Virtual Research Environments (VREs) to enable access to confidential data for scientific purposes

David Schiller*

* Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB), Nuremberg, Germany, david.schiller@iab.de

Abstract. The FP7 funded project "Data without Boundaries" (DwB) showed the need for transnational research and gave first hints for needed developments. Main characteristics for future research projects in the Social Sciences (and beyond) are: need for detailed data on individuals (persons, households, institutions), need to link different data sources (sometimes without having them physically in one place), interdisciplinary approaches, new methods of data analysis, and irrelevance of national boundaries. Data used may range from open data to highly confidential data; whereby confidentiality arises from different reasons. In modern research projects data can be confidential because of disclosure risk or/and intellectual property. While new data sources and research methods ask for effective, flexible and understandable statistical data protection methods; the irrelevance of national boundaries asks for secure and flexible infrastructures to support research projects and data owners. Virtual Research Environments (VREs) are made for modern infrastructure networks; but at the same time they focus on open and non-confidential data. According to that adjustments to the VRE concept are needed that bring together researchers and data sources from different locations and offer security levels that enable work with open as well as with confidential data. This paper will highlight first steps and concepts towards a secureVRE for Europe.

1 Introduction

This paper focuses on enabling modern research in the Social Science and highlights some of the future challenges in this area. It states that those future challenges can only be overcome when appropriate infrastructures to access and to work with research data are in place. Virtual Research Environments (VRE) represent such needed infrastructures. Their concept will be described as a kind of development line. In addition some of the crucial areas where Statistical Disclosure Control (SDC) methods and future VRE infrastructures need to be linked-up are highlighted.

The FP7 funded project "Data without Boundaries" (DwB), runing from May 2011 to April 2015, evaluated the needs for Social Science research in Europe and proposed first concepts for future improvements. As the name of the project implies, the focus was on supporting research projects that don't want to care about borders of any kind but will concentrate on working with the best data available in order to proof hypothesis and explain phenomenons. Hence producing high-level knowledge for Europe. Some of the ideas in this text are based on the concepts and discussions emerging from the DwB project.

The following bullet points name only a small number of the main characteristics for future research projects in the Social Sciences that are relevant for the topic of this paper:

- The need for detailed data on individuals (persons, households, institutions) to apply modern methods of data analysis;
- the need to link different data sources, to find the best basis for valid findings;
- evaluation of interdisciplinary approaches, because there should be no borders between disciplines, if exchange between them enriches the research findings;
- new methods of data analysis, or at least new combinations of data analysis (computer sciences and survey methodologies) to handle new data sources and new data source combinations; and
- irrelevance of national boundaries, due to the fact that they often limit the possibility of producing the best possible results.

Do support research a variety of research data types, ranging from open data to highly confidential data is needed. Thereby combination of open and confidential data may typically result in open data becoming restricted and not the other way round. Sophisticated systems to secure confidential data and ensure the freedom of science are needed and have to be developed.

In addition it has to be noted that data may not only be confidential because of disclosure risks. During the research process researcher create confidential data, e.g. intellectual property like program code. Such "by-products" of research should, in the pure meaning of science, be freely available to reproduce research and to build on it; but on the other hand they are a value in it self for the researcher that created them and this needs to be reflected as added value in the accounts of the specific researchers (or research projects).

While new data sources and research methods ask for effective, flexible and understandable statistical data protection methods; the irrelevance of national boundaries and the conflicting priorities of protecting data and enabling research ask for secure

and flexible infrastructures to support research projects and data owners within modern technical environments.

Virtual Research Environments (VREs) are made for modern infrastructure networks; but at the same time they are build to handle open and non-confidential data. Nevertheless if the concept behind VRE-infrastructures is well understood, it becomes obvious that they can be adjusted to the actual needs of handling confidential data as well. That includes the possibility to create secureVREs that bring together researchers and data sources, based on agreed on security levels that enable work with open as well as with confidential data. The talk will highlight first steps and concepts towards a secureVRE for Europe.

First the challenges of Social Science Research will be discussed, before short introduction into the Virtual Research Environment (VRE) concept is given. Afterwards the adjustments needed to build a secureVRE are introduced and topics regarding interactions between Statistical Disclosure Control (SDC) methods and secureVRE infrastructures are discussed. The text closes with final conclusions and an outlook into the future.

2 Challenges in Social Science Research

Research in the Social Science is facing a high number of challenges; no matter if it is about quantitative or qualitative research. For modern modes of analysis in quantitative research, detailed microdata on individuals are needed. Tables (aggregated information) or Public Use Files (completely anonymized data on individuals) are normally not enough. Even Scientific Use Files (factually anonymized[1] data on individuals; direct identifiers, like name or addresses, are removed and methods of statistical disclosure control have been applied) may not serve sophisticated methods of analysis. Secure Use Files[2] (confidential data for scientific purposes, only allowing for indirect identification; direct identifiers like name or addresses are removed,

[1] "Factual anonymity means that the data can be allocated to the respondent or party concerned only by employing an excessive amount of time expenses and manpower" (Knoche 1993).

[2] European Union Commission Regulation No 557/2013 (17 June 2013) is about accessing confidential data. It defines secure-use files as "confidential data for scientific purposes to which no further methods of statistical disclosure control have been applied;" whereby "confidential data for scientific purposes means data which only allow for indirect identification of the statistical units, taking the form of either secure-use files or scientific-use files;" scientific-use files finally "means confidential data for scientific purposes to which methods of statistical disclosure control have been applied to reduce to an appropriate level and in accordance with current best practice the risk of identification of the statistical unit;". Statistical disclosure control methods "means methods to reduce the risk of disclosing information on the statistical units, usually based on restricting the amount of, or modifying, the data released;". In general it is stated that "a risk management approach should be the most efficient model with a view to making a wider range of confidential data available for scientific purposes while preserving the confidentiality of respondents and statistical units".

no further methods of statistical disclosure control have been applied) are in many cases the source that enables sophisticated and up to date research. According to that confidentiality and privacy issues become more and more important. Securing data by a sophisticated portfolio (Lane et al. 2008) and risk management approach is therefore a crucial task for every data provider.

Qualitative data on the other hand has his own challenges. Data is not that easy to store because it comes in many different formats; content is extremely specific (audio, video, interview transcription, etc.) and disclosure control methods are hard to apply without destroying the scientific value of the data (and these are only some of the issues when giving access to qualitative data). A portfolio and risk management approach is also needed for this kind of data in order to make it access-able for researchers.

Data for modern Social Science research comes not only from surveys. Data sources of other origins, like administrative data, statistical data, "Big Data", are more and more often used in the Social Sciences. New methods of analysis need to be taught to students, especially regarding specific types of Big Data. In addition data sources from different disciplines (e.g. health data or geo-spatial data) are compared to create the best basis for research findings. Those merged data source need expert knowledge on data linkage techniques in order to produce meaningful data sources. One important topic thereby is the ownership and the country of origin of data. Data may not be allowed to be moved from one country to the other, or even from one institution to another. On the other hand also Open Data can be used for research and merged to confidential data sources. Also cross national research (researcher from abroad likes to work with data from a foreign country) or international research (researcher wants to carry out comparative research or use sample frames that cross borders) becomes more and more important. According to that, borders or restrictions of data ownership should not limit the possibilities of producing high level research findings. Once again not a easy task to be solved.

Finally the contrariness of intellectual property and reproducibility of research findings during the research process has to be managed. Program code produced to analyse data and new adjusted data sets resulting from analysis (sometimes just as a interim step in the research process) represent the intellectual property of the researcher and are therefore a valuable and confidential good in itself. On the other hand those confidential information are needed to enable other researchers to carry out reproducibility studies in order to proof the correctness of the results and/or build on the achieved findings. Keeping in mind that publishing new results in important journals is the main currency for researchers, giving away their intellectual property would simply be harming to their scientific career (at least as long as not everybody is doing so). According to that enabling intellectual property and reproducibility at the same time is still an unsolved challenge.

3 Virtual Research Environments

This section is about introducing the concept of Virtual Research Environments (VRE). It will show that a VRE is more than a simple workspace and that such infrastructures are a crucial component for modern and efficient research processes.

Research Environment

Research is not done by a single individual sitting in a office without any exchange with the surrounding world (in any case it should not be like that). Modern projects are carried out within a network of partners. Efficient Research Environments are needed to support the processes of research projects and ensure quality standards. This is done by providing a physical environment on the one hand and an intellectual environment on the other hand.

Physical environment:

It is made of the things that surround the researcher and the tools and services he needs to carry out his work. The physical environment could include for example:

- laboratories or other facilities, furniture;
- hardware, and other specific equipment needed;
- unique databases, samples, bio-banks;
- specialised literature, user manuals, guidelines;

Intellectual environment:

It consists of humans as resource for knowledge and support for the research project. The intellectual environment could include for example:

- Funding bodies providing the basis for the resources needed to carry out the research;
- Project Partners constituting the core research group working on the project;
- Students as supporting partners of the project;
- Participants as basis for the data used in the specific research, e.g. people asked a questionnaire and now going for an interview;
- Research Network to support the project in specific topics and thereby not part of the core research group (e.g. labour market specialists in the world);
- Mentors, and advisor to provide high-quality expertise and guidance;

- Independent Experts to support the project (e.g. with topics like research methods, underlying theories and assumptions, procedures of Statistical Disclosure Control (SDC), etc.);

- Peer review groups to give feedback to research results and ensure quality of findings and good scientific practice;

Summary

Together the physical and the intellectual environment build the research environment needed for any research project. If a surrounding like this is not provided, high quality outputs cannot be ensured. According to that some universities ask for a description of the research environment used even before the project is allowed to start; e.g. the Macquarie University, Sydney.[3]

Virtual Research Environment

A Virtual Research Environment (VRE) is a online system to support researchers. It brings together the environments needed, by using virtualization tools. They especially ease the work of research teams working in different locations (Carusi and Reimer 2010). From the perspective of the physical environment VREs offer tools and services needed. Quite common are forums and wikis but also document hosting, collaborative text editing and discipline specific tools (e.g. special tools for visualization or data analysis). One advantage of the VRE concept is that such tools do not need to be installed at every team members workspace. User are provided with a customized workspace tools may run on a centralized server and can be used by researchers sitting in different locations and using their personalized virtual workspace. From the perspective of the intellectual environment, VREs work as networks supporting communication and collaboration between team members and other actors within the project specific intellectual environment.

Said so, the VRE can be seen as a network accessed through a virtual workspace (available, e.g., via a browser on the users work station; located in a physical room and therefore part of the physical environment as well) that provides tools (and services) as well as communication with human beings. A general approach to a VRE architecture is shown in figure 1 (p. 7).

Thereby a VRE can function as a cloud solution with shared services; no mater if it is a public, private or hybrid cloud. This decision depends on the research carried out and the confidentiality of the used and produced data. Of course, a VRE can run within on single institution, actually a lot of universities provide VREs for their members, but it can also operate as a multinational network with a high number of involved partner institutions (Candela et al. 2010).

[3]Some of the content of the lists is taken from their web page.

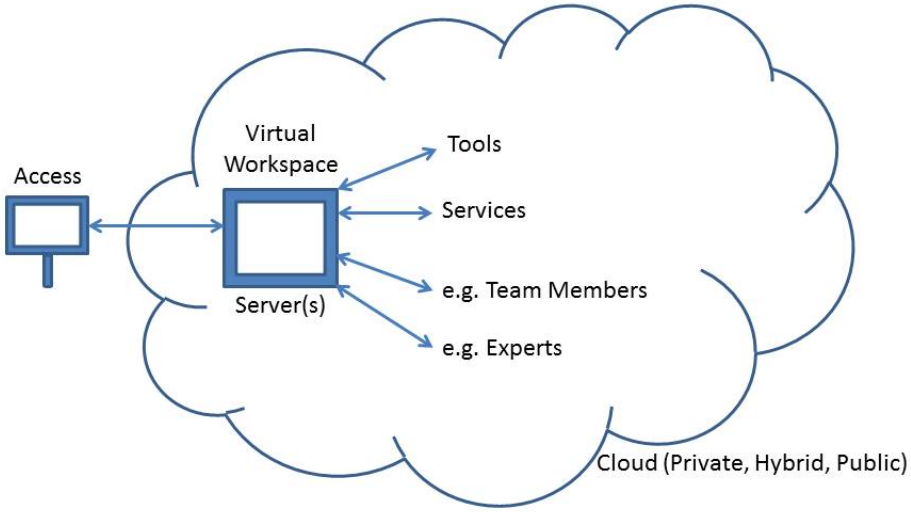

Figure 1: VRE, general architecture

Standardized and secure interfaces become important when developing such a network, especially when different organisations are involved. The need to develop such infrastructures as networks reflects the needs of modern research. The "Data without Boundaries" (DwB) project has published a number of texts on this issue (Bender et al. 2014; Schiller 2013; DwB deliverable 4.2;). Next steps need to be done.

4 A secureVRE to handle confidential and open data

The concept of a secureVRE is simply about adding security mechanisms to the VRE approach. Thereby the focus is on supporting the whole life cycle of a research project. The concept is shone in figure 2 (p. 14) and described within this section.

VRE are made for open data and free interaction between users. Working with confidential data and producing intellectual property results in the need for security measures. Those measures have to correspond with the security levels needed. According to that the secure Virtual Research Environment (secureVRE) is structured in five levels that are defined as followed:

Open data (BLUE AREA) Public data (or open data) & metadata (data docu-

mentation) at data holding institutions & elsewhere. Important for the project work, data in the BLUE AREA is available without restrictions. It can be found on the internet by manual and automated search. At the same time such data is not stored in the secureVRE in the first place. There is no need to store it in a secured environment (it es freely available anyway) but it may be more practical to store a customized copy within the secureVRE network.

sVRE project public space (GREEN AREA) The GREEN, YELLOW and ORANGE AREA are the project areas within the secured secureVRE network. The GREEN AREA represents an public access able space. Anyone can have a look at the content of this project related space (read access), while the project members are responsible to populate and delete content (read/write access).

sVRE project group space (YELLOW AREA) This is the virtual workspace for the project members. It is secured, so that only users belonging to the project are able to access it. Within this area the project members can interact and cooperate.

sVRE individual space (ORANGE AREA) Once again a virtual workspace; this time limited to one individual user (who is likely a member of a bigger project team; on the other hand he or she can also work alone on a project or work in more than one project team). This user may have special rights, e.g. to work with confidential data and takes therefore care about special tasks within the research process. In addition this workspace can be used to store user specific information and content.

Secure data in data holding institutions (RED AREA) Due to legal issues and organisational structures some confidential data sources have to stay physically in the facilities of the data owner or data provider. Therefore such data can only be accessed via special secured interfaces. The data owner or data provider is thereby not a part of the organisational secureVRE structure; it is a kind of third party interacting with the secureVRE (although, from a VRE architecture perspective the RED AREA is a part of the VRE because it is able to communicate with the VRE; but for data protection reasons this communication is restricted and based on special agreements).

The role and tasks of those areas within the architecture of the securevRE will become clearer when introducing the phases of an generalized project life cycle; but already at this point the importance of standardized and secure interfaces between those layers has to be highlighted and underlined. The interfaces of a secureVRE can be regarded as a kind of air lock between laboratories placed in special containers. Interfaces need to be: standardized so that containers can be connected and secured;

so that nothing can escape or pour in. Thereby two layers of standardization and security have to be considered when talking about the secureVRE interfaces. First IT security measures like encryption techniques or digital rights management; second statistical disclosure control (SDC) methods that ensure that only the allowed information move through the secured interfaces.

Security measures implemented and interfaces used need to support the complete project life cycle. Major steps of such a life cycle are:

Resource Discovery Before every research project, team members have to be aware of the available resources to proof there hypothesis and carry out their research. Some of the important information are: data documentation, availability of research data, costs for access, possibilities of merging data. Only if all of this information are available, team members can decide, if the research can be done with the available resources; e.g. number and knowledge of team members, funding, expected duration of research. Information will be collected in the sVRE project group space.

Develop Proposal In order to access data and receive funding project members have to prepare proposals, i.e. review literature, calculate budget, plan dissemination etc. This needs to be done by all the team members within the sVRE project group space.

Gather & Generate Resources, Request Access This is maybe the most important (or at least time intensive) phase of a research project. Resources have to be incorporated into the project workspace, they have to be adjusted (e.g. data editing) and sometimes generated (e.g. new merged data sets), first descriptive analysis are carried out, and all the time a close interaction between team members and maybe external experts is needed. If needed, access to sources has to be requested. During this phase data from the BLUE AREA is moved into th secureVRE workspace (YELLOW or ORANGE AREA), project generated sensitive data is created in an interaction between individual project members and the project group, access is requested from external data providers (RED AREA), and first analysis with those data sources are done.

Analyse & Experiment This part of the research process has to be understood as an iterative process. Models have to be adjusted, first findings have to be discussed, data sources as well as experimental settings may have to be modified, and steady interaction between team members and external data providers is needed.

Publish & Disseminate Now the most important part for the researchers has arrived. They are going to publish their findings and receive the credits that

are so important in the scientific world. Before this happens papers have to be written and reviewed. Once again interaction between team members and maybe external experts is needed. Papers could be made available via the secureVRE public project space or could be send into the BLUE AREA as Open Access publications. Of course, also interfaces to publishers and editors can be used in order to publish peer reviewed papers in high ranged journals. Beside the publication of papers also other project outputs could be moved into public spaces, e.g. generated non-sensitive data, program code, descriptions of statistical models.

Data Stewardship & longterm storage Publishing papers is of high value for individual researchers, project groups, and research institutions. From the point of view of the scientific community the last phase supported by the secureVRE architecture is even more important. This phase is about good scientific practice, reproducibility and building on achieved findings. Project output has to be moved into the BLUE AREA, if it is non-confidential, and into the RED AREA, if it is confidential. In this case BLUE AREA means freely available data repositories and RED AREA data archives with controlled access to data. Thereby project output is more than just papers. It also includes program code, new data generated, description of the research process not relevant for the final paper but of high value for researchers working on a similar topic (e.g. data quality issues). It has to be highlighted that such longterm storage is only useful, if the project output is documented in a meaningful and standardized way (e.g. by using DDI[4]).

Especially the interactions between the YELLOW and ORANGE AREA are important in order to support research projects working with confidential data and having team members located in different institutions. This was also one of the major findings of "researcher view" workshops that were carried out by the European "Data without Boundaries" (DwB) project.[5]

One of the reasons for that is the sophisticated access system when accessing confidential data. Especially when different data sources from different institutions and countries are analysed by researcher from different institutions and countries. It may for example be that one project member is allowed to work via, e.g., remote desktop solutions (Schiller and Welpton 2014) with confidential data stored at a external data provider (RED AREA). He can see the secure use files stored in the RED AREA within his individual workspace in the ORANGE AREA. The remaining project partners are only allowed to look at output files after disclosure control that are created by the individual researcher and afterwards moved into the YELLOW AREA, the project workspace. Within this environment interim results can

[4]See also ddialliance.org.

[5]See also dwbproject.org.

be discussed between all team members. In addition comparing confidential data with freely available open data makes the procedures even more complicated. Legal permissions, user contracts and data source specific rules have to be supported and consistency has to be ensured. This is not easy when a project has to be set up and, because the secureVRE architecture has to support this framework of permissions and restrictions, a highly sophisticated and trustworthy digital rights management has to be in place.

Beside the technical architecture of a secureVRE, there are also organisational aspects that need to be taken into account. A secureVRE needs staff members with different areas of knowledge; the architecture of a secureVRE has to be implemented into an existing legal framework, cooperation within the VRE structure including users and external partners like data providers, journals, legal advisory boards all have to be based on agreements and contracts, finally the security of the secureVRE has to be checked by external and regular audits.

The described architecture of a secureVRE will allow to support modern research by enabling the usage of open data as well as confidential data. It will ease work-flows when working with a team of experts located all around the world by caring for reproducibility and longterm storage of research findings, and by setting standards in project work-flows and data documentation.

5 Statistical Disclosure Control (SDC) methods and secureVRE infrastructures

Even the most sophisticated digital rights management system will not ensure security and usability when working with confidential research data. The high number of interfaces and data exchange will only work when modern Statistical Disclosure Control (SDC) techniques are implemented. Interfaces that need to be armed with SDC methods are, e.g.:

- the analyses of data stored in the RED AREA (at the facilities of data providing institutions) may it be via remote desktop or job submission techniques (Schiller and Welpton 2014).
- when moving generated data source form the ORANGE to the YELLOW AREA or to the GREEN AREA.
- when moving research outputs for longterm storage into the BLUE AREA.

Thereby SDC methods could be automated, work as guidelines for staff members of the secureVRE, or act as hybrid solution between automated and manual solutions. According to that research projects brining together experts in secure VRE

structures and SDC methods are needed. If both groups only work within their common area of expertise the support needed for modern research in the Social Science will not be provided.

6 Conclusion and Outlook

Future needs in Social Science research can be meet by modern infrastructures that need to be developed as networks. Thereby the findings relevant for modern societies can only be produced when the era of silos will be brought to an end. Not an easy task - because silos can be found in many places and manifestations, e.g.: as national borders, as not comparable data sources, as missing interdisciplinary work (including SDC and IT work as well as single-disciplinary research projects), as limits in publish-able research outputs, or as limitations in funding strategies. No matter what the nature of those silos are, they always limited the capabilities of modern research.

According to that more network activities including a general openness for other approaches and longterm funding strategies are needed. Thereby smaller and huger projects may appear as useful, as long as a general development strategy exists and coordinates in the background. A first project proposal for an secureVRE architecture including 19 productive and 10 reviewing partners from all over the world was submitted but did not receive funding. More initiatives are on the way.

References

Bender, S., Burghardt, A. and Schiller, D. (2014): International access to administrative data for Germany and Europe. *Facing the future. European research infrastructures for the humanities and social sciences*, 75-86.

Candela, L., Castelli, D. and Pagano, P. (2010). Making Virtual Research Environments in the Cloud a Reality: the gCube Approach, *European Research Consortium for Informatics and Mathematics, ERCIM NEWS*, 83, 32-33.

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

Knoche, P (1993). Factual Anonymity of Microdata from Household and Personrelated Surveys - The release of Microdata Files for Scientific Purposes, *Proceedings of the International Symposium on Statistical Confidentiality*, 407-413.

Lane, J., Heus, P. and Mulcahy, T. (2008). Data Access in a Cyber World: Making Use of Cyberinfrastructure. *Transactions on Data Privacy*, 1(1), 2-16.

Schiller, D. (2013). Proposal for a European Remote Access Network (Eu-RAN) - main components. *UNECE Working paper*, 10.

Schiller, D. and Welpton, R. (2014). Distributing access to data, not data - providing remote access to European microdata. *IASSIST quarterly*, 38/3, 6-14.

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

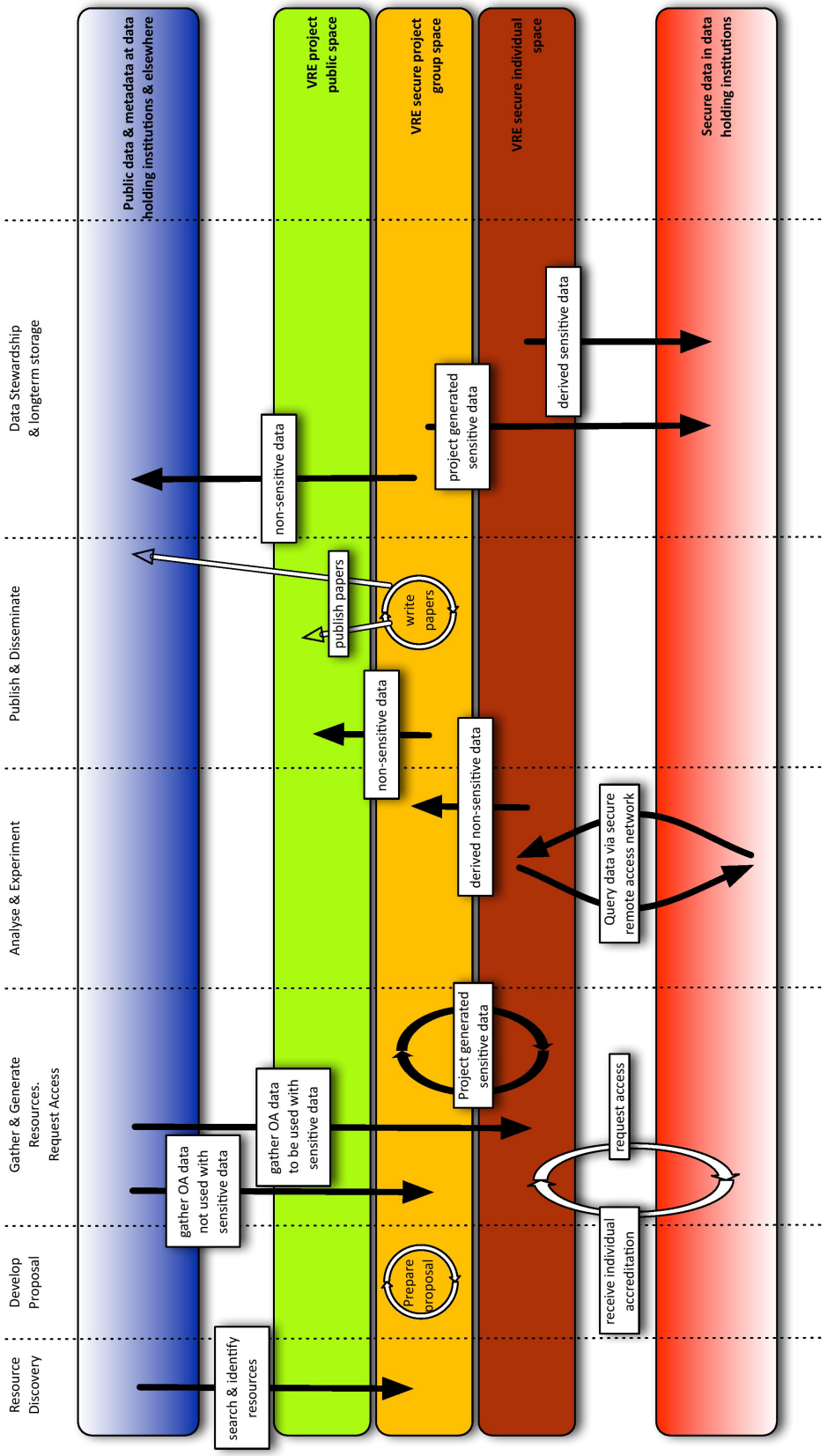

Figure 2: sVRE, linear life cycle; created by Mike Priddy, DANS, 2014