RatSWD Working Paper Series

www.ratswd.de

RatSWD.
German Data Forum

229

International Access to Administrative Data for Germany and Europe

Stefan Bender, Anja Burghardt, and David Schiller

January 2014

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

International Access to Administrative Data for Germany and Europe[1]

Stefan Bender, Anja Burghardt, and David Schiller (all IAB)

Introduction

In the last years access to research data has made a lot of progress in EU countries. Nevertheless transnational access to confidential microdata – although there are some developments like "Data without Boundaries" – is still complicated and needs improvement.

The first part of the paper describes the (international) access to highly sensitive German administrative labour market data and how this international access is expanded within the research data centre in research data centre (RDC in RDC) approach.

In the second part a broader view of international access to official microdata in the EU will be given. Starting with a brief overview of the EU-funded project "Data without Boundaries" (DwB) a possible roadmap for international access in Europe and beyond is presented.

1 International access to German microdata

The German "research data centre movement" is quite a recent development (see KVI 2001 or Bender et al. 2011) with little more than 10 years of experience. Other countries, often with less stringent or different data protection legislation, have a much longer tradition in operating Research Data Centres (RDCs). Nevertheless, Germany is a very interesting example, because it progressed from nearly no access to a systematic access to microdata in less than 15 years. The German Data Forum (RatSWD) has and will play a decisive role in this development (www.ratswd.de/en).

This independent body of empirical researchers and representatives of important data producers has succeeded in opening and improving access to existing data, as well as creating an increased synergy between researchers and data producers. For example, until the end of 2013 the RatSWD has accreted 27 RDCs.

1.1 Access through the FDZ of the BA at the IAB

German administrative labour market data is of mutual value for research in the fields of economy, sociology, and related disciplines. In order to provide access to such data the

[1] This work was made possible by the European Commission Seventh Framework Programme (grant agreement No. 262608) funded project "Data without Boundaries" (DwB).

Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) was established in 2004 (fdz.iab.de/en.aspx).

The FDZ facilitates access to survey and administrative labour market data for non-commercial empirical research. Data are coming from the social security notifications and the internal processes of the BA as well as from surveys carried out by the IAB. Combining administrative data with other data (like Big/Smart Data) will enrich possible research. Therefore the German Record Linkage Center (GRLC) was established in 2011 as a service centre using record linkage techniques to link different data sources and produce innovative research data (www.record-linkage.de). Funded by the the German Research Foundation (DFG) until 2014 the German Record Linkage Center has a service facility located at the FDZ and is performing research on technical solutions mainly at the University of Duisburg-Essen. Provided services are for example individual advice during planning and realization stages of data linkage projects, conducting data linkages as commissioned work and updating and maintaining the record linkage software Merge Tool Box (Schnell et al. 2004).

There are different ways to make microdata accessible via the FDZ. Those access ways correspond with different security measures and are therefore enforced to protect data sources with different levels of disclosure risk. For non-disclosing data sources like Campus/Public Use Files a free download of the data is possible. Scientific Use Files are more detailed and already used for many research projects; those data files are available under restricted download for a specific time-restricted non-commercial research project. The highly detailed data sources that are needed for a sophisticated research project with a high impact level are only accessible under strongly restricted circumstances. One option is that the researcher has to come to a location of the data holder (for example: Nuremberg) to work with the data in a specifically secured room (on-site access or guest stay). The other option is to use job submission solutions in order to access from a distant location. When using job submission the user sends his or her enquiries to the data holder. The calculations are carried out on the servers of the data holder and results are sent back after output control was applied.

1.2 The Research Data Centre in Research Data Centre Approach

The highly detailed data of the FDZ has to be stored in the facilities of the BA in Nuremberg due to security reasons. Therefore a researcher wanting to work with these data sources has to travel to the location of the data holder. The same issues occur not only for the FDZ but for every confidential data sources all over Europe.

The central idea of the Research Data Centre in Research Data Centre (RDC-in-RDC) approach is to enable data access from other RDCs or institutions (guest-RDC) which share comparable standards as the RDC where the data are actually stored (data-RDC), but which are located at different sites. In doing so it does not matter whether the guest-RDCs are located in Nuremberg, Germany or abroad. All locations have the same standard in accessing the data. The only difference is that the guest researcher's room is not at the local data-RDC (for instance in Nuremberg) but at another guest-RDC. The guest-RDCs can be institutions which fulfil the security requirements of the FDZ (Bender and Heining 2011). The guest-RDC

cares for physical access control to a secure room, so that only researchers with a valid contract are able to access the facility. In addition the guest-RDC staff monitors the researcher in the secure room and makes sure that nothing prohibited happens. The secure room is a separated room only made for accessing confidential data. It is only equipped with devices to access the distant data sources; no access to any additional sources of information is possible. From the secure room within the guest-RDC a secured and encrypted connection through the internet to the secured servers behind the firewalls of the data-RDC is opened. Within this secure remote desktop approach, researchers access the secured servers of the data holder from the guest-RDC and can work with the confidential research data (browse, modify, run calculations, get output). The microdata always stay on the secured servers and only a live stream of the used graphical user interface is transferred to the device (screen) of the user.

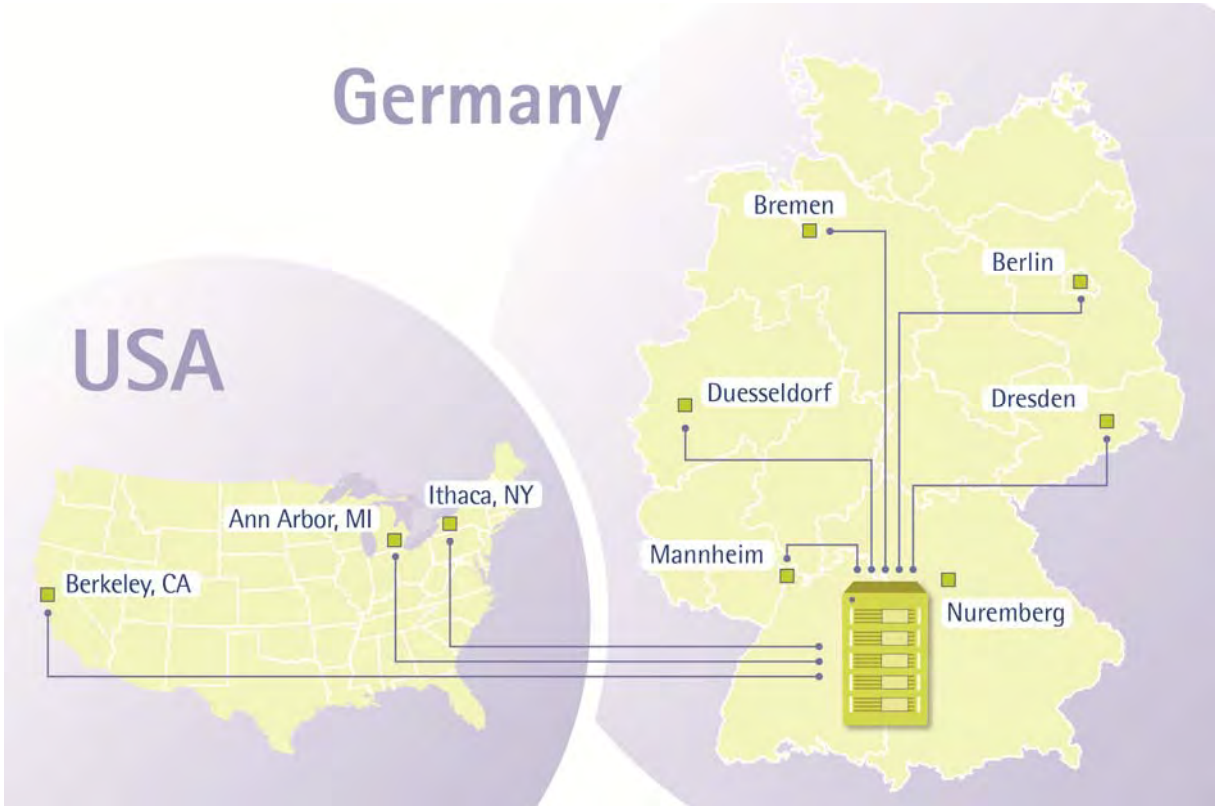

Figure 1: The RDC-in-RDC network of the FDZ (end of 2013)

At the end of 2013 (see figure 1), the RDC-in-RDC-network of the FDZ had eight guest-RDCs and this number is continually increasing (for example with guest-RDCs at the UKDA at the University of Essex and the Harvard University).

In the future this approach could be used as a template and starting point for a network of guest-RDCs (Safe Centre). Such Safe Centres, based on agreed on security standards, could be located in every bigger region. Accordingly, researchers would not have to travel far. A secure network could connect those Safe Centres to different data-RDCs that are using the Safe Centres as access point without setting up access points by themselves (Brandt and Schiller 2013).

2 International access to decentralized European microdata: "Data without Boundaries" (DwB) project

The European Commission Seventh Framework Programme (grant agreement no. 262608) funded project "Data without Boundaries" (DwB) works on improvements for European research in the Social Sciences. The focus is on discussing, describing and promoting concepts, solutions and frameworks. 29 partners from the European Statistical System (10 National Statistical Institutes or statistical departments), the Council of European Social Science Data Archives (11 Data Archives) and the Research Community (7 universities and 1 private company involved in methodological research) are working together.

Currently, comparative research projects based on microdata from different countries need to go through a process of multiple accreditations and deal with quite different technical and methodological environments. So, there is an existing wealth of official microdata, currently under-used and held behind national, legislative, technical, and cultural borders. These can be crossed over with cooperation and political will. The main goal of DwB is to establish an equal and easy access to official microdata for the European Research Area, within a structured framework where responsibilities and liabilities are equally shared. DwBs work will result in proposing concepts and improvements for a European research accreditation and a European distributed remote access to confidential microdata for national datasets. DwB also takes part in the discussion about metadata standards (SDMX/DDI) with the aim of building a single point of information on research data in Europe.

Under the umbrella of DwB it is possible to address the European need for a comprehensive and easy-to-access research data infrastructure which enables continuous cutting-edge research and reliable policy evaluations. Thereby DwB enhances transnational access to microdata. DwB is in contact with and tries to accommodate the needs of existing infrastructures (Council of European Social Science Data Archives (CESSDA) and the European Statistical System (ESS)).

3 A roadmap for a developed international access to decentralized European microdata

DwB produces evaluated concepts and pilots deriving from discussions and project work, but because of its conceptual character, there will be no real implementation and therefore no live and running improvements towards a comprehensive and easy access for researchers using the European research data infrastructure. According to that the next logical step is the implementation of the findings of DwB within an European framework.

Such a framework was proposed in work package 4 (improving access to microdata) of DwB. This framework is designed not only to connect researchers to different RDCs all over Europe, but also to connect researchers and project teams all over Europe. The conceptualization of this European infrastructure, called the European Remote Access Network (Eu-RAN), also incorporates the models and proposals of other work packages in DwB like a European Service Centre as information platform about research data in Europe.

3.1 Future: European Remote Access Network (Eu-RAN)

The Eu-RAN will bring together researchers and research groups from all over Europe with data sources from all over Europe (Schiller 2013). The Eu-RAN is segmented into the following main components: The access points to use the network; the Single Point of Access (SPA) with additional services adhered and the secured remote access network itself (see figure 2).

Regarding the access points the Eu-RAN will support three different security levels with different access restriction measures (see figure 2): The first access point, which is granted after a log-in process, is access from „Anywhere“ and allows the use of communication tools, to browse metadata and to work with freely available data. The second access point are universities and research institutes. The final and most restricted access point is a „Safe Centre“. Thus, the Eu-RAN infrastructure is able to support different needs depending on the security level of the available microdata.

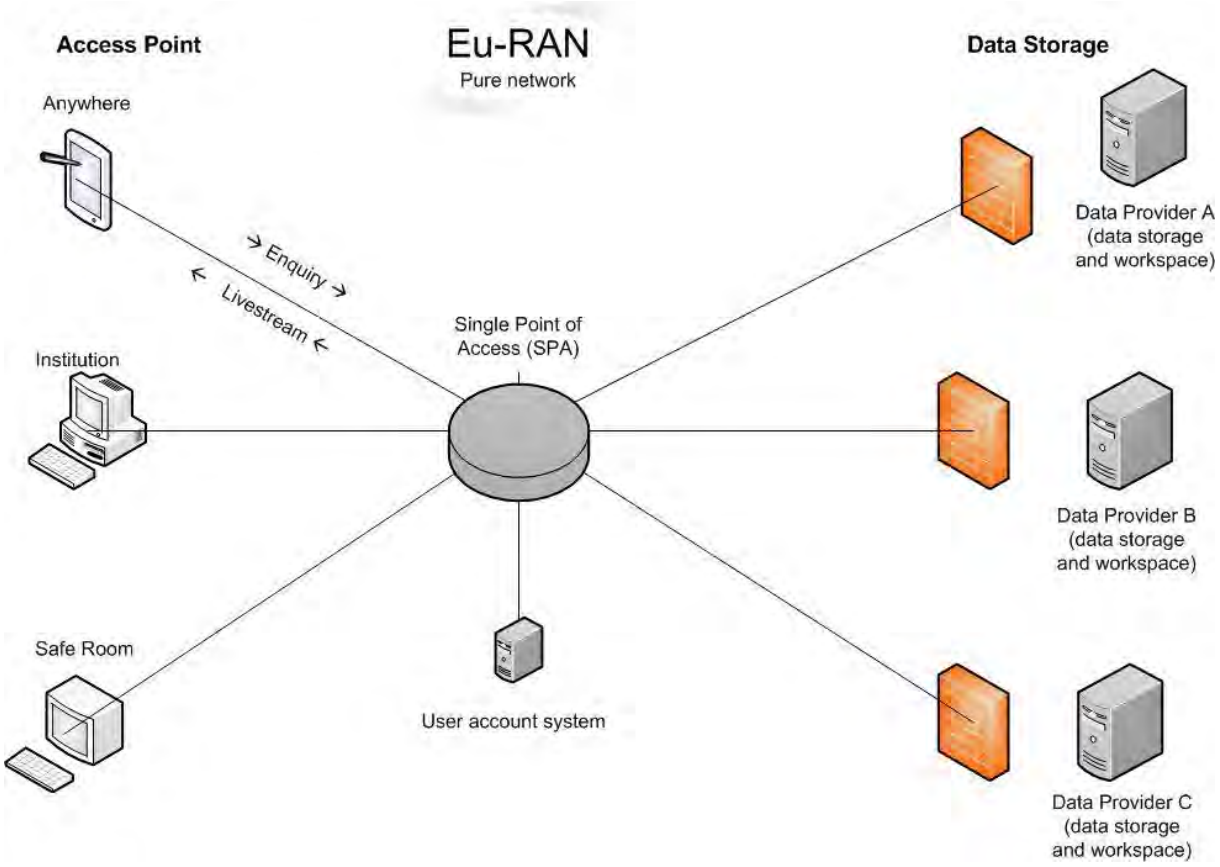

Figure 2: The EU-RAN architecture

The main task of the SPA is the user authentication check. When wanting to access the Eu-RAN from one of the access points, the enquiries of the user are routed to the SPA where the authentication of the user (including location of access point) will be checked. After being logged into the system, the user has access to the data and services his or her contract is valid for.

Access points, SPA and data storage servers are connected by a secured remote access network. By using encrypted tunnels through the internet all information are secure at any point of time. The data itself (confidential research data, outputs and information about the

users) are not moved, they stay behind the firewalls on secured servers. Only enquiries from the access devices and pictures for the graphical user interface at the access device are transmitted through the encrypted tunnels. When using the RDC-in-RDC approach, a closed and secure working environment for research on confidential data is in place.

Beside the technical network, an organisational network of partners (data-owner, data-holder, data-user) is required and the Eu-RAN infrastructure needs to be based on contracts, agreements and above all trust between the partners.

3.2 Future: Single Point of Access (SPA) and incorporated service hub

There are already a number of remote desktop solutions running in Europe (Report on the state of the art of current SC in Europe 2012; DwB deliverable 4.1). Those solutions enable access to the data sources of one specific data owner (National Statistical Institute, Data Archive etc.). Establishing the Eu-RAN in the first place only joins those existing solutions in a Single Point of Access (SPA) and enables the users to access different data sources from one access point (see figure 2). The power of the central node, the SPA, is only fully used when services to support the researchers are added. Such services are controlled by a service hub working as a central device that host numbers of different services. Only a few of the potential services will be mentioned in order to give an impression of the service hub (see also figure 3).

One of the services will be the web portal that functions as the graphical user interface to use the whole Eu-RAN infrastructure; an additional service is the user account management system that deals with the authentication of users and where users can manage their contracts and data holders can manage the access rights of the user. In addition tools, which support research, like editors, statistical packages, wikis, calendars, (data) documentations or interfaces, are provided services within the SPA.

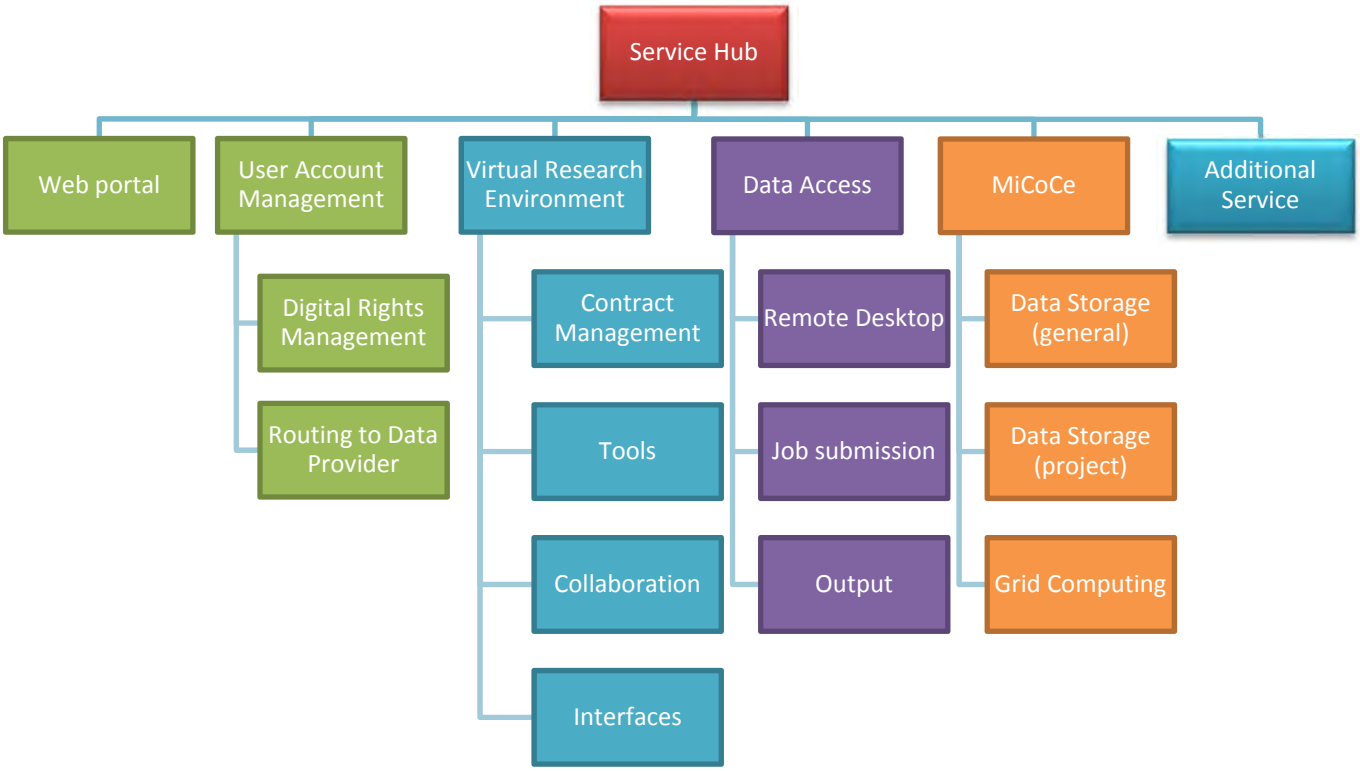

Figure 3: Services of the service hub

When working with data a good documentation of the data is needed in the first place. Therefore a European Service Centre for official statistics, as proposed by DwB (Report on concept for and components of European Service Centre for official statistics 2012; deliverable 5.1), should be hosted by the service hub. When talking about transnational contexts, two services (respectively groups of services) are of crucial importance in order to enable collaborative work of research groups with researchers from different locations and comparative analysis with data from different data sources. Those two crucial services are the Virtual Research Environment and the Microdata Computation Centre. Of course all those services will not be reachable without a running Eu-RAN in the background.

3.2.1 Virtual Research Environments (VRE)

In general Virtual Research Environments (VRE) are web portals that provide given services to users. Those services are connected to underlying databases. The whole construct can be secured and access can be restricted, if need be. In a VRE researchers use -for example- editors, calendars, wikis, forums or statistical software for their daily work. VREs are shaped to carry out scientific research in a community and they are able to be used as a platform for exchange between different disciplines or countries (Allan 2009; Carusi and Reimer 2010). By supporting the work of users with tools, a possibility for standardization and archiving (reproducibility of research) is opened up in the back end of the VRE infrastructure.

From the perspective of accessing confidential research data, VREs can also incorporate interfaces to sensitive microdata via job submission or remote desktop solutions.

3.2.2. Microdata Computation Centre for de-centralized data sources (MiCoCe)

In Europe, most of the confidential microdata has to be stored in the country, where the data was collected. Even if there are no explicit definitions in the law, security measures force data holders to keep the microdata in the country of his origin (Tubaro et al. 2013). The RDC-in-RDC approach and related solutions based on secure remote access allow analysing data even from locations abroad. Additionally there is a huge demand on a European level to analyse data from different countries simultaneously. The challenge is therefore to enable analysis with distributed data sources and fulfil the requirement of not moving the data at the same time. The solution could be to set up a Microdata Computation Centre for de-centralized data sources (MiCoCe). In the MiCoCe enquiries to the distributed data sources will be send from a central point and the single results per data set will be combined to a final result. According to that, no disclosing microdata will be moved in the MiCoCe, only non-disclosing (part-) results are moved to a secured central node. Solutions can come from the areas of statistics, grid/parallel computing and federated databases.

4 Summary and Outlook

Data access in Germany is now well established because there was and is a strong will from data providers, researchers and ministries (sponsoring and legal support) to give access to highly sensitive microdata. By establishing the German Data Forum as an institution for supporting and developing the data infrastructure and the access to these data, Germany can be seen as a blueprint for other countries.

German administrative data, as provided by the FDZ of the BA at IAB, is a powerful resource of knowledge discovery that can already be enriched by linking it to survey or – in the near future – to Big Data. With the RDC-in-RDC approach the FDZ of the BA at the IAB has shown that it is possible to give transnational access to researchers outside Germany, although there are restrictions by law barriers.

Transnational access to microdata - although there are solutions like the RDC-in-RDC approach or secure remote access with less restrictions regarding the access points - is still to be regarded in its infancy. Here, DwB plays an important role.

DwB is the starting point for transnational access to microdata, but not its end point. The concept of a European Remote Access Network that builds the basis for transnational research in Europe, by offering different security levels to serve the needs of data holders all over Europe and by providing tools like the VRE and the MiCoCe that support transnational collaborative research teams and enable the use of distributed data source, will unleash the power of European data for cutting edge research in Europe.

Furthermore there should be a continuous dialog, which will lead to a roadmap for an international access to sensitive microdata. Infrastructures like the Eu-RAN, VREs and the Microdata Computation Centre will support future research. Having an infrastructure in place that can also deal with all kinds of data (from survey data to administrative data to Big Data) enables the next level of research in the social sciences.

The next logical step is to bring the proposed concepts and models into life. The established culture of communication between archives, NSIs and the research community is the basis for the development and guarantees a promising real implementation of the proposed infrastructures based on the state of play in Europe and aiming on the needs of European research.

Establishing a trustworthy working organisational network, which is supported by technical solutions, will build the needed environment for high-quality research on sensitive microdata in Europe and worldwide.

References

Allan, R. (2009): Virtual Research Environments: From Portals to Science Gateways. Oxford: Chandos Publishing.

Bender, S./Himmelreicher, R./Zühlke, S./Zwick, M. (2011): Access to Microdata from Official Statistics. In: German Data Forum (Ed.): Building on Progress - Expanding the Research Infrastructure for the Social, Economic and Behavioural Science, Budrich UniPress, Opladen & Farmington Hills, MI, 215-230.

Bender, S., and Heining, J. (2011): The Research-Data-Centre in Research-Data-Centre Approach: A First Step Towards Decentralised International Data Sharing. IASSIST Quarterly 35 (3), 10-16.

Brandt, M., and Schiller, D. (2013): Safe Centre Network - Need for Safe Centre to enrich European research. In Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality Ottawa.

Carusi, A., and Reimer, T. (2010): Virtual Research Environment – Collaborative Landscape Study.

KVI- Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001): Nomos Verlagsgesellschaft, Baden-Baden.

Report on the state of the art of current SC in Europe (2012, September): www.dwbproject.org/deliverables (deliverable 4.1).

Report on concept for and components of European Service Centre for official statistics (2012, April): www.dwbproject.org/deliverables (deliverable 5.1).

Schiller, D. (2013): Proposal for a European Remote Access Network (EU-RAN) - main components. In Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa.

Schnell, R./Bachteler, T./Bender, S. (2004): A Toolbox for record linkage, Austrian Journal of Statistics 33 (1–2) 125–133.

Tubaro, P./Cros, M./Silberman, R. (2013): Access to Official Data and Researcher Accreditation in Europe: Existing Barriers and a Way Forward. IASSIST Quarterly 36 (1), 22-27.