Huschka, Denis

Working Paper

Why should we share our data, how can it be organized, and what are the challenges ahead?

RatSWD Working Paper, No. 216

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Huschka, Denis (2013) : Why should we share our data, how can it be organized, and what are the challenges ahead?, RatSWD Working Paper, No. 216, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin

This Version is available at:
https://hdl.handle.net/10419/75339

RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

216

Why should we share our data, how can it be organized, and what are the challenges ahead?

Denis Huschka

April 2013

Federal Ministry of Education and Research

Working Paper Series of the German Data Forum (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Contact: S C I V E R O Press | German Data Forum (RatSWD) | Mohrenstr. 58 | 10117 Berlin | office@ratswd.de

Why should we share our data, how can it be organized, and what are the challenges ahead?

Denis Huschka
Managing Director of the German Data Forum (RatSWD)

Abstract:

The paper is based on a keynote talk held at the international conference "Opening data services in the social sciences", Belgrade, Serbia, 20/21 March 2013. The author wishes to thank the SERSCIDA project[1].

Empirical social sciences strongly contribute towards a better understanding of societies, especially of those societies that undergo rapid social changes. Empirical analyses are fed into the steering processes that are shaping a Europe of Nations. But data are also essential for the support of social and economic developments in national contexts. I was asked to reflect on three questions in my talk, namely:

Why should we share our data?

How can data sharing be organized?

And what are the challenges ahead?

[1] https://www.serscida.eu/en

1 Why should we share data?

A basic fact is: more and more potentially interesting data on almost every aspect of life are becoming available. This is why everything is digital nowadays and storage and handling of vast amounts of digital data are not a problem anymore - at least not regarding computer performance and storage capacity.

To illustrate how much digital data are out there, I always resort to the following image: in 2010 alone, the capacity of memory space sold worldwide was an estimated 5.000+ petabytes. This is a rather impressive figure! 5000 petabytes of memory – this equals about fourty stacks of floppy disks, which we all used in the early nineties, piling up from the earth to the moon.

Seeing as this capacity was actually sold, something must be on all those memory sticks and hard drives. Certainly not everything is relevant to social scientists, but even if a fraction of those data are relevant, this is much more research data then we had only a few years ago.

So the point is: potentially, we have more data available than ever before. This is the good news. The bad news is: The more data is out there, the harder it is to find a relevant data set! Most of the potentially interesting data sits on computers of solitary researchers or rots in archives and eventually just disappears – and we cannot make use of it. This has to change! The goal is to make finding data as easy as finding a book on Amazon or any information on Google.

Starting with the first question: Why should we share data?

a) There is a case to be made for sharing data as a precondition for scientific work: Science always implies the possibility of replication of analyses. As simple as it sounds: it is a basic requirement for scientific work. Only results that can be replicated are truly scientific results. If there is no chance to replicate research results, they can be regarded as no more then personal views in the opinion or review section of a daily newspaper.

   Replication, as a basic requirement, is also part of the codes of conduct of most science funders and other scientific organizations. Yet the reality is different in some fields and scientific disciplines. (Although, personally, I think this is about to change).

For others to be able to replicate empirical analyses means making the data available to colleagues or to whoever is interested in scientific results which are, after all, a public good.

Some American publishers in the field of economics even require their authors to hand over the data used for their papers. It is questionable whether publishers are a good place to store these data – however – the point is to ensure that replication is possible.

b) There are also economic arguments to be made**:** making the most of limited resources. This is especially relevant since scientific data production often relies on tax money – here it is even required by law to use tax money as effective as possible.

This can be boiled down to the following: we should share our data – especially since the data can be used for way more and different analyses then the original data producer or primary researcher might have intended.

Also, by sharing data generally and generously, we could save money by avoiding to collect data twice on the very same or very similar topics. This money could then be spent on new and innovative data collections.

Another way to use data more efficiently is to combine two or more datasets thus creating a new data set with clever merging techniques. Often additional data collections can be avoided if people share their data and eventually combine data in innovative ways.

c) A further scientific argument is: Those who share data are cited more and therefore gain higher reputation in their fields of expertise. Citations are the currency in science. Being cited as a data producer is an easy way of collecting citations. There is also a higher chance for not only your data work to be cited and acknowledged by the research community but also your scholarly work based on that data. Share your data – and your work will become much more visible!

The same argument also applies to exclusive or administrative data producers such as statistical offices: the reputation within the scientific community, but also among the general public, increases if data are used more by as many researchers as possible for as many analyses as possible. The data producer’s work becomes more. Another side effect is that the statistical offices get free analyses by top-level researchers

which then can also be used to advance their own analyses and to improve their own data collections.

d) The fourth argument aims at the fact that sharing data has a positive impact not only on science but also on society as a whole. Sharing data boosts competition among researchers thereby improving quality and quantity of scientific results. More scientific analyses also produce more knowledge. This knowledge are very important for policy planning and policy making. If politicians were able to rely more on better analyses, chances for wrong decisions could be reduced. Better decision-making contributes to a better and more prosperous society.

2 How can data sharing be organized?

I believe that everybody would agree with the statement that it is a very bad idea to just put empirical data containing information about individuals and companies on a website somewhere. Data protection is a very important and very serious issue!

First and foremost, data protection means protecting the respondents - namely the people or the companies who kindly participated in a survey or data collection. I think that the confidentiality of such information is a very valuable good indeed.

So how can we assure the confidentiality of data while, at the same time, making it available and sharing it?

A solution would be to aggregate the sensitive individual data in a way that makes re-identification of individuals or companies impossible. This is how open access or the open.data or data.gov initiatives often work. But such aggregated data would be of limited value for cutting edge scientific research.

Researchers require access to the raw data and data on the micro or individual level. And researchers in the social sciences are usually not interested in results for single individuals or companies. That’s why analysing individual level data and generating aggregated, generalized results from individual data does not pose a real security problem. But we need a structured and organized way to let these things happen in accordance with laws and ethical rules.

So, how to do it? We need a research data infrastructure that guarantees both: data protection and free access for researchers and maybe also for an interested public. Such an infrastructure must offer solutions for researchers to access

sensitive information in a way that is easy and perfectly safe at the same time. These infrastructures actually already exist. They are called data archives, research data centres or data service centres.

Whatever they are called: their common characteristic is that they offer access to data in a safe way. However, there are differences between the existing infrastructures in terms of service for secondary researchers, in terms of technical support, and also in terms of quality standards and compliance standards.

So how do we organize high quality access to research data exactly? Should we just go ahead and establish more data archives? Certainly this is not the worst idea. But as always the world is not that simple, because classic data archives cannot do everything.

The needs of the stakeholders, which have to be taken into an account and dealt with in the process, are manifold. That is why we need a flexible infrastructure to meet different needs and wishes: of both data providers and data users, and all this while keeping data security and research ethics in mind, too!

In Germany, we have an established infrastructure which comes in many forms: data archives, research data centres and data service centres. Each of these solutions has its advantages and disadvantages which have to be discussed in detail. However, the main question regarding the advancement of a research data infrastructure is always: How centralised do we want it to be? The scale ranges from having a central national archive, on the one hand, to a loose network of data centres on the other. In Germany, we rely on a decentralised infrastructure comprised of archives and specialized data centres under the auspices of an umbrella organization which is the German Data Forum.

So why don’t we have just one archive in Germany? Indeed, traditional central archives do have some advantages. They offer data to interested researchers and the public in a single place. Centralization has positive effects in terms of standardization of procedures, of methods and dataset features like meta-data. They can also facilitate financial efficiency; at least this is what politicians and the civil servants responsible for paying for the data archives like to hear.

However, central archives come with some serious disadvantages, the most problematic of which is the inability of big central archives to cater for specialized needs regarding data handling, tailored user support and data analyses by researchers.

Modern datasets that enable researchers to carry out cutting-edge research are much more sophisticated and much more complex then they were years ago. Modern datasets have longitudinal features and can also include bio-markers, geo-codes, etc. They can be merged with other datasets, for instance, combined employer-employee datasets. Such complex structures call for a very high level of expertise among data archivists for supporting data users. But, nine times out of ten, data archivists do not have this knowledge and forcing them to acquire such wide-ranging and in-depth knowledge of many datasets would not be efficient at all.

I believe that these special requirements call for a more specialized solution: infrastructures which, in Germany, are called research data centres. There are currently 25 of these specialized data centres in Germany and the number is growing. They range from very small RDCs, basically offering access to one very complex dataset, to RDCs with several thousand datasets such as the RDC of the Federal Statistical Office.

The main advantage is: In research data centres, the data is being provided by the people who produce it. All the questions that researchers have when re-analysing data can best be answered by those who actually produced the data. One could say: Service is being provided *by users for users*. Specialists for cataloguing and storing data do not necessarily have the expertise to help others to handle and analyse data.

RDCs also create a direct link between the producers and the secondary users because data producers deal directly with their academic peers. This brings about tangible advantages in terms of a direct and very effective knowledge exchange and also the data producers can benefit from the experiences of secondary researchers using their data.

So a decentralized but specialized infrastructure has some very important advantages. However, a central authority that safeguards quality of procedures, data quality, adherence to data protection regulations and other important policies is crucial. In the case of the German research infrastructure for the social, behavioral and economic sciences, this is ensured by the German Data Forum (hereafter: GDF).

Together with all research data centres the GDF formulates the “terms and conditions” for data sharing, and also safeguards compliance with these rules. This is accomplished by an “accreditation process” which data centres can apply for.

On the other hand, a GDF accreditation has something to offer for those who stick to the rules: the GDF serves as a platform to bring all relevant

stakeholders together to discuss common strategies. They thereby gain a louder voice and are thus able to bring forward the needs and problems of data centres and archives to policy makers and research funders.

The German Data Forum also serves as a central clearing house and is the body that makes sure that researchers, policy makers and data producers are in a constant dialogue over arising issues.

The German Data Forum is set up as a governmental advisory committee. It consists of 16 members, eight of which represent important data producing facilities in Germany. The other half consists of researchers who are elected by their peers. The result is that scientists from a range of disciplines and data producers form a common voice. The GDF has included many fields of research among which are sociology, economics, psychology and demography.

All in all, I believe that this set up has been very successful in Germany.

What have we reached so far?

- We improved access to data, in some cases our work even made the data accessible for the first time.
- We contributed towards much better empirical grounds for policy making.
- We made our data production more efficient, also in an economic sense.
- We contributed towards high-quality standards of German research data and also of transparency and safety of dissemination procedures.
- We also contributed to better statistical education at universities and towards improved and more widespread knowledge in the field of survey methodology.
- Last but not least we made research more innovative by enabling ground-breaking projects enabling the social sciences to produce better answers to questions from the public and policy makers.

3 What are the challenges ahead?

There is much to praise about what has been achieved in Germany and other countries so far – not to forget supranational bodies like CESSDA and projects like Data without Boundaries - but there are still many challenges ahead.

- First of all: We should make more high-quality data available for research! Besides all the positive developments in so many countries and the improvements for so many data sets – there is still a long way to go! And making data available also implies making it much easier to find that data. The ultimate goal should be: finding relevant data for any kind of research must be as easy as looking something up on Google!
- We have to secure and further develop free, fast, transparent and easy ways of access to data!
- We should provide access to data from home / work computers via secure and reliable remote access.
- We should work on Metadata! And Metadata schemes must be made much simpler to handle then the ones currently used.
- We should develop further and implement persistent identifiers for data sets and researchers!
- We should link datasets, publications based on this data, researchers who wrote the paper and the producers of the data. This “social network of science” has to be made visible.
- We should think about long-term preservation of data! Putting data on CDs and flash sticks is NOT long term preservation!
- We should think more about privacy protection and research ethics. This is becoming more and more relevant since new dataset merging techniques also produce new means of potential re-identification of individuals.
- And last but possibly most important: We must synchronize our national research data infrastructures. There are no national boundaries for research – so why are there national boundaries for research data?