Open Research Data: The current and future use of repositories by the Swiss research community

Markus von der Heyde[1]

[1]vdH-IT, Weimar, Germany, ORCID: 0000-0002-6026-082X, info@vdh-it.de

Keywords
Open Research Data, Data Repository, Data Sharing Practice, Data Reuse Practice.

1. ABSTRACT

One way to share the results of scholarly production is to upload publications, in conjunction with the research data underlying them, into a data repository. As of 2017, the Swiss National Science Foundation (SNSF) has mandated making data from funded projects accessible. This provided a reason to examine the sharing and reuse behavior of researchers in the Swiss community in 2018. A "landscape survey" across the complete Swiss research community collected information from 2,384 scientists about their data sharing and reuse practices. In addition, a "repository survey" added the perspective of 208 international repositories and their plans for future development. The results were analyzed using statistical methods.

Generally, the motivation and concerns for data sharing and reuse in the Swiss community are not different from other scientific communities. Overall, about a third of the Swiss research community share data in repositories. The drivers of sharing and reuse that are summarized in this paper, in the sense of motivations and intentions to share and reuse, replicate the main findings in previous literature. We also compare and correlate these drivers with actual data reuse and sharing practice.

The Swiss research community uses international repositories extensively: 75% are based in the EU or internationally. Switzerland provides institutional or financial support for only 13% of all repositories mentioned. Future requirements for services from the Swiss community are not yet met by the international repositories' plans.

This paper focusses on recommendations to local support units and IT service providers at the university level. Suggestions for local measures as well as collaborative approaches are anchored in the results gained from the original data analysis. A change of perspective is the main point: If organizations encourage reuse of data, sharing will consequently rise.

2. INTRODUCTION

Up to now, sharing and reusing data has been perceived as a core necessity in some disciplines and regarded as superfluous in others (Borgman, 2012; Hahnel et al., 2018; Linek, Fecher, Friesike, & Hebing, 2017; Pasquetto, Randles, & Borgman, 2017). At the same time, data analytics and quantitative methods are used by individual scientists in the majority of scientific fields (Stratmann, 2018; von der Heyde, Hartman, Auth, & Erfurth, 2018). Therefore, one might assume that sharing and reusing data will be considered the norm at some point in the future, in the same way as scientific publications are nowadays (von der Heyde, Auth, Hartman, & Erfurth, 2019). The recent past have shown a variety of views about the question, if data which is shared is also used (Wallis, Rolando, & Borgman, 2013).

Today, different disciplines use different means to publish research papers, most probably driven by cultural and individual factors. One might thus expect the disciplinary communities to also develop new standards on how to publish the corresponding research data (Kim, 2017; Kim & Stanton, 2012, 2016; Rat für Informationsinfrastrukturen (RfII), 2016). Concerns about privacy and intellectual property rights are sometimes higher regarding the publication of medical or social data; sometimes, this rightly prevents open access to sensitive data (National Institutes of Health (NIH), 2003).

In fact, most data on which publications are based can and probably should be shared within the scientific community. The majority of scientists believe in this view and about two thirds shared their data from publications in 2018 in some way (Hahnel et al., 2018).

Funding agencies around the world at least recommend, or more strongly demand, making access to data (historical artefacts, simulations, empirical research data, concepts, and primary literature) obligatory, if no data privacy or other major concerns preclude it. One way to share the results of scholarly production is to upload publications, in conjunction with the research data underlying them, into an Open Data Repository. In Switzerland, the Swiss National Science Foundation (SNSF) and swissuniversities are encouraging this and plan to support the research community with appropriate funding (Yılmaz, 2018). Therefore, a study was mandated by the SNSF to examine the sharing and reuse behavior of researchers in the Swiss community, as well as which repositories they use (von der Heyde, 2019c).

The data (von der Heyde, 2019e, 2019f, 2019d) and initial analysis (von der Heyde, 2019b, 2019a) from the two surveys are published on Zenodo, as required by the contracting authorities (SNSF and swissuniversities). This paper concentrates on the perspective of the higher education community, specifically on the (IT-)service providers within and around research-focused disciplines.

The key questions of this article are:

- What can we learn from a nationwide survey in Switzerland for other countries?
- What are the potential drivers for more sharing and effective reuse?
- How can organizations support local researchers?
- What are effective measures for collaboration in the support of data literacy?

The next section describes the background of data and methods in a minimal way, while later sections concentrate on the answers to the core questions. The paper concludes with open questions and an outlook on the future development of data reuse and sharing in the scientific context.

3. BRIEF DESCRIPTION OF THE SURVEYS

As the data and methods have already been published on Zenodo, we limit the background to the minimum required to understand the validity and consequences of the additional analysis done for this paper.

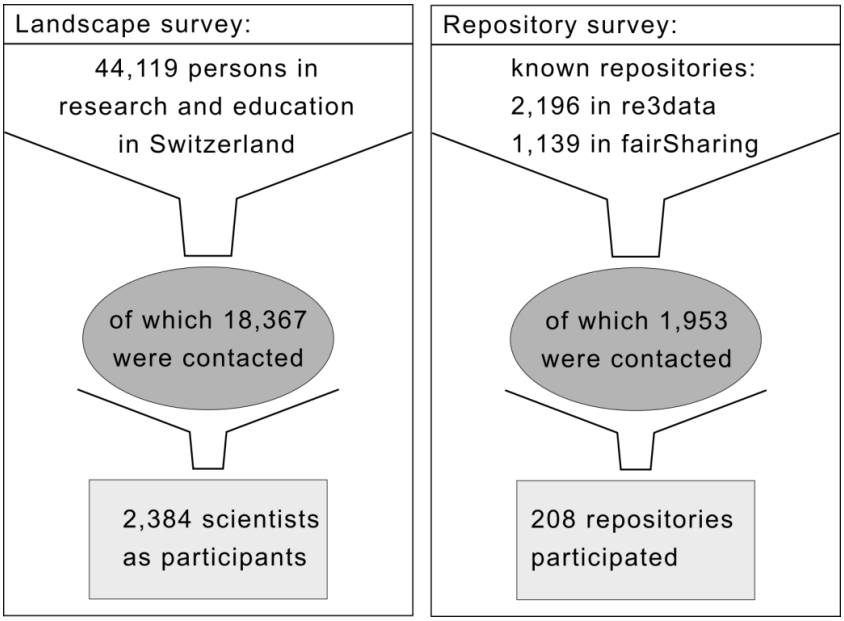

Figure 1: Overall size of scientific community and number of repositories in comparison to survey participation.

The overall participation (see Figure 1) was satisfactory. Both survey designs were based on prior work. The landscape survey mostly used questions validated earlier in the specific scientific context (Ajzen, Fishbein, Lohmann, & Albarracín, 2005; Fecher, Friesike, & Hebing, 2015; Kim, 2017; Kim &

Stanton, 2012, 2016; Kim & Yoon, 2017; Linek et al., 2017). Questions addressing motivation and practices about scientists' behaviors were changed as little as possible, preserving the meaning. They were posed symmetrically for reuse and sharing. In addition, the actual number of publications for which data was either shared or reused was assessed. General individual factors like age, gender and scientific seniority were also collected and correlated with the overall data.

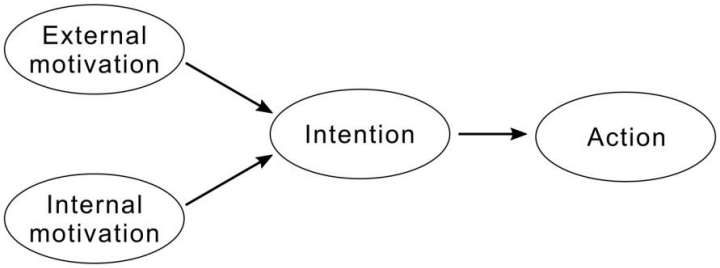

Figure 2: Extended research design.

While Kim and colleagues had mainly investigated the links between motivation and intention, the Swiss survey also included the resulting action. At the same time, sharing and reuse were assessed with the same set of questions. Thus, the survey design (shown in Figure 2) extended Kim's et al. framework on data sharing and the underlying factors to the link between data reuse and sharing from motivation to action.

4. THE SWISS COMMUNITY AS AN INTERNATIONAL EXAMPLE

The Swiss scientific community is internationally well networked. Schneider and Sørensen display in their Figure 1 that Switzerland is in the lead, in the sense that 70% of all publications in 2014 involved international partners (Schneider & Sørensen, 2015). For most countries, the rate of internationally produced papers is increasing; therefore, one can assume that other countries will reach the Swiss level at some point. On a global level, publications involving scientists from Switzerland cover about 2% of the global scientific output. However, the total size of the Swiss community is small in comparison to the USA or Germany.

The validity of the survey was demonstrated at the level of the scientific design, as well as at the participatory level, in the data papers (von der Heyde, 2019b, 2019e). The comparison to other international surveys with similar research designs were shown in the initial analysis and have also been published on Zenodo (von der Heyde, 2019c). Due to a larger data set, some of the effects which had been only hypothesized earlier by Kim and colleagues could be confirmed. Also, all major effects seen in previous work by Kim and colleagues (Kim & Stanton, 2012; Kim & Yoon, 2017) were replicated.

Since we could not identify a survey including a higher overall number and proportion of scientists for any other country, we consider the collected data set a typical example of a national scientific community. Moreover, given the highly developed collaborative environment within Europe, including Switzerland, we assume that the identified effects would not be limited to the Swiss community.

However, one main effect seems to be linked directly to the Swiss community: When scientists were asked to state repositories they are using, the Swiss repositories were mentioned, particularly at the institutional and national level. Therefore, Zenodo and FORSbase were mentioned more often than would be probably expected in a balanced international group of scientists.

In the survey, Zenodo appeared as the biggest not-for-profit general purpose repository. Given the high reputation of CERN, which is the main development partner, this does not come as a surprise. FORSbase covers data from many disciplines within the social sciences. However, it should not be considered as an alternative to general purpose repositories like Zenodo. It can be better classified as a method-based approach to collect data from a variety of social disciplines and humanities. Unlike other international repositories (e.g., GEO, OSF, ENA), the intensive usage of FORSbase cannot be explained by pure discipline-specific effects, since the overall occurrence rate with respect to the overall participation is much higher.

While these two Swiss repositories are among the top three mentioned, the Swiss research community does not solely focus on national resources. On the contrary, they extensively rely on international repositories: Institutions from Switzerland are involved in only 25% of the repositories mentioned. The other 75% are based within the EU or internationally. Switzerland provides institutional or financial support for 13% of all repositories mentioned. Therefore, Swiss scientists use 87% of repositories in which the Swiss community does not have a direct financial investment.

5. DRIVERS FOR SHARING AND REUSE

Various drivers for data sharing and reuse were systematically evaluated by Fecher et al., who described the process from the researchers' perspective (Fecher et al., 2015). Kim and colleagues systematically studied the internal and external motivation which drives the sharing and reuse intention (Kim & Stanton, 2012, 2016; Kim & Yoon, 2017). How to assess the link between intention and action was studied by Ajzen et al. (Ajzen et al., 2005). Our research in part used identical questions in order to replicate the known baseline. While this anchored the approach and expected results, we focused on the application on the Swiss research community.

The labels of all variables in this chapter are in reference to the original studies and subsequent evaluations of the data. See also Table 2 and Table 3 of the Appendix for the rated statements.

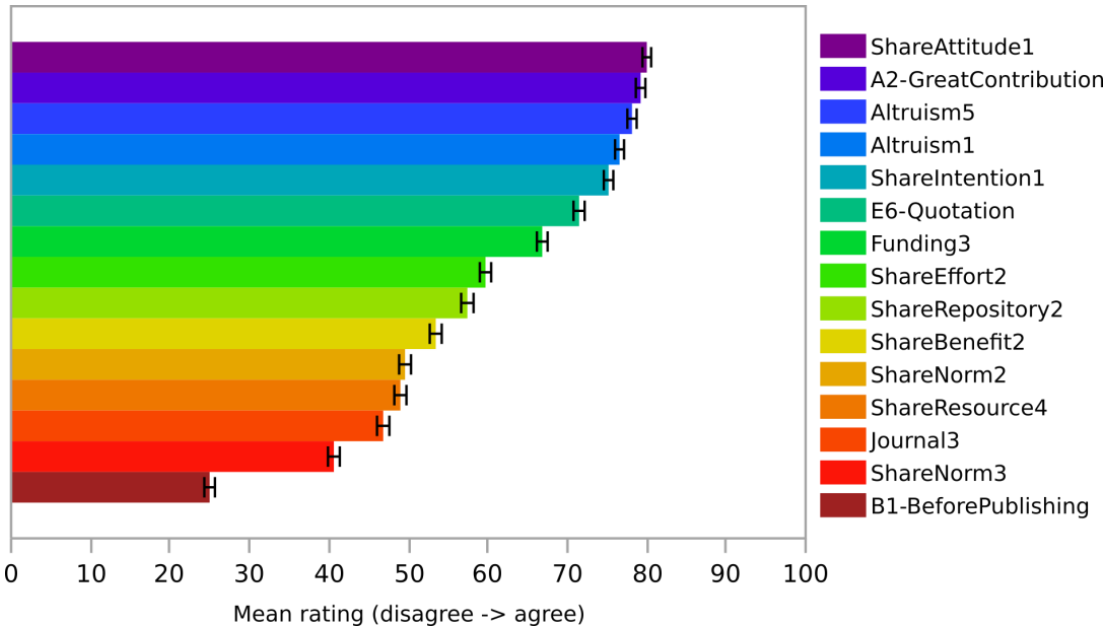

Figure 3: Mean values of the drivers for sharing data. Error bars depict one standard error from the mean.

Within the initial analyses, known drivers from other international surveys (e.g., Kim et al.) were confirmed by using principal component analysis and factor analysis methods. In addition to the common (altruistic) motivations and intention, the future citation of published data is the most prominent reason for sharing data (see Figure 3). The requirement to publish data by the funding agency is rated astonishingly low, if compared to the SNSF policy.

The factors regarding the potential reuse of data were assessed across all disciplines. Again, the international findings were confirmed. Figure 4 shows the mean ratings of all assessed variables concerning reuse of data. Key drivers are the perceived usefulness of the reused data. Altruism and the discipline-specific climate of reuse are important to forming the intention to reuse. As well, the effort connected to reusing data is perceived as important by the participants.

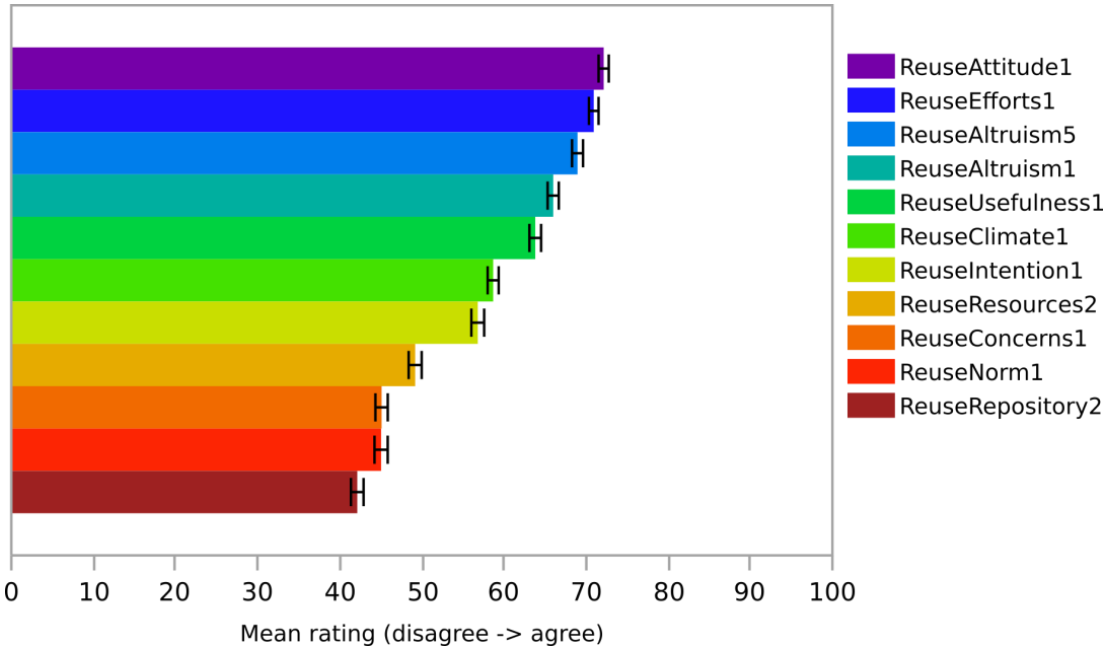

Figure 4: Mean values of the factors for reuse of data. Error bars depict one standard error from the mean.

Additional correlation analysis revealed the most prominent factors for the actual sharing activity of scientists. See Table 1 for the average sharing frequency of scientists, which was highly correlated with the average reuse frequency. The most highly rated statement was "In my discipline, researchers can easily access data repositories to reuse data." These results are consistent with the hypothesis that scientists who successfully reuse other researchers' data are in fact sharing their data more often.

Table 1: Correlations of factors from actual sharing and reuse with the ratings of the statements in connection with intended sharing and reuse.

Factor	AvgSharing Frequency	Rated statements
AvgReuseFrequency	0.63	calculated as: papers with reused data / total papers
SharingRatio	0.43	calculated as: papers with shared data / total papers
ReuseRepository2	0.39	In my discipline, researchers can easily access data repositories to reuse data.
ShareNorm2	0.39	In my discipline, researchers care a great deal about data sharing.
ShareIntention1	0.38	I am likely to share my data from future research.
ReuseNorm1	0.36	In my discipline, it is expected that researchers reuse other researchers' data.
Journal3	0.35	Journals require researchers to share data.
ReuseIntention1	0.34	I am likely to reuse other researchers' data for my future research.
ShareRepository2	0.33	In my discipline, data repositories are available for researchers to share data.

6. POTENTIAL ACTIONS

As the original report to the SNSF contained recommendations on the policy level, this paper focusses on the perspective of the IT service units and universities responsible for the direct support of researchers.

6.1. LOCAL SUPPORT

Both scientists in the landscape survey and representatives of the repositories participating in the international repository survey rated the need for future services. The initial analysis identified major demand for enhancements in data security and support for legal issues.

Part of this requirement is the introduction of community-wide authorization and authentication. One option for repositories would be to join a federation like eduGAIN to implement access mechanisms. As well, general measures of data protection have to be implemented by the data repositories. Extensive support for legal issues, on the other hand, is part of the local policy and therefore needs support by the local organization (see chapter 9 in (von der Heyde, 2019c)).

While asking scientists to consider (or require) sharing data, organizations need to establish a better understanding of the policies and regulations around it. This is also tightly linked to the overall knowledge of existing resources for reuse and sharing of existing data.

6.2. COLLABORATIVE APPROACHES

Both national initiatives and international approaches have been put forward in the recent past to address extensive support of repositories for open data. Clearly, there seems to be no alternative to public funding of scientific data repositories. The current revenue streams often stay below 15% of the current costs. Paying for the use of already funded project results is not yet an established practice. It might be attractive for big publishers like Elsevier and Springer to offer platforms for storing research data. Currently, they do not charge users for a retrieval. However, this could change without notice to all original data owners.

Overall, there seems not to be a general shortage of either discipline-specific or general purpose repositories. On the contrary, the landscape is highly fragmented (see chapter 5 in (von der Heyde, 2019c)). It potentially needs consolidation, because the cost for running a repository has increased over time. Whether and how bigger clusters of repositories could reduce financial burdens remains open. Also, the question remains of how funds could be spent most effectively for the community.

The overall user priorities for future services are clearly not met by the priorities repositories report in their future planning (see chapter 6 in (von der Heyde, 2019c)). Coordination between user demand and overall planning could provide a benefit. Ideally, a large community which is not limited to discipline-specific channels should address this matter.

In order to monitor this benefit, a survey across the international research community could be performed repeatedly. Long-term documentation of the joint efforts should be included into any funding program.

Finally, as sharing is a prerequisite for meaningful reuse, but does not make sense in and of itself, we should revise our focus. Only if data is reused – that is, only if it *can be* reused – is the additional work of data curation worth the effort. The solution suggested by the survey data is to train students and young scientists how to reuse data. This ultimately will lead to increased data sharing (and better reuse) in the future. Any initiative which supports the scientists and eases reuse would be of collaborative benefit if transferred to other higher education institutions.

7. SUMMARY

The use of data repositories is one way to share (provide and reuse) the data which underlies published scholarly papers. The use of repositories promises long-term preservation and access, in comparison to the typical personal copy of the researcher. Therefore, funding agencies recommend and sometimes expect the use of structured data storage in data repositories. In the light of the Swiss National Science Foundation (SNSF) decision for funded researchers to provide a data management plan, which could involve a data repository, a study with a multilevel approach was

conducted and evaluated. As the original report addresses the policy and funding agency level, this paper focusses on recommendations to local support units and IT service providers at the university level. Local measures like the implementation of community-wide authorization and authentication, as well as collaborative approaches like consolidation of repositories, are anchored in the results gained from the original data analysis. One central change of culture concerning the researchers' responsibility is grounded in a statistical finding: encouraging people to reuse (e.g., while writing their master or PhD thesis) should also increase the intention to share data, since the current research indicates that these are correlated.

8. REFERENCES

Ajzen, I., Fishbein, M., Lohmann, S., & Albarracín, D. (2005). The influence of attitudes on behavior. *The Handbook of Attitudes*, *173*(221), 31. Retrieved from https://psycnet.apa.org/record/2005-04648-005

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. https://doi.org/10.1002/asi.22634

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *PLOS ONE*, *10*(2), 25. https://doi.org/10.1371/journal.pone.0118053

Hahnel, M., Fane, B., Treadway, J., Baynes, G., Wilkinson, R., Mons, B., … Osipov, I. (2018). *State of Open Data 2018*. Retrieved from figshare website: https://doi.org/10.6084/m9.figshare.7195058.v1

Kim, Y. (2017). Fostering scientists' data sharing behaviors via data repositories, journal supplements, and personal communication methods. *Information Processing & Management*, *53*(4), 871–885. https://doi.org/10.1016/j.ipm.2017.03.003

Kim, Y., & Stanton, J. M. (2012). Institutional and individual influences on scientists' data sharing practices. *Journal of Computational Science Education*, *3*(1), 47–56. https://doi.org/doi:10.1234/12345678

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799. https://doi.org/10.1002/asi.23424

Kim, Y., & Yoon, A. (2017). Scientists' data reuse behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *68*(12), 2709–2719. https://doi.org/10.1002/asi.23892

Linek, S. B., Fecher, B., Friesike, S., & Hebing, M. (2017). Data sharing as social dilemma: Influence of the researcher's personality. *PLOS ONE*, *12*(8), e0183216. https://doi.org/10.1371/journal.pone.0183216

National Institutes of Health (NIH) (Ed.). (2003, March 5). *NIH Data Sharing Policy and Implementation Guidance*. Retrieved from https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). *On the Reuse of Scientific Data*. *16*, 8. https://doi.org/10.5334/dsj-2017-008

Rat für Informationsinfrastrukturen (RfII). (2016). *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Retrieved from http://www.rfii.de/?wpdmdl=1998

Schneider, J. W., & Sørensen, M. P. (2015). Measuring research performance of individual countries: the risk of methodological nationalism. *Paper for the ECPR General Conference in Montreal, Aug*.

Stratmann, M. (2018). Herausforderungen durch die Digitalisierung – Forschung und Kommunikation im Umbruch. *Chemie Ingenieur Technik*, *90*(9), 1133–1133. https://doi.org/10.1002/cite.201855002

von der Heyde, M. (2019a). *International Open Data Repository Survey: Description of collection, collected data, and analysis methods* [Data paper]. Retrieved from https://doi.org/10.5281/zenodo.2643450

von der Heyde, M. (2019b). *Open Data Landscape: Repository Usage of the Swiss Research Community: Description of collection, collected data, and analysis methods* [Data paper]. Retrieved from https://doi.org/10.5281/zenodo.2643430

von der Heyde, M. (2019c). *Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future* [Report to the SNSF]. Retrieved from https://doi.org/10.5281/zenodo.2643460

von der Heyde, M. (2019d, April). *Data and tools of the landscape and cost analysis of data repositories currently used by the Swiss research community*. Retrieved from https://doi.org/10.5281/zenodo.2643495

von der Heyde, M. (2019e, April). *Data from the International Open Data Repository Survey*. Retrieved from https://doi.org/10.5281/zenodo.2643493

von der Heyde, M. (2019f, April). *Data from the Swiss Open Data Repository Landscape survey*. Retrieved from https://doi.org/10.5281/zenodo.2643487

von der Heyde, M., Auth, G., Hartman, A., & Erfurth, C. (2019). Skalierung von Plattformen in der disruptiven Digitalisierung der Forschung. In T. Barton, C. Müller, & C. Seel (Series Ed.), *Angewandte Wirtschaftsinformatik*: *Vol. 5. Digitalisierung in Hochschulen*. In print. Retrieved from http://springer.com/series/13757

von der Heyde, M., Hartman, A., Auth, G., & Erfurth, C. (2018). Forschung in der disruptiven Digitalisierung von Hochschulen - Faktoren der Skalierung und ein Zukunftsszenario. *Informatik Spektrum*, *41*(6), 359–368. https://doi.org/10.1007/s00287-018-01126-1

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE*, *8*(7), e67332. https://doi.org/10.1371/journal.pone.0067332

Yılmaz, A. (2018, June). Call for tenders for a landscape and cost analysis of data repositories - SNF. Retrieved 15 June 2018, from http://www.snf.ch/en/researchinFocus/newsroom/Pages/news-180611-call-for-proposals-call-for-tenders-for-a-landscape-and-cost-analysis-of-data-repositories.aspx

9. APPENDIX

In the landscape surveys, the participants rated the statements shown in Table 2 for sharing (presented in random order) on a continuous scale between 0% (disagreement) to 100% (agreement) using a slider.

Table 2: Rated statements in the landscape survey on data sharing.

Reference / Variable	Rated sharing statement on scale 0%=disagree ... 100%=agree
Motivation and data sharing behavior	
Altruism1	I am willing to help other researchers by sharing data.
A2-GreatContribution	Freely available research data is a great contribution to scientific progress.
Altruism5	Sharing data contributes to better scientific research.
Perceived career benefit	
ShareBenefit2	Data sharing would enhance my academic recognition.
E6-Quotation	I would share my data if I were cited in publications using my data.
Perceived career risk	
B1-BeforePublishing	I would share my data even if other researchers could use my data to publish before me.
Perceived effort	
ShareEffort2	I need to make a significant effort to share data.
Attitude toward data sharing	
ShareAttitude1	Sharing data is valuable.
Normative pressure	

Reference / Variable	Rated sharing statement on scale 0%=disagree ... 100%=agree
ShareNorm2	In my discipline, researchers care a great deal about data sharing.
ShareNorm3	In my discipline, researchers share data even if not required by policies.
Perceived availability of data repositories	
ShareRepository2	In my discipline, data repositories are available for researchers to share data.
Perceived pressure by funding agencies	
Funding3	Public funding agencies require researchers to share data.
Perceived pressure by journals	
Journal3	Journals require researchers to share data.
Resources	
ShareResource4	In my organization (e.g., university), information technologies are available to support my data sharing.
Intention to share data	
ShareIntention1	I am likely to share my data from future research.

Also the participants rated statements shown in Table 3 for reuse in a separate block, again in random order on a continuous scale between 0% (disagreement) to 100% (agreement) using a slider.

Table 3: Rated statements in the landscape survey on reuse of data.

Reference / Variable	Rated reuse statement on scale 0%=disagree ... 100%=agree
Motivation and data sharing behavior	
ReuseAltruism1	I am willing to reuse others' data for my research.
ReuseAltruism5	Reusing others' data contributes to better scientific research.
Perceived Usefulness	
ReuseUsefulness1	Reusing other researchers' data improves the quality of my research.
Perceived Concern	
ReuseConcerns1	If I reuse other researchers' data I worry that I might misinterpret the data.
Perceived Effort	
ReuseEfforts1	Reusing other researchers' data requires time and effort to locate data sets.
Attitude towards data use	
ReuseAttitude1	Reusing other researchers' data is valuable.
Subjective Norm	
ReuseNorm1	In my discipline, it is expected that researchers reuse other researchers' data.
Availability of data repositories	
ReuseRepository2	In my discipline, researchers can easily access data repositories to reuse data.
Organizational Resources	
ReuseResources2	In my organization (e.g., university) information technologies are available to support my data reuse.
Disciplinary Climate	
ReuseClimate1	In my discipline, researchers cooperate well.
Intention to Reuse Other Researchers' Data	
ReuseIntention1	I am likely to reuse other researchers' data for my future research.

10. AUTHOR'S BIOGRAPHY

Dr. Markus von der Heyde received his PhD in Computer Sciences from the University of Bielefeld, Germany for his work at the Max Planck Institute for Biological Cybernetics Tübingen in 2000. His approach to adopt biological principles into distributed computer applications in order to enhance stability and robustness was applied in DFG and EU funded research projects. Between 2003 and 2011 he served as ICT director of Bauhaus University in Weimar and focused on topics such as information security, service management, strategy and governance. Since 2011, he has been the CEO of vdH-IT and a management consultancy firm specializing in IT governance and digital transformation in higher education. In cooperation with various partners, he has conducted the German CIO studies since 2014. Recently, he conducted the Open Data Repository Landscape Analysis for the Swiss National Science Foundation (SNSF).
See more details on https://www.researchgate.net/profile/Markus_Von_Der_Heyde3.