gesis Leibniz-Institut für Sozialwissenschaften

GESIS *Papers* 2017|26

Implementing Open Science: The GESIS Perspective

Christof Wolf

GESIS Papers 2017|26

Implementing Open Science: The GESIS Perspective

Talk given at Institute Day of GESIS, 28 September 2017

Christof Wolf

Support from Bernhard Miller in preparing this talk is greatfully acknowledged.

GESIS – Leibniz-Institut für Sozialwissenschaften 2017

GESIS Papers

GESIS Papers

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: +49 (0)621 / 12 46 -149
Telefax: +49 (0)621 / 12 46 -100
E-Mail: christof.wolf@gesis.org

ISSN: 2364-3781 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

1 Introduction

The concept of Open Science has recently experienced quite a career. Open Science is described as the most profound change in the scientific process in a long time. “Openness” is embraced by many researchers as *the* way to collect, process and generate new knowledge. And it is certainly a rare case in which funders and other stakeholders outside the scientific community have taken such interest in *how* science produces knowledge.

However: Open Science has had a difficult time to make progress. Open Access, for example, has been around since the 1990s and yet we are far from universal open access. Or take Open Data. Evidence suggests that data is often not shared despite journal and / or funder policies.

To address the issue of "Implementing Open Science" we therefore need to address this puzzle: Why is something that is welcomed by so many so difficult to get going?

And if there are useful answers: How can an infrastructure providers like GESIS help overcome these problems?

2 Puzzle

Let me back up the puzzle with some more data:

Since the Budapest Declaration on Open Access in 2002 many more such declarations, national and international in scope have been published. The EU [1] demands publications from its Horizon 2020 projects to be Open Access, as does the Swiss National Fund. While less unequivocal, there is further backing: The German Ministry of Education and Research (BMBF) demands - but doesn't sanction – Open Access publication; for DFG-funded projects, open access is recommended. German copyright law has been changed to facilitate Open Access. Now German publications can be "republished" (to translate the impossible "Zweitveröffentlichung").

Undoubtedly this backing and support alone is a success. It shows that the appeal of Open Access has convinced many, within but also beyond the scientific community.

And there is progress: According to the EU Open Science Monitor, 32% of all publications referenced by the Web of Science in 2015 were published in green Open Access – up from 7% in 1991! That's a 4.5-fold increase in the publications deposited in an open access repository by their authors.

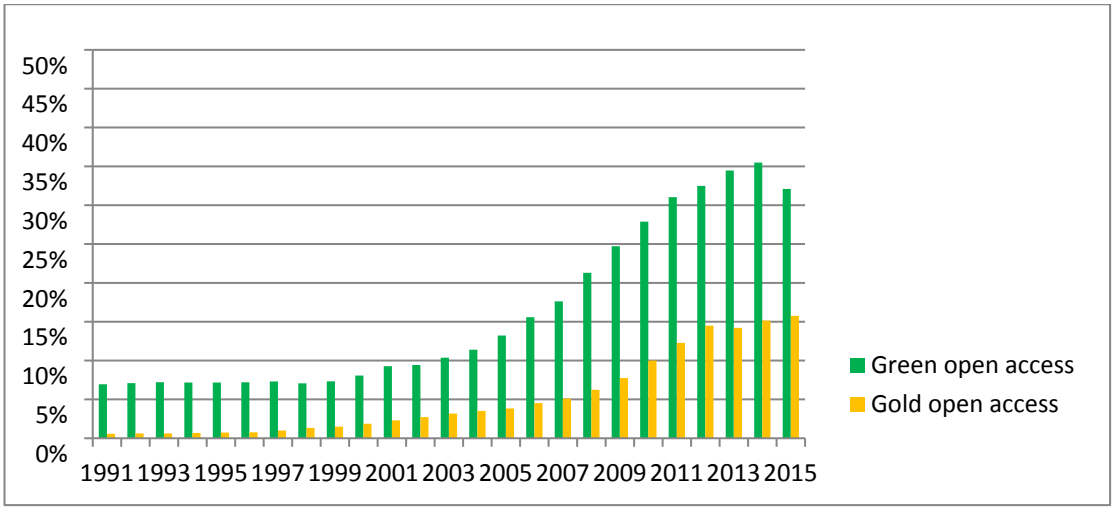

Figure 1: Progress made. The Growth of Open Access as measured by the percentage of publications in the Web of Science which are openly available
(Source: European Commission: Open Science Monitor)

Yet, put in perspective a 4.5-fold increase is not that impressive. By comparison, the number of Facebook users multiplied by 20 between 2008 and 2017. Or, maybe a fairer comparison: between 2009 and today ResearchGate increased its user base from 25,000 to 3 million: a 120-fold increase in less than a decade. [2]

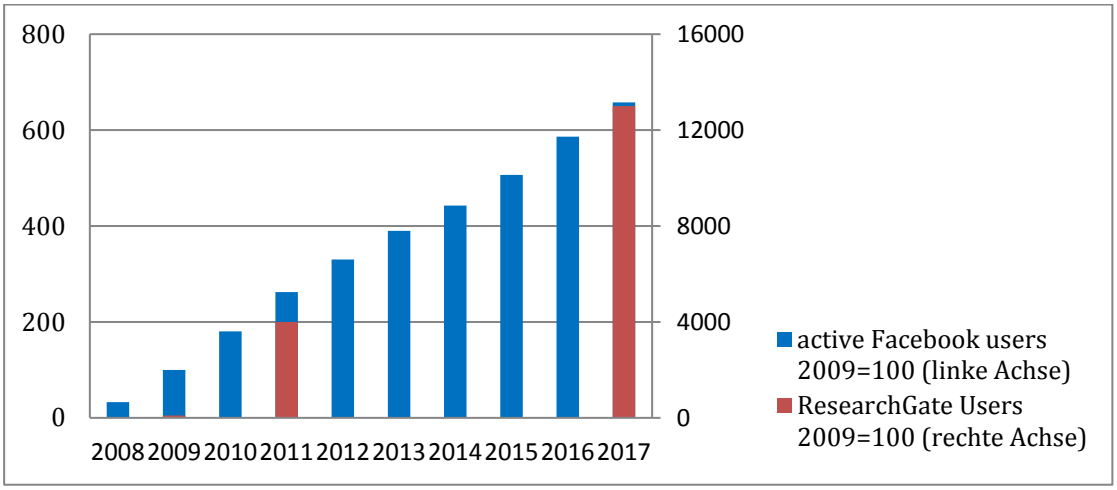

Figure 2: Putting it into perspective. The Growth of Social Networks (Source: Statista, Research Gate, [2])

Let's look at Open Data or data-sharing: There is one impressive example: The Human Genome Project, the effort established in the 1990s to decode human DNA, was a huge step forward for science. The project was able to complete its mission more quickly and more comprehensively than planned because of Open Data. All data collected in the research alliance had to be documented according to an agreed-upon standard and shared online within 24 hours. Impressive!

The DFG's long-term program for the Social Sciences offers some analogy. Consortia, e.g. the German Longitudinal Election Study or TwinLife collect data for entire disciplines, and make them openly and timely available.

But yet again: Open Data is far from being the rule - or even widely accepted practice. According to a study by the Publishing Research Consortium (PRC) in 2010 access to datasets, data models, algorithms and programs was almost universally ranked important or highly important; however, only 38% felt that data were in fact easily accessible.

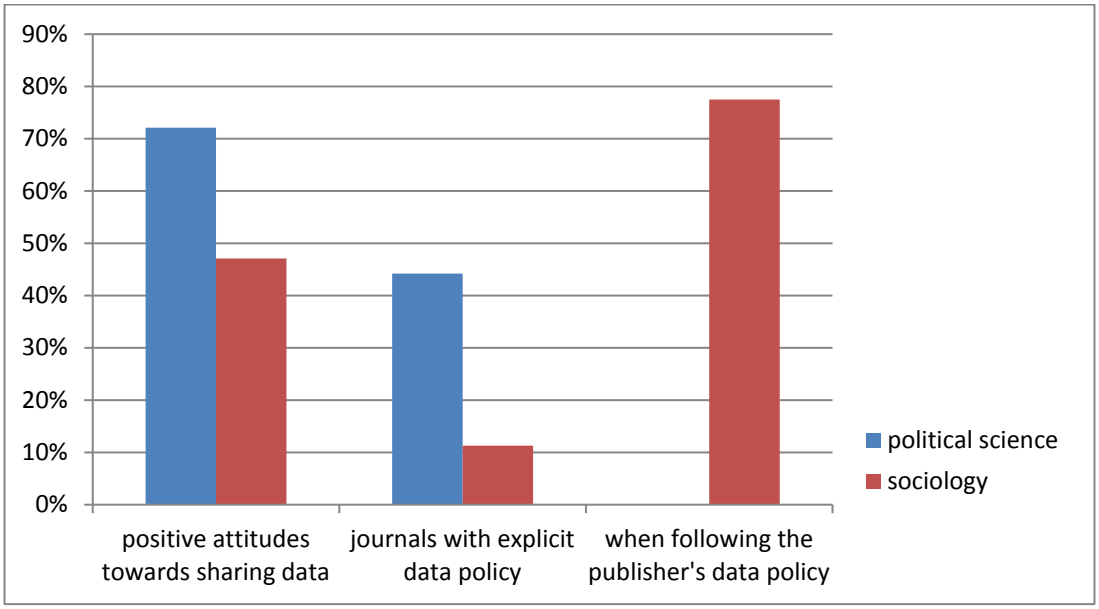

Figure 3: Journals' Role in Data Sharing (Zenk-Möltgen et al., forthcoming)

A forthcoming study by our colleagues Wolfgang Zenk-Möltgen, Alexia Katsanidou, Esra Akdeniz and Verena Naßhoven among other things look at data-sharing policies in journals. [3] Clear rules, it would seem, are a good way to encourage sharing. Of 142 journals in sociology only 16 (11.3%) had an explicit data policy in 2016. If, however, we also count journals which follow their publishers' data policies the number increases to 110 of 142 journals (77.5%). In political science 53 of the 120 selected journals (44.2%) had some sort of data policy for their authors. These numbers are encouraging. And, Wolfgang and colleagues have established that there is a significant correlation between the existence of policy and research data published.

However, from the authors of empirical articles in the study just over half (56.5%) stated that their data are available to other researchers. More sobering still: only for one third (36.5%) of the articles in the study's sample the data could actually be found. This hints at yet another problem - findability. When we did a survey of all Social Scientists in Germany in the course of our regular portfolio analysis last year we found that the second most prominent way to "archive" research data (after GESIS) in the Social Sciences are (private) homepages. As we all know there is no guarantee that these data will be available in the long-term. Therefore, the puzzle remains: When Open Science is concerned there is a clear gap between what is desired and the actual situation.

3 Analysis and Basic Principles for OS Infrastructure

My attempt at an analysis will be brief - and focus on two levels of explanation. Individual and institutional factors:

In a recent paper Benedikt Fecher and co-authors (2017) have looked at how individual considerations trump systemic arguments.[4] The most prominent concern cited for not sharing data: "others could publish before me". Furthermore, 59% of respondents (across disciplines) said that "considerable" effort would deter them from sharing.

This leads me to institutional factors. In a 2011 survey among environmental scientists there was much support for the proposition: "Many organizations do not provide support to their researchers for data management both in the short- and long-term."[5]

GESIS has therefore thought about how to address both the institutional and the individual challenges that the broad adoption of Open Science principles still face.

Generally, we assume that Open Science can be promoted by

- offering easy-to-use technical solutions for documenting and sharing,
- establishing standards for documentation,
- establishing technical standards for interfaces to ensure interoperability of Open Science repositories,
- offering training for scholars on engaging in Open Science practices,
- creating incentives for scholars to make their scientific result available openly.

We are convinced that progress in these areas will lower the effort needed to make a paper, a dataset or source code openly available.

4 The GESIS Perspective

Our concept for Open Science at GESIS rests on four pillars:

- Open Access,
- Open Data,
- Open Source and
- Open Methodology.

4.1 Open Access

As we saw, there are some structural problems that help understand why Open Access has not fully met expectations. From a research organization's perspective there is another one: at least until now and particularly in the Social Sciences maximizing impact and maximizing Open Access are (at least partly) conflicting targets. Currently, there are no high-impact journals in the Social Sciences that are also fully Open Access. I do not only say this as President of an institute that faces its next evaluation in eight months. Despite being an utterly useless measure of any particular contribution, the impact-factor and its cousins are still widely used criteria to assess scientific output. So if I want to demonstrate that our research is noticed in the community, there is a dilemma. It might only be a problem of the choice of indicator – but currently it is a highly risky strategy to aim only or mainly for Open Access contributions because this would mean avoiding high-impact journals. GESIS has not resolved this dilemma.

We have, however, decided to focus our efforts on green Open Access. As some of you know, GESIS is home to the Social Science Open Access Repository SSOAR (and, in fact, also hosts the LeibnizOpen-Server), Germany's sixth-best repository and second only to ZBW's repository EconStor within Leibniz, according to a recent worldwide ranking. SSOAR scores particularly well in the indexation in Google-Scholar and the size/number of full-texts (2nd place in Germany).

To contribute to Open Access GESIS will therefore create a declaration of consent for its employees, which will facilitate the automatic re-publication on SSOAR. GESIS will also systematically expand SSOAR on the basis of its research data bibliographies (publications citing data archived with GESIS). To further increase the visibility of Open Access contributions GESIS is examining the integration of altmetrics into SSOAR.

4.2 Open Data

Sharing data spawns new work through alternative analysis and updates; it reduces overall research costs through reusing and repurposing existing data; it enables verification and replication.

Since its foundation in 1960 the Data Archive has been committed to Open Data. One key contribution to Open Data therefore will be to continue doing what we have proven to be good at. But, we aim at increasing institutional support to further lower barriers to data-sharing. An important measure in this regard is that our archive will make its services for self-archiving data even more visible and easier to use.

We think - maybe unsurprisingly - that GESIS is a most attractive place to make your research data available. The contents of our archive are well documented, easily found and mostly directly accessible through the Web.

Data-sharing also needs incentives. That is why journals and their policies are so important. GESIS offers *replikationsserver.de* as a joint initiative with the *Zeitschrift für Soziologie* and the *Soziale Welt*, two leading journals of social research in Germany. We plan to convince further journals to use this service and thus increase the reach of open data and open methodology.

We ascribe to the FAIR principles, i.e. research data must be Findable, Accessible, Interoperable, and Re-usable. [6] We currently explore operationalizations of these principles and will develop indicators reflecting the FAIRness of the data archived at GESIS.

In order to support Open Data within GESIS, we will discuss a voluntary self-commitment of employees and doctoral candidates to share their research data openly.

4.3 Open Source

Strategically, GESIS focuses on Open Source software that it develops and provides itself. Yet, it is clear that there is great potential as well in Open Source applications developed by third parties (like LimeSurvey, R or Libre Office). Thus, as a service to the community GESIS will increasingly offer consultancy and training in order to expand the use of social scientifically deployable open source software in everyday work and research.

GESIS uses open source software in its projects or services wherever feasible. For example, the backend technology for SSOAR is the open source software DSpace. Also, the source code for the GESIS Data Catalog (DBK) is open, as is the GESIS Research Information System (GRIS). To encourage code produced at GESIS to be shared, code-publications will be made more visible (they can be recorded in our Research Information System). GESIS will also draw up publication guidelines on the basis of best practice.

4.4 Open Methodology

To GESIS, Open Methodology is the Open Science pillar that currently deserves most attention. Open Methodology refers to the transparent and comprehensive documentation of data collection and data analysis. Its roots, if you will, are to some extent in Open Notebook Science in the natural sciences. The aim – in the laboratory context at first – is to document all relevant decisions and observations made in the process of research practice – including unsuccessful ones – and thus to maximize transparency.

That all – or all important – steps of a research process are retraceable is essential: Open Data can only be understood if it is verifiable how they were collected. Published results require that each step in the analysis can be reproduced. Conversely, methodological transparency is hardly conceivable without Open Data.

Given the importance of transparent documentation GESIS wants to assist in unifying the various ideas for documenting the research process. We envision "Open Methodology Guidelines", developed by leading experts in the respective fields. They should each reflect a consensus and be practical enough for every-day use. Our Survey Guidelines, we think, are an ideal blueprint for this project for which we hope we will receive support from our communities.

Very importantly, GESIS will use those Open Methodology Guidelines as a basis for trainings for researchers and provide them for teaching purposes.

When we drafted our Open Science strategy one of the first objections was that the effort invested in Open Methodology would likely decrease our publication output. This objection of course echoes the

concern cited earlier on the efforts for documenting data and it highlights that we need to gather more experiences with Open Methodology.

The experience drawn from documentations required by journals surely suggests that documentation entails substantial effort. But effort is likely to be limited if documentation accompanies the research process and becomes an integral part of it.

Many of us in the Social Sciences have been trained, e.g. in introductory statistics courses, that do-files and the like should be well documented to allow others to understand the analyses. Also, there are proposals for how to ensure reproducibility; some of them predating the Open Science discussion. Perhaps most notably is Gary King's engagement for replication. On a very practical stance Scott Long has set standards for Open Methodology with his "Workflow of Data Analysis using Stata" in 2009. [7]

To help establish how much time documentation takes up, GESIS will conduct two to three pilot projects during the next year. We can probably agree that ultimately we need guidelines that strike a balance between reproducibility and effort. Only if we – GESIS – but also we as a discipline – find a convincing balance here will we achieve more openness.

The question about the trade-off between publications and documentation highlights a possible conflict: As other research institutions GESIS sets internal targets for publications by its staff. If, however, we insist on publication of all program code and/or statistical syntax alongside journal publications output is likely to dip. Starting this fall GESIS will record Open Methodology documentation as a research activity in its own right. In addition our research information system will allow us to record open source contributions. Contributions to Open Science will therefore be more visible. To what extent GESIS might count such activities as performance criteria we have not yet decided. Establishing a data basis, however, will help make more informed decisions.

5 Conclusion

Where does this leave us? Taken together, we hope that the message our measures on the four pillars send is one of commitment to the idea of Open Science.

To be sure many questions still need answering. To touch upon one example: should there be limits to effort invested on openness relative to e.g. data usage? In principle, probably not. But should we, for example, invest the same rigor and resources that we invest in a highly used survey in cleaning up, documenting and making available contact data for a survey? Those data are rarely directly relevant to substantive results and are not used by more than a handful of colleagues (who will, of course, be given these data and support in understanding them). If the answer is yes, this would come at substantial cost.

Despite such open questions, however, we are confident that these ideas on Open Science will help foster the data-, code- and text-sharing culture at GESIS and beyond - probably the most important motive of all.

To conclude on a more general note: each time we open Google Scholar we are told to be "standing on the shoulders of giants". Those giants, arguably, are but a mosaic of dwarfs, molded together by science as its methodological glue. I find it hard to argue that without openness there will be no progress or even no greatness. Clearly history has proven this perspective wrong. But the glue that holds the giant together should be renewed. This is where Open Science can help. And this is why Open Science deserves support.

Sources cited

[1] European Commission (undated): Open Access to scientific information. https://ec.europa.eu/digital-single-market/open-access-scientific-information

[2] Mark Crawford (2011): Biologists Using Social-networking Sites to Boost Collaboration. BioScience, 61 (9), p. 736. https://doi.org/10.1525/bio.2011.61.9.18.

[3] Wolfgang Zenk-Möltgen, Alexia Katsanidou, Esra Akdeniz, Verena Naßhoven, Ebru Balaban (forthcoming): Factors Influencing the Data Sharing Behavior of Researchers in Sociology and Political Science.

[4] Fecher, Benedikt, Sascha Friesike, Marcel Hebing, Stephanie Linek (2017): A Reputation Economy: How Individual Reward Considerations Trump Systemic Arguments for Open Access to Data. Communications, 3, 2017. Available at SSRN: http://dx.doi.org/10.1057/palcomms.2017.51.

[5] Tenopir, Carol, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, Mike Frame (2011) Data sharing by scientists. PLoS One. Practices and Perceptions. http://doi.org/10.1371/journal.pone.0021101.

[6] Wilkinson, Mark D. et al., 2016: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3: 160018. http://doi.org/10.1038/sdata.2016.18

[7] King, Gary (1995): Replication, Replication. PS: Political Science and Politics, 28, Pp. 444-452. http://j.mp/2oSOXJL. Long, Scott J. (2009): The Workflow of Data Analysis Using Stata. Stata Press.