Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (iii): Modes of access to microdata

Proposal for a European Remote Access Network (Eu-RAN) - main components

Prepared by David Schiller, Institute for Employment Research, Germany (david.schiller@iab.de)

Proposal for a European Remote Access Network (Eu-RAN) - main components

David Schiller*

* Institute for Employment Research, Nuremberg, Germany, david.schiller@iab.de

Abstract. The European Commission Framework Program 7 (FP7) funded project Data without Boundaries (DwB) works on improvements for access to confidential microdata from the European Member States (MB). One of the twelve work packages deals with the technical aspect of improving access to European microdata. A concept for a European Remote Access Network (Eu-RAN) was proposed and a limited pilot, which should work as a proof of concepts for some of the functionalities of Eu-RAN, is currently carried out. Eu-RAN is build around a Single Point of Access that allows connecting researchers from different locations with data sources located in different European countries via secure Remote Access solutions. Additional tools like a sophisticated user account management, a Virtual Research Environment (VRE) and a Microdata Computation Centre (MiCoCe) complete the services offered by the Eu-RAN. While the goal of this work package is to build and proof a technical concept, technical aspects cannot be addressed without taking care of legal, organizational and financial issues. This paper outlines the architecture of Eu-RAN, highlights important legal, organizational and financial topics and describes some of the components in more detail.

1 Introduction

The European Commission Framework Program 7 (FP7) founded Data without Boundaries (DwB) project runs for four years; from May 2011 to April 2015. 29 Partners from eleven European countries work on improvements for transnational research in Europe. Within the project twelve work packages (WP) deal with topics like resource discovery, data documentation and accreditation.[1]

WP 4, termed *Improving Access to OS microdata*, focuses on technical developments to improve access to confidential microdata. The focus is thereby on secure Remote Access (RA), respectively on Remote Desktop solutions, defined as followed: *any kind of terminal or desktop solution that uses a secured connection to the servers of the respective data provider, whereby the user can see and work live with the real, highly-detailed and confidential microdata.*

[1] For more information visit www.dwbproject.org

This paper highlights the main findings of DwB deliverabel 4.2 (*Feasibility study on the organizational architecture for managing pan European access*, 2013).[2] This feasibility study was build on a survey of European Remote Access centres.[3] The feasibility study resulted in a proposal for a European Remote Access Network (Eu-RAN).[4] The paper gives a basic description of the proposed architecture and should function as a basis for further discussions.

2 Architecture of the European Remote Access Network

Within this section the basic architecture of Eu-RAN will be described. To enable collaborative transnational research in Europe a solution with secure and efficient interfaces that brings the best of the existing isolated applications together is needed. Therefore the central innovation of the Eu-RAN is a Single Point of Access (SPA) that connects researchers located all over Europe with research data stored all over Europe and additionally offers a broad portfolio of services to ease transnational access. The description of the Eu-RAN infrastructure starts with the connection between users and data before the services attached to the SPA are introduced.

2.1 Single Point of Access (SPA)

The Single Point of Access (SPA) is a conceptual picture rather than a technical implementation. It is the central point of the network and functions as home for the provided services. While Eu-RAN stands for the complete product, SPA stands for the central point to access information about research data in Europe and the research data itself.

2.1.1 Central Service HUB

The service HUB (figure 1) is the control centre behind the SPA. It nodes together the different service and makes them access able. As shown in figure 1 (Additional Service), the number of services is not limited and can be extended, if new services are needed to serve the scientific community. Before explaining the currently

[2]For more information see www.dwbproject.org/about/deliverables.html.

[3]The findings of this survey can be found in DwB deliverable 4.1; www.dwbproject.org/about/deliverables.html. The survey was carried out among CESSDA archives (www.cessda.org) and partners of the European Statistical System (ESS) (www.cros-portal.eu).

[4]The deliverable was prepared by Iris Dieterich and David Schiller (IAB), Maurice Brandt and Christopher Gürke (Destatis), Eric Debonnel, Philippe Donnay and Kamel Gadouche (CASD), Leo Engberts and Jan Mol (CBS), Frederique Cornuau, Marie Cros and Roxane Silberman (CNRS), Atle Alvheim (NSD), Steve Bond and Felix Ritchie (ONS), Richard Welpton (SDS). The authors of the deliverable gratefully acknowledge feedback from all DwB partners and external experts. Primarily versions of this report have been discussed during the WP4 workshop of NSIs and CESSDA members on SC in Europe in Wiesbaden, the EDAF in Luxembourg and at a workshop connected to the conference on NTTS in Brussels.

suggested services, the embedding of the central service HUB into the Eu-RAN infrastructure will be explained (see figure 2).

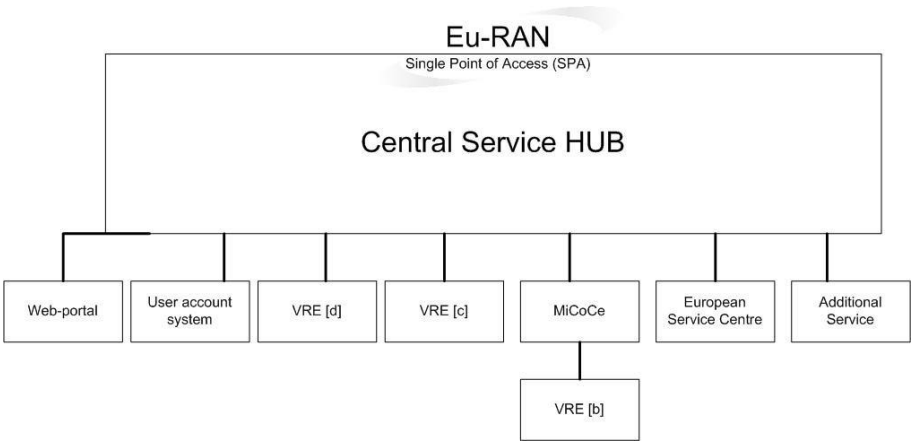

Figure 1: Eu-RAN/SPA Central Service-Hub

2.1.2 Secure Remote Access

All connections between the users and the research data are conducted through secure Remote Access (Schiller & Welpton, in press). Modern encryption techniques and secure Virtual Private Network (VPN) tunnels are used to transfer information between ths users access device, the SPA and the data storage servers. Unlike to common RA solutions in the social sciences (*Report on the state of the art of current SC in Europe*, 2012) where only one secure connection between the user and the data repository is needed, the Eu-RAN has to establishing two secure connection: one between the user and the Single Point of Access (SPA) and one between the SPA and the secured data storage serves at the facilities of the data owners. Thereby the interface between the secured connections provided by the Eu-RAN and the secured data storage facilities (servers) of the different RDCs involved into the Eu-RAN infrastructure is a crucial point for the network. The interface must be easy applicable by the data providers; otherwise it will be inefficient to become a part of the Eu-RAN.

2.1.3 Access Points

From a technical point of view working with data remotely is only limited by the possibility to use a device that provides access to a network, usually the internet. Due to data protection regulations some of the current remote access solutions are limited to the following specific access points (*Report on the state of the art of current SC in Europe*, 2012):

1. Highly-sensitive data sources are usually only available during on-site access in the facility of the data owner. The researcher works with data in a special

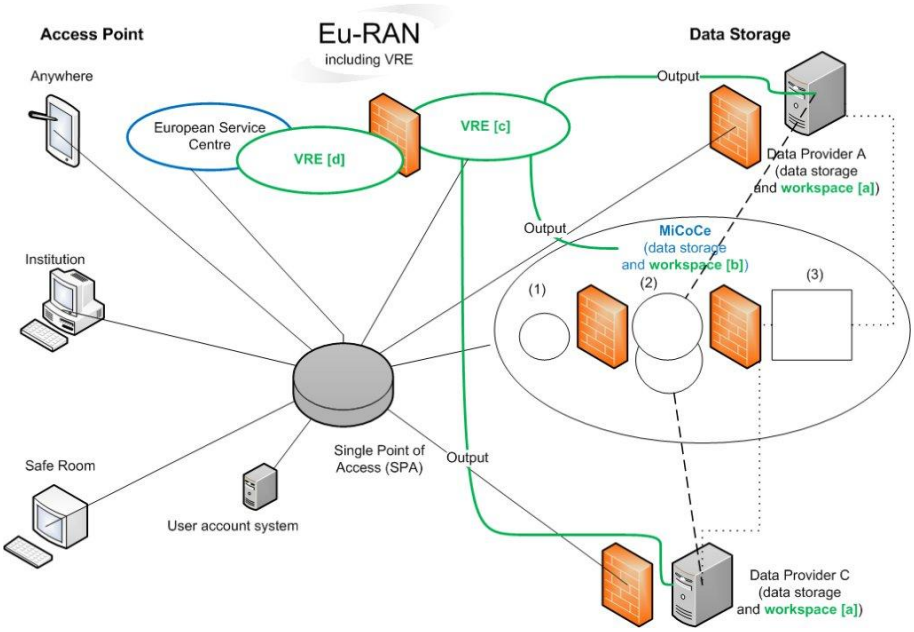

Figure 2: Eu-RAN including with SPA-services

secured room with a physical access control system (known as a Safe Centre). Within this room, only devices needed to work with research data are available. No access to the internet or to any other information, except the research data itself, is possible. Travelling to the home location of the data provider to use such data source, results in high travel expenses for the researchers. By using a remote access solution, this burden can be scaled-down (Bender & Heining, 2011). Researchers can work in "Safe Centres" closer to their home location (they may even be located within the researchers' home institution). These "Safe Centres" provide the same security level as the room in the facility of the data provider; a secure connection enables the researcher to work with the distant data source. To unburden data access in a significant way, this approach needs a number of "Safe Centre" with an agreed-upon security standard (Brandt & Schiller, 2013).

2. Allowing access from a given room within an institution (e.g. a university) is a less restricted approach that also limits access to a predefined location with some agreed-upon security conditions (for instance lock-able room). The researcher contractually agrees on accessing data from this room only. In parallel, a given IP-range (or specific IP addresses for terminals) is needed to be able to reach the distant data source.

3. Finally, there are also remote access solutions that are not restricted to a given

access point and researcher can use data via the internet from anywhere. In these cases only the identity of the user is checked and no validation check of the access point is carried out.

The Eu-RAN has to be flexible enough to cover all these solutions and to enable access from each of those three different access points.

2.1.4 Data Storage and work space

The Eu-RAN should knit different access points together with different storage facilities in Europe. Those data storage facilities are located at existing RDCs (NSIs, Archives and other data providers). The network needs to access the respective secured data storage servers. In addition to the data storage, the remote servers at the existing RDCs will provide analytic processing for the data including some basic software applications that will be needed by researchers. Basic provision must be made to allow researchers on the same project to access the same working directory through the Eu-RAN. This may be known as a "project area" and allows researchers on the same project to access the data approved for the project.

2.2 Web-portal

The following services (2.2 to 2.6) are supported by the Central Service HUB shown in figure 1. The web-portal offers the needed graphical user interface to communicate with the Eu-RAN services provided via the SPA. When logging into the web-portal the users are provided with a clear interface to submit enquires and view the regarding outputs. The first service the users are communicating with is the user account system; in the next step the users find themselves in a work space or VRE. Once again the web-portal offers the graphical interface while the underlying services provide information and tools.

2.3 User account system

When about to use the Eu-RAN infrastructure, users must be directed to an account management system that verifies their authentication and allows them to use the provided tools, functionalities and data. An integrated user account system manages the access rights of each user. By logging into the portal the users authentication is checked. This will be done by a two-factor authentication consisting of something the user knows (e.g. a password) and something the user has (e.g. a token, a fingerprint or a specific keystroke). For data sources that are only allowed to be accessed from a predefined location (institution, safe room) an additional check of the access point is needed (done through a given IP-range, hardware-certificate etc.). The basis for this twofold validation check is: first, the contract between the data owner (usually not the operator of the Eu-RAN service) and the user; and second the access point permitted to access the given data source. While the contract results in a user account to use the Eu-RAN infrastructure and to access the specific data

source ; the combination of access point and data source is predefined (between the data owner and the Eu-RAN provider) and results of country- and data specific restrictions. After successful checks the users will be automatically routed from the web portal (SPA) to the relevant data source via a secure connection. For the user the forwarding is performed in the background and not visible. He or she only sees the log-on screen and in the next step the accessible data within the work space of the regarding data provider.

It would be desirable for the Eu-RAN, to have a facility for managing online applications. This would provide convenience to researchers submitting applications to access the data, and for data producers who grant permission to access the data. The regarding work flow could look like that: After users have gathered the needed information within the European Service Centre (2.4) they can use data provider specific online application forms. All needed information will be filled-in the form and transmitted to the regarding data provider. The data provider carries out the usual accreditation procedure. In the end user account information are sent to the user account system of Eu-RAN. From that point on the user can access his research data via the Eu-RAN Infrastructure.

With regard to the current work of DwB WP 3 "Enhancing legal, information security and researcher accreditation frameworks for access to data", a permanent user account for recognized European researchers (maybe with reference to the recognized "research entities" mentioned in the new Commission regulation (EU) 557/2013) can be imagined. This account would permit use of designated services and contains a kind of researcher biography. In addition the account could be used as starting point for applications to use research data via the Eu-RAN. Resulting in a more comprehensive user account during the runtime of the given project.

2.4 European Service Centre for Official Statistics (ESS-OS)

Working with data is not the beginning of a research project. Before accessing data, researchers need a variety of preliminary information. First, they need to know which data sources are available, and from where. Then they need detailed information about the datasets and the variables (preferably standardized for all datasets, collected amongst all countries). On the basis of these sources of information, the researchers can decide whether their project is doable with the existing data. If so, information about the conditions of access to data are needed, including information about the accreditation process, contracts, legal restrictions, statistical disclosure control of results and other research outputs, and technical access modalities. The researchers are only likely to apply for accessing the Eu-RAN, if the procedures be considered to be manageable and suitable in terms of both time and finding.

An opportunity for building and sustaining a information platform that contains such information is the creation of a European Service Centre for Official Statistics (ESS-OS). The concept and contents of such a centre are described in DwB deliv-

erable 5.1 (*Report on concept for and components of European Service Centre for official statistics*, 2012). Providing this information portal as one of the services of the SPA (driven by the Central Service HUB, see 2.1.1) reduces the invest in establishing and maintaining the service and opens up possibilities for interactions between the different services.

2.5 Virtual Research Environment

The two main goals of the Eu-RAN Virtual Research Environment (VRE) are: first, providing a suitable and consistent working environment for each researcher (single researcher and research teams) included tools are analytical, text editing software and metadata information systems and second, providing tools for working collaboratively (Allan, 2009; Carusi & Reimer, 2010; Schiller et al., in press).

While using Eu-RAN VREs are needed in different locations (see also the corresponding VREs in figure 2):

(a) Regularly the data storage facilities of existing data providers are used and the research environment is offered by them. This will not be consistent across all providers, which means that a researcher carrying out comparative research will have to learn different environments.

(b) When use of MiCoCe is made, a single and consistent VRE will be provided and customized to needs of users.

(c) The Single Point of Access as part of Eu-RAN offers a secured VRE for researchers working together in a given project. The main focus of this VRE is in providing tools to work collaboratively. Examples are joint work on program code and publications, messaging services to interact with the project members and with the data provider, and metadata information systems. Within the SPA, the collaborative tools may only work with data/output that has been cleared by the data providers.

(d) Another VRE might be located at the European Service Centre and only available via a secured interface to the VRE [c]. This VRE would be accessible with only minor entrance restrictions, and would include tools supporting communication within the scientific community.

Due to the fact that different VREs host data/output with different confidentiality levels the VREs have to be strictly separated. In addition functionalities should be implemented that allow checked information to move from one VRE to another. The needed checks are comparable with the checks needed for disclosure control of outputs. For all types of VREs, it is important to keep the needs of the researchers in mind. They are already part of a number of social networking systems and they will only participate in collaborative VRE functionalities if they receive some benefit

from doing so. This applies especially for the VRE for the scientific community [d]. As mentioned, also relevant for all levels of VREs is security. The more restricted information are available the higher the security levels. At the same time the VREs should be user friendly and easy to use. Exchange between the different levels (e.g. program code from the personal user VRE level [a; b] to the VRE level for joint projects [c] and, in an next step, to the VRE level for the scientific community [d]) is a crucial functionality that generates a number of technical challenges to ensure security. At the same time unrestricted information (e.g. metadata) should be available for the user within the restricted VREs as well.

2.6 Microdata Computation Centre

The Microdata Computation Centre (MiCoCe) offers further options for transnational comparative research. It provides storage space, statistical software and above all computational power in order to bring together data from different countries for comparative analyses in a secure IT environment. Thereby the MiCoCe offers three levels of services (see figure 2):

1. **MiCoCe as an archive:** It functions as a storage and computation centre for new RDCs that do not want to invest or do not have the resources to build the whole infrastructure by themselves. If allowed in their country, these RDCs can use the capabilities of the MiCoCe instead of developing a complete infrastructure by themselves.

2. **MiCoCe as a data host for projects:** The MiCoCe services can also only be used for a restricted time period, for example the runtime of a research project. Based on contracts one or more data providers may store data in the MiCoCe secure environment for a given research project; the researchers conduct their analysis and the data is automatically deleted after the project is completed. Programme codes might be archived at the MiCoCe facilities, while the original data source is stored in the facilities of the involved data providers. Putting both together enables replication studies.

3. **MiCoCe for analysis "on the fly":** Finally the MiCoCe could be the basis for computations that do not need the dataset to be transferred physically (Hibbert et al., 2007; Kamm, Bogdanov, Laur, & Vilo, 2013; Muilu, Peltonen, & Litton, 2007; Wolfson et al., 2010). One approach uses **federated databases**. A database located within a central facility (MiCoCe) forwards enquiries to the relevant distant data storage facilities where they communicate with adjusted databases. Results are put together in the central facility where they can be shown without disclosure control, due to the secured environment provided by the Eu-RAN. Some of these solutions are even able to deal with different data storage formats. One additional scenario derives from the the

area of **Grid-computing**. While this approach is normally used due to high needs on computation power, the Eu-RAN MiCoCe would make use of it for disclosure control and data protection reasons. Two approaches are possible: first, data stays in the facilities of the data provider, enquiries are send from MiCoCe to these facilities, results are send back and statistically put together with results coming from other data provider facilities; second, data stay in the facilities of the data provider, small non-disclosive parts of the data are send to MiCoCe working memory, immediately after the calculation they are deleted and the next small non-disclosive parts are send.

3 Topics to addressed when proposing the Eu-RAN

Legal topics: The proposed Eu-RAN should offer technical solutions to serve different data providers with different microdata located within different legal systems. To comply with legal regulations and data protection rules all over Europe security is the main focus for the Eu-RAN. The proposed system must become a trusted partner for data providers in Europe. DwB WP3 deals with "Enhancing legal, information security and researcher accreditation frameworks for Access to Data". The regarding findings will find their way into the architecture of the Eu-RAN.

Organisational topics: A number of organisational topics have to be discussed when building the Eu-RAN. Within this paper these topics should only be touched. Appropriate solutions for those topics have to be found in a lively exchange between the partners of WP4 and the relevant stakeholders in Europe. Topics are: physical location of the SPA and the MiCoCe; funding for establishing and the live running Eu-RAN; needed staff; work flow, responsibilities and agreements with Eu-RAN partners.

Technical topics: The sophisticated step from a written proposal to a implemented service should not be underestimated. Extensive discussions with IT experts and small pilots as proof of concept are needed to build an functioning and useful infrastructure for European Research.

4 Summary and Outlook

The Eu-RAN as proposed improvement for transnational research in Europe is already discussed with stakeholders from all over Europe. Building on existing and proofed solutions next steps based on technical developments and the needs of researchers and data owners are suggested. Three partners of WP4 are already involved in a pilot to discuss technical, legal and organisational issues. In addition more feedback about the proposed concept is needed and will find his way into the further development of the Eu-RAN infrastructure. Beside this discussions limited

technical pilots as proof of concept are needed. Because the concept for the EuRAN ar only the fist step on the way to a live running solutions that supports both researchers and data owners.

References

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

Bender, S., & Heining, J. (2011, Fall). 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., & Schiller, D. (2013). Safe Centre Network - Need for Safe Centre to enrich European research. In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality.*

Carusi, A., & Reimer, T. (2010). *Virtual Research Environment - Collaborative Landscape Study.*

Feasibility study on the organizational architecture for managing pan European access. (2013, July). www.dwbproject.org/deliverables.

Hibbert, M., Gibbs, P., O'Brien, T., Colman, P., Merriel, R., Rafael, N., & Georgeff, M. (2007). The molecular medicine informatics model (mmim). In *Medinfo 2007: Proceedings of the 12th world congress on health (medical) informatics; building sustainable health systems* (p. 1230-1234).

Kamm, L., Bogdanov, D., Laur, S., & Vilo, J. (2013). A new way to protect privacy in large-scale genome-wide association studies. *Bioinformatics*, *29*, 886-893.

Muilu, J., Peltonen, L., & Litton, J.-E. (2007). The federated database a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Europe. *European Journal of Human Genetics*, *15*, 718723.

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

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

Schiller, D., Alvheim, A., Mulcahy, T., Pohl, C., Priddy, M., Silberman, R., & Welpton, R. (in press). The need for Virtual Research Environment for Research in the Social Sciences. *IASSIST Quarterly*.

Schiller, D., & Welpton, R. (in press). Providing Remote Access to European Microdata. *IASSIST Quarterly*.

Wolfson, M., Wallace, S. E., Maca, N., Rowe, G., Sheehan, N. A., Ferretti, V., ... Burton, P. R. (2010). DataSHIELD: resolving a conflict in conemporary biosience - performing a pooled analysis of individual-level data without sharing the data. *International Journal of Epidemiology*, *39*, 1372-1382.