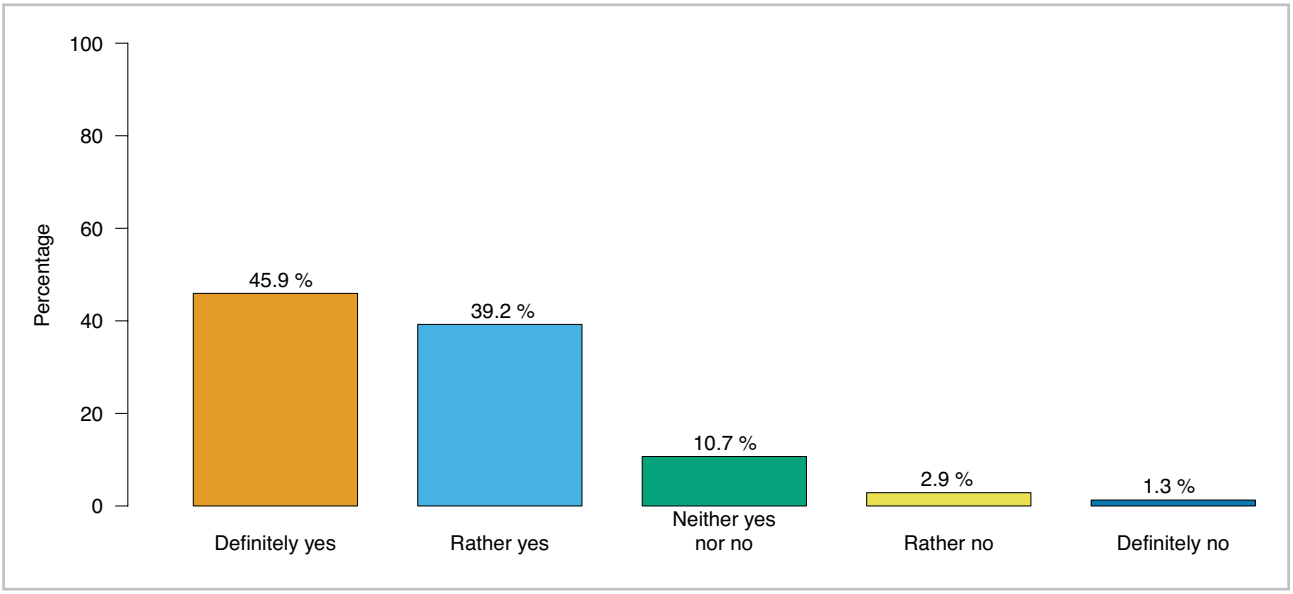


Fig. 46. If the decision concerning whether or not to share your research data lied with you, would you be ready to share it with researchers conducting commercial research?

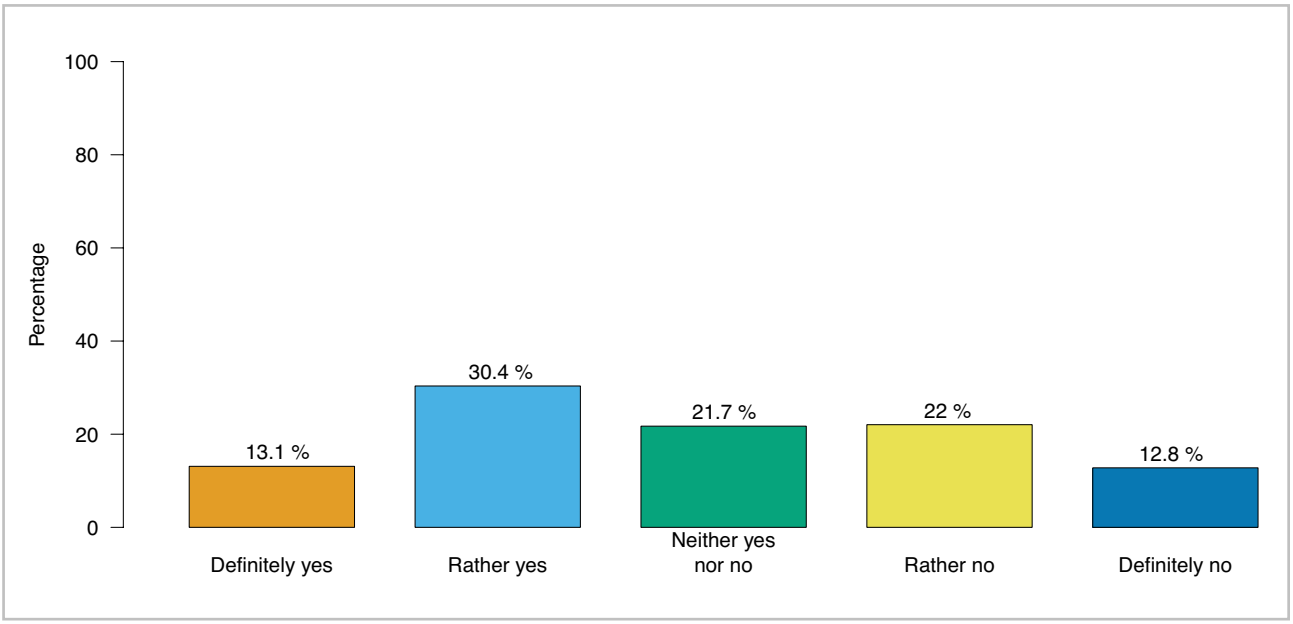


Fig. 47. „If the decision concerning whether or not to share your research data lied with you, would you be ready to share it with researchers conducting commercial research?" by „Scholarly degree or title"

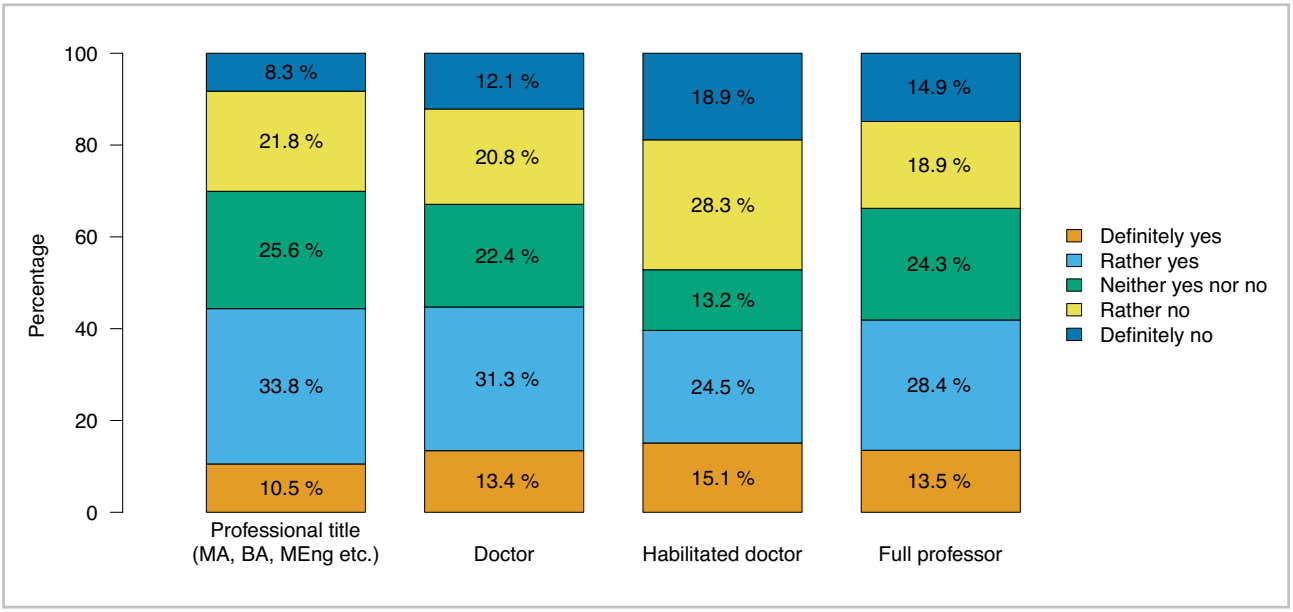
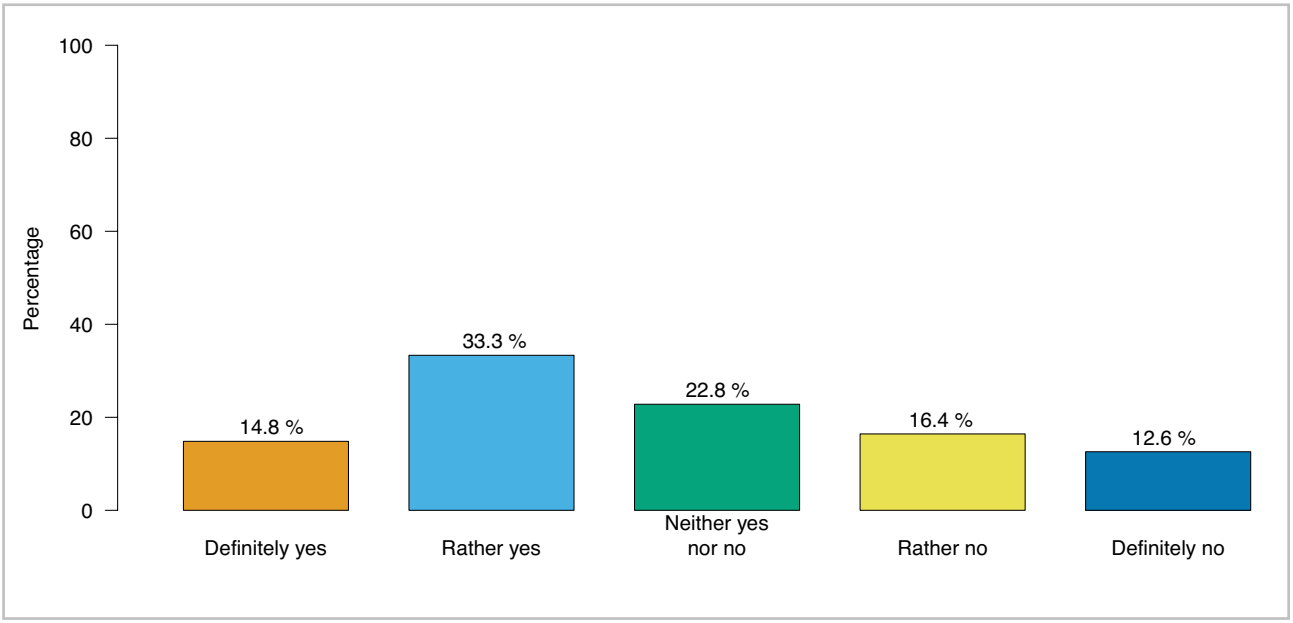


Fig. 48. If the decision concerning whether or not to share your research data lied with you, would you be ready to share it publicly with everyone without any exceptions?



Using data shared by others

The respondents were also asked about the importance of various factors they may take into account when deciding whether to use data shared by others. Three factors were considered especially important by the respondents.

Nearly all respondents (98.4%, Fig. 49) considered important that it is clear how one can legally use the data (85.1% - very important, 13.3% - rather important). Almost the same fraction of respondents (98.2%, Fig.

50) acknowledged the importance of data being produced by reliable people and institutions (83.6% - very important). For 95.5% (Fig. 51) it is important that a dataset has exhaustive documentation (71.5% - very important).

Further, over 90% (Fig. 52) of the respondents declared that it is important for them that the data they are considering to use would let them arrive at new, original research results (69.6% - very important, 25.7% - rather important).

In the rest of the cases, although still a vast majority of respondents confirmed the importance of the proposed factors, the proportion of answers "very important" and "rather important" was much more balanced. And so, 82.8% of the respondents answered that it is important that there is a contact person for the dataset, but here the number of answers "very important" was almost the same as "rather important" (41.1% and 41.7%, respectively, Fig. 53). Also the easiness of using a dataset is important, but not to everyone (Fig. 54): 28.3% chose very important, 47.8% - rather important, while 18.9% picked a neutral answer to this question, and 5% stated that it is not important for them.

Within this group of questions, although the joint fraction of answers pointing to the importance of certain factors was very similar among men and women, the latter usually gave about 10-15 percent more answers indicating the strong importance of the discussed factors (see Fig. 57-64).

Fig. 49. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that it were clearly stated how the data can be used according to the law?

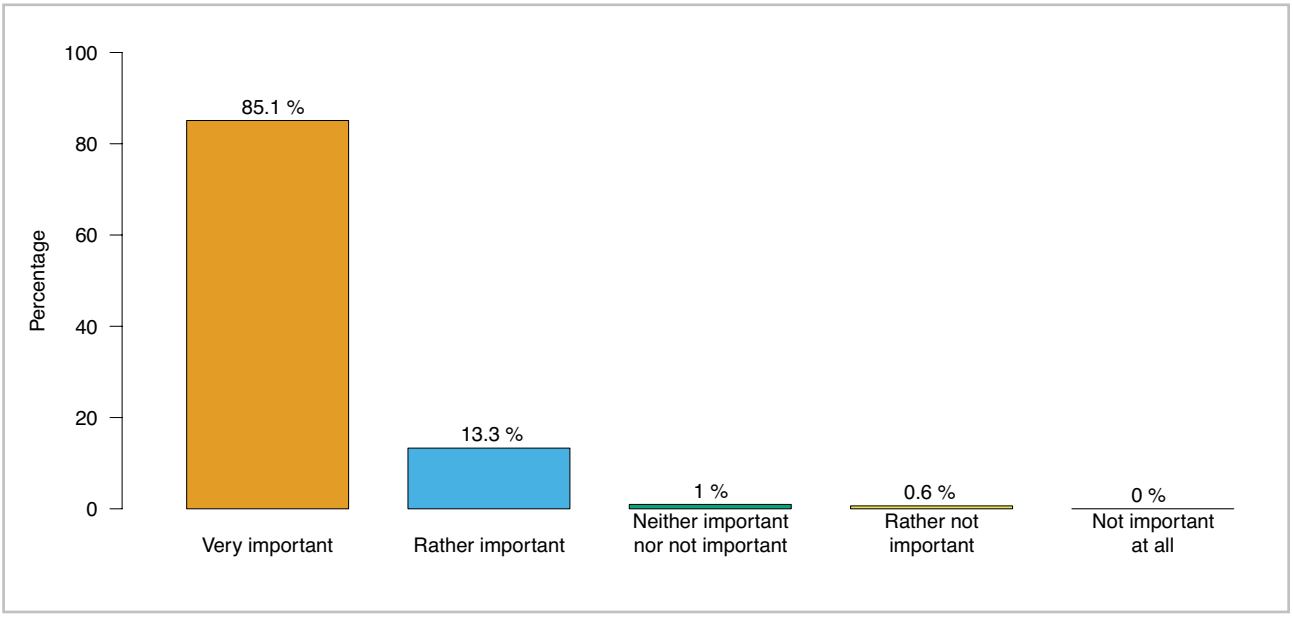


Fig. 50. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you to be sure that the data was produced by reliable people and institutions?

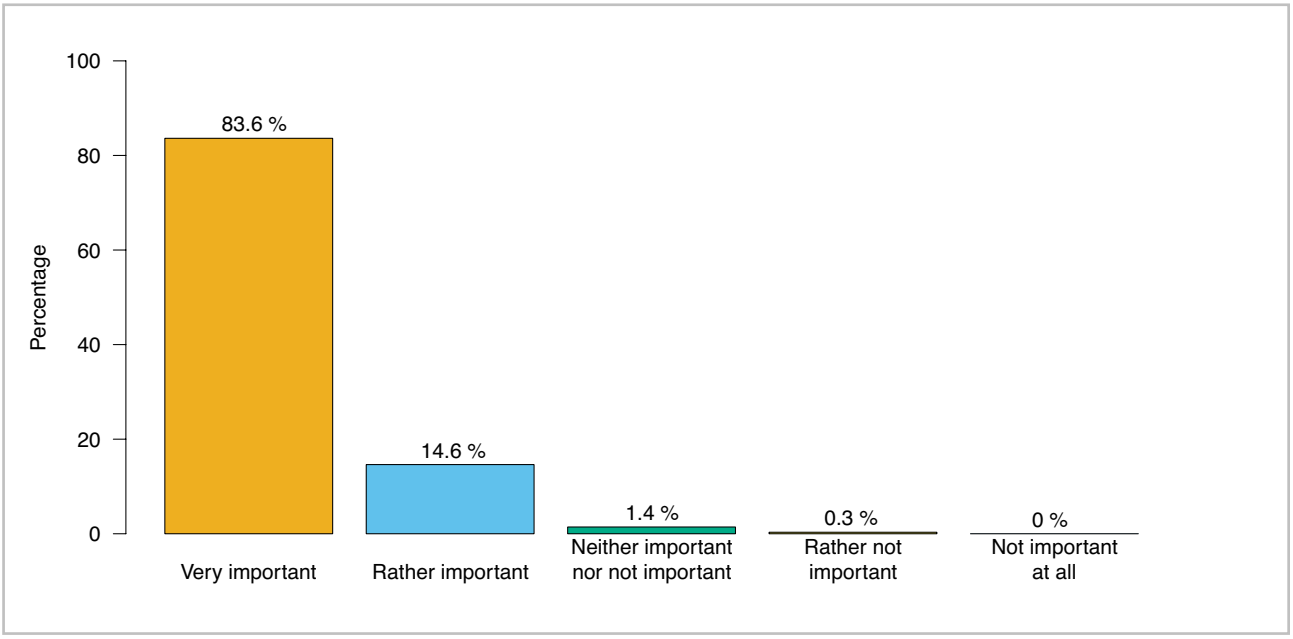


Fig. 51. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that the data were well documented?

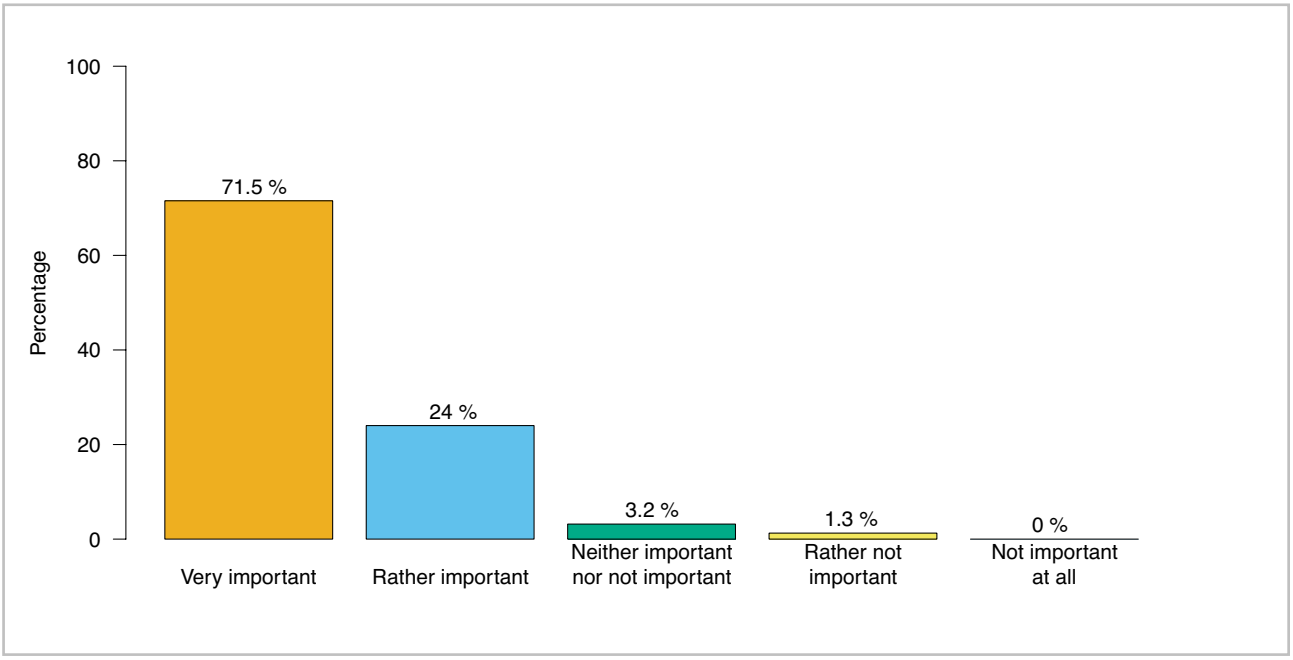


Fig. 52. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data would let you get new, original research results?

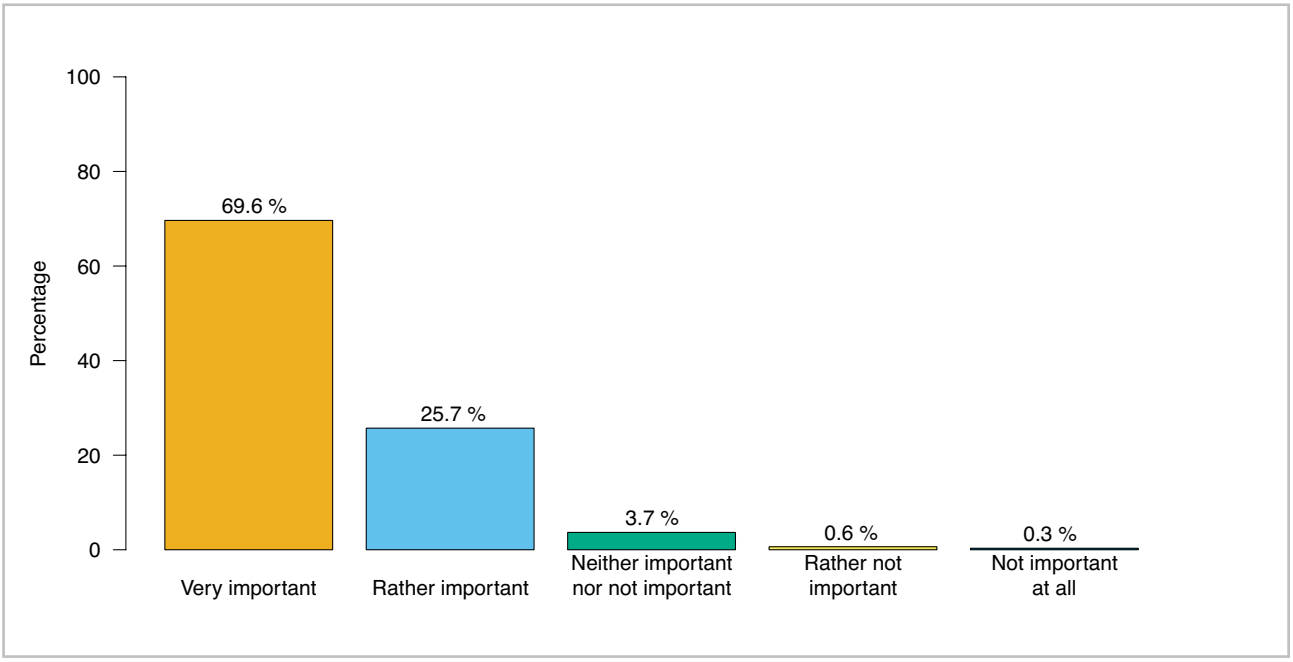


Fig. 53. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that there was a designated person that could be contacted in case of questions or doubts?

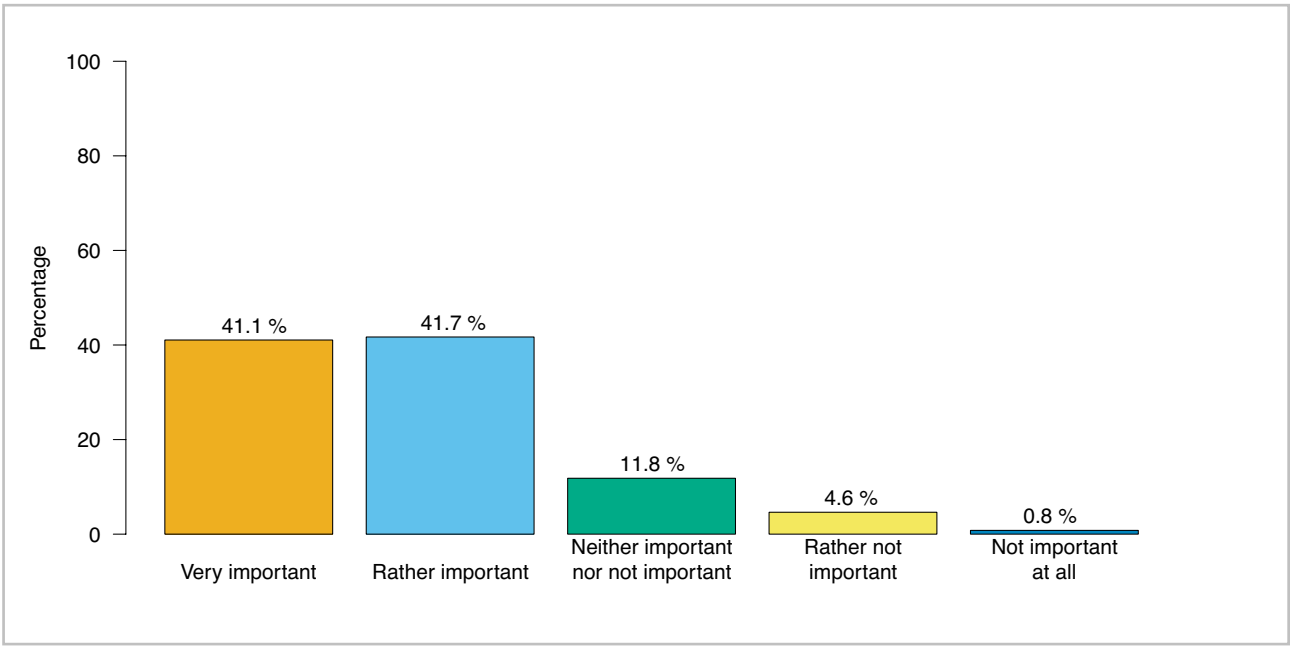


Fig. 54. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data was easy?

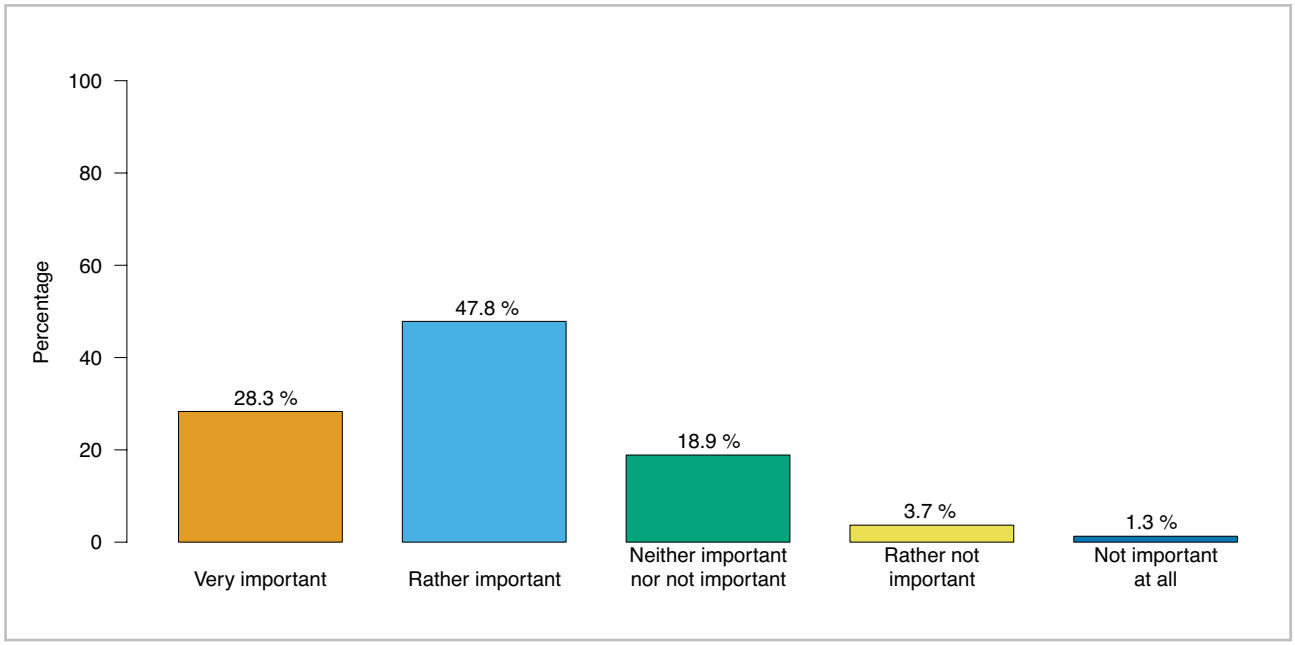


Fig. 55. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data would let you save time?

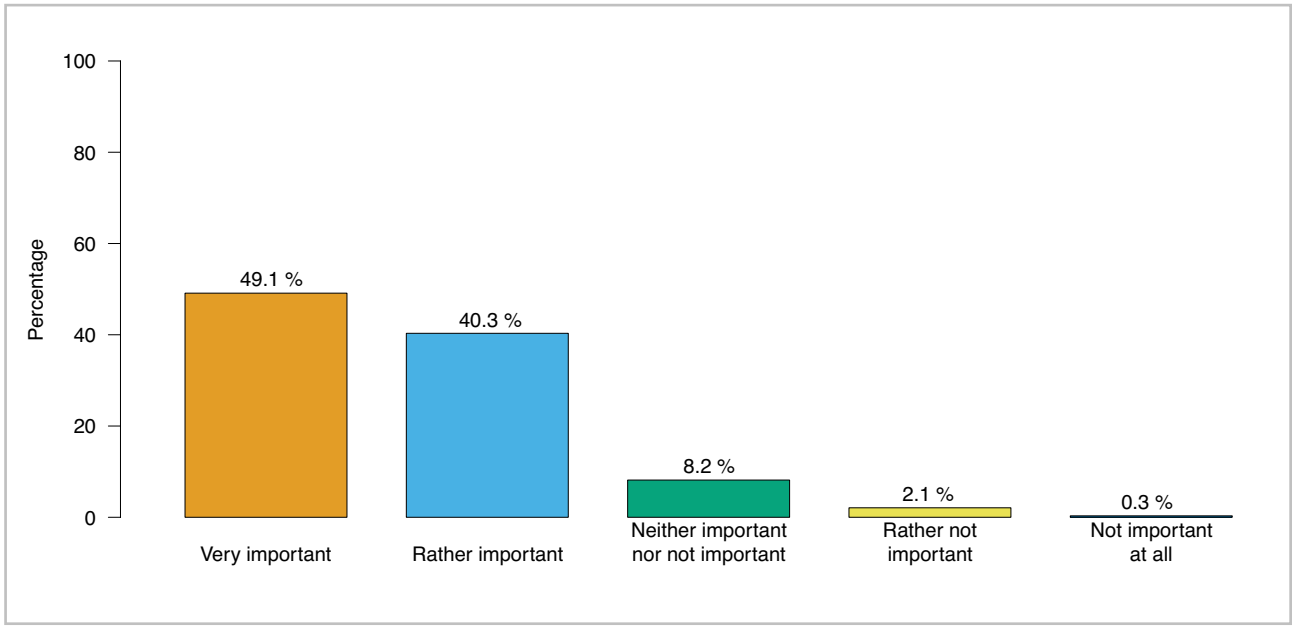


Fig. 56. If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data would let you reduce costs?

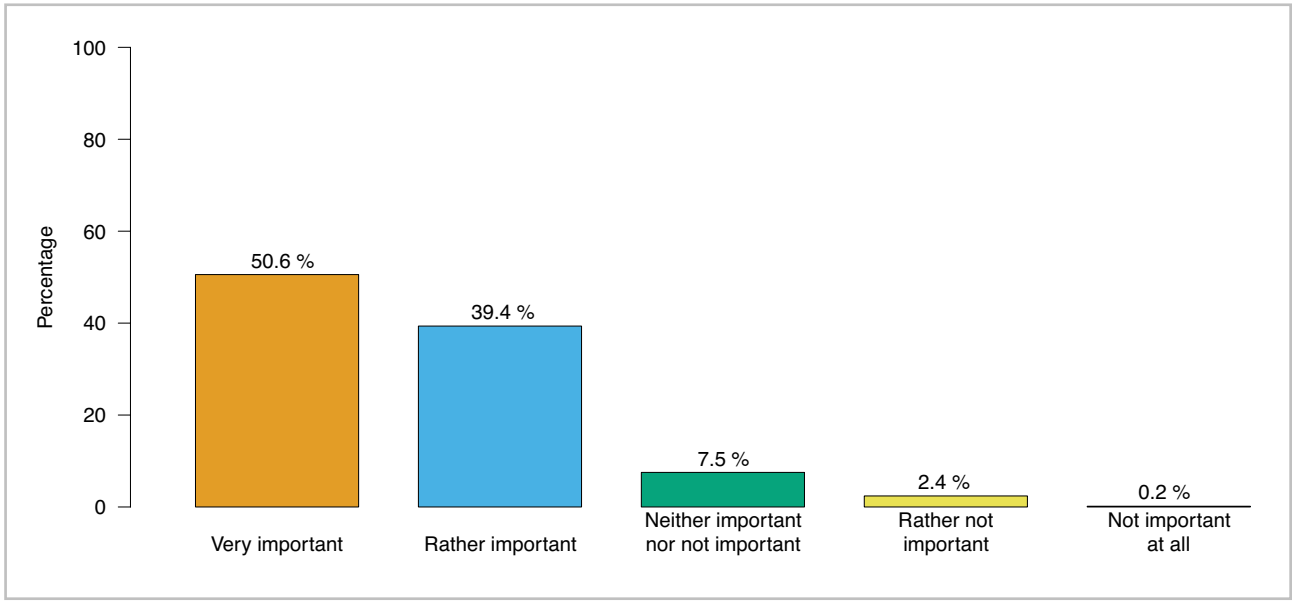


Fig. 57. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data would let you save time?" by „Gender"

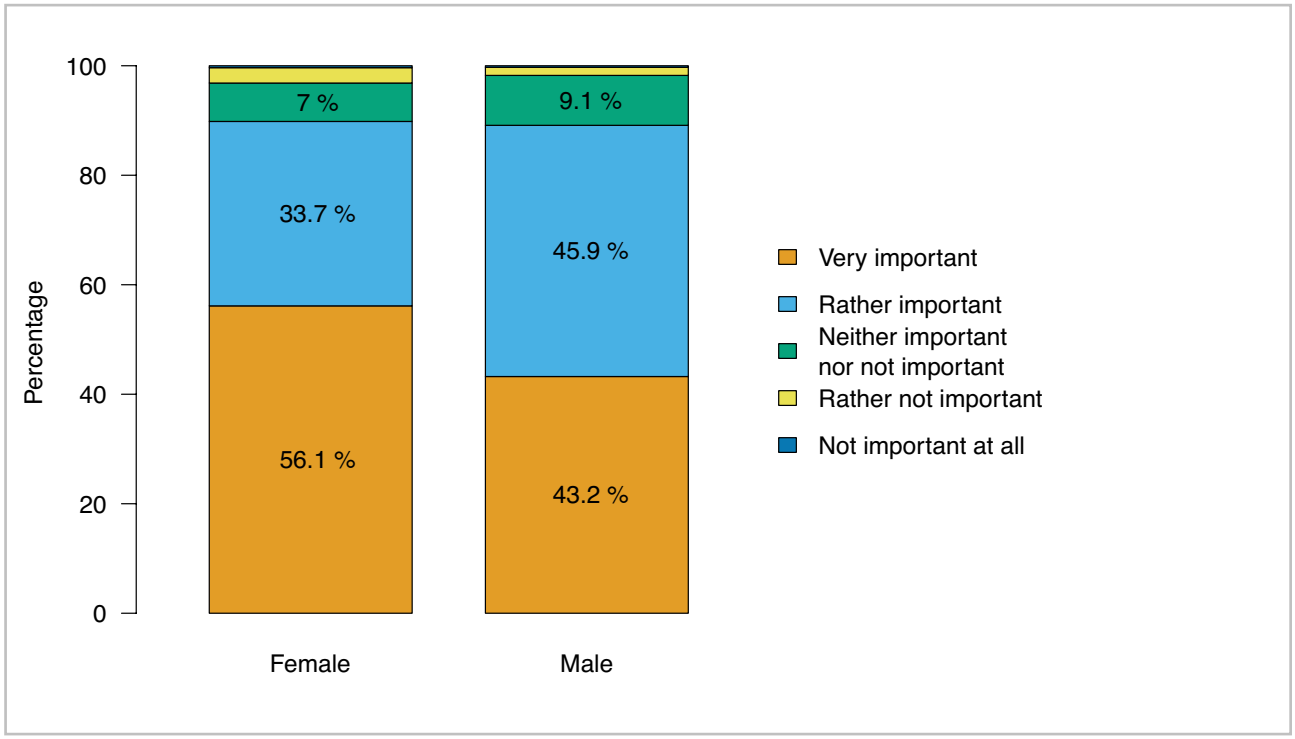


Fig. 58. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data would let you reduce costs?" by „Gender"

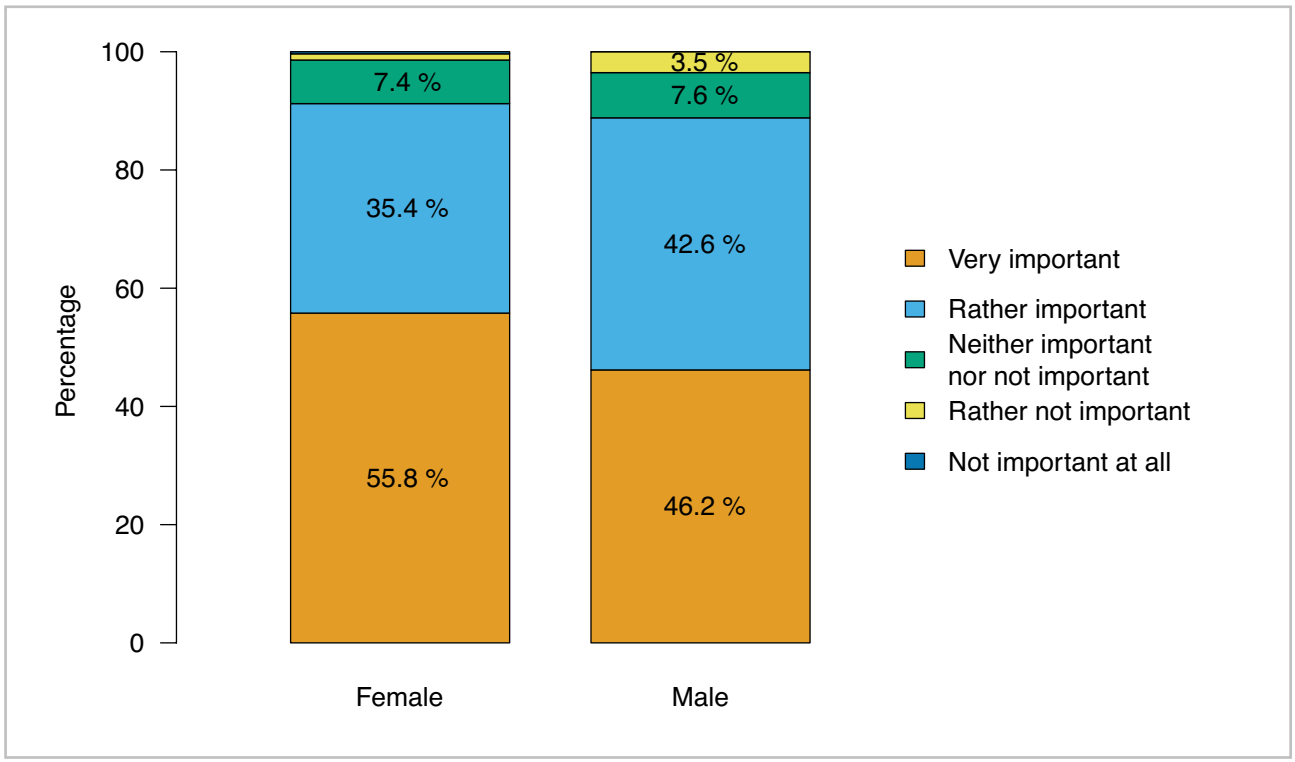


Fig. 59. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data would let you get new, original research results?" by „Gender"

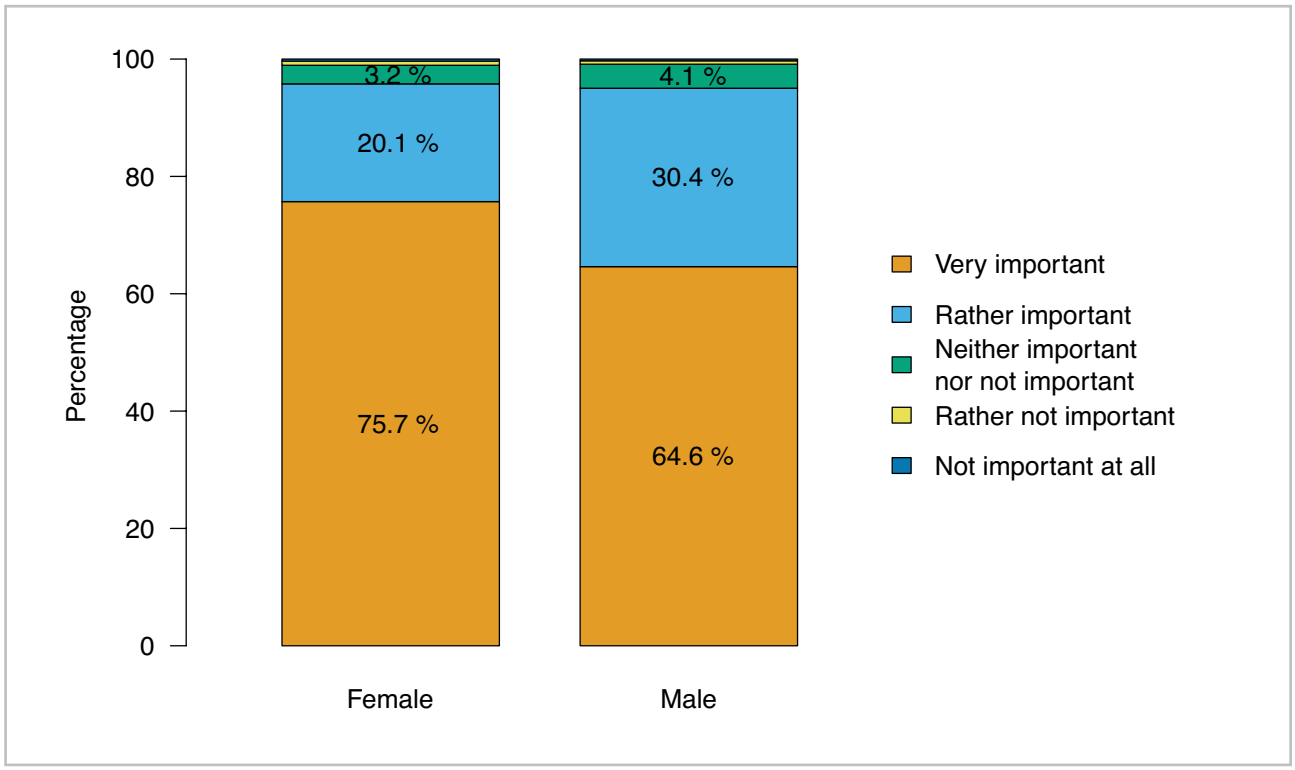


Fig. 60. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you to be sure that the data was produced by reliable people and institutions?" by „Gender"

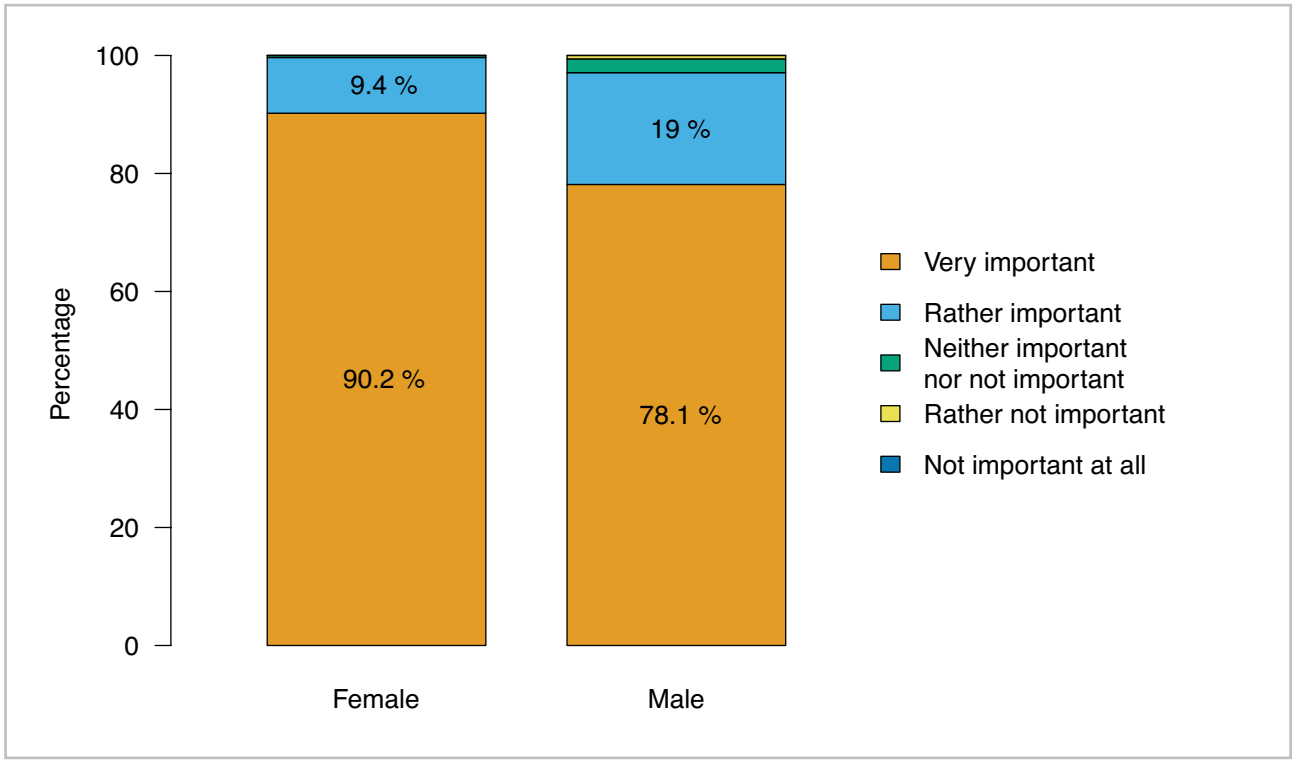


Fig. 61. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that it were clearly stated how the data can be used according to the law?" by „Gender"

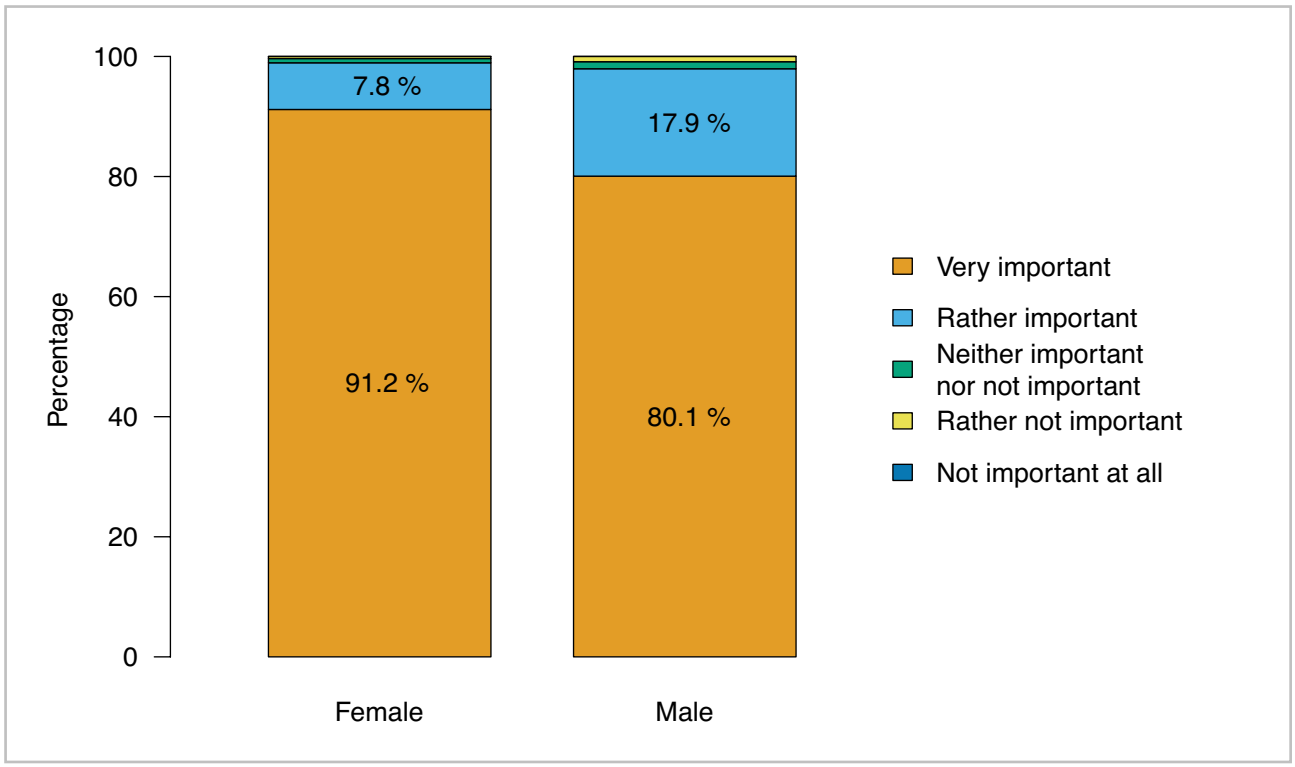


Fig. 62. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that the data were well documented?" by „Gender"

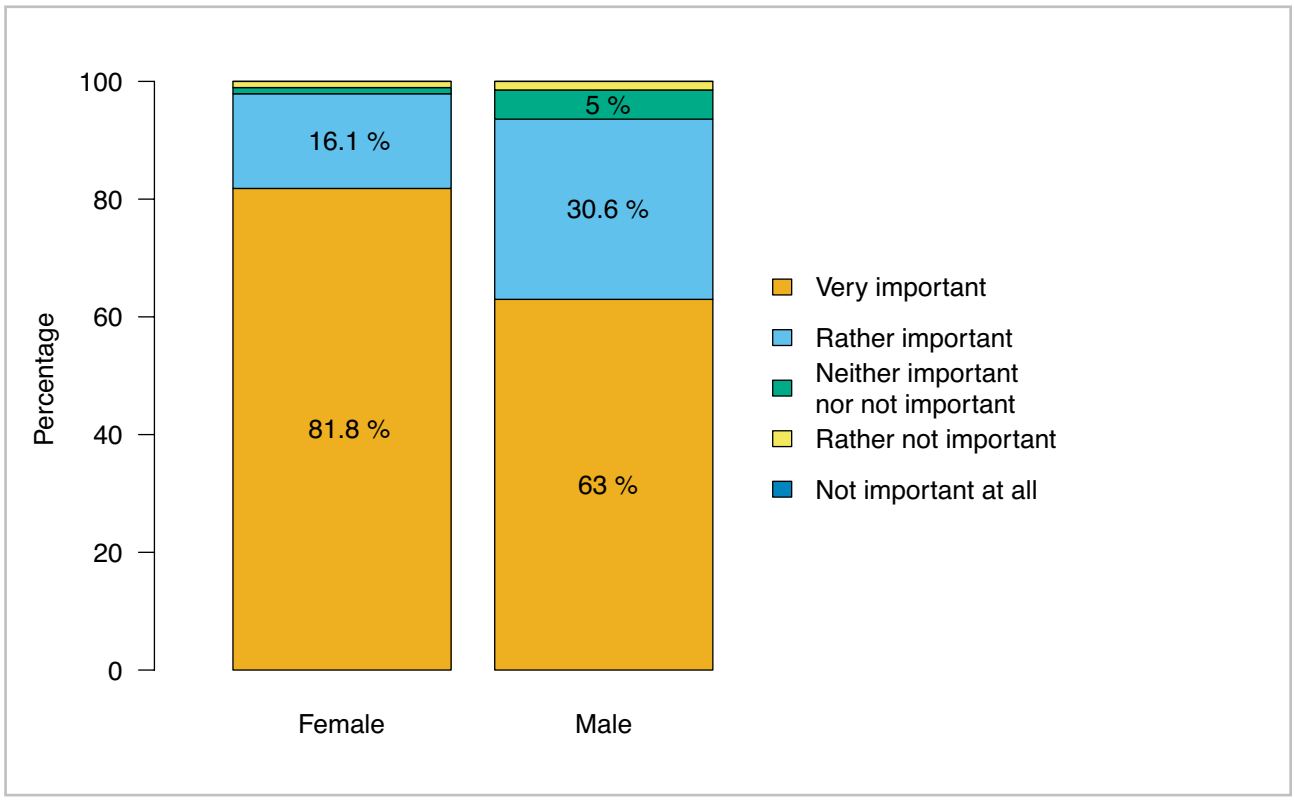


Fig. 63. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that using the data was easy?" by „Gender"

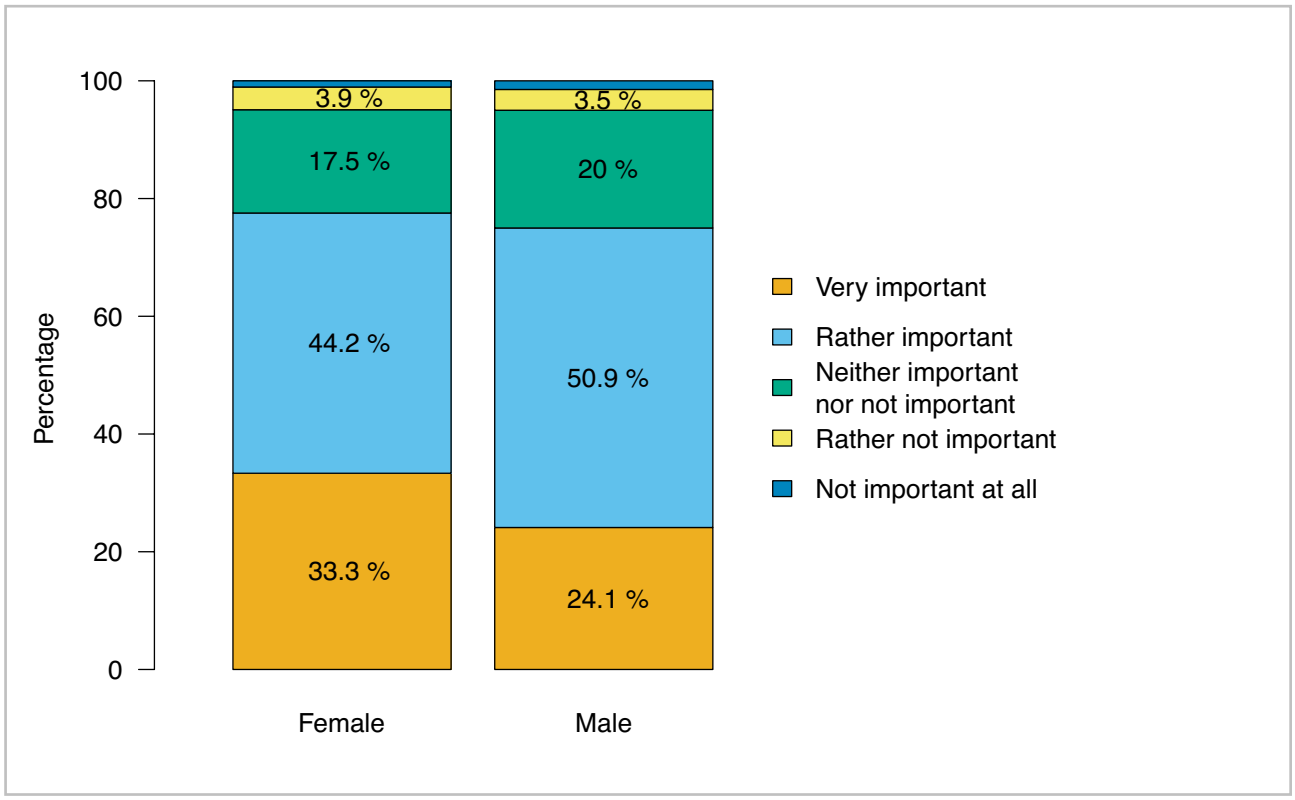
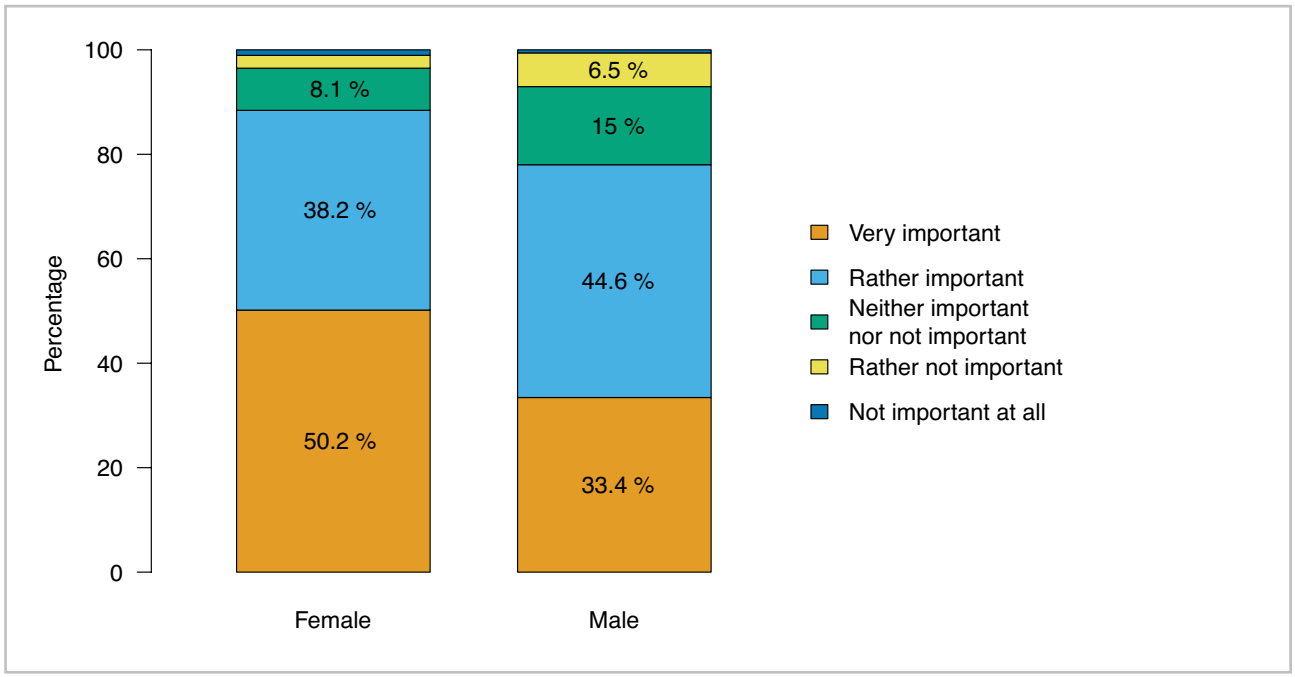


Fig. 64. „If you were making a decision whether or not to use research data shared by other researchers, how important would it be for you that there was a designated person that could be contacted in case of questions or doubts?" by „Gender"



Past experience with sharing data and using data shared by others

Almost three fourths of the respondents had some experiences with sharing data; a similar fraction had in the past used data shared by others (Fig. 65 and 66). About one fourth (25.2%, Fig. 65) had no past experiences with sharing data. A very similar fraction (25.8%, Fig. 66) had no experiences with using data shared by others. It must be noted that the question asked about sharing data in general; this should not be considered the same as open sharing or public sharing. 14.1% of the respondents were researchers who had no experience with sharing nor with using data shared by others. The opposite situation, where a researcher had experience with sharing data and with using data shared by others, constituted 64.4% of all cases.

Answers indicating past experiences with sharing, as well as those indicating experiences with using data shared by others, were most frequent among researchers with full professorship and least frequent among young researchers, with MA degrees. This shows that data sharing is not only a question of whether the researcher is willing to share or to use things shared, but also whether he or she has the opportunity and power to do so. It's understandable that advanced researchers had more opportunities and power than relative newcomers (Fig. 67 and 68).

Fig. 65. Have you ever in the past shared your research data with others?

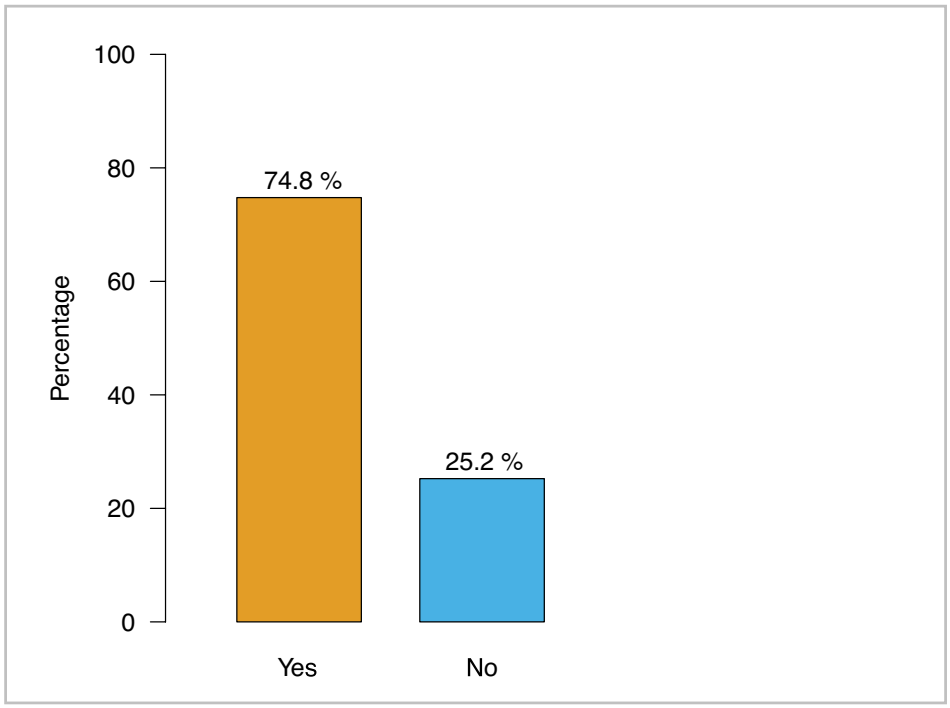


Fig. 66. Have you ever in the past used data shared by other researchers?

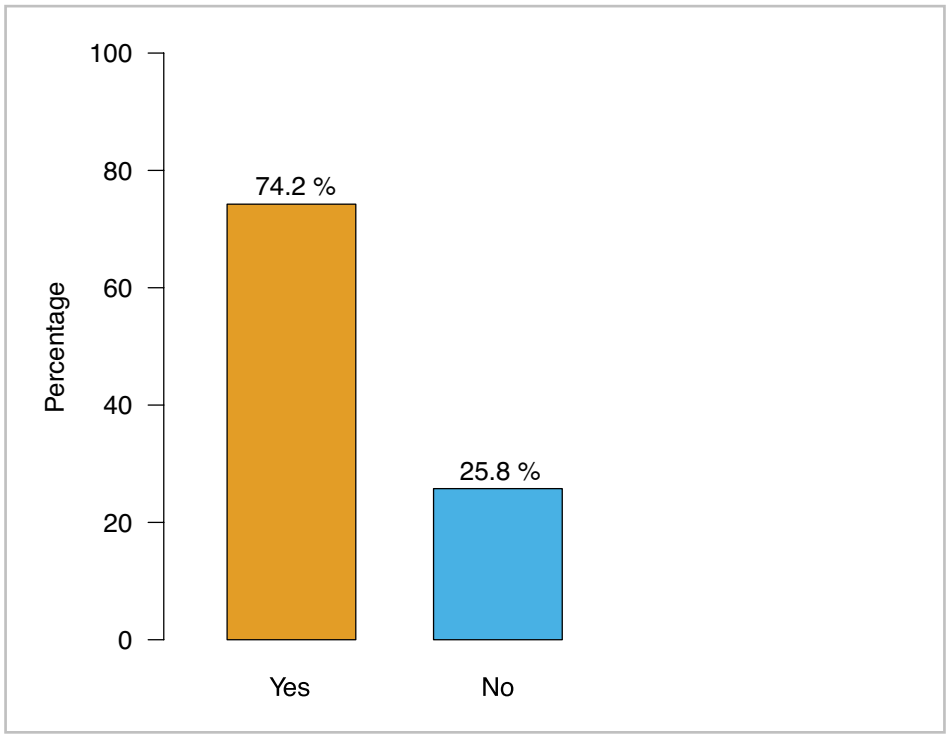


Fig. 67. „Have you ever in the past shared your research data with others?" by „Scholarly degree or title"

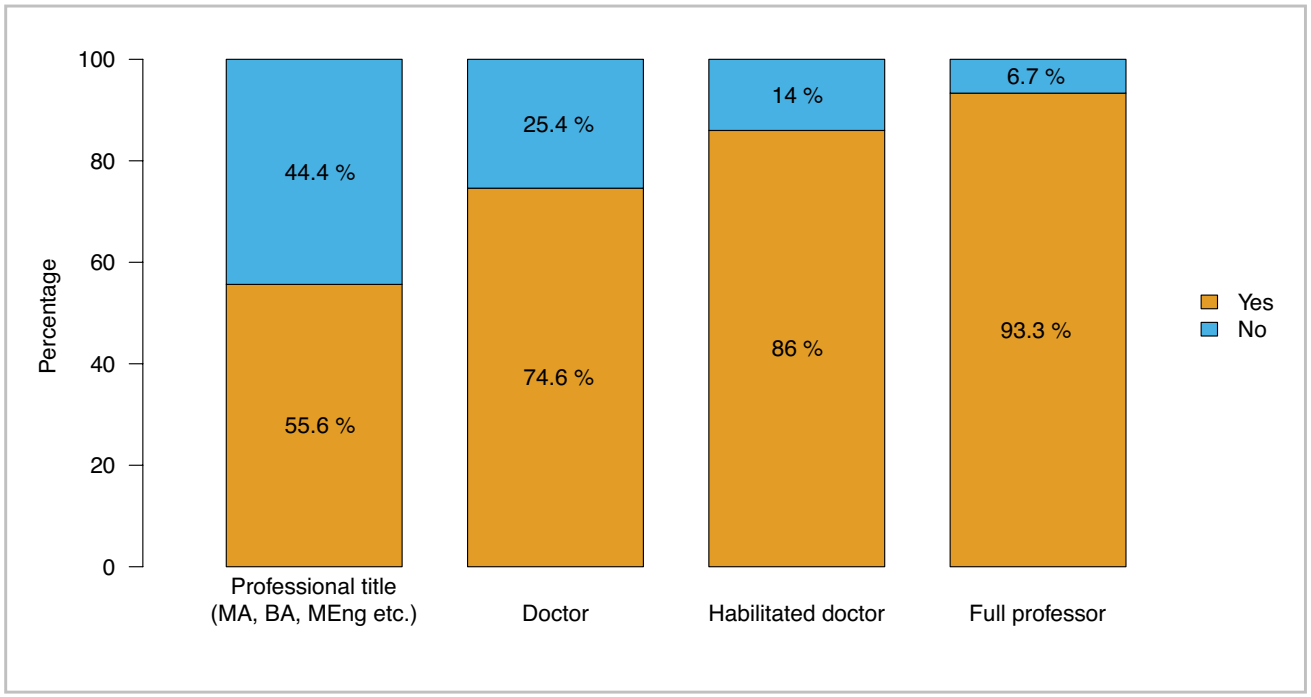
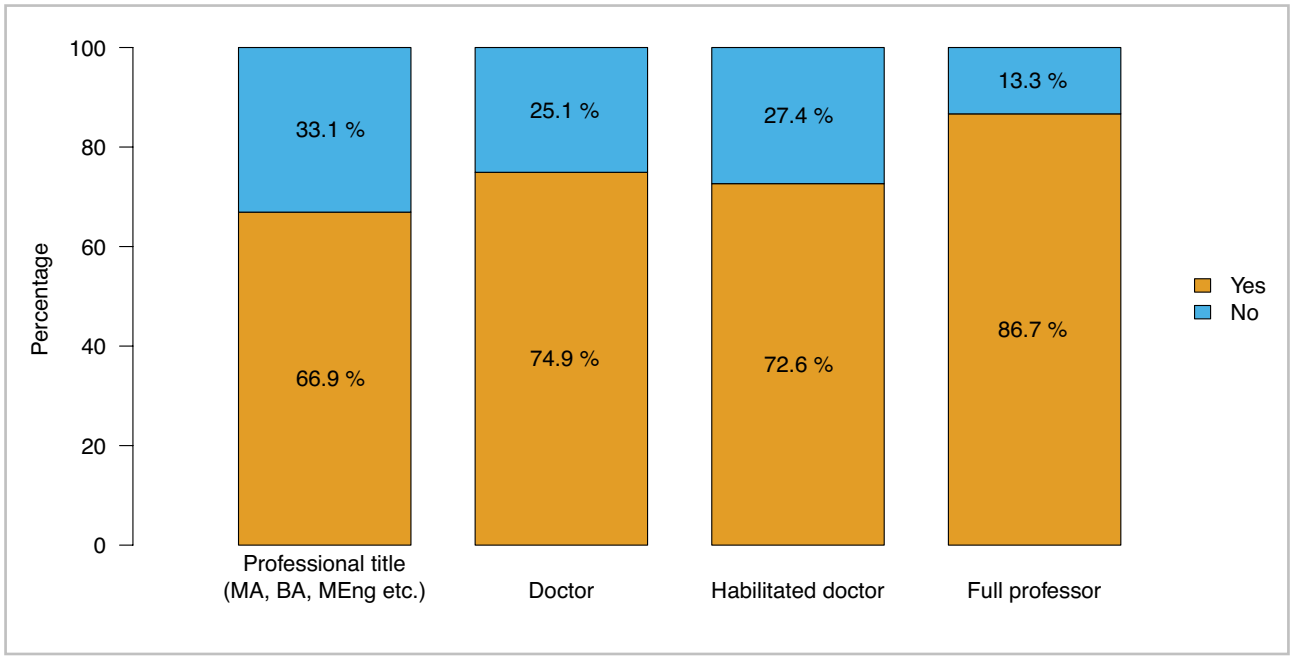


Fig. 68. „Have you ever in the past used data shared by other researchers?" by „Scholarly degree or title"



Main results

The vast majority of respondents turned out to be in favor of sharing data, as well as of using data shared by others, although one has to remember that the selection of the respondents was not random and the results cannot be extrapolated to the community of Polish researchers as a whole. Moreover, when these general statements are replaced by more refined questions, the respondents become more cautious and selective in

terms of with whom to share and to what purpose the data could be used, so that general support for sharing should not be easily equalised with enthusiasm for open or public sharing.

The majority of participants also claimed that they know where on the Internet they could share their own data with others and where they can find data shared by others. On the other hand, answers to basic questions concerning legal issues revealed that in this area a large fraction of respondents wrongly assessed the correctness of the respective statements testing their knowledge. At the same time, legal issues were among the most important factors when it comes to making a decision whether to use a dataset shared by others or not. This may mean that a large fraction of participants is aware of the importance of these issues, but at the same time a lot of them lacks the necessary knowledge. Also, the fact that the highest fraction of wrong answers appeared among professors, who are role models and teachers for younger adepts, should be noted.

Career-related factors were most frequently chosen as very important when making a decision about sharing or not sharing data. These were factors such as citations or having enough time to finish all planned publications before a dataset becomes available to others. Direct financial benefits were considered important by a far smaller fraction of respondents. Thus, career-related factors seem to be more important than direct financial benefits when considering proper rewards and incentives for sharing data.

At the same time, the majority of researchers considered it important that they would be able to decide who and for what purpose will be allowed to use their data. This may mean that some ways of sharing data (in terms of the scope of potential users and potential ways of using a dataset) are more acceptable than others.

Also among the factors that would prevent a researcher from sharing data the most important were those that could impede a scholarly career, such as being outrun with future publication by other researchers benefiting from the data.

The survey participants wanted sharing to be simple and effortless and a large fraction of them would probably resign from sharing if it would require a significant effort. Interestingly, when it comes to the effort already made when producing the data, the fraction becomes much lower, so it is rather an additional effort that could prevent them from sharing.

The respondents had no objections towards sharing data with researchers whom they know personally, who work in the same institution, or even who conduct noncommercial research in general. But when it comes to researchers who perform commercial research, the situation changes: here more than a half of the respondents are undecided or even oppose sharing.

The qualitative study

For the purpose of this report, between May and November 2015, a qualitative study consisting of 10 individual, deepened interviews was conducted. Nine of the interviewees, representing a wide spectrum of

academic disciplines, were researchers engaged in sharing their own data and using data shared by others, and one of the participants was a member of the management of an institution involved in a data sharing initiative in the field of social sciences and humanities.

The study was based on the assumption that while open data may be a new and unknown concept for researchers, the custom of sharing data (without any direct, conscious references to the idea of openness) is much more common. Thus, only in some cases researchers that are not working within the open data paradigm would be intentionally closing and protecting their data. Much more common would be a situation where a researcher is sharing his or her data with others, although with some limitations. Between the two extremes of open data and closed data stretches a gray scale of sharing data.

Variety of data

The interviewed researchers worked with a wide spectrum of data: from large databases measured in terabytes, through photos and images, public statistics and video recordings, to transcripts and relatively small spreadsheets. Different kinds of data create different challenges, but sometimes even sharing of a very small dataset may be beneficial for other researchers as well as for the society. When data is hard to collect (for instance due to relative rarity of a given phenomenon), even opening a small dataset - which usually would not require an enormous effort - may significantly increase the amount of data available for analyses. An example given by one of our interviewees describes this kind of situation:

> In some cases, for instance when it comes to rare genetic diseases, even, let's say, very simple data characterising these diseases, based on a relatively small number of patients, may be extremely valuable, because it is very difficult to obtain. [10][11]

On the other hand, the rapid increase in the size of the datasets creates new circumstances. Researchers sometimes try to apply old ways of thinking and managing data to this new situation, which may lead to trouble. The same researcher describes the actions of one of his colleagues regarding an unwontedly large dataset:

> In any case, there was no discussion about how to process it. I mean, I think that there was an idea that they will put it on their desktop and they will open it somehow. But it would simply not fit. I mean, there is no awareness that a gigabyte, a terabyte - whether there is any difference. That a slightly bigger hard drive [will do the job]. But it is simply not so. [10]

The colleague was about to conduct a study resulting in one terabyte of data per week, but she was unaware of the fact that this amount of data will create problems she had not met on her path before. She had no clear vision of how to collect, read, store or analyse a dataset of this size, somehow assuming that a big dataset is just like a small dataset, only bigger.

The fact that different kinds of data may generate different kinds of problems results in differences in effort required to share data in various disciplines. The disciplines may differ because they deal with

11 The number at the end of a quote stands for the code of a respondent.

different types of data, but also because in some cases the complexity of data types may create additional challenges, and because different journals apply different norms concerning data:

> What is always painful for us is the comparison between [subdiscipline A] and [subdiscipline B]. [Researchers from subdiscipline B] really want significantly more open data. And this is a consequence of the policies of journals, which from some point on started to require that sequence data must be stored in open databases. As a result, more and more data was being gathered and more and more meta-analyses have appeared, which has somehow contributed to the development of the discipline. In [subdiscipline A] [...] on every level there are different issues that are important and different data that is needed. [10]

The first of the two disciplines is much more coherent in terms of the types of data that are involved. Moreover, important disciplinary journals have decided to adopt policies that require depositing datasets related to published results, paving the way for changes in the way data is shared and used within the field. In the second case, the situation is much more complex, because the researchers working in this field need various kinds of data that are related to various levels of the studied phenomena.

Incentives and benefits

Contrary to the publication of a paper, in Poland the publication of a dataset is not recognized as an achievement important for evaluation procedures or career development. At the same time, a publicly available dataset might attract attention to the publications of those who share it and bring citations to the publications that describe it - a key issue for researchers even under the current circumstances. Thus, citation of a publication related to a dataset was most commonly perceived as a proper and readily available (i.e. not requiring any systemic amendments) reward for those who share their data.

Sharing data was also perceived as a factor facilitating cooperation with others, especially other researchers. It may work as a particular kind of advertisement, showing the scope of expertise of the authors and attracting external peers interested in a mutually beneficial collaboration:

> Here is the story: I was attending a conference [...] and I was conducting a demonstration of this tool. [A researcher from overseas] was passing by and he saw it. He said: listen, you know, we have such and such data [...], it would be nice to process it. Is it possible? I said - OK. Let's do it this way. We will process it for you for free and you will make it open. And we signed an agreement, we have a cooperation agreement, it happened that we had funding, so we could pay people. There is no such thing as a free lunch: when someone does a job, you have to pay him. At least from my point of view, this situation where I can pay the people here for processing the data, if it is beneficial for us in terms of citation and for the community in terms of usability, this kind of situation suits me. It seems to me that this is a good model for everyone. And it also suited well [the researcher from overseas]. So like I said, he shared the data, we processed it, and now there is a licence and it's all in the agreement. [10]

Moreover, sharing data may also work as an attractor to academic newcomers (such as PhD students) who are looking for an opportunity to affiliate with an institution that conducts interesting studies. It may also become a source of new information on what can be done with a particular dataset, which may enrich the scholarly kitchen of those who share it. In their case, the fact that someone did something new with the data - something they did not do themselves or even did not expect could be done - may lead them to new methodologies and/or new fields of interest. The well-known proverb that "one man's trash is another man's treasure" turns out to insufficiently describe the case of data sharing, because shared data may become a treasure even to those who at first mistakenly considered it to be trash.

One of the interviewed researchers, when asked about the possibility of including the producers of shared data into the author list of a publication based on the shared dataset, gave an interesting justification for inclusion:

> A lot of studies and a lot of highly cited articles with high impact on the development of science is created by large research groups. There are different roles then: some people are programming, others create measurement tools, conduct statistical analyses, model the data. And now, if a whole team is publishing, then what? Should we take something away from some of them, because their role wasn't equivalent? That would have no sense. So, if it was splitted into separate processes, then why not? [5]

When a whole team works on a study and separate subgroups of the team are responsible for different tasks - such as data production and data analysis - including data producers as authors may not seem controversial. Thus, according to the interviewee, inclusion of producers of openly available datasets into the authors' list would only mean that well-established rules are applied and adjusted to new circumstances, where data producers and data analysts do not constitute two subgroups of the same team, but two separate teams.

The rule that data producers should be included as authors was not shared by others, or at least not without qualification. When in a given field an additional, creative input is required to become an author of a publication, researchers become reluctant to extending the author's list:

> This is also possible, but we usually expect that an author would contribute to the work, not only bring the data. Maybe here it is hard to tell what this should look like. It depends on the situation. If I really made some input and it would be possible, then yes, of course. [9]

From the point of view of the researcher, the citation of a dataset and usually also of a publication describing it should be a sufficient reward for the data producer. On the other hand, it isn't hard to imagine a situation in which the data producer in fact provides additional input to the publication, because he or she actually knows the data best and is likely to be invited to collaborate on the topic.

When there is will, know-how and required infrastructure (hardware as well as software), then other benefits may arise from digitization and sharing:

> So our idea was to help our colleagues and replace these [materials]… Well, I understand that the evidence must remain, but just to make for them a digital version of these [materials], so that they could browse them. [10]

In this case, a researcher being mainly an analyst and not a data producer decided to help the experimentalists by digitizing their material samples. This allowed them to hope for citations of the dataset - a thing much harder to achieve in the case of material samples, which hitherto remained closed, just as many digital datasets do.

This indicates that - similarly to how things are with scholarly publications - the issue of openness concerns not only digital data (things stored on optical discs, pendrives or hard drives hidden in researchers' drawers and lockers), but also data encapsulated in physical, tangible samples or preparations. Although digital data may be as closed as non-digital data, only digital data may become fully open, and for this reason digitization may be a necessary precondition of openness. Material samples - although sharable - do not let for comparable easiness and availability.

Moreover, the fate of the data that became open may be tracked by its producers. This can be done with digital tools, as well as through more traditional channels of communications:

> Recently [my colleague] has been at a large conference [...]. He says that many of them approached him. Some Spaniard approached him and said, I quote, that this site saved his butt, because he couldn't find this kind of 3D data anywhere. And it was perfect to finish the article or something. He declared that there will be a citation. So, we receive this kind of signals, through publications, through direct feedback or through our server logs. [10]

Some interviewees invoked an argument that sharing data should be (and often is) beneficial not only to the people in academia, but also to the general public. This argument was supported by the well-known obligation to make reverse input to the society which financed the very research necessary to create data:

> Americans do it by promoting some kind of ethical attitude, that we are solving problems and not, so to say, making an investment in our own careers. Of course in the meantime they make an investment in their own careers. These two things are not in conflict. It's unethical to use public money to, so to say, develop one's career without giving anything back. [8]

The scientist sees a major difference between the way scientific research is done in Poland and in the USA, where he spent some time as a postdoc. According to his view, in the USA science is made to solve particular problems, and systemic solutions are aimed at achieving this goal. When individual, egoistic motivations and the possibility of solving a problem come into conflict, it is resolved in favour of the latter. In Poland, science is made more to the benefit of individual careers: scientists are still more concerned with gaining scientific degrees paving the way for sinecures than with solving problems. Thus the doubts concerning opening data and the lack of requirements concerning sharing, similar to those applied in the USA.

For researchers, in many cases raw data is very important, as it often allows to corroborate previous scientific claims, helps establish new findings and create new, refined software:

> You can reanalyse it and, for instance, find in the structure even more things than the authors have found. Or, for instance, you could find out that they've solved some part, but they haven't

> noticed something else. For instance, having the newest versions of the software, new algorithms, you can take this as an argument for depositing these data. Having raw data, we have the exact initial data. [9]

In addition to this, another researcher pointed to the need of sharing data in such a way that would be beneficial to targeted groups in the society. Decision makers, other researchers, professionals, entrepreneurs or the local community - all of them may be interested in open access to scientific data, but they may have different needs and different skills allowing to grasp and analyse the data. Access to raw data may be of paramount importance to other researchers, but it will not help in solving the problems of the general public. The public may have an interest in receiving the information embedded in the raw data, but at the same time it may be lacking the skills necessary to understand it in raw form or to analyse it:

> It is, in a way, when you say that business needs the results of scientific research. But often what entrepreneurs understand as science and what scientists would like to do as science differs a lot. Entrepreneurs usually need simple, effective solutions, that are not sublime. (...) And from my perspective, in the public sector the situation is similar. They sometimes need simple analyses of statistical data, for us that isn't even science, and for them, if it is well done and fine indicators are calculated, it is like - wow! So contrary to what you would think, this kind of data together with simple but persuasive visualisation is super attractive for them. [6]

Thus, data should be offered in a form and manner adjusted to the needs and abilities of the target users. (This means something exactly opposite than the manipulation of data intended to misinform its recipient.) Without these adjustments, the data may be virtually useless and serve not as a tool able to bring benefits, but as a blown egg or a tombac - something made to resemble something else, but in fact very far from it.

For some, the raw data will be everything that is needed. For others, only the executive summary of a final report or popular science journal will contain assimilable information. This indicates that open access to scholarly data and open access to scholarly publications are two closely interconnected issues, as on different levels they both serve the same needs of the informed society.

Obstacles, disincentives, and challenges

The interviewees perceived various obstacles to sharing their data. One of them are unclear expectations regarding data. One of the interviewees invoked a formal question asked by the Ministry of Science and Higher Education about the datasets owned by scientific institutions and the possibility of their commercialisation:

> The majority of scholarly institutions in my area, or at least all those that I know, received a ministerial order, or request, to prepare a list of data that may be commercialised. It was to somehow show the competitiveness of science in Poland, something like that. So what kind of data you produce that could be commercialised. And now, to my horror, my management said that we could commercialise everything. [4]

Apart from the conclusion that everything could be commercialised, the institution had no direct plans for the commercialisation of its datasets. But just this simple question was enough to create an overprotective reaction resulting from concerns that sharing would destroy the purely hypothetical market value of the data. This indicates that lack of a clear governmental policy regarding data sharing may lead to a situation where data owned by academic institutions will remain closed "just in case", as the institutional decision makers would not want to risk taking actions which could waste its market value. The abovementioned request was not followed by any decisions or actions taken by the Ministry, but was enough to evoke suppositions meant to precede expected governmental actions:

> And I hear now that many institutions have done the same as my management. That is: here you are, we will commercialise everything, because we are able to show a theoretical recipient for every piece of information. [4]

Thus, one should not treat the hypothetical possibility of commercialisation as a sufficient reason for not sharing. It may be considered a valid argument if there are some reasonable plans for commercialisation (but even then this does not mean that reasons of this sort should automatically prevail over those in favour of sharing). Otherwise the argument is not only empty, but also dangerous, as it may serve as a general excuse for those who for some other reasons oppose sharing data.

Some of the interviewees have had discussions with researchers who doubt or oppose data sharing and have tried to persuade them to the idea. This may occur in research teams where different groups of researchers are responsible for collecting and analysing the data:

> We have long discussions, especially with professor [A], […] who in fact is the person who got me into [the discipline]. And he says: "If I've spent a few years putting so much effort into collecting this data, why should I share it with anybody? A theoretician will come, take the data and publish an article and I will have nothing out of it." So I do not agree that there is nothing out of it. Because when I share the data, then, first of all, there will be a link to the original publication describing the data. So there is benefit in the form of a citation. Not an additional publication, but a citation. That's the first thing. Second of all, if someone wants to get a deeper understanding of what is inside, he will have to meet with the experimentator and this opens up… I think there is a greater chance of establishing a solid collaboration with an author of a publication when the related data are open rather than not. It's just that I would not make contact with each and every person whose article I am reading and ask if they have data, so that we could have a joint publication. I don't want to do that. I simply have things to do. But if I get the data and something interesting will start to come up, then there is much greater chance that I will contact that person. [10]

Those researchers who are responsible for collecting or producing data are reluctant to opening it because they feel that they have put a significant effort into producing the dataset, and they fear that if the data were openly published, then other researchers may use it as a basis for their own papers and announce their discoveries earlier. Researchers responsible for data analysis, who are more willing to open the datasets they work with, try to persuade the data producers by indicating potential rewards related to data sharing.

It is worth noting that this divide between the reluctant experimentalist and the eager analyst may not be accidental, as the researchers playing each role receive different direct benefits from the open scholarly environment. The experimentalists may receive citations and use the data as a "honey pot" attracting external collaborators, but the analysts - thanks to data shared by other experimentalists - may produce new discoveries resulting in new papers. In a scholarly system where publications are much more important than citations - such as the Polish system - benefits for data analysts are more attractive and tangible. Moreover, experimentalists will also have to struggle with blurry rules of data citation - yet another factor hindering obtainment of due benefits.

Nevertheless, the same researcher indicated that it is important to share different kinds of data, as the synergy between them helps reveal their full potential:

> I mean, usually there is no way that we could share the data [with people from outside of the research group] at an early stage. We can talk about it when we are finalizing a publication, when we see that the project is closed, that it is time to finish the publication, then we start to talk about it with them [i.e. scientists responsible for producing data] about whether to share and what kind of data. We usually strive to share the software that we develop and we often would like to provide some exemplary data. [10]

The researcher is responsible for developing applications helping to analyse and visualise the data produced by experimentalists. He is used to publicly sharing his software, but he is aware of the fact that even the fanciest "empty box" will not be attractive. Also for this reason he tries to persuade those who created the input data to share it in a way that will not threaten the planned publication of a paper.

In disciplines where there is a sharp distinction between data producers and data analysts, but the rewards related to publications are received mostly by the latter, this may lead to problems with data provision. As one of the interviewed researchers put it:

> And now these publications are required, but in the meantime you are required to juggle with codes that can do billions of different things, that deal with terabytes of data. And now, of course, the development of such codes may take a year, and in the meantime it is not possible to publish an article if you are not employed in several projects at the same time. And you are out of circulation momentarily. Because people think: well, he does not fit the bill, because in this whole club there is a lot of papers. It does not matter that after three, four years - because he produced all these codes - suddenly it may turn out [...] that everything is ready, that he can do lots of projects based on that simulation. But the time for reckoning in Academia is much shorter. Usually two, three years. You have to look for a new contract, and - secondly - you need to build a formidable base of articles. [1]

According to the researcher, a young adept of his discipline is expected to be a skillful programmer, as this kind of expertise is necessary to conduct research within the field. Programming tasks may take few years and are not financially beneficial in comparison to what the market offers to job seekers with similar skills. During this few-year-long period a young scientist has no data to publish upon (as the data is still emerging), as well as no

time to prepare a paper (as he or she is preoccupied with the necessary programming tasks). As a result, the achievements of the researcher at the end of their PhD are usually considered insufficient to provide an academic position. This kind of setting where the data producers are necessary for the field to develop, but are not receiving rewards allowing them to stay within the field, is not sustainable in the longer perspective. It is so because this path of career can hardly be described as attractive (one can even argue that "career" is not a proper word for this kind of relatively short-term activity), especially when the market pressure is constantly growing.

Similar concerns may also emerge in multidisciplinary areas of interests, where researchers from different fields of science make use of the same data at a different pace:

> It has been discussed for a few years now. It seems that something has started, that they can start building their careers on this basis - that people cite their data. Because this is typical cry when it comes to our, let's say, [scholars from discipline A]. They are collecting data very quickly, but to publish something [...], a year is not enough. They need to have a tendency from 10 years, or show some changes, or something like this. On the other hand, we, [scholars from discipline B] continuously take data from them, as it is glued to our our [...] data. If we have good relations with a colleague, then we add him as an author. If the relations are difficult, we pay him 1000 zlotys, he has a contract for the work, and we even leave out the fact that this measurement was conducted by such a gentleman, so he is not benefiting in the academic sense. But they cry that they are not able to use the data that quickly, and everybody needs this data. [4]

In the situation described above, some data collected by scholars from the discipline A are consumed by scholars B much faster than scholars A themselves are able to consume it. This gratitude may be repaid with money (which does not add anything to the academic career of a scholar A) or with the co-authorship of a paper. After a few years, when the scholar A is finally ready to publish his or her paper, the fact that it is based on a certain dataset will not attract readers, because the dataset itself was available much earlier thanks to the papers published by scholars B. Moreover, scholars A risk that someone else will publish on the same or similar topic before them, as the data is continuously being published by scholars A. The proposed solution is to include data citations as a factor allowing to build one's career.

Another factor that may discourage open sharing of data is related to the large differences in the funding available to different groups of researchers. An interviewee who works in an institution responsible for collecting data was reluctant to the idea of open sharing of the datasets with researchers. The institution had been collecting the data for years as part of its mission, using for this purpose its scarce statutory funding which allowed for only relatively low wages for the employees.

> And it is known that we do it. And now, for instance, here comes someone who has a grant and would like to use this data. So a question arises: wouldn't it be possible to collaborate on this? And maybe, on the other hand, if the salaries in research were higher, maybe we would all look at this from a different perspective? If I knew that my work is being well paid for. [7]

The fact that their data could be readily used by other researchers having ample grants was perceived as unfair, as collecting the data required a significant effort. Nevertheless, in this case the researcher was reluctant not to the very idea of sharing data, but to the idea of sharing it in an open manner. She would gladly discuss the terms of sharing and get involved in a collaboration. Without it, she perceives a risk that the division between her and other researchers will deepen in various dimensions.

One of the interviewed researchers indicated that different kinds of publicly funded institutions should use different embargo period policies. If an institution is responsible only for the provision of raw data and its faculty is not expected to analyse the data and publish the results of such analyses, it should not embargo its datasets at all. If it is responsible for analyses and is expected to provide publications, it may apply an embargo period sufficient for conducting and publishing the study.

The same researcher noticed that embargo periods are less harmful in disciplines where analyses are based on long data series. In such cases one or even two years of embargo will not change the situation of researchers waiting for the data to become available:

> So for us it is more important to have as much data as possible, in the geographical sense, than to have it extremely up to date. […] They do not lose value that way.
>
> Question: I get it. After ten years, the fact that nine years ago the data was closed doesn't…
>
> Yes, it has no meaning. And it is important that the data - it doesn't get old. This is data that is still interesting. No matter if we collected it fifty or a hundred years ago - we are very happy with it. Of course these are rare cases. [4]

On the other hand, in dynamic disciplines, where the situation changes rapidly and data - just like publications - are rapidly getting old, even short embargo periods may be very harmful.

In some cases, technical issues related to the size of the dataset or optimal access to resources enforce the establishment of a dedicated infrastructure. In cases where scientific data is so large that it is very difficult to store and download, an application enabling online access and stating queries to the database may be a necessary condition of accessibility. A hard scientist planning to open his several hundred terabytes of data decided to engage in creating a special infrastructure just to enable access to the dataset. Thus, the software will enable subsetting and downloading of selected parts of the database directly in the scope of interest of a given scientist:

> All kinds of data are shared, raw data too, but the mechanism of sharing raw data is usually somehow restricted. It is impossible to download everything, because it would blow up wires. So, one has to contact directly to arrange some kind of other form of transfer, for instance show up with a hard disk. Or it is shared piece by piece. So that if someone would like to download all the raw data, then, well, it would be possible, but it would take a proportionally long time. It's just that there are technical confines, so the data is available, but with some restricting mechanisms, because otherwise the raw data could clog servers for months. [1]

The technical infrastructure itself may create new opportunities and pave the way for new projects that seize them. This may be illustrated by the case of a database that was initially created mainly to store data that had previously existed only in analog form, but was later digitized. Soon, the database became a kind of ark for other intramural and extramural projects that were looking for a place to store their data:

> And then [institution A] had financial problems, so that it was very difficult for it to archive things from those projects and, willy nilly, although initially there had not been such an assumption, more and more of these [data] and projects were hustled into [institution B], under the banner of this database. So very quickly the boundaries between the program and the database blurred [...]. So this [...] database has become a place not only for archiving and distributing, but also a place that generates new [data]. [3]

The interviewee described the database as a project in motion, that during the years has been trying to adjust to the changing environment. In the case of long-term projects this kind of flexibility is very important, as the abovementioned factors will sooner or later enforce changes - one size never fits all and forever.

Flexibility may be forced either by technological or by "soft" factors, such as those resulting from changes in expectations of researchers, related to what should be done with data. The statements cited below given by one of the interviewees adhere to both these factors. A researcher and his colleagues, knowing that within their field there is an ongoing discussion concerning sharing additional types of data that hitherto remained closed, wanted to find a proper place to store it and make it openly available:

> This is a vast amount of data, gigantic amount. And it will grow constantly, because of new technology, every new kind of detector that shows up on the market, it usually has better parameters resulting in larger files. And it grows exponentially. [9]

> And then there was searching - where to place it. First, we requested our library - I called our IT center and found out that our library manages [a repository], and a lady from the library sent me to [an institution], claiming that there is such a project, an open data repository, and this is the place where I can do it. So I went there and this is how it began. [9]

Appropriate database or archive must develop because new datasets are being added, but it must also grow because newly added datasets are of much larger size resulting from new devices, new formats and better quality (sampling, resolution etc.). But as it is visible in the second citation from above, the change may also be stimulated by expectations of users. The researcher wanting to make a deposit of his dataset, in the first place asked if it is possible to do it in his institutional repository devoted to sharing publications. The repository was established before open data became a hot topic and was not prepared to serve the expectations related to storing data that emerged in the meantime.

On the other hand, lack of resources to provide sustainability of a project will result in difficulties also for the accompanying initiatives. When financial constraints exclude temporarily a possibility of extension of the infrastructure, the projects must be put on hold or reduced:

> But, like I say, we are restricting it a lot, because we have no means and conditions. But it is still happening and still, if someone comes to us and says that there is a very interesting [data to be acquired], we will still try to [acquire] it. We do not rebuff it. But we do not do any big events, only if it is something specific... For instance, we have collaborators and we have such a model that I always encourage, that they use us as a kind of base, a partner, and they try to get funding for [data acquisition] and so on. Then we take part as collaborators. [3]

In a situation where there is little space left to store the data, each new possibility that under more favourable circumstances would have been welcome, requires a difficult decision whether to seize it or not. In other cases the necessary infrastructure may exist, but it remains empty, as there are no requirements concerning open access to data, little financial support and no direct rewards for researchers who put effort to provide it, and no incentives that could persuade those who hesitate. Moreover, where there are no rewards and no incentives, the effort put into opening data can be perceived as a waste:

> The safest thing to do for a researcher is to write an article in English and publish it in a journal having a lot of ministerial points. And if there is concomitant software, that is great, but it is not crucial that he provides new tools, which in larger perspective could have much stronger influence on science, if several hundreds of researchers applied it in their research, than one article, which will be read and cited by ten, or even twenty people. [5]

As "the safest way is to go by the rules", when the rules do not encourage and facilitate sharing data, a researcher who nevertheless pays attention to making his or her datasets available consumes resources - such as time and money - that could be applied to activities more beneficial to his or her career. When these resources are scarce, sharing data is not a priority:

> A practical barrier is that it takes time. I mean, we could share much more data, because we have lots of it, but we only managed to share the amount of data that was budgeted in a given program. Then we pay the informatician, who sits for two weeks and formats it in a feasible manner, puts it in one place and shares it. Unfortunately, only few programs or grants had this type of expenses secured and, as a result, it is not so common. We try to do it on a small scale. I think that nobody has doubts that data should be shared. [4]

Thus, the costs involved in data sharing should be considered eligible by funding agencies, as without this kind of support data will not become available even if the necessary infrastructure exists and researchers are willing to make their data available to others. Similar problems may occur within public institutions, which have no technical background, know-how and manpower to make dispersed data stored within institutional infrastructure openly available:

> And so we exchanged perspectives, that we want to have specific data, we are hassling them to give us the data, and what do they meet with? Sometimes they do not even have the technical base to share the data, for example in the form of computer files saved within some services. But - again, according to common sense - if someone did it once,

> prepared it, even in a csv format, then we wouldn't have to keep sending requests, keep asking them to pull it out of the system. It would save them a lot of work. But there is still a lot to be done here, to be pushed through all these offices. Because there is no such thinking that it could be useful to someone and serve a good cause. [5]

Large amounts of data of significant scholarly value are collected and stored within public institutions. The interviewee indicates that because of the lack of technical infrastructure and the ignorance towards the potential behind opening data, access is complicated and time-consuming for both sides. An employee sent to serve the needs of researchers may be very competent, helpful and amiable, but it is the lack of strategic planning and actions on the level of a ministry, a borough or an institution that are the real issue here.

Problems resulting from lack of long term institutional planning concerning data infrastructure are visible when researchers encounter difficulties not only with opening data, but even with intramural sharing and collaboration:

> Now there is a man, the best in our team, [...] who has to get at the data. He is on postdoc [overseas] at the moment. So I wanted to make it available to him. Now, if it is to make sense, he has to have at least one experimental set and one referential set. Control and experiment, so I would have to share with him at least 130 gigabytes. So I went to my colleagues and said that I would like to use our cloud, because maybe it would be the simplest solution. But it doesn't fit. The cloud [is too small]. So far nobody has had needs of this size. So in practice we have put it on a PC, the one right here, and the colleague remotely logs in here. It is his PC for computations anyway, so it is the most practical solution anyway. [10]

When it comes to data policy, the most important gap in this field is visible in the Ministry of Science and Higher Education: researchers partaking in the interviews often invoked the lack of recognition of sharing data by the academic system in Poland. The Polish scholarly community is not obliged to share their data in any manner. Those who share cannot count on any direct benefits related to promotion and career. One of the interviewees described two possible strategies serving as an answer for this lack of data policy:

> We can accept that the system is as it is, so if we have nothing to gain - we will not be doing it. Or we can do the opposite - we can - bottom up - try to include citations to software and to datasets. And some day, when the system catches up sufficiently to recognize that this is happening, it is the reality, that people notice these things and cite them, then, maybe, then it will start to be recognized and the respective fields will show up. [5]

On the one hand, researchers may simply adapt by not sharing. On the other, they may nevertheless share their data, hoping that the circumstances will change and the efforts will bring benefits in the future. The fact that sharing data, as well as the utilization of available datasets is not recognized by the system, creates an environment extremely harmful to data sharing itself: not only there are no direct incentives for sharing, but even the indirect benefits resulting from using data shared by others are weak, as others are not encouraged to do it.

Legal and ethical issues

Although the majority of interviewees admitted that they do not feel strong when it comes to legal aspects of sharing data and using data shared by others, this does not mean that these activities are chaotic and random. On the contrary - the interviewed researchers were often able to invoke informal norms that regulate who is allowed to make decisions concerning a dataset, in particular a dataset that is about to be opened.

In cases where an interdisciplinary research team was divided into experimental and analytical subgroups, the experimentalists would be responsible for all decisions concerning the data, as they were also responsible for its production. The analysts would be allowed to use the data (which would usually result in a joint publication) and could try to persuade the experimentalists to share certain datasets, but the final decision would not belong to them:

> It is assumed, at least I assume, that the decision is with [the data producer] whether he wants to share it or not. If he wants to share it, I can help with the practical side of the process. Because we can counsel when it comes to how it should be done or what kind of databases are the best. If he does not want to do it, I can try to persuade him why he should do it, but the decision is still with him. [2]

Another interviewee gave a justification based on an analogy with who is responsible for deciding about the publication of a paper:

> For sure, in the funding agreement there are no provisions of this sort. But there are provisions that the results of this study should be published in some sort of renown journals. The higher the quality of a journal, the better. And what will be included in a publication, which elements from our work, our studies - this is our decision. Because we share basically all of the results of our research: the processed models, structures, also the data used for all calculations. So the raw data would be only one step backwards. So if we wanted, we could publish much more information and it would be perfectly legal, but we publish only the things that could be important for somebody and are meaningful for this work or for this project. [9]

According to this view, both papers and datasets are research results. Since the granting agreement requires the results to be published, the question concerning the scope of this mandate arises. This particular research group publishes results that could be important for other researchers and are significant for the journal article. Thus a dataset becomes a part of or an extension of a paper, and decisions concerning how the dataset will be used are made in the same way.

Aspects concerning open licenses were also described as vague and hard to understand:

> Usually no one reads this. And here it would be really useful… There are like five or six types of these licenses as far as I remember. And, for instance - a short description of these licenses and a comparison in a few sentences or bullets. What possibilities do these licences give. Some kind of short characteristics of their distinctive features. [9]

The interviewee was confused when it came to choosing a license for the dataset he wanted to deposit in a repository. Although the repository used internationally recognised Creative Commons licenses, he did not know what provisions stand behind them. It is worth mentioning that some researchers who choose more restrictive licenses for their datasets (such as CC-BY-NC, forbidding commercial use) do not exclude the possibility of sharing the data beyond the scope of the license. One of our interviewees was using CC-BY-NC for his datasets as in the first place he wanted to make it available for other researchers conducting noncommercial research. This does not mean that he is strictly against commercial use of his data, but rather that he wants to separately establish the terms of such usage with potentially interested parties:

> I try to share all my publications, data and software, with everyone, for free. Again, basically for non-commercial use. That's my vision. But if there is a company that would like to make money out of it, then I see no reason why it should not be beneficial to the people who collected data or created software. [10]

When the same researcher wants to use a dataset that is publicly available, but has no proper license attached, he tries to take into consideration the supposed goals of sharing this particular dataset:

> But also if there are datasets that seem to suggest that they are open and that do not explicitly prohibit data processing, then we also try to have them here. [10]

As a lack of proper licence is often not the result of a deliberate decision, but rather of a lack of knowledge about what licences are, how they operate and why it is important to use them, a potential user stands before a dilemma whether to err on the side of legal purity or on the side of the interest of science (and often of his personal interest measured by career development). In such cases, the interviewee tries to guess what was the purpose of sharing of the data. If he finds sufficient reasons to come to the conclusion that a given kind of use will not go against the will of those who made a particular dataset publicly available, he decides in favour of using the data.

According to one of the interviewed researchers, legal issues, if not tackled properly from the very beginning, may become an obstacle to sharing data. While considering the possibility of opening data related to projects commissioned by public institutions, she saw a potential problem with getting back to old agreements allowing only limited use of the data.

> But then we have to take a good look at the provisions in the agreements we sign. We have it enforced, that the contractor, i.e. our team, can use the data in academic work, for further analysis, or something like this. I'm citing from memory now. And then probably we would have to change this, because we are now talking about public sharing of the results of these studies, such as scholarly publications or our reports. […] So to be correct with the client, we would have to receive permission. And now, the question arises: would they be willing? We get back to the problem of full openness and transparency. But it would seem - for instance to me - that looking at the development [of my institution], it would be interesting. […] But revisiting closed

> agreements in the public sector is generally hard, isn't it? The thing is already accounted for, checked, and posted. I don't know. I am not a lawyer, but the renegotiation of an agreement that is closed, accounted for - this may pose a problem. [6]

The interviewee is aware of the constraints resulting from already signed agreements, but sees some benefits that may result from opening of the data. The potential obstacles, such as the lack of understanding of why someone would want to reopen issues that had been closed a long time ago, the reluctant approach of more traditional lawyers, and the problematic issue of reopening an already settled project, may be treated as arguments against even considering this task. At least some of these problems are much easier to resolve if tackled properly at the earliest possible stage of the project.

Main conclusions from the qualitative study

- The wide spectrum of data used by researchers representing various disciplines creates different challenges. Whenever it is possible, appropriate infrastructure and funding should be planned and secured on the earliest possible stage of the research project. It is much easier to actually share data when it is taken into account in advance, so the research activities are conducted in a way that makes further open sharing much easier and allow to avoid the renegotiating of deals that have already been made or the amending of datasets that have already been created.

- The infrastructure for open data should be open also in the sense that it is never a thing that could be whole or complete. First of all, devices and instrumentation constantly develop, leading to new kinds of data, enrichment of old kinds of data and datasets of a size unthinkable by many researchers and software developers ten or twenty years ago. Also today's ways of local storing of data will pass just like minidiscs, DATs or streamers have passed, which in a few years will start to create problem with its reading.

- These issues may vary even within one and the same discipline, where specific subdisciplines may use data of a different kind, size and susceptibility to aging. Thus, planning of an infrastructure for a specific institution always requires careful identification of needs, so that the resulting system could be useful for all stakeholders, not just those, whose needs are well known and researched.

- In Poland, where only a restricted set of achievements is taken into account during the evaluation of scholarly units that is conducted every four years, scholarly units as well as faculties have no direct interest in sharing scholarly data they own. In other words, sharing data will not bring any direct, formal benefits to the unit (measured in its score in the evaluation and a resulting increase of the basic subsidy), as well as it will not push further a career of a researcher who shares data (in terms of doctoral degree, habilitation or full professorship). Moreover, as units and researchers are more and more often expected to include a commercial factor in their research, they try to be cautious and - just in case - keep their data closed, as opening is perceived as threatening commercialisation. Although in some rare cases this attitude may be justified, as a *de facto* dominant rule it is harmful in the majority of cases where there are

no specific plans concerning commercialisation and the possibility of monetisation (of the data itself or of other things related to data in such a way that making data open would destroy the commercial potential of this other thing).

- As a result, there is a lack of institutional interest in sharing data and no appropriate policies follow. Researchers willing to share data are in a vast majority of cases left with no support from their home institutions in terms of financial, technical or legal backup.

Appendix: Recommendations for the Polish Academia

In order to foster and facilitate the further development of the open research data landscape in Poland, we provide the following recommendations:

1. Research funding institutions are recommended to:

- require Data Management Plans from their grant applicants (delivered either at the point of application or at a later stage during the project);
- ensure that costs related to data management and data sharing are eligible in grant applications;
- launch open data pilots in selected research areas.

2. Scientific institutions are recommended to:

- introduce institutional research data policies, which should address topics such as (a) incentives and obligations for open sharing of data, (b) strategies for access to the necessary infrastructure, (c) strategies for safe, long-term data preservation, (d) legal aspects, and others;
- build competencies and skills of professional support staff (data managers, data curators);
- educate students and researchers on data management and sharing (by including data topics in curricula, offering workshops, etc.).

3. Publishers of academic journals are recommended to:

- introduce policies concerning data underlying published articles.

4. Researchers are recommended to:

- develop their data competencies and skills;
- plan management of research data at the early stage of their projects;
- archive valuable data in suitable repositories;

- open valuable data when possible.

5. The Ministry of Science and Higher Education is recommended to:

- support further deepened research concerning sharing data and its benefits, including the identification of best practices, strategies, policies, etc.;
- raise awareness among members of the academic community concerning data management and sharing;
- gradually include and reward data sharing in evaluation processes;
- recognise data management and sharing as part of research career development;
- provide a funding stream for projects and initiatives aimed at data sharing;
- launch a public discussion involving all stakeholders concerning various aspects of a national research data policy;
- develop and implement an (open) research data policy on the national level.

6. General recommendations for all stakeholders involved:

- ensure interoperability of infrastructure;
- ensure flexibility of infrastructure;
- ensure usability of infrastructure;
- ensure long-term preservation;
- provide training and education (including legal issues).

Bibliography

Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities, http://openaccess.mpg.de/Berlin-Declaration.

Dryad, Frequently Asked Questions, http://datadryad.org/pages/faq#deposit.

Fecher B., Friesike S., Hebing M., Linek S., Sauermannn A., *A Reputation Economy*: *Results from an Empirical Survey on Academic Data Sharing*, 2015. DIW Berlin Discussion Paper No. 1454. Available at SSRN: http://ssrn.com/abstract=2568693 or http://dx.doi.org/10.2139/ssrn.2568693.

Fecher B., Friesike S., Hebing M., *What Drives Academic Data Sharing*? PLoS ONE 10(2): e0118053. doi: 10.1371/journal.pone.0118053.

Horizon 2020 Annotated Model Grant Agreements, Version 2.1, 30 October 2015, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf.

Heinz P., Dallmeier-Tiessen S., *Open Research Data: From Vision to Practice*, 2014, http://book.openingscience.org/vision/open_research_data.html.

Judgement of the Court of 9 November 2004, The British Horseracing Board Ltd and Others v William Hill Organization Ltd., C-203/02, ECR I-10415, http://curia.europa.eu/juris/liste.jsf?num=C-203/02.

Murray-Rust P., Neylon C., Pollock R., Wilbanks J., *Panton Principles. Principles for open data in science*, 2010, http://pantonprinciples.org/.

OECD, *Declaration on Access to Research Data from Public Funding*, http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157.