Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782   Start Date of Project: 01/01/2019   Duration: 40 months

Deliverable 5.10 White Paper on Remote Access to Sensitive Data in the Social Sciences and Humanities: 2021 and beyond

Dissemination Level	PU
Due Date of Deliverable	30/04/22
Actual Submission Date	29/04/22
Work Package	WP5 Innovations in Data Access
Task	Task 5.4 Remote Access to Sensitive Data
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.72

Abstract:

This White paper provides foundational context and recommendations for future infrastructure investment for remote access to sensitive data in the social sciences and the humanities. First essential definitions are established, followed by an overview of the European remote access landscape in the social sciences and humanities and a set of minimal requirements for the provisioning of secure access.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.0	10/03/2021	Creating a report based on a template	Yevhen Voronin
0.1	23/03/2022	Completed draft for external review	All
0.2	07/04/2022	Addressed reviewer's comments	All
1.0	11/04/2022	Final edits	Libby Bishop and Yevhen Voronin

Author List

Organisation	Name	Contact Information
GESIS	Libby Bishop	ElizabethLea.Bishop@gesis.org
CLARIN ERIC	Daan Broeder	d.g.broeder@uu.nl
RU	Henk van den Heuvel	Henk.vandenHeuvel@ru.nl
FORS	Brian Kleiner	Brian.Kleiner@fors.unil.ch
UKDS	Beate Lichtwardt	blicht@essex.ac.uk
GESIS	Deborah Wiltshire	Deborah.Wiltshire@gesis.org
GESIS	Yevhen Voronin	Yevhen.Voronin@gesis.org

Executive Summary

This white paper provides the necessary basis for understanding the requirements and specifications for remote access to sensitive data (data with potentially harmful effects in the event of their disclosure) in the social sciences and the humanities (SSH). It is result of the work implemented in SSHOC Task 5.4 Remote Access to Sensitive Data. It is intended to provide guidance and recommendations to the EOSC stakeholders for future infrastructure investment for remote access to sensitive data in the SSH. To ensure that this guidance can, in fact, be implemented, the recommendations are based on the knowledge of numerous data professionals who have direct experience planning, implementing, managing, and sustaining diverse forms of remote access and secure facilities. In doing so, our goal has been to maintain the vision of expanding such infrastructure, while remaining grounded in the practicalities of operating such facilities in a sustainable manner.

In this domain, it is now recognized that the ideal of "open data" needs to be balanced with privacy and other factors that can require moderating access to sensitive data, as reflected in the EU Commission's (2016) stance of "as open as possible, as closed as necessary." Developments in the past five years have advanced data access, primarily through "safe enclaves", i.e., physical rooms that provide security for data access (see Glossary). This represents a major improvement for data accessibility, but international, comparative, efficient research requires augmenting the research infrastructure by enabling remote access to data from a researcher's desktop. Solutions have operated for several years (e.g., UK Data Archive Secure Lab, ICPSR Virtual Data Enclave), but most of these still face limitations on the scope of data available, geographic limitations, etc. More recently, new infrastructures are being developed, some spanning several countries. These efforts are commendable and represent major improvements. However, limited resources, and complex legal variations (national implementations of GDPR), as well as other factors, have prevented implementation of a broader solution.

As countries across Europe look at the emerging multi-national infrastructures, it is crucial to address the need for a European answer, at scale, with sustainable funding. The recommendations offered here are guided by our observations that most successful infrastructures embody two features: 1) they are human as well as technical, and 2) they are neither purely centralised nor decentralised, but well-crafted hybrids.

Abbreviations and Acronyms

AAI	Authentication and Authorization Infrastructure
ACA	Academic Use
BA	Federal Employment Agency
CASD	Secure Access Data Center
CASRAI	Consortia Advancing Standards in Research Administration Information
CBD	Convention on Biological Diversity
CESSDA	Consortium of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
CSC	Center for Science Information Technology (Finland)
DANS	Data Station Social Sciences and Humanities
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DIW	German Institute for Economic Research
DWB, DwB	Data without Boundaries
DZHW	German Centre for Higher Education Research and Science Studies
ECRIN	European Clinical Research Infrastructure Network
EHRI	European Holocaust Research infrastructures
ELRA	European Language Resources Association
EOSC	European Open Science Cloud
ERAN	European Remote Access Network
ERIC	European Research Infrastructure Consortium
EUDAT	European Association of Databases for Education and Training
FDZ	Research Data Centre
FDZ-DZHW	The Research Data Centre for Higher Education Research and Science Studies
FORS	The Swiss Centre of Expertise in the Social Sciences
FSD	Finnish Social Science Data Archive
GDPR	General Data Protection Regulation
GESIS	Leibniz Institute for the Social Sciences

IAB	Institute for Employment Research
ICPSR	Inter-university Consortium for Political and Social Research
IDAN	International Data Access Network
IEC	International Electrotechnical Commission
IPR	Intellectual Property Right
ISO	International Organisation for Standardisation
KNAW	Royal Netherlands Academy of Arts and Sciences
LAT	Legal Assessment Tool
LDC	Linguistic Data Consortium
LIfBi	Leibniz Institute for Educational Trajectories
LIS	Luxembourg Income Study
MONA	Microdata Online Access
NeIC	Nordic e-Infrastructure Collaboration
NEPS	National Education Panel
NSD	Norwegian Centre for Research Data
NSI	National Statistics Institute
PDE	Physical Data Enclave
PUB	Public Use
RA	Remote Access
RAIRD	Remote Access Infrastructure for Register Data
RASD	Remote Access to Sensitive Data
RatSWD	German Data Forum
RDC	Research Data Center
RES	Restricted Use
RI	Research Infrastructure
RSA	Remote Secure Access
SDC	Secure Data Centre
SOEP	Socio-Economic Panel
SSH	Social Sciences & Humanities

SSHOC	Social Sciences and Humanities Open Cloud
TRE	Trusted Research Environments
TSD	Services for Sensitive Data (University of Oslo)
UKDA	UK Data Archive
UKDS	UK Data Service
VDE	Virtual Data Enclave
VM	Virtual Machine

Table of Contents

Annex 2. Access Levels and Dissemination Practises at CLARIN, UKDA and GESIS

Annex 3. Use Case from FORS

1. Introduction

1.1 Motivation

Open Science[1] and FAIR data[2] initiatives place considerable emphasis on direct access to research data. However, strong legal and ethical constraints exist where this concerns confidential or personal data which are invaluable for the social sciences and some subdomains of the humanities. When creating or advancing research infrastructures and recommendations for their use, which is an important part of the SSHOC project, it should be our task, while still fully respecting these constraints, to optimise access for the end user.

There is a well-established need for researchers to be able to use sensitive data that cannot be openly shared, i.e., sensitive data, in domains such as health, genetics, and environmental science. This need extends to social science and humanities research. The breadth of problems is great, spanning topics in electoral fraud, inequality, human rights abuses, migration, clinical language observations, and many more. Where data can be modified, e.g., by removing personal identifiers, options are available so that in many situations researchers can access the data by downloading a copy to a local machine (or server) for ready use. However, not all data can be handled this way. Often, the reduction in data quality from de-identification (or anonymization) is too severe, removing analytical value. Examples include research where extreme values are the focus of interest (inequality), non-numeric data formats such as audio and video, and geo-referenced data.

In recent years data access to sensitive data has expanded primarily through “safe enclaves”, i.e., physical rooms that provide security for data access (See Glossary). By making formerly unavailable data shareable, this represents a major improvement for data accessibility, but international, comparative, efficient research requires augmenting the research infrastructure by enabling remote access to data from a researcher’s desktop. Solutions have operated for a number of years (e.g., UK Data Archive Secure Lab, ICPSR Virtual Data Enclave), but most of these still face limitations on the scope of data available (e.g., only official statistics), geographic limitations (e.g., data available only to researchers in one country), etc. More recently, infrastructures such as Tryggve[3] (a Nordic platform for collaboration on sensitive data in Denmark, Finland, Norway, and Sweden) (Kvalheim and Myhren 2016). These efforts are commendable and represent major improvements. However, limited resources, and complex legal variations (national

[1] FOSTER Consortium 2018

[2] Wilkinson et al. 2016

[3] NeIC: https://neic.no/tryggve; [Accessed April 1, 2022]

implementations of GDPR), as well as other factors, remain major challenges for the social sciences and humanities.

The COVID-19 pandemic has rapidly shifted both expectations and capacities for secure remote desktop work, including research. Ability to access data from a wider range of locations is becoming the default expectation among researchers. Research infrastructures are responding rapidly as well. Research infrastructures that were initiated with exclusively on-site access are beginning to introduce remote desktop (e.g., CASD, a member of IDAN). This white paper supports and extends these initiatives by clarifying terminology, assessing existing platforms and services, and providing recommendations regarding how to develop remote access infrastructures for sensitive data.

1.2 Goal

The primary goal of this white paper is to provide guidance and recommendations to the EOSC stakeholders for future infrastructure investment for remote access to sensitive data in the SSH. This means making such access as easy as possible, using procedures and technologies that satisfy legal and ethical responsibilities. Where access can be provided to researchers' desktops, this is a laudable goal for ease of access. However, in the foreseeable future, expanding access via individual and networked secure facilities will remain essential. Fortunately, establishing secure facilities is often a useful step toward this goal. While such infrastructure exists in other domains, such as genetic data, the focus here is on, or alternative, requirements for the SSH.

Moreover, this white paper bases its findings on the knowledge of numerous data professionals who have direct experience planning, implementing, managing, and sustaining diverse forms of remote access and secure facilities. In doing so, our goal has been to maintain the vision of expanding such infrastructure, while remaining grounded in the practicalities of actually operating such facilities in a sustainable manner.

1.3 Scope and Limitations

Growing pressures to open data run up against equally strong requirements for data security and protection. This document makes several contributions to these discussions. It finds common ground and clarifies differences between the social sciences and the humanities. The more closely the requirements of the two domains can be aligned, the more cost-effective future research infrastructure investments will be. An essential second step has been to clarify terms used in the area. This has been done by using (and adapting where necessary) existing standard glossaries and providing an Annex of selected terms used in this white paper. Finally, recommendations are offered which, because they are informed by current practices and developing technology, offer a realistic path for the expansion of remote access for sensitive data, whether through on-site, remote execution or remote desktop.

There are, inevitably, remaining limitations on what this white paper can achieve. Different practices and needs remain between the social sciences and the humanities. After diligent efforts have been made toward integration, a sensible program will acknowledge these differences, design infrastructures to support them, and not build an idealised system ill-suited to either domain. Second, while great efforts were made to be inclusive, resource constraints prevent it from being comprehensive. Similarly, all information is as current as possible, but this is a highly dynamic field.

1.4 Organisation of this Document

Because there are very few agreed definitions of even basic terminology in this domain, this paper begins in Section 2 with definitions, a "discursive glossary", defining the key terms used, especially in two areas – modes of data dissemination and access classifications for data, and explaining interrelationships among these concepts. Section 3 presents a brief overview of the European remote access landscape in the social sciences and humanities, with a glance at other disciplines.

Section 4 proposes a set of minimal requirements for the provisioning of secure access for sensitive data. Three dimensions are defined: organisational/administrative, legal, and technical. Section 5 reviews, compares, and assesses several solutions that are currently in use by major organisations active in the SSH. Recommendations for future research infrastructure development are offered in Section 6, and Section 7 concludes.

2. Remote secure access for sensitive data in SSH

Remote secure access is best understood in the broader context of data access management. Typically, a dataset held in a repository has an *access level* determined by the data provider, that depends on the detail, confidentiality, and sensitivity of the data. Differences in how a specific dataset can be accessed vary according to these levels and is often governed by a Data Access Policy. In this section, definition of sensitive data against the background of various disciplines (2.1) will be presented. Next, definition of sensitive data will be put in context by defining various access levels (2.2). Then, the dissemination modes related to these access levels (2.3) will be distinguished and this will be concluded with an overview and comparison for these modes and the terms used in both the social sciences and the humanities (2.4).

2.1 Key Definitions: Data and Sensitive Data

Research data can be categorised in various ways. In general, "Research data are data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content has the potential of becoming research data"[4].

Table 1 presents an overview of definitions of selected data types from the ISO-5127 terminology registry[5]. These definitions are formulated in liaison with ISO/TC 46: standardisation of practices relating to libraries, documentation and information centres, publishing, archives, records management, museum documentation, indexing and abstracting services, and information science, demonstrating the wider scope of the definitions.

Table 1: Definitions of data types, ISO 5127 2017

Type	**Definition**
research data (3.1.10.10)	data (3.1.1.15) collected, observed, or created, for purposes of data analysis (3.1.11.18) to produce original research information (3.1.1.16) and results
open data (3.1.10.13**)**	data (3.1.1.15) available (3.1.11.03)/visible to others and that can be freely used, re-used, re-published (3.3.4.01) and redistributed by anyone
sensitive data (3.1.10.16)	data with potentially harmful effects in the event of disclosure or misuse
personal data personally identifiable information PII (3.1.10.14)	data (3.1.1.15) relating to an identified or identifiable individual
classified data (3.1.10.17)	data (3.1.1.15) to which access (3.11.1.01) is restricted by administrative means varying according to the degree of data protection (3.13.5.01) or information (3.1.1.16) protection sought
confidential data (3.1.10.18)	data (3.1.1.15) to which only a limited number of persons have access (3.11.1.01) and which are meant for restricted use

The ISO definition of "sensitive data" is a broad one. For sensitive data specifically related to persons, ISO also uses the term "sensitive PII" which it defines as a "category of personally identifiable information (PII), either whose nature is sensitive, such as those that relate to the PII principal's most intimate sphere, or that might have a significant impact on the PII principal"[6].

[4] CASRAI: https://casrai.org/term/research-data/; [Accessed January 27, 2022a]

[5] ISO 5127 2017: https://www.iso.org/obp/ui/fr/#iso:std:iso:5127:ed-2:v1:en

[6] ISO/IEC TS 20748-4 2019, 3.18: https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:20748:-4:ed-1:v1:en:term:3.18

Personal and sensitive data in the sense of the GDPR are data which contain personal information in the sense described in Article 4[7]:

'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

This therefore also includes sensitive data in the perspective of the GDPR[8]:

"The following personal data is considered 'sensitive' and is subject to specific processing conditions:

- personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs;
- trade-union membership;
- genetic data, biometric data processed solely to identify a human being;
- health-related data;
- data concerning a person's sex life or sexual orientation."

This type of sensitive data is also called special category data in Recitals 51-56. In Recital 52[9] exceptions for the processing of special categories of personal data are defined among which a derogation for "scientific or historical research purposes or statistical purposes" can be found.

In SSH, sensitive data is often personal data. In other domains more general definitions are used for information that may also warrant strong protection measures for instance from an ethical perspective e.g., some bio-diversity information is sensitive from the perspective of avoiding species extinction.

> *In this white paper e the term sensitive data will be used in a broader sense than is common in the social sciences. It will denote any type of research data that cannot be openly shared and hence need special safeguards in terms of access restrictions to prevent their unregulated distribution.* The sensitivity of the data may be caused by the personal information contained in the data, the nature of its content (e.g., business related data) or the Intellectual Property Rights associated with it.

For sensitive data other classifications and subcategories are in use which are less related to their nature (as being personal or IPR protected data), but rather reflect their vulnerability and demand for protection. These are the terms that have been typically encounter when access levels of data were revised. Although

[7] EU General Data Protection Regulation (GDPR) 2018a: https://gdpr-info.eu/art-4-gdpr/

[8] European Commission: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en; [Accessed February 7, 2022]

[9] EU General Data Protection Regulation (GDPR) 2018b: https://gdpr-info.eu/recitals/no-52/

not legally precise, in practice, sensitive data often have a higher disclosure risk than data that is more openly shared.

2.2 Access Levels to Data

Data access categories can be based on different factors - for example, the type of usage (research or for profit), the category of user (professor, student, other), or the sensitivity of the data itself. In practice, many archives use combinations of factors to assign data access (UK Data Archive, Australian Data Archive, GESIS, and ICPSR). One robust system is called "Five Safes" and is a framework for assessing overall risks by considering projects, people, settings, data, and outputs (Ritchie 2017). The distinction is not always clear, but the point is that access categories are metaphorical boxes into which datasets are placed to give some indication of the restrictions attached to the dataset[10].

Data access conditions are often broad categories into which datasets are allocated depending on either the sensitivity level of data, on depositor requirements on access, or on stipulations in the participants' informed consent forms. In SSH the most common approach[11] to create data access categories is to define three broad categories: open, accountable, and restricted access.

Table 2: Three categories of data access conditions according to Horton and Perry 2020

Open	Data are usually deemed to have no or only negligible disclosure risk. This can be because data have no identifiers, or they have been removed, or there is clear and documented consent to share any personal data.
Accountable	In this category, there are some conditions or limitations imposed on access. These range from minimal (online registration, agreement to terms of use) to more complex conditions (for example, prior approval of the data producers, proven affiliation with a research institution). Usually, these conditions are in place because disclosure risks of the data are "small, but not negligible"[12].
Restricted	Data can require restricted access for multiple reasons. Most typical for social science research is significant disclosure risk, or the actual inclusion of personal data. But there are other reasons such as intellectual property limitations, as explained above. Moreover, some data cannot be altered (de-identified) without destroying value for research purposes. In other cases, there can be restrictions imposed by data producers/owners. These restrictions can include access only via a safe enclave and mandatory training for researchers.

[10] Horton, Perry, and Bishop 2020

[11] See Horton and Perry 2020

[12] Other terms for this access level are "Safeguarded" (UKDA), "Academic" (CLARIN).

Interestingly ISO-5127 distinguishes a somewhat deviant set of access types[13]. These are presented in Table 3 below.

Table 3: Access types of definitions, ISO 5127 2017

Type	Definition
free access (3.11.106)	open access (3.11.1.05) that does not require financial compensation
open access (3.11.1.05)	unrestricted access (3.11.1.01) to information (3.1.1.16), documents (3.1.1.38) or information services (3.2.1.33)
closed access (3.11.1.07)	access (3.11.1.01) to information (3.1.1.16), documents (3.1.1.38) or information services (3.2.1.33) limited by general or specific regulations
remote access (3.11.1.04)	use of an electronic resource (3.3.3.06) stored on a server through a computer network

ISO's definition of remote access is preferred to consider as pertaining to a mode of dissemination (access control in ISO terms[14]), which will be addressed next[15].

2.3 Modes of Dissemination

Another view on access to data focuses on the modes of data dissemination (adapted from RatSWD):

Direct Access – The data are copied (can include licence, and conditions) from a repository to the user's computer. The data are sometimes referred to as 'downloadable data' and are data that are considered to be sufficiently anonymised so that the risk of confidential information disclosure is negligible[16].

Securing direct access to sensitive data can take various shapes. In the humanities, download facilities for sensitive data are very common. Access is arranged through user authentication and a licence to be signed by the receiving party in which the conditions for the processing of the data are outlined. Examples of this approach are the CLARIN data centres Talkbank and The Language Archive.

In the domain of language and speech technology, there are also commercial data providers such as ELRA and LDC that have cleared the privacy issues of the data which they offer and provide these via a licenced download. These organisations take the burden of data protection measures away from their

[13] ISO 5127 2017: https://www.iso.org/obp/ui/fr/#iso:std:iso:5127

[14] ISO 5127 2017, 3.11.1.08: https://www.iso.org/obp/ui/fr/#iso:std:iso:5127

[15] In environments with complex needs (e.g., medical field), more fine-grained categories are used: see Wiki: http://ecrin-mdr.online/index.php/Object_access_types; [Accessed February 8, 2022]

[16] Dalenius 1977

customers by making the appropriate arrangements with their data providers or their own staff with respect to the required informed consent procedures and/or contractual coverage.

An important reason for preferring this download access solution is the direct contact with the data that the researcher obtains. Browsing and analysing the data through a variety of (self-selected) tools permits a more flexible examination of the data than remote access facilities can typically offer.

When direct access is not possible, due to specific privacy or disclosure risks, the data may not be transferred. In all the modes below, the common element is data remaining under control of the data source (e.g., repository).

Data Enclave – (also Safe Rooms, Secure Labs, Safe Enclave, Safe Haven, Safe Pods, Physical Data Enclave (PDE), Trusted Research Environments (TRE)). This offers researchers the opportunity to work on site at the premises designated by the data producers. Typically, before any results of statistical analysis are physically released to the researchers for further use, they are checked by the data producers or a delegated service to ensure that statistical confidentiality is maintained, and data corruption avoided (output control). Data enclaves can have access facilities that can connect users to sensitive data located at different physical locations or organisations.

Clearly, such facilities have a major added value, by enabling access to data that cannot be downloaded. The network of enclaves is growing. An example is the International Data Access Network (IDAN), a collaboration which facilitates the use of "controlled access" data between six European Research Data Centres across four countries. However, two major limitations remain. First is the time and costs for researchers to travel to the enclave, which can be very high. Second, limited capacity (e.g., seating, and local expertise for disclosure checks on outputs) limits the availability of places in these facilities.

Remote Execution (or Remote Data Processing) – This solution is remote in that from another location researchers send a scripted analysis specification to the repository, where it is executed. Researchers are not able to view or browse data or results on the screen. The (interim) results are sent to the researchers after a data protection check. This may be useful for some studies where the data and their structure is precisely known, but otherwise a serious limitation is that researchers do not have the desired control over the process of analysis and the correctness of its output. An example is the remote execution system of LIS, now the Cross-National Data Center in Luxembourg, which gives access to comparative data on income, wealth, employment, and demographics.

Remote Desktop – With the remote desktop procedure, data are stored and processed exclusively on the servers of the data-retaining organisation. The user interface is transferred to the researchers' screen via a secure connection (virtual desktop). The researchers' access device is only used to communicate with the server and does not hold a local copy of the data. The applications and data are located exclusively on the repository or data owner's server, whereby viewing, and browsing the results and data is possible within a familiar desktop environment[17] (Virtual Data Enclave at ICPSR, Alter), but downloading the data is not possible. In a remote desktop environment, the users are not able to copy (combinations of) indirect identifiers and to google for a person.

Within the category of remote desktop, there can still be variability. In some cases, the system is totally location-independent; researchers can work from anywhere. In others, it is “remote” from the data source, but might require the researcher to work in a formal institution, such as a university, or a safe room, where additional security measures can be present and confirmed (such as no internet connection, and no permission to bring a mobile phone or camera).

In this white paper, *remote access* is defined as [18] *any access mechanism securing that dataset cannot be copied, but remain on the secure servers of the data provider and can only be in location A and are accessed through a secure internet connection from specific other location(s) B*. No physical transfer of the sensitive data ever occur; all browsing and analysis of data are undertaken remotely from location B, on the secure servers that are based in location A.

Ideally remote access can be established from any computer at office and home locations. However, some data owners favour retaining the controls of an institutionally regulated location, e.g., an office or lab.

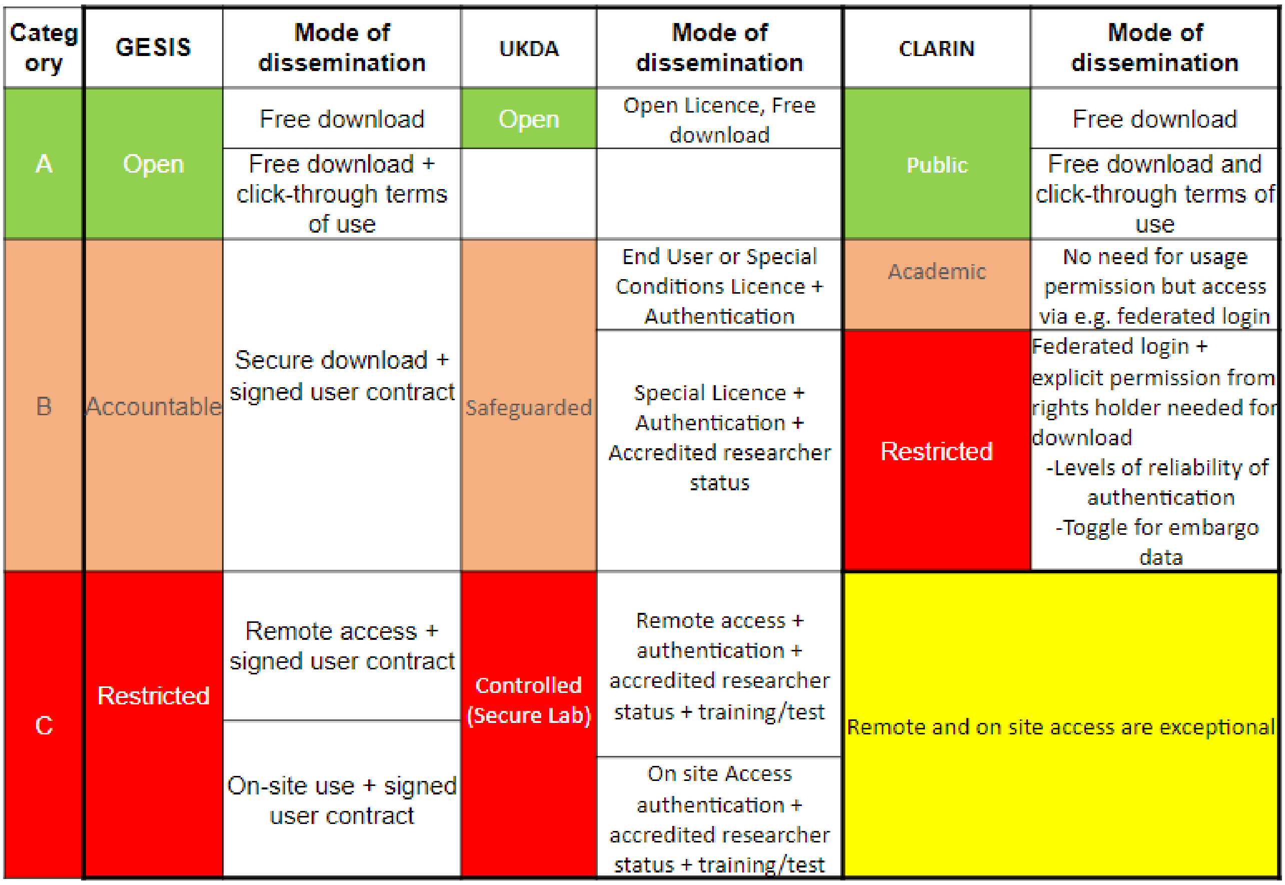
2.4 An Illustration: A Schematic Overview

This section brings together the access levels (2.2) and modes of dissemination (2.3). The diagram below presents the spectrum of access levels of data resources varying from totally open and directly downloadable resources up to extremely restricted through on-site access only, exemplified by practices at GESIS, UKDA and CLARIN. As for CESSDA, there is no standardised approach for access to sensitive data, but this is delegated to the associated archives such as GESIS and UKDA. The practices of GESIS and UKDA are exemplary for the social sciences, and CLARIN for the humanities, especially the humanities using and providing language resources. These organisations cover most of the options in Europe regarding access to data seen in the social sciences and the humanities. However, it should be kept in mind that there is a qualitative difference between CLARIN ERIC on the one hand and UKDA and GESIS on the other. UKDA and GESIS are separate organisations that provide access to data in their holdings. CLARIN ERIC is a coordinating organisation that provides recommendations for metadata interoperability, a platform for metadata harvesting and some Authentication and Authorization Infrastructure (AAI) for its member data-centres to organise access to their data.

Figure 1: The spectrum of access levels of data resources

[17] Alter and Vardigan 2015

[18] Adopted from Woollard et al. 2021



Category	GESIS	Mode of dissemination	UKDA	Mode of dissemination	CLARIN	Mode of dissemination
A	Open	Free download	Open	Open Licence, Free download	Public	Free download
		Free download + click-through terms of use				Free download and click-through terms of use
B	Accountable	Secure download + signed user contract	Safeguarded	End User or Special Conditions Licence + Authentication	Academic	No need for usage permission but access via e.g. federated login
				Special Licence + Authentication + Accredited researcher status	Restricted	Federated login + explicit permission from rights holder needed for download -Levels of reliability of authentication -Toggle for embargo data
C	Restricted	Remote access + signed user contract	Controlled (Secure Lab)	Remote access + authentication + accredited researcher status + training/test	Remote and on site access are exceptional	
		On-site use + signed user contract		On site Access authentication + accredited researcher status + training/test		

In the diagram Categories A, B, C correspond to "open", "moderate", and "restricted" respectively, as described in section 2.2. The modes of dissemination for each of the organisations are represented as well with some detail of the realisation.

Note that although both “Free download” and “Free download via clickthrough” are considered open, in legal terms the responsibility and liability of the hosting organisation for users breaching the licence conditions can differ. The table only expresses the differences between the access processes.

With respect to downloadable data, restricted data in the terminology of CLARIN ERIC[19] are data with the label RES[20]. Comparable to this category at UKDA is data access marked as Safeguarded with Special Licence[21]. According to a procedure that is similar to that of CLARIN ERIC for such data, the user needs a

[19] Which classification is a generalisation of the individual resource centres constituting the CLARIN infrastructure

[20] See CLARIN: https://www.clarin.eu/content/licenses-and-clarin-categories; [Accessed February 8, 2022]

[21] See UKDS: https://ukdataservice.ac.uk/help/access-policy/how-to-download-and-order-your-data/; [Accessed February 8, 2022c]

login and has to sign a licence agreement. At GESIS this level of access permitting download is termed Accountable and clearly limited[22].

Relevant in the context of remote access are the categories Restricted and Controlled Access[23].

Remote access is offered at UKDA via its Secure Lab facility. The access level for this type of sensitive data is termed Controlled, and access is subject to a number of conditions and agreements[24]. Remote (secure) access is currently not provided in the CLARIN infrastructure. A new system is being developed at GESIS which will allow remote access that will be live in the future.

On-site access means that the data can only be accessed within the physical data enclave of the data provider. At UKDA the access level for this type of sensitive data is also termed Controlled, and access is organised as another of the options in the Secure Lab facility. GESIS provides on-site data access via its Secure Data Center (SDC). Here data can be analysed by appointment and on signing a contract for on-site use at a Safe Room guest workstation in Cologne. This mode of access is typically not offered by CLARIN data centres.

More detailed information about access levels and modes of dissemination at the three organisations can be found in Annex 2.

3. Current Solutions in Access to Sensitive Data in SSH

3.1 Social Sciences

The current landscape for remote access to sensitive data in the social sciences in Europe is generally fragmented but slowly still evolving, with a history that goes back for about two decades. It was the National Statistical Institutes (NSIs) that took the lead in the early 2000s in developing physical safe rooms for making available their more sensitive microdata for research purposes. Already at the end of that decade many NSIs had in place physical safe rooms where researchers could come to access sensitive data (reference).

Other actors followed suit in establishing safe rooms for sensitive social science data, including several national data archives (e.g., UKDA, PROGEDO, and more recently Sikt - formerly NSD), research institutes (e.g., RemoteNEPS), or for specific projects (e.g., SOEPremote). In this way physical safe rooms became

[22] GESIS 2018

[23] We refrain from categorising embargoed data since by definition this is a temporary status, and once an embargo ends, data will move into one of the other levels.

[24] See UKDS: https://ukdataservice.ac.uk/help/access-policy/types-of-data-access/; [Accessed February 21, 2022e]

more common in the social sciences in Europe, but still with significant gaps – many NSIs, archives, and institutes had no such dissemination mode, and still do not for different reasons. On the other hand, there are new initiatives to create safe rooms in Europe for sensitive social science data. For example, a recent informal poll indicated that several national data archives within CESSDA have concrete plans to develop safe rooms for sensitive data (e.g., DANS, CSDA, and FORS). At FORS in Switzerland a safe room was conceived and developed during the course of the SSHOC project as part of task 5.4 and is currently being tested. The process is described in detail in Annex 3 to this white paper.

Over time many NSIs began to recognise that their safe rooms were generally inconvenient for researchers, who had to take time off to travel long distances. Consequently, they began to explore and develop alternative technical solutions that could allow access to data remotely, without significantly increasing the risk of disclosure of personal information. Soon remote access solutions started to emerge as an alternative[25].

By 2013, remote access was already an available dissemination mode for official microdata at NSIs in Sweden, Denmark, the United Kingdom, Slovenia, and the Netherlands (cite DwB 4.1). Also, official statistical data were disseminated remotely on behalf of NSIs through organisations such as the UKDA (United Kingdom), CASD (France), and the Research Data Centre of the German Federal Employment Agency at the Institute for Employment Research (FDZ-IAB) (Germany). By 2019, other NSIs with remote access to microdata could be added to the list – Finland, Norway, Estonia, Slovenia, Ireland, France, Croatia, Latvia, and Lithuania[26]. In 2021 the European Statistical System Committee decided to develop remote access to European secure use files, which previously could only be accessed by travelling to the Eurostat safe centre in Luxembourg[27].

Already in the 2000s there were a few small-scale initiatives to *connect* physical safe rooms in order to reduce travel distances and facilitate access to data across borders. In this way, researchers could visit a safe room in one country in order to access data remotely from a safe room in another country. An example is the availability of data from IAB in Germany on-site at the ICPSR, beginning in 2004. Such initiatives became more common however in the 2010s. This was advanced most significantly in several Nordic countries where efforts have been made to harmonise practices and facilitate cross-national access to NSI's microdata. Currently this is possible in Denmark, Finland, and Sweden. A second example is the International Data Access Network (IDAN), a collaboration between six research data centres from

[25] The interest in facilitating access to official microdata and the role of data archives in this caught the attention of the European Commission, which funded the project Data without Boundaries (DwB) from 2011 to 2015. DwB aimed to consolidate and advance access to official statistics for the benefit of the research community in Europe. Out of this work arose a vision for an easier access to official microdata for research use across borders within a connected network of nodes, where data did not move. Even though there was no follow-up to DwB (rejected application), there were initiatives to carry things forward, especially regarding remote access (Cornuau et al. 2013).

[26] Rat Für Sozial- Und Wirtschaftsdaten (RatSWD) 2019

[27] Eurostat 2021

France, Germany, Netherlands, and the United Kingdom. The overall goal of the network is to facilitate transnational access to confidential microdata across the partners institutions. The partners work towards promoting the available microdata, developing common knowledge and standards, and establishing bilateral agreements across all their institutions. The longer-term aim is to have a framework for a multilateral agreement that would replace the bilateral agreements. Just as the final version of this paper was being drafted, the UK Data Service will be making selected UK controlled data available to researchers abroad via the IDAN network[28]. This access relies on safe facilities (not remote desktop), however, any international expansion of access to sensitive data is a significant step.

Despite these initiatives and developments over the previous two decades, it is clear that the promise of remote access to sensitive data in the social sciences in Europe remains unfulfilled. The landscape remains fragmented, and change is slow. This can be attributed to several factors. First, large-scale European investment in remote access for the social sciences was mostly limited so far to the EC-funded DwB project[29]. While CESSDA has recognised the potential importance of building up a European remote access network (ERAN), few concrete steps have been taken collectively to make this a reality among European national service providers.

As a result, most of the infrastructure development for remote access has occurred in a bottom-up fashion, mostly on an individual institutional level or else with small-scale bilateral collaborations. Further, NSIs have little incentive to invest in infrastructure for social science researchers (this usually not being in their mandates), and national data archives generally have other priorities and limited resources to invest in remote access.

Second, there are still many legal and political challenges that slow down or prevent the establishment of remote access solutions and networks across European countries. Notably, recent years have also brought the issue of data privacy and security to the forefront of public policy debate. Data breaches have raised public fears and concerns, and GDPR has taken effect demonstrating some “teeth” for enforcing data protection.

Nonetheless, the growing list of working facilities and the corresponding demand from researchers should be seen as a positive. While they do not yet meet the comprehensive list of essential requirements for SSH, they do provide proof cases that both the demand for the service is high and that the necessary technologies exist. In addition, the existence and expansion of partial facilities (remote desktop for limited data collections, “enclave-to-enclave” links) has made the demand for sharing of secure data far more visible. The growth in use is high, even when researchers continue to have to overcome significant

[28] UKDS: https://ukdataservice.ac.uk/about/research-and-development/international-data-access-network-idan/; [Accessed March 7, 2022d]

[29] The recently funded Horizon Europe “eRImote” project, intended to advance remote access across all scientific disciplines in Europe, is an exception.

challenges of travel, waiting times, etc[30]. It is also clear that there is a growing need for linked data, which typically links less-sensitive sources (scientific use file) with sensitive data.

Further, the changes in data protection regulations cut in two directions regarding remote access infrastructure. While fears are generally higher among data owners as well as lawyers who advise institutions that disseminate sensitive data, there is a growing appreciation that data are in fact better protected when held in secure infrastructures, even when these enable sharing. Moreover, while there are risks of a fragmented regulatory environment as individual country derogations of GDPR diverge, it is also possible, as the Norway example shows, to use the GDPR to push for cross-national legal harmonisation.

3.2 Humanities

The humanities comprise a widely varied set of (sub) disciplines, where researchers for different purposes using different methodologies will also need access to data that is classified as sensitive. However, except in the case of CLARIN infrastructure, which is described below, based on available recommendations and infrastructures, there seems to be no large need and no common systematic approach in the humanities and the humanities research infrastructures, such as DARIAH[31] to manage access to confidential data. There are also highly specialised humanities infrastructures such as the emerging European Holocaust Research Infrastructure (EHRI)[32], which by nature of its subject matter needs extra attention for its data processing and publication.

Those humanities fields that are involved with access to sensitive data often already border on or collaborate with disciplines where access to sensitive data is well understood and practised, such as in the medical sciences. Sensitive personal information as part of research data is usually involved when observing and studying individuals, especially behaviour and psychological characteristics. In the domain of behavioural sciences (although these are seen as part of the social sciences) many of these are with respect to their methodologies and research data management[33] more in line with the humanities (personal historical accounts) or medical sciences. Examples are psychology, pedagogy, behavioural aspects of biology but also economic and political sciences. See also table 4. In general, whenever there is a medical angle, the data access policies and strategies will be always dictated by the different medical data access policies[34] e.g., Data Enclave type restrictions, beyond restrictions imposed by the GDPR. If the data pertain to child subjects, even with parental consent, things become even more complicated.

[30] Rat Für Sozial- Und Wirtschaftsdaten (RatSWD) 2019, p. 6

[31] DARIAH: https://www.dariah.eu; [Accessed February 8, 2022b]

[32] EHRI: https://www.ehri-project.eu/; [Accessed February 8, 2022]

[33] e.g., the use of interviews and recordings

[34] Dutch government travel information for entrepreneurs: https://business.gov.nl/regulation/medical-records/; [Accessed February 18, 2022]

The general situation in the humanities is that sensitive data is made available via downloads through restricted access regulations. CLARIN-ERIC is a typical and prominent example of this approach as it was illustrated above in section 2.4 by highlighting CLARIN's RES restricted access label. Access is regulated through licences outlining the purposes for which the data can be used and the stipulation of warranties and (accountability) provisions installed to prevent data leaks. This approach is backed up by the strategy of desensitising data by using explicit consent from the data subjects to share their data[35].

In Europe also DARIAH-EU is very active in sharing data for the humanities. For the DARIAH community DANS[36] is offering Dataverse as a download method to share sensitive data that can be anonymised under a restricted access licence[37]. This recipe for data re-use is therefore similar to CLARIN's. According to DANS, Dataverse is not suitable for sensitive data that cannot be anonymised or pseudonymised. In that case DANS recommends a more secure environment. The intention is that the DANS Data Stations will eventually be linked to a more secure environment, in which you can store very sensitive data.

For access to confidential data via download, there is a strong need for the correct licences. This pervades all disciplines, e.g., in Biomed the aforementioned ECRIN services for sharing sensitive data[38] and a Legal Assessment Tool (LAT)[39]. For the broad humanities the "DARIAH Pathfinder to Data Management Best Practises in the Humanities"[40] suggests the Creative Commons Licence picker[41] and the CLARIN Lindat Licence selector tool[42] (CLARIN's restricted access label system). There is also advice on use and obtaining of consent[43] from individual subjects and data contributors.

There are also specific project and infrastructure initiatives that address the access to confidential data in different subdisciplines of the humanities and social sciences. For instance, for criminology, there is an option to create 'safe-to-use' sanitised datasets[44]. The associated site also provides references (although US centric) with respect to specific legal restrictions for this type of data. Other websites /

[35] See e.g. the ELDAH consent form wizard to help GDPR compliant consent form creation: DARIAH 2020: https://www.dariah.eu/2020/09/15/sshoc-workshop-putting-data-protection-into-practice-the-gdpr-and-the-dariah-eldah-consent-form-wizard/

[36] DANS-KNAW: https://dans.knaw.nl/en/; [Accessed March 2, 2022]

[37] See e.g. Wittenberg and Király 2020 and Dataverse: https://guides.dataverse.org/en/latest/user/dataset-management.html#restricted-files; [Accessed February 8, 2022]

[38] ECRIN: https://ecrin.org/news/proposed-categorisation-system-published-resources-toolbox-sharing-sensitive-data; [Accessed February 8, 2022b]

[39] BioMedBridges: http://www.biomedbridges.eu/supporting-researchers-sharing-sensitive-data-identifying-requirements; [Accessed February 8, 2022]

[40] Tóth-Czifra 2019

[41] GitHub: https://choosealicense.com; [Accessed February 8, 2022]

[42] Kamocki, Straňák, and Sedlák 2015

[43] DARIAH: http://consent.dariah.eu; [Accessed February 8, 2022a]

[44] Wheeler 2020: https://andrewpwheeler.com/2020/10/25/open-source-criminology-related-network-datasets/

information sources concentrate on the use of specific types of data e.g. (Willams and Burnap) for twitter[45].

The highly topic-specific and specialised EHRI infrastructure was already mentioned. Its data includes personal testimonies (recorded speech or in writing) containing sensitive data. It provided plans and recommendations on access and protection of sensitive data in several project documents e.g., "EHRI-3 Data Management plan"[46] and the older "Digital Handbook on Privacy and Access"[47].

Humanities researchers still need access to sensitive data that cannot be shared by the method of download via a restricted access licence. In fact, there are many situations in which they have to resort to and comply with solutions that have been implemented for medical and social sciences, including RSA. However, often they do not find the tools in such environments that they need for their research.

This need is observed for sensitive interviews such as patient doctor consults, clinical data of e.g., dysarthric speakers, police, and court interviews. The table below presents an exemplary overview.

Table 4: Exemplary overview for sensitive data

Domain	**Type of data**	**Expected restrictions beyond GDPR**	**Organisations**
Medical	Interview doctor/patient: Psychology; pathological speech,	Doctor patient confidentiality	Hospitals
Behavioural, development studies	Children speech,	Parental permissions	Special schools
Criminology, sociology	Police/lawyer interviews	Confidentiality lawyer client	Police

If RSA solutions could be implemented for these data types, a wealth of additional data would open for humanities research.

[45] Williams and Burnap: https://www.britsoc.co.uk/media/24899/using-twitter-for-criminology-research.pdf; [Accessed February 2, 2022]

[46] Bryant et al. 2021

[47] Luyten and Boers 2013

3.3 General Solutions and Other Disciplines

Outside SSH the domains with the most mature development concerning access to sensitive data are the Biomedical sciences (Biomed). In Biomed, there is the additional challenge of managing the confidentiality of medical data, for instance with clinical trials where work was done in the ECRIN[48] research infrastructure and in different EU projects. For instance, a proposed categorisation system for sharing sensitive data[49] and a Legal Assessment Tool (LAT) [50].

Otherwise, in Biomed, there is an effort on managing protocols and security aspects to access repositories with human bio samples (biobanks)[51].

The SSHOC sister-project for the life sciences cluster EOSC-Life[52] is working on capabilities and identification of suitable cloud providers for secure hosting of sensitive data and enabling multi-RI applications and workflows that comply with sensitive data regulations[53]. As a follow-up of the earlier assessment and recommendation tools for handling sensitive data, now a "Toolbox for sharing of sensitive data"[54] is developed, with a scope that includes intellectual property considerations, biohazard concerns, or the Nagoya Protocol[55] concerning access to genetic resources.

There have been initiatives by e-Infrastructures to inventories sensitive data management solutions such as in the EUDAT project[56] and look for possible common approaches, and currently there are infrastructure organisations that provide general information to researchers on how to manage sensitive data e.g., OpenAIRE[57]. Next to this there are quite a few specific solutions that are more broadly applicable, also including application to commercial data as banking, telecom, insurance data where there exists a legal obligation for high levels of protection or even within a single company, keep departmental information systems separate (Chinese wall[58]).

[48] ECRIN: https://ecrin.org; [Accessed February 8, 2022a]

[49] ECRIN: https://ecrin.org/news/proposed-categorisation-system-published-resources-toolbox-sharing-sensitive-data; [Accessed February 8, 2022b]

[50]BioMedBridges: http://www.biomedbridges.eu/supporting-researchers-sharing-sensitive-data-identifying-requirements; [Accessed February 8, 2022]

[51] See, for instance, Nordberg (2021) who states that existing EU regulation applicable to biobanks and biobank research is dispersed through a number of areas of law, including data protection, clinical trials and tissue regulation.

[52] EOSC-Life: https://www.eosc-life.eu; [Accessed February 8, 2022]

[53] EOSC-Life 2021: https://www.eosc-life.eu/d7/

[54] Boiten et al. 2021

[55] CBD 2015: https://www.cbd.int/abs/about/

[56] Kuchinke and EUDAT Sensitive Data Working Group 2017

[57] OpenAIRE: https://www.openaire.eu/sensitive-data-guide; [Accessed February 8, 2022]

[58] Wikipedia 2021a: https://en.wikipedia.org/wiki/Chinese_wall

Some tools worth mentioning are the DataSHIELD[59] solution for secure bioscience collaboration that can also be applied in the social sciences[60]. The commercial Aircloak software[61] offers advanced anonymizing functionality and has use-cases in healthcare, banking, and telecommunication.

Large national organisations with a research data management mission also provide general solutions that fit requirements from different communities as for example TSD[62], ePouta[63]. These are tools that were specifically developed and offered to manage sensitive data. TSD follows the remote access approach, whereas ePouta is a virtual private cloud that provides secured VMs and storage resources directly to your organisation's network. Developing such solutions are costly investments for public organisations that are driven by the national interests to provide facilities for high-profile research, mostly in the Biomed domain, although the facilities can be more broadly applied. The fact that services are developed at a national level should not prevent their availability at the European level, both ePouta and TSD are also provided as the sensitive data services in the EOSC-hub project[64] and are registered in the EOSC Marketplace[65].

The biodiversity and humanities disciplines also have requirements for data that may not be sensitive in the legal definition but are nevertheless considered as such. For instance, some information on endangered species populations cannot be shared freely. And anthropologists' recordings of indigenous peoples' religious and initiation rites, similarly, may not be directly protected by law, but such data are considered sensitive for ethical reasons. Generally, such information is stored in repositories with controlled access.

4. Minimal Requirements for Remote Access to Sensitive Data

4.1 Scope and goal

Due to certain factors (e.g., risk of disclosure), sensitive data can be accessed using a remote access system of some kind that does not allow the data to be downloaded directly. With such systems, researchers can work remotely in a secured and controlled environment. Although the downloading

[59] DataSHIELD: https://www.datashield.org; [Accessed February 8, 2022]

[60] Gaye et al. 2014

[61] Aircloak: https://aircloak.com; [Accessed February 8, 2022]

[62] Øvrelid, Bygstad, and Thomassen 2021;
University of Oslo: https://www.uio.no/english/services/it/research/sensitive-data/; [Accessed February 8, 2022]

[63] EOSC: https://marketplace.eosc-portal.eu/services/csc-epouta; [Accessed February 8, 2022b]

[64] EOSC-hub: https://www.eosc-hub.eu/services/Services%20for%20sensitive%20data; [Accessed March 17, 2022]

[65] EOSC: https://marketplace.eosc-portal.eu; [Accessed March 17, 2022a]

solution is common in the humanities, even for sensitive data, researchers may encounter barriers (as in the case of copyright restrictions). When Direct access is not possible, the remote access alternative is a preferred solution.

The scope of this section is not to set out requirements for “basic” remote access solutions, but only the additional or different elements needed for sensitive data. The focus will be on the particular minimal requirements for remote access solutions to sensitive data.

This document discusses and modifies the data access categories used in several existing systems. Multiple sources for these categories on remote access to sensitive data were consulted[66]. In the next subsections, the minimal requirements will be categorised into three main categories: organisational/administrative, legal, and technical. The main minimal requirements have been included in a spreadsheet that was used for the assessment of existing platforms[67]. For some requirements, there is more detail provided here than in the comparative spreadsheet.

4.2 Organisational/Administrative Requirements

Organisational/administrative requirements comprise procedures and resource aspects for initial work planning for remote access services. Among the key elements here are the definition of eligible users, user-management specifications and user/staff training, documentation, specification of costs involved and general workflow. Some basic requirements are met by all services, so their inclusion in the spreadsheet was of no additional value. Detailed items are available in the list below:

Procedures

- Define eligible user base (e.g., students).
- Specify user account management:
    - account set up,
    - user to sign the agreement – need to validate the signature,
    - authenticate user identity,
    - ID user device (identity and location).
- Provide access request form, specifying user info and affiliation, data collections requested, purpose of research, access duration).
- Describe and issue specific roles to staff members.
- List of accredited access points and registering their technical particulars e.g., IP addresses, certificates etc.
- Manage appointments.
- Set up working environment, including requested secure data.
- Output checking workflow and staff (ranging from self-check to double review).
- Provide training for users in security requirements (various modes possible and consider recertification).

[66] Eurostat 2021; Rat Für Sozial- Und Wirtschaftsdaten (RatSWD) 2019; Schiller et al. 2017

[67] Bishop 2021

- List software supported.
- Define interface language – English (mandatory; additional local language interface is recommended).

Resource Requirements (planning phase)

- Specify hardware costs.
- Specify software licences.
- Specify need for administration staffing, support, service level.

4.3 Legal Requirements

Legal requirements pertain to compliance with (inter)national legal aspects, specifying necessary elements in user agreements and the possibility of sanctions in case of a breach of the agreement.

- The RA facility should have some form of Secure Access User Agreement – legal document specifying:
    - services of parties outlining the tasks and responsibilities of Party A (RA provider) and Party B (RA user),
    - the period of the agreement and options for modification and termination,
    - specifications regarding available research data,
    - the occurrence of fees,
    - application process for Users, (not sure if this is in the User Agreement),
    - data access that is based on appropriate legislative framework (GDPR and national laws),
    - a pledge on data secrecy / a Secure Access Agreement (to be signed by the user and their organisation),
    - reporting obligation for user to provider in case of a breach (of information security or procedure),
    - sanctions in case of a breach (of information security or procedure).

- Data are handled in compliance with legal and intellectual property rights requirements of the data owner.
- The RA facility complies with national, and international legal requirements.
- Ability to prosecute individuals and institutions in case of a breach is ensured.
- Contract and sanctions are sufficient to replace physical security checks of Safe Rooms.
- Description of the obligations of the research entities hosting the access points is provided.

4.4 Technical Requirements

Technical requirements specify applied preconditions for remote access from user and service perspectives, including the internet connection, authentication procedure, restrictions to prevent users from copying/printing the data, metadata in standard formats, disclosure review, logging facility and others.

- The internet (or private network) provides the remote connection between researcher and data.
- Connection uses standardised technologies for encryption of communication.
- Two-factor authentication is provided.
- Terminal server technical is provided (e.g., CITRIX, VPN).
- Software for input devices to communicate with servers is provided (i.e., mouse and keyboard).
- Confidential data does not leave the host repository.
- Restrictions to prevent download, copy, or printing of data are made.
- Data are only accessible for a defined period.
- Access to other software needed by a user is enabled (SPSS, Stata, R, python, tools for tables and graphics etc.).
- Ability to import external data supplied by the user is provided and subject these data to disclosure risk assessment.
- There is an ability to support data linking between data provided by a user and sensitive data (screening of external data for security risks) – optional, not mandatory.
- Personal workspace for end-user is provided (own storage space for result files, code libraries and other user-created files).
- Ability for an authorised group of researchers to access the same data is provided.
- Ability to submit results for disclosure review is provided.
- There is a capacity to scale, as remote access increases demand – servers, support simultaneous users, service processing capacity.

- Logging facility of user activity is implemented:
    - Logon Duration.
    - Failure time.
    - Session state.
    - Session change time.
    - Session type.
    - Associated user Display Name.
    - Application Name.
    - Published application Name.
    - Application start time.
    - Application end time.

- Metadata in standard formats are provided and at least Dublin Core is available for datasets in the RA facility.
- User needs to have: web browser, internet connection (other needs will vary with the system deployed, e.g. use of thin clients).

5. Assessment of current solutions

5.1 Methodology

The focus of this white paper with respect to existing solutions for accessing sensitive data is on describing requirements and general approaches rather than inventorying and evaluating the technical (software) implementations that are currently available and used. This is to avoid detailed comparisons and discussions about technologies, such a comparison was not possible with the resources available. Nevertheless, this sketch is not complete without at least listing some of the solutions that are currently in use by the major organisations active in the SSH RI landscape. And although the effort behind this white paper does not permit any testing or extensive user satisfaction checks, there can be a comparison based on published data sheets and other documentation.

Criteria for selection:

- SSH domain - SSH data and SSH researchers are the primary user community
- Provision of data via remote access (enclave, execution, or desktop)
- A complete service package (account management, data, output controls, etc.), i.e., offerings that are only technical platforms were excluded

In the context of providing lists of solutions, the limitation was to those implementing RSA or other special secure environments, and do not mention specifically the different repository systems used in the Humanities that facilitate the licence agreement and download workflow. One such solution i.e., Dataverse was mentioned in chapter 3.2.

After making the final selections, research was conducted using the provider websites and other secondary sources to complete the spreadsheet[68]. Then every provider was sent its information and invited to comment. A majority replied and all suggestions were added to our information.

5.2 Assessments

The purpose of these assessments is decidedly not to rank, but rather to understand how actual infrastructures are evolving, where obstacles remain, and to gain insights into how to best support future developments. Brief synopses are provided below, with the key insights gained at the end of this section.

Centre d'Accès Sécurisé aux Données (CASD), Secure Access Data Center

[68] Bishop 2021

CASD[69] is a paid service funded by the Investment in the Future (Investissements d'Avenir) program and managed by the National Research Agency that allows researchers to work remotely and securely with the microdata from the French National Statistical Institute, the Ministries for Justice, Education, Agriculture, and Finance, and also health microdata. To access CASD services, users need to sign a preliminary contract agreement that consists of CASD Executive Service Agreement and CASD Terms and Conditions of Use, accompanied by the Pricelist of the services, the technical form to be completed online for hosting an SD-Box™, Access Point Hosting Voucher(s) and Purchase Order(s). After that, users are required to complete the enrolment session (valid for 4 years). Getting access to data and to one's secure work environment is performed via a secure terminal called an SDBox™, by means of the individual access smartcard (issued at the enrolment session) and biometric authentication. In the remote environment, standard software is offered: R Studio, Stata, Microsoft Office, etc. To collaborate with other members, CASD provides shared workspaces. Data linking and installing extra software can be done by request. For export, there are two procedures: manual and automatic (not all projects qualify for the automatic export procedure). When manual exports are requested, they are examined by CASD beforehand. Automatic exports are not examined prior to transmission but may be subject to an a posteriori check by the data depositor.

Key features: remote access via an SD-Box™, paid service, mandatory enrolment sessions, data linking and adding custom software possible by request.

DataSHIELD – Population Health Sciences Institute at Newcastle University, the United Kingdom

DataSHIELD[70] is open-source software (a modified R statistical environment linked to an Opal database) designed to work with sensitive data in different domains (mainly biomedical, healthcare, and social science). It is widely implemented in international projects: EUCAN-CONNECT, LifeCycle, RECAP preterm, InterConnect and others. The workflow follows the primary strategy: "Analysis is taken to the data – not the data to the analysis"[71] that minimises disclosure risks. As a result, data stay secure behind the firewalls on the system where they reside. DataSHIELD wiki page[72] provides a detailed description of workflows, functions, updates as well as disclosure checks. Automatic disclosure control is also available – for example, "each data computer tests any contingency table it creates and will only return a full table to the analysis computer if all cells are empty or contain at least five observations"[73]. Although there is

[69] Bouhari 2018;
CASD 2022;
CASD: https://www.casd.eu/en/; [Accessed March 3, 2022]

[70] DataSHIELD: https://www.datashield.org/; [Accessed February 8, 2022]

[71] Gaye et al. 2014

[72] DataSHIELD wiki: https://data2knowledge.atlassian.net/wiki/spaces/DSDEV/overview?homepageId=12943453; [Accessed March 6, 2022]

[73] Gaye et al. 2014, p. 1934

no mandatory training, DataSHIELD offers a range of various training resources such as beginners' workshops, annual conferences, and R Statistical Language courses.

Key features: open-source solution (R with Opal), widely used in international projects, a rich database of documentation and training materials.

ePouta and Sensitive Data Services – Center for Scientific Information Technology (CSC), Finland

CSC provides two services to support remote secure access to sensitive data. The first one, ePouta[74], is an infrastructure as a service – a cloud computing environment. It does not provide any specific software or licences, and functions mainly as a link between storage resources and the provision of virtual machines. Hence, the use of this service is more relevant for research organisations/communities/groups rather than individual researchers. Besides, ePouta is currently used mainly with Finnish institutions, as a specific VPN connection must be set up with one's home organisation that is easier to do with Finish partners. User policy and terms of usage specify compliance with national and international legal requirements as well as possible security actions. ePouta is not a classic remote access solution but rather offers secure provisioning of VMs and storage directly into the own organisation.

Key features: infrastructure as a service, relevance for research communities, limited scalability to the international level (mainly available for Finnish academic organisations/institutes), rich set of trainings.

The second set of services, called Sensitive Data Services[75], is web-user interfaces for managing sensitive data. This is a new set of services, since the open beta version was released in 2021. Completion is planned for the end of 2022. It consists of four components: (i) Sensitive Data Connect which imports, stores and shares data, (ii) Sensitive Data Desktop with a prebuilt computing environment and pre-installed open-source software (e.g., Python, R), (iii) Sensitive Data Submit which publishes data under controlled access and (iv) Sensitive Data Apply which ensures data re-use. Before uploading sensitive data, they are required to be manually encrypted. Automatic encryption has not been implemented yet.

Key features: virtual desktop, manual data encryption, complete web-interface solutions, rich set of short interactive tutorials on YouTube channel.

Microdata Online Access (MONA) – Statistics Sweden[76]

[74] CSC: https://research.csc.fi/en/-/epouta; [Accessed March 6, 2022a]

[75] CSC: https://research.csc.fi/sensitive-data-services-for-research; [Accessed February 28, 2022b]

[76] MONA: http://www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/mona--statistics-swedens-platform-for-access-to-microdata/about-mona/; [Accessed March 1, 2022a];
MONA: http://www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/mona--statistics-swedens-platform-for-access-to-microdata/rules-and-regulations/terms-of-use/; [Accessed March 1, 2022b];
Rat Für Sozial- Und Wirtschaftsdaten (RatSWD) 2019

This is the platform for enabling remote secure access to microdata from Statistics Sweden. It has operated since 2005, growing every year, and now has over 750 users. Affiliated researchers from authorised research institutions should register, sign the agreement, accept terms of use, and be authenticated with a one-time password. After that, Sweden Statistics data can be analysed via remote access using some common programs for data analysis (e.g., R, Stata, SPSS, QGIS or FreeMat). Usage of MONA is possible outside of Sweden, in countries within the EU or EEA, or third countries approved by the European Commission for transmission. Statistics Sweden requires the requesting institution to ensure that the researchers comply with the data protection requirements and the contractual terms (Terms of Use). In such cases, outputs are checked manually.

Key features: remote desktop; growth in demand; data available outside of home country; no training required; manual output check for intl users.

Microdata.no – Statistics Norway and Norwegian Social Science Data Services (NSD), Norway

Microdata.no[77] is a remote execution solution that provides an opportunity to analyse (confidential) register data from Statistics Norway – i.e., Norwegian National Registry, National Education Database, Register for Personal Taxpayers, Labour market data, FD-Trygd (event history database on welfare grants). Affiliated researchers from approved institutions and students can apply to get access to the service. Institutions that have signed an agreement can manage their own users. Registration of users should be done with (electronically) signed end-user agreements. For authentication, an electronic ID is required. Although the service supports neither traditional statistical software (e.g., Stata, SPSS) nor open-source solutions (e.g., R, Python), it offers an alternative built-on Python interface for importing, processing, and analysing and visualising register data (both progress data and status data). Data linking with personal data and the use of own statistical packages or libraries are not possible. All output is subject to confidentiality measures[78]. Some examples are not allowing to define populations with fewer than 1000 people, smoothing graphic hexbin plots, hiding tables with too many low values, using winsorisation to prevent extreme observations and using randomised noise to check confidentiality. For new users, Microdata.no holds introductory courses in the English language. Additionally, the user guide provides complete information for Microdata.no users. Courses on specific advanced topics are available in the Norwegian language.

Key features: remote execution with alternative built on Python interface, student access possible, extensive user guide, no data linking, no own software or libraries, automatic confidentiality measures.

[77] Microdata 2022;
Microdata: https://microdata.no/en/about/; [Accessed March 12, 2022a];
Microdata: https://microdata.no/en/faq/; [Accessed March 12, 2022b];
RAIRD 2019;
Risnes et al. [Accessed March 12, 2022]

[78] Microdata 2022, p. 138-146

RemoteNEPS – Research Data Center (RDC) of the Leibniz Institute for Educational Trajectories (LIfBi), Germany

LIfBi provides secure remote access to NEPS microdata in a controlled environment via virtual desktop[79]. Among eligible applications are researchers in possession of a university degree employed by a scientific institution (recognized research institution). To connect to a remote desktop, registration with an additional biometric authentication system (keystroke biometrics) as well as a signed Data Use Agreement (for a project) and supplement document are required. The Data Use Agreement[80] complies with national and international legal regulations (e.g., GDPR etc.); indicates accessed datasets and duration of the access; specifies terms of extraordinary termination and consequences resulting from breach of the agreement. According to the rules, all new users are required to complete a training course before using virtual desktops and NEPS data. Access can be provided to groups via common directories at their own workstations. The connection to the service is encrypted and can be done through any operating system with the Internet. Main software for data analysis is available in the controlled environment (e.g., R, Stata, SPSS). Additional software can be provided under request (for the corresponding payment). The output from the data analysis undergoes a manual check for compliance with regulations before it is available for data users. All exports and imports of files on RemoteNEPS are logged and saved.

Key features: virtual desktop, access to NEPS data, detailed Data Use Agreement, manual output check, mandatory training, single study, biometric authentication, linking, and groups permitted.

Secure Data Lab – UKDS[81]

Data is made available through both a physical facility and via remote desktop. Researchers must be affiliated with authorised institutions, and they must undergo training. Business, administrative, cohort and longitudinal data from multiple sources is provided. Non-sensitive data can be added to individual projects, subject to review and approval by SecureLab staff. Group projects are possible if all participants have signed a joint data usage agreement. Specialised staff apply statistical control techniques to ensure the delivery of safe statistical results.

Key features: remote desktop; data from multiple sources; mandatory training; support for data linking and group projects, some data available to non-UK Researchers.

[79] NEPS: https://www.neps-data.de/Data-Center/Data-Access/RemoteNEPS; [Accessed March 7, 2022b]; Rat Für Sozial- Und Wirtschaftsdaten (RatSWD) 2019

[80] NEPS: https://www.neps-data.de/Data-Center/Data-Access/Data-Use-Agreements; [Accessed March 7, 2022a]

[81] UKDS: https://ukdataservice.ac.uk/help/secure-lab/securelab-faqs/; [Accessed February 23, 2022b]

SOEPremote – Research Data Center of the Socio-Economic Panel (SOEP)[82]

The Socio-Economic Panel (SOEP) is one of the largest and longest-running multidisciplinary household surveys worldwide. Access is possible via thin client for the secure facility as well as with remote execution of user-provided syntax. The applicant must provide documentation that his/her institution conducts independent scientific research. The direct use of SOEP data is subject to the strict provisions of German data protection law. Output checking is manual, with some automation used for remote execution outputs. In case of a successful security check, individual upload of external data to the remote environment is possible.

Key features: student access possible (with supervisor approval), remote desktop, remote execution (mail-based) enabled, eligibility depends on documenting conduct of independent research.

The Research Data Centre (FDZ) – Federal Employment Agency (BA) at the Institute for Employment Research (IAB)[83]

The Institute for Employment Research (IAB) conducts research on the labour market in order to advise political actors at all levels. Projects must be in the domains of the labour market and occupational research. Both on-site use and remote execution are possible. To register, the usage agreement should be signed with an electronic or handwritten signature. Regarding statistical software, STATA, R and Octave are available for users. They can submit scripts (e. g. Stata do-files) and access the approved results after verification of compliance with data protection legislation via a personal account. Checking is primarily manual, with some automation for remote execution. Some options for additional data may be possible.

Key features: remote execution, student access possible (with supervisors), manual and partially automatic output check.

The Research Data Centre for Higher Education Research and Science Studies (FDZ-DZHW), Germany

FDZ-DZHW[84] provides remote access to qualitative and quantitative sensitive data (including DZHW data) from the field of higher education research and science studies. The data are anonymised to a certain extent, depending on the mode of access and purpose of use. The service is open for academic institution members and students (with supervisors). A prerequisite for registration is the signing of the Data Usage

[82] DIW Berlin: https://diw.de/en/diw_01.c.615551.en/research_infrastructure__socio-economic_panel__soep.html; [Accessed March 11, 2022b]
DIW Berlin: https://diw.de/en/diw_01.c.601584.en/data_access.html; [Accessed March 11, 2022a]

[83] FDZ IAB: https://fdz.iab.de/en.aspx; [Accessed March 11, 2022]

[84] FDZ-DZHW: https://www.fdz.dzhw.eu/en/data-usage; [Accessed March 6, 2022];
Rat Für Sozial- Und Wirtschaftsdaten (RatSWD) 2019

Agreement that defines the accessed data, time period, security actions and compliance with national and international legal requirements. If a group of researchers connects to the service, each researcher must sign the agreement. To connect to the remote desktop, it is sufficient to use an ordinary web browser. The connection is encrypted via VPN. Various standard programs for data analysis are available in a secured environment (e.g., Stata, SPSS, R, MAXQDA). After users request an export of results of data analysis, FDZ-DZHW checks the output for compliance with data protection. To introduce the service to new users, DZHW offers various video training sessions.

Key features: virtual desktop, a wide range of available softwares, access to DZHW data, detailed Data Usage Agreement, manual output check.

Virtual Data Enclave (VDE) – ICPSR

ICPSR provides both a secure physical facility and remote desktop access via its VDE[85]. The VDE is open to both students and non-academics. There is also capacity for data linkage, that is, for users to link their own data with data provided by ICPSR. All users must undergo training. All outputs are checked before publication is permitted, and sanctions exist for any violations. Checking procedures and sanctions vary by the specific data source, and sometimes by the provider institution.

Key features: high level of service, flexible. In order to do this, much customization is required, e.g., for output checking procedures.

5.3 Insights from Assessments

One strong pattern emerging from the assessments is that nowadays substantial legal frameworks are in place addressing the rights and responsibilities among data owners, data providers, related hosting institutions, and users. This includes the need to make legalities explicit and visible, e.g., by publishing them on websites, and having the organisation in place and specifying the different staff roles. It is perhaps obvious, but worth emphasising, that this advances more quickly within a single regulatory domain, usually one country. Another distinction is between "generalists" (e.g., UKDA Secure Lab) which offer data from diverse data owners or providers, and "specialists" where data from one or a more limited number of sources is offered, sometimes from a single project (e.g., RemoteNEPS). It is clear that the typical progression is to establish data enclaves, interconnect them, and also expand functionality by adding remote desktop, even if this offering is not possible for all data. Additional requirements that researchers are increasingly seeking are for linkage between restricted and non-restricted data, and for the ability to support group projects with multiple researchers accessing the same data. Where remote

[85] ICPSR: https://www.icpsr.umich.edu/VDE/; [Accessed March 10, 2022]

access is offered, training is mandatory, though it varies whether this can be taken online or must be in-person. Key bottlenecks are the time needed to establish all the necessary legal approvals. At the opposite end of the workflow, the main obstacle is time needed for output controls. This is still labour intensive, even when supported by emerging automated tools.

6. Recommendations

These recommendations are guided by our observations that most successful infrastructures embody two features: 1) they are human as well as technical, and 2) they are neither purely centralised nor decentralised, but well-crafted hybrids.

Infrastructures must, of course, provide essential "plumbing" - hardware, software, platforms, resources, and so on. However, without adequate human support (FTEs, skills, training, etc.), too often infrastructures are built but not adopted, embraced but not established, or started but not sustained. And by nature of their specialist niche, secure data professionals tend to be widely distributed, sometimes only a single individual within a large institution. Improving human networks is essential.

Similarly, for those more familiar with high-powered computing facilities, small-scale growth can appear messy, even chaotic. It can be tempting to believe that the solution is to select "best practice", centralise, standardise and roll-out. Belief is that a hybrid-model combining centralised and decentralised elements, is most likely to succeed in the SSH context. Some centralised elements are essential, especially regarding coordination and sustainability. However, flexibility is equally necessary to address not only diversity between social sciences and humanities, but also within each domain (e.g., FORS must address sensitive data from a wide range of disciplines when establishing its service).

Recommendations

1) Explore and assess if the idea of "Remote Access to Sensitive Data as a Service" as a Service Hub within EOSC is viable. At least two areas of work seem amenable to being supported in this manner.

- Better provisioning of legal opinions and services needed to establish remote access.

A robust finding of this task has been the complexity of finding legal solutions for accessing sensitive data. Whether for bilateral agreements to connect safe enclaves in two countries or scaling up within a single country, creating legal frameworks is laborious. Once solutions are developed, effective dissemination to others within the same legal system could expedite development for others. This work should build on the strong foundation established in Work Package 8 of SSHOC[86].

One central obstacle often involves the prohibition of certain data from leaving the country and consequently the obligation to leave them on national servers. This is a point on which it should be possible to change laws and practices, at least within the EU. Cross-boundary access has been negotiated, for example by IAB, CASD, and most recently between GESIS and UKDS, but a more generic legal solution would greatly ease the process for future data access.

- IT services hub and perhaps a "basic" certified provider.

The situation is not dissimilar regarding IT. Solutions are now available (e.g., CASD), but finding, assessing, and determining what works with existing local infrastructures remain major challenges. A central hub offering IT consulting could be a solution. Factors to consider include: certification for a designated provider and certification criteria for organisations choosing to implement customised solutions. There is also the matter of costs: where services are provided through data archives, they have typically been free to the user. However, the sustainability of this model needs to be reviewed. Where solutions can be found that work for both social sciences and humanities, these are especially worthy of support.

2) Expansion through Networking of Networks

- Recent years have seen substantial growth, albeit fitful and at times fragmented. Safe rooms are being added (FORS) and facilities that already had operational safe rooms are adding remote desktop access. From there, growth often takes the form of connecting hubs, such as with the established IDAN network, and the recently launched RDCnet, which is connecting data centres throughout Germany. The recommendation is to support this type of growth, and make it more efficient (e.g., templates for legal documents, IT advice, coordinating expert advice. In short, to stop reinventing the same wheel. Even if legal barriers prevent extensive cross-border access for the time-being, the vision is still worthy. Moreover, even when data itself cannot be made accessible, much know-how and development ideas can still be profitably exchanged.

- Networking the (human) community of secure data access professionals

  The demand for access to sensitive data, in ever more convenient modes, is growing and this growth will continue. A central challenge is to provide the community of experts committed to meeting this demand with resources and institutional support to deliver. Subtask 5.4 within the SSHOC project has successfully established just such an international community. While two national groups (e.g., in the UK and in Germany) are currently viable, a sustainable model should be found for this new international network. Also, the continuation of the collaboration started

[86] Hansen et al. 2022

with the SSHOC project offers the prospect of establishing a knowledge hub to keep providing expertise with respect to RSA set-up and operation.

7. Conclusion

In the health domain, it is now widely recognized that there is no “old normal” to which we can return. Whatever its other effects, the pandemic has reshaped the digital landscape, advancing capabilities at a pace previously thought impossible. While there is relief and celebration at renewed opportunities for face-to-face meetings, digital capacities are now if not a default, then an expected service. Researchers, who in their private capacity have been able to conduct many aspects of life, including sensitive domains of personal health and financing, online, will have even greater expectations for secure remote access to data, and not only via safe facilities, but from desktops in home offices and elsewhere.

This work should continue to take on these challenges. There is a need to consider the opportunities and initiatives related to technologies that allow packaging and bringing computation to the data. With the advent of containerization technologies e.g., Docker and Kubernites and software library approaches such as Jupyter Notebooks, it has become easier and more attractive to flexibly move (customer developed) special algorithmic (e.g., statistics) processing facilities close to the hosted data. The algorithms can be programmed by researchers and, after proper inspection, can be offered for deployment in a secure facility that hosts sensitive data. Approaches such as these have already for some time been tested with various types of data such as copyrighted newspaper articles. Challenges with the container approach are intrinsic security aspects of the technologies themselves and there is also awareness of dangers with respect to information from the hosting environment and other containers leaking away via malicious or badly programmed containers since its introduction[87]. But the use of containers for managing access to sensitive data adds new requirements for additionally checking the logic for data access and disclosure within a single container. Despite these challenges, there is considerable interest in these approaches since it can make the basic RSA approach far more flexible with respect to offered data analysis tools.

There are also challenges in addressing novel data types. To date, the type of data handled in SSH facilities has been very largely limited to structured, quantitative, tabular data. But interest is growing exponentially in other data types and formats, such as social media data and qualitative data. Often, a key hurdle with such data is the process of checking outputs for disclosure risk. Such risks are more complex and less quantifiable than with, for example, survey data. A working group, organised with support of the public body UK Research and Innovation, has convened several times, and produced a blog[88] exploring these challenges in more detail.

[87] De Benedictis and Lioy 2019

[88] Green et al. 2021

Finally, this work has demonstrated that whatever differences exist in the needs and practices between the social sciences and humanities, the optimal, indeed the only, successful way forward is continued collaboration. Given that sharing by download is the general practice in humanities, remote access is a relevant additional alternative to be considered as dissemination mode for sensitive data in the humanities. It can open up data in areas such as criminology and clinical language studies which are inaccessible at present. This is the vision: to make SSH data as open as possible, while addressing the practical requirements of providing protection when necessary.

8. References

Aircloak. n.d. "Peace of Mind – Immediate Insights." Aircloak. Accessed February 8, 2022. https://aircloak.com/.

Alter, George C., and Mary Vardigan. 2015. "Addressing Global Data Sharing Challenges." *Journal of Empirical Research on Human Research Ethics* 10 (3): 317–23. https://doi.org/10.1177/1556264615591561.

BioMedBridges. n.d. "Supporting Researchers Sharing Sensitive Data: Identifying Requirements." Accessed February 8, 2022. http://www.biomedbridges.eu/supporting-researchers-sharing-sensitive-data-identifying-requirements.

Bishop, Elizabeth. 2021. "Milestone 28 Assessment of Existing Platforms." Report. SSHOC. https://doi.org/10.5281/zenodo.5914390.

Boiten, Jan Willem, Christian Ohmann, Ayodeji Adeniran, Steve Canham, Monica Cano Abadia, Gauthier Chassang, Maria Luisa Chiusano, et al. 2021. "EOSC-LIFE WP4 TOOLBOX: Toolbox for Sharing of Sensitive Data - a Concept Description." Zenodo. https://doi.org/10.5281/zenodo.4483694.

Bond, Steve, Maurice Brandt, and Peter-Paul de Wolf. 2014. "Guidelines for the Checking of Output Based on Microdata Research." Data without Boundaries. Work Package 11. https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf.

Bouhari, Yacine El. 2018. "Secure Remote Access to Confidential Data in France," June. https://doi.org/10.5281/ZENODO.3775834.

Bryant, Mike, Elly Dijk, Rebecca Dillmeier, René van Horik, Michael Levy, and Rachel Pistol. 2021. "D12.1 EHRI-3 Data Management Plan." EHRI. WP12 JRA4 New Approaches to Holocaust Research and Archiving. https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/D12.1%20EHRI-3%20Data%20Management%20Plan.pdf.

CASD. 2022. "CASD User Guide." https://www.casd.eu/wp/wp-content/uploads/casd_user_guide-5.pdf.

———. n.d. “Secure Data Access Centre.” CASD. Accessed March 3, 2022. https://www.casd.eu/en/.

CASRAI. n.d. “Research Data Management Glossary.” CASRAI. Accessed January 27, 2022a. https://casrai.org/rdm-glossary/.

———. n.d. “Research Data (Term).” CASRAI. Accessed January 27, 2022b. https://casrai.org/term/research-data/.

CBD. 2015. “About the Nagoya Protocol | Convention on Biological Diversity.” Secretariat of the Convention on Biological Diversity. June 9, 2015. https://www.cbd.int/abs/about/.

CLARIN. n.d. “Federated Identity.” CLARIN ERIC. Accessed March 25, 2022a. https://www.clarin.eu/content/federated-identity.

———. n.d. “Licenses and CLARIN Categories.” CLARIN ERIC. Accessed February 8, 2022b. https://www.clarin.eu/content/licenses-and-clarin-categories.

Cornuau, Frédérique, Marie Cros, Silberman Roxane, Iris Dieterich, David Schiller, Maurice Brandt, Christopher Gürke, et al. 2013. “Feasibility study on the organizational architecture for managing pan European access, deliverable D4.2.” *Data without Boundaries*. https://doi.org/20.500.12210/1209.

CSC. n.d. “EPouta - Services for Research - CSC Company Site.” EPouta. Accessed March 6, 2022a. https://research.csc.fi/en/-/epouta.

———. n.d. “Sensitive Data Services for Research - Services for Research.” Sensitive Data Services. Accessed February 28, 2022b. https://research.csc.fi/sensitive-data-services-for-research.

Dalenius, Tore. 1977. “Towards a Methodology for Statistical Disclosure Control.” *Statistik Tidskrift* 15: 429–44.

DANS-KNAW. n.d. “DANS | Centre of Expertise & Repository for Research Data.” Data Archiving and Networked Services - Royal Netherlands Academy of Arts and Sciences. Accessed March 2, 2022. https://dans.knaw.nl/en/.

DARIAH. 2020. “SSHOC Workshop: Putting Data Protection into Practice: The GDPR and the DARIAH ELDAH Consent Form Wizard.” October 13, 2020. https://www.dariah.eu/2020/09/15/sshoc-workshop-putting-data-protection-into-practice-the-gdpr-and-the-dariah-eldah-consent-form-wizard/.

———. n.d. “DARIAH ELDAH Consent Form Wizard (CFW).” Accessed February 8, 2022a. https://consent.dariah.eu/node/2.

———. n.d. “Digital Research Infrastructure for the Arts and Humanities.” DARIAH. Accessed February 8, 2022b. https://www.dariah.eu/.

DataSHIELD. n.d. “DataSHIELD. A Software Solution for Secure Bioscience Collaboration.” DataSHIELD. Accessed February 8, 2022. https://www.datashield.org/.

DataSHIELD wiki. n.d. “DataSHIELD - Confluence.” DataSHIELD Wiki. Accessed March 6, 2022. https://data2knowledge.atlassian.net/wiki/spaces/DSDEV/overview?homepageId=12943453.

Dataverse. n.d. “Dataset + File Management.” Accessed February 8, 2022. https://guides.dataverse.org/en/latest/user/dataset-management.html#restricted-files.

De Benedictis, Marco, and Antonio Lioy. 2019. "Integrity Verification of Docker Containers for a Lightweight Cloud Environment." *Future Generation Computer Systems* 97 (August): 236–46. https://doi.org/10.1016/j.future.2019.02.026.

DIW Berlin. n.d. "Data Access. SOEP Research Data Center (RDC SOEP)." DIW Berlin. DIW Berlin. Accessed March 11, 2022a. https://diw.de/en/diw_01.c.601584.en/data_access.html.

———. n.d. "Research Infrastructure 'Socio-Economic Panel (SOEP).'" DIW Berlin. DIW Berlin. Accessed March 11, 2022b. https://diw.de/en/diw_01.c.615551.en/research_infrastructure__socio-economic_panel__soep.html.

Domingo-Ferrer, Josep. 2009. "Disclosure Risk." In *Encyclopedia of Database Systems*, edited by Ling Liu and M. Tamer Özsu, 848–49. Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-39940-9_1506.

Dutch government travel information for entrepreneurs. n.d. "Keeping and Sharing Medical Records." Business.Gov.Nl. Accessed February 18, 2022. https://business.gov.nl/regulation/medical-records/.

DWB. 2015. "Data Without Boundaries (DWB) Project." 2015. https://web.archive.org/web/20180402175416/http://www.dwbproject.org/.

ECRIN. n.d. "Facilitating European Clinical Research | ECRIN." Accessed February 8, 2022a. https://ecrin.org/.

———. n.d. "Proposed Categorisation System Published for Resources in the Toolbox for Sharing Sensitive Data | ECRIN." Accessed February 8, 2022b. https://ecrin.org/news/proposed-categorisation-system-published-resources-toolbox-sharing-sensitive-data.

EHRI. n.d. "European Holocaust Research Infrastructure." Text. EHRI. Accessed February 8, 2022. https://www.ehri-project.eu/.

EOSC. n.d. "EOSC Marketplace. Welcome to the EOSC Portal Catalogue and Marketplace." EOSC Marketplace. Accessed March 17, 2022a. https://marketplace.eosc-portal.eu/.

———. n.d. "EPouta Virtual Private Cloud." EOSC Marketplace. Accessed February 8, 2022b. https://marketplace.eosc-portal.eu/services/csc-epouta.

EOSC-hub. n.d. "Services for Sensitive Data." EOSC-Hub. Accessed March 17, 2022. https://www.eosc-hub.eu/services/Services%20for%20sensitive%20data.

EOSC-Life. 2021. "Demonstrator 7: Accessing Human Sensitive Data from Analytical Workflows Available to Everyone in EOSC-Life." EOSC Life. January 19, 2021. https://www.eosc-life.eu/d7/.

———. n.d. "European Open Science Cloud - Life." EOSC Life. Accessed February 8, 2022. https://www.eosc-life.eu/.

EU General Data Protection Regulation (GDPR). 2018a. "Art. 4 GDPR – Definitions." General Data Protection Regulation (GDPR). 2018. https://gdpr-info.eu/art-4-gdpr/.

———. 2018b. “Recital 52 - Exceptions to the Prohibition on Processing Special Categories of Personal Data.” General Data Protection Regulation (GDPR). 2018. https://gdpr-info.eu/recitals/no-52/.

European Commission. 2016. “Guidelines on FAIR Data Management in Horizon 2020.” Version 3.0. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

———. n.d. “What Personal Data Is Considered Sensitive?” Text. European Commission - European Commission. Accessed February 7, 2022. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_en.

Eurostat. 2021. “Remote Access to European Microdata. Report on Legal, Organisational, Methodological and Technical Aspects.” Preliminary access. Unit A-5: Methodology; Innovation in Official Statistics.

———. n.d. “Statistical Confidentiality and Personal Data Protection - Access to Microdata.” Eurostat. Accessed March 10, 2022. https://ec.europa.eu/eurostat/web/microdata/statistical-confidentiality-and-personal-data-protection.

FDZ IAB. n.d. “Das Forschungsdatenzentrum Der BA Im IAB (The Research Data Centre (FDZ) of the Federal Employment Agency at the Institute for Employment Research).” FDZ IAB. Accessed March 11, 2022. https://fdz.iab.de/en.aspx.

FDZ-DZHW. n.d. “Data Usage and Research.” FDZ-DZHW. Accessed March 6, 2022. https://www.fdz.dzhw.eu/en/data-usage.

FOSTER. n.d. “Open Science Definition | FOSTER.” FOSTER. Accessed March 10, 2022. https://www.fosteropenscience.eu/foster-taxonomy/open-science-definition.

FOSTER Consortium. 2018. “What Is Open Science?,” November. https://doi.org/10.5281/ZENODO.2629946.

Gaye, Amadou, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M. Jones, Joel Minion, et al. 2014. “DataSHIELD: Taking the Analysis to the Data, Not the Data to the Analysis.” *International Journal of Epidemiology* 43 (6): 1929–44. https://doi.org/10.1093/ije/dyu188.

GESIS. 2018. “Usage Regulations.” GESIS. https://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/_bgordnung_bestellen/2018-05-25_Usage_regulations_GESIS_DAS.pdf.

GESIS. n.d. “Secure Data Center.” GESIS. Accessed March 14, 2022. https://www.gesis.org/en/services/processing-and-analyzing-data/research-visits/secure-data-center-sdc.

GitHub. n.d. “Choose an Open Source License.” Choose a License. Accessed February 8, 2022. https://choosealicense.com/.

Granda, Peter, and Emily Blasczyk. 2016. “Cross-Cultural Survey Guidelines.” 2016. https://ccsg.isr.umich.edu/chapters/data-dissemination/.

Green, Elizabeth, Felix Ritchie, Libby Bishop, Deborah Wiltshire, Simon Parker, Allyson Flaster, and Maggie Levenstein. 2021. "What Are the Output Disclosure Control Issues Associated with Qualitative Data?" *DRAGoN: Data Research, Access and Governance Blog* (blog). November 23, 2021. https://blogs.uwe.ac.uk/dragon/what-are-the-output-disclosure-control-issues-associated-with-qualitative-data/.

Hansen, Mathilde Steinsvåg, Ina Nepstad, Marita Helleland, Håkon Jørgen Tranvåg, Trond Kvamme, Siri Tenden, Ingvild Eide Graff, et al. 2022. "*Deliverable 8.4 Report on Ethical and Legal Issues and Implications for EOSC*." WP8 - Governance/Sustainability/Quality Assurance. Task 8.3 Legal and Ethical Issues. Version 1. SSHOC.

Horton, Laurence, and Anja Perry. 2020. "Access Some Areas: Reforming Access Categories for Data in a Social Science Data Archive." *International Journal of Digital Curation* 15 (1): 5. https://doi.org/10.2218/ijdc.v15i1.708.

Horton, Laurence, Anja Perry, and Libby Bishop. 2020. "Open Where Possible, Closed If Necessary: Reforming Access Categories for Social Science Data Archives," February. https://doi.org/10.5281/ZENODO.3670943.

ICPSR. n.d. "Virtual Data Enclave (VDE) Technical Guide." ICPSR. Accessed March 10, 2022. https://www.icpsr.umich.edu/VDE/.

ISO 5127. 2017. "ISO 5127:2017(En), Information and Documentation — Foundation and Vocabulary." International Organization for Standardization. 2017. https://www.iso.org/obp/ui/fr/#iso:std:iso:5127:ed-2:v1:en.

ISO 22059. 2020. "ISO 22059:2020(En), Guidelines on Consumer Warranties/Guarantees." International Organization for Standardization. 2020. https://www.iso.org/obp/ui/#iso:std:iso:22059:ed-1:v1:en.

ISO 24622-1. 2015. "ISO 24622-1:2015(En), Language Resource Management — Component Metadata Infrastructure (CMDI) — Part 1: The Component Metadata Model." International Organization for Standardization. 2015. https://www.iso.org/obp/ui/#iso:std:iso:24622:-1:ed-1:v1:en.

ISO 27789. 2021. "ISO 27789:2021(En), Health Informatics — Audit Trails for Electronic Health Records." International Organization for Standardization. 2021. https://www.iso.org/obp/ui/#iso:std:iso:27789:ed-2:v1:en.

ISO/IEC 11179-7. 2019. "ISO/IEC 11179-7:2019(En), Information Technology — Metadata Registries (MDR) — Part 7: Metamodel for Data Set Registration." International Organization for Standardization. 2019. https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-7:ed-1:v1:en.

ISO/IEC 20546. 2019. "ISO/IEC 20546:2019(En), Information Technology — Big Data — Overview and Vocabulary." International Organization for Standardization. 2019. https://www.iso.org/obp/ui/#iso:std:iso-iec:20546:ed-1:v1:en.

ISO/IEC 20889. 2018. "ISO/IEC 20889:2018(En), Privacy Enhancing Data de-Identification Terminology and Classification of Techniques." International Organization for Standardization. 2018. https://www.iso.org/obp/ui/#iso:std:iso-iec:20889:ed-1:v1:en.

ISO/IEC TS 20748-4. 2019. "ISO/IEC TS 20748-4:2019(En), Information Technology for Learning, Education and Training — Learning Analytics Interoperability — Part 4: Privacy and Data Protection Policies." International Organization for Standardization. 2019. https://www.iso.org/obp/ui/#iso:std:iso-iec:ts:20748:-4:ed-1:v1:en:term:3.18.

ISO/IEC/IEEE 15939. 2017. "ISO/IEC/IEEE 15939:2017(En), Systems and Software Engineering — Measurement Process." International Organization for Standardization. 2017. https://www.iso.org/obp/ui/#iso:std:iso-iec-ieee:15939:ed-1:v1:en.

ISO/TR 18307. 2001. "ISO/TR 18307:2001(En), Health Informatics — Interoperability and Compatibility in Messaging and Communication Standards — Key Characteristics." International Organization for Standardization. 2001. https://www.iso.org/obp/ui/#iso:std:iso:tr:18307:ed-1:v1:en.

ISO/TR 21797. 2019. "ISO/TR 21797:2019(En), Reference Data for Financial Services — Overview of Identification of Financial Instruments." International Organization for Standardization. 2019. https://www.iso.org/obp/ui#iso:std:iso:tr:21797:ed-1:v1:en.

Kamocki, Pawel, Pavel Straňák, and Michal Sedlák. 2015. *Public License Selector* (version 0.0.6). ÚFAL. https://github.com/ufal/public-license-selector.

Kuchinke, Wolfgang, and EUDAT Sensitive Data Working Group. 2017. "How Can E-Infrastructures Deal with the Sensitive Data Challenge (Working Paper)." https://doi.org/10.23728/B2SHARE.3D1DFB9B889C4022AE7B308DF009FCC9.

Kvalheim, V, and M Myhren. 2016. "New Legislation – a Unique Opportunity for Harmonising the Legal Framework for Research in the Nordic Countries. NSD Position Paper." NSD - Norwegian Centre for Research Data.

Law Insider. n.d. "Research Community Definition." Law Insider. Accessed March 10, 2022. https://www.lawinsider.com/dictionary/research-community.

Luyten, Dirk, and Hans Boers. 2013. "D.3.2. Digital Handbook on Privacy and Access." EHRI. https://www.ehri-project.eu/sites/default/files/downloads/Deliverables/Deliverable%203%202%20original%20june.pdf.

Microdata. 2022. "User Guide for Microdata.No." Microdata. https://microdata.no/brukermanual-en.pdf.

———. n.d. "About Microdata.No." Microdata. Accessed March 12, 2022a. https://microdata.no/en/about/.

———. n.d. "FAQ." Microdata. Accessed March 12, 2022b. https://microdata.no/en/faq/.

MONA. n.d. "About MONA." Statistiska Centralbyrån. Accessed March 1, 2022a. http://www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/mona--statistics-swedens-platform-for-access-to-microdata/about-mona/.

———. n.d. "Terms of Use." Statistiska Centralbyrån. Accessed March 1, 2022b. http://www.scb.se/en/services/ordering-data-and-statistics/ordering-microdata/mona--statistics-swedens-platform-for-access-to-microdata/rules-and-regulations/terms-of-use/.

NeIC. n.d. “Tryggve - NeIC Web.” Nordic e-Infrastructure Collaboration. Accessed April 1, 2022. https://neic.no/tryggve/.

NEPS. n.d. “Data Use Agreements.” NEPS. NEPS. Accessed March 7, 2022a. https://www.neps-data.de/Data-Center/Data-Access/Data-Use-Agreements.

———. n.d. “RemoteNEPS.” NEPS. NEPS. Accessed March 7, 2022b. https://www.neps-data.de/Data-Center/Data-Access/RemoteNEPS.

Nordberg, Ana. 2021. “Biobank and Biomedical Research: Responsibilities of Controllers and Processors Under the EU General Data Protection Regulation.” In *GDPR and Biobanking: Individual Rights, Public Interest and Research Regulation across Europe*, edited by Santa Slokenberga, Olga Tzortzatou, and Jane Reichel, 61–89. Law, Governance and Technology Series. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-49388-2_5.

Office for National Statistics. n.d. “Accessing Secure Research Data as an Accredited Researcher.” Office for National Statistics. Accessed February 23, 2022. https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme.

OpenAIRE. n.d. “How to Deal with Sensitive Data.” OpenAIRE. Accessed February 8, 2022. https://www.openaire.eu/sensitive-data-guide.

Øvrelid, Egil, Bendik Bygstad, and Gard Thomassen. 2021. “TSD: A Research Platform for Sensitive Data.” *Procedia Computer Science*, CENTERIS 2020 - International Conference on ENTERprise Information Systems / ProjMAN 2020 - International Conference on Project MANagement / HCist 2020 - International Conference on Health and Social Care Information Systems and Technologies 2020, CENTERIS/ProjMAN/HCist 2020, 181 (January): 127–34. https://doi.org/10.1016/j.procs.2021.01.112.

RAIRD. 2019. “The RAIRD Project.” April 1, 2019. https://web.archive.org/web/20190401131357/http://raird.no/.

Rat Für Sozial- Und Wirtschaftsdaten (RatSWD). 2019. “Remote Access zu Daten der amtlichen Statistik und der Sozialversicherungsträger.” *RatSWD Output Paper Series*. https://doi.org/10.17620/02671.42.

Risnes, Ørnulf, Bjørn Roar Joneid, Eirik Alvær, Kenneth Schnelle, Ivar Refsdal, Vassilios Kalantzakos, Kjetil Thuen, et al. n.d. “FAIR Data Versioning in Microdata.No.” Microdata. Accessed March 12, 2022. https://microdata.no/FAIR_Data_Versioning_in_Microdata.pdf.

Ritchie, Felix. 2017. “The ‘Five Safes’: A Framework For Planning, Designing And Evaluating Data Access Solutions,” September. https://doi.org/10.5281/ZENODO.897821.

Schiller, David H., Johanna Eberle, Daniel Fuß, Jan Goebel, Jörg Heining, Tatjana Mika, Dana Müller, Frank Röder, Michael Stegmann, and Karsten Stephan. 2017. “Standards Des Sicheren Datenzugangs in Den Sozial- Und Wirtschaftswissenschaften - Überblick Über Verschiedene Remote-Access-Verfahren.” 261. *RatSWD Working Papers*. RatSWD Working Papers. German Data Forum (RatSWD). https://ideas.repec.org/p/rsw/rswwps/rswwps261.html.

Tóth-Czifra, Erzsébet. 2019. “DARIAH Pathfinder to Data Management Best Practices in the Humanities. Version 1.0.0.” *DARIAH-Campus*, May. https://campus.dariah.eu/id/yR8mHfs3eW-ibu58LerCt.

UKDS. n.d. “Apply to Access Controlled Data in SecureLab.” UK Data Service. Accessed February 23, 2022a. https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/.

———. n.d. “General SecureLab FAQs.” UK Data Service. Accessed February 23, 2022b. https://ukdataservice.ac.uk/help/secure-lab/securelab-faqs/.

———. n.d. “How to Download and Order Your Data.” UK Data Service. Accessed February 8, 2022c. https://ukdataservice.ac.uk/help/access-policy/how-to-download-and-order-your-data/.

———. n.d. “International Data Access Network (IDAN).” UK Data Service. Accessed March 7, 2022d. https://ukdataservice.ac.uk/about/research-and-development/international-data-access-network-idan/.

———. n.d. “Types of Data Access.” UK Data Service. Accessed February 21, 2022e. https://ukdataservice.ac.uk/help/access-policy/types-of-data-access/.

———. n.d. “Who Can Apply to Access SecureLab?” UK Data Service. Accessed February 23, 2022f. https://ukdataservice.ac.uk/help/secure-lab/am-i-eligible-to-apply-to-access-securelab/.

University of Oslo. n.d. “Services for Sensitive Data (TSD).” Accessed February 8, 2022. https://www.uio.no/english/services/it/research/sensitive-data/index.html.

Wheeler, Andrew P. 2020. “Open Source Criminology Related Network Datasets.” Andrew Wheeler. October 25, 2020. https://andrewpwheeler.com/2020/10/25/open-source-criminology-related-network-datasets/.

Wiki. n.d. “Object Access Types - ECRIN-MDR Wiki.” Accessed February 8, 2022. http://ecrin-mdr.online/index.php/Object_access_types.

Wikipedia. 2021a. “Chinese Wall.” In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Chinese_wall&oldid=1062729781.

———. 2021b. “Data Dissemination.” In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Data_dissemination&oldid=1062901043.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18.

Williams, Matthew L, and Pete Burnap. n.d. “Using Twitter for Criminology Research.” British Sociological Association. Ethics Case Study | 2. Accessed February 2, 2022. https://www.britsoc.co.uk/media/24899/using-twitter-for-criminology-research.pdf.

Wittenberg, Marion, and Péter Király. 2020. “Requirements of DARIAH Community for a Dataverse Repository.” September 28. https://doi.org/10.5281/zenodo.4054847.

Woollard, Matthew, Beate Lichtwardt, Elizabeth Lea Bishop, and Dana Müller. 2021. "D5.9 Framework and Contract for International Data Use Agreements on Remote Access to Confidential Data," January. https://doi.org/10.5281/zenodo.4534286.

List of Figures

List of Tables

Annex 1. Glossary

This glossary defines the main terms used in the paper.

Term	**Definition**
Access level	level of authority required from an entity to access a protected resource (ISO/IEC 11179-7 2019, 3.1.1).
Access policy	definition of the obligations for authorizing access to a resource (ISO 27789 2021, 3.2).
Anonymized data	personal data modified in such a way that direct reference to data subjects is eliminated (ISO 5127 2017, 3.1.10.15).
Archival system	organized collection of hardware, software, policies, procedures, and people, which maintains, stores, manages and makes available records over time (ISO 5127 2017, 3.1.13.35).
Authentication	the process of confirming the identity of a principal entity (CASRAI [Accessed January 27, 2022b]).
Classified data	data to which access is restricted by administrative means varying according to the degree of data protection or information protection sought (ISO 5127 2017, 3.1.10.17).
Click-through license, click-wrap license	a legal agreement to which one indicates acceptance by clicking on a button or link.
Closed access	access to information, documents or information services limited by general or specific regulations (ISO 5127 2017, 3.11.1.07).
Confidential data	data to which only a limited number of persons have access, and which are meant for restricted use (ISO 5127 2017, 3.1.10.18); confidential data may lead to the identification of a specific unit of observation (individual, economic entity) (DWB 2015).
Data access condition	conditions to be met before access can be granted.
Data accessibility	degree to which users can access data.

Term	Definition
Data archive	an archival service providing the long-term permanent care and accessibility for digital objects with research value (CASRAI [Accessed January 27, 2022b]).
Data archiving	digital preservation process that is moving data into a managed form of storage for long-term retention (ISO 5127 2017, 3.1.11.19).
Data dissemination	the transmitting or distribution of statistical data, or other, to end users (Wikipedia 2021b)
Data download	transfer of data from a remote facility to a local computer through a computer network.
Data enclave, safe enclave, safe rooms, secure labs, safe haven, safe pods, Physical Data Enclave (PDE), Trusted Research Environments (TRE)	special rooms that offer researchers the opportunity to work on site at the premises designated by the data producers. Typically, before any results of statistical analysis are physically released to the researchers for further use, they are checked by the data producers or a delegated service to ensure that statistical confidentiality is maintained, and data corruption avoided (output control).
Data provider	individual or organization that is a source of data (ISO/IEC/IEEE 15939 2017, 3.5).
Dataset	identifiable collection of data available for access or download in one or more formats (ISO/IEC 20546 2019, 3.1.11).
De-identified data	data resulting from personally identifiable information after the process of removing or altering one or more attributes so that the (direct or indirect) identification of the relevant person without knowledge of the initial information is either impossible or requires an unreasonable amount of time and manpower (ISO/TR 18307 2001, 3.56).
Digital content	data which are produced and supplied to a consumer in digital form (ISO 22059 2020, 3.4).
Digital repository	facility that provides reliable access to managed digital resources (ISO 24622-1 2015, 2.27).
Direct access	data are copied (can include licence, and conditions) from a repository to the user’s computer.

Term	Definition
Disclosure risk	as the risk that a user / intruder can use the protected dataset to derive confidential information on an individual among those in the original dataset (Domingo-Ferrer 2009).
Data dissemination mode	the manner of data dissemination.
Downloadable data	data that is offered for download by a data provider.
Embargo(ed) data	data not to be made available before or after a specific date.
FAIR data	Findable Accessible Interoperable Reusable data.
FAIRness of data	the degree towards data conforms to the FAIR principles.
Free access	open access that does not require financial compensation (ISO 5127 2017, 3.11.106).
General Data Protection Regulation (GDPR)	European Union legislation for data protection.
IPR protected data	data which access and usage is regulated by Intellectual Property Rights (IPR) laws.
License to use	a licence to use an intellectual property.
Microdata	dataset comprises records related to individual data principals (ISO/IEC 20889 2018, 3.23).
Nordic model	in the context of this document the "Nordic model" indicates concentration of legislation to foster improved data access.
Open access	unrestricted access to information, documents, or information services (ISO 5127 2017, 3.11.1.05).
Open data	data available without restrictions from copyright, patents or other mechanisms of control or costs, regardless of access, or use (ISO/TR 21797 2019, 3.5); data available/visible to others and that can be freely used, re-used, re-published and redistributed by anyone (ISO 5127 2017, 3.1.10.13).

Term	Definition
Open Science	the practice of science in such a way that others can collaborate and contribute, where research data, lab notes and other research processes are freely available, under terms that enable reuse, redistribution and reproduction of the research and its underlying data and methods (FOSTER [Accessed March 10, 2022]).
Output checking	the process where research results (tables, models, estimations etc.) that researchers have created and want to take out of the controlled environment of the RDC are checked for possible disclosure. Only when found non-disclosive, they are then sent to the researchers (Bond, Brandt, and de Wolf 2014, p. 36).
Personal data	any information relating to an identifiable natural person (data subject) (EU General Data Protection Regulation (GDPR) 2018a)
Personal data, personally identifiable information	data relating to an identified or identifiable individual (ISO 5127 2017, 3.1.10.14).
Remote access (RA), remote secure access (RSA)	use of an electronic resource stored on a server through a computer network (ISO 5127 2017, 3.11.1.04); any access mechanism securing that dataset cannot be copied, but remain on the secure servers of the data provider and can only be in location A and are accessed through a secure internet connection from specific other location(s) B. No physical transfer of the sensitive data ever occurs (adopted from Woollard et al. 2021).
Remote desktop, virtual desktop	data is stored and processed exclusively on the servers of the data-retaining organization. The user interface is transferred to the researchers' screen via a secure connection.
Remote execution, remote data processing	execution of (data analysis) software in a remote facility.
Research community	the community of researchers that work for, and provide research to, academia and not for profit organisations (Law Insider [Accessed March 10, 2022]).
Research data	data collected, observed, or created, for purposes of data analysis to produce original research information and results (ISO 5127 2017, 3.1.10.10); data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are

Term	Definition
	used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results (CASRAI [Accessed January 27, 2022a]).
Research Data Centre	a common term for the part of an organisation that is responsible for providing access to their microdata for research purposes (Bond, Brandt, and de Wolf 2014, p. 36).
Researcher's desktop	local desktop computer a researcher uses.
Secure location	location that is properly secured against unwanted physical access.
Sensitive data	any type of research data that cannot be openly shared and hence needs special safeguards in terms of access restrictions to prevent its unregulated distribution.
Statistical confidentiality	means that data on individuals (or businesses) may be used only for statistical purposes and that rules and measures must be applied to prevent the disclosure of information on an individual or business entity (Eurostat [Accessed March 10, 2022]).
Workstation	computer configured for a specific task.

Annex 2. Access Levels and Dissemination Practises at CLARIN, UKDA and GESIS

Downloadable data

Restricted data in the terminology of CLARIN ERIC[89] are data with the label RES[90]. Here is the relevant information on this website:

> **RES language resources** have additional restrictions which require permission from the rights holder. These resources may contain material whose usage is limited due to copyright and/or personal data protection issues. In practice, these language resources require both using federated login[91] to authenticate the end-user and sending a separate application to the rights holder for authorization, possibly including a research plan with the resource.
>
> Examples of RES licences (and their CLARIN category labels)
>
> **Categories recommended by CLARIN with approved licences**

CLARIN RES+BY+NORED	CLARIN RES
CLARIN RES+BY+NC+NORED	CLARIN RES-NC
CLARIN RES+FF+BY+LRT+NORED	META-SHARE Commercial No Redistribution For a Fee
CLARIN RES+FF+BY+NC+LRT+NORED	META-SHARE Noncommercial No Redistribution For a Fee
CLARIN RES+FF+BY+LOC+LRT+NORED+*	ELRA licences

Comparable to this category at UKDA is data access marked as Safeguarded with Special Licence[92]. Also, for this data the user needs a login and to sign a licence agreement. According to a procedure that is similar to that of CLARIN ERIC:

[89] which classification is a generalisation of the individual resource centres constituting the CLARIN infrastructure

[90] See CLARIN: https://www.clarin.eu/content/licenses-and-clarin-categories; [Accessed February 8, 2022b]

[91] See CLARIN: https://www.clarin.eu/content/federated-identity; [Accessed March 25, 2022a]

[92] See UKDS: https://ukdataservice.ac.uk/help/access-policy/how-to-download-and-order-your-data/; [Accessed February 8, 2022c]

> **To request access to Special Licence data**:
>
> 1. Add the required dataset(s) to a project via your account.
> 2. The status will be shown as 'Request access'. Click this button to view the steps required to gain access to this dataset.
> 3. Click each of the steps in turn and follow the instructions displayed, e.g., accept any special conditions, or download the forms that are required.
> 4. When you have downloaded and completed the Special Licence form it must be returned to the UK Data Service Helpdesk. The form will then be checked, and we will inform you of any changes that are required. Once complete the form will be sent to the data owner for approval.
> 5. The steps displayed will be marked with a tick once complete.
> 6. Once your application is approved you will be notified that the data are available for you to download, and the dataset status will be 'Active'.
> 7. Follow the 'Action' button to download the data.

At GESIS this level of access is termed Accountable. Usage regulations[93] describe the current access classes.

Remote Access or Remote Secure Access

This level of access classification is not provided in CLARIN ERIC access level terminology (but needs to be implemented).

At UKDA this type of sensitive data is termed Controlled and detailed as Secure Remote Access[94].

> The UK Data Service also provides access to data that are deemed too confidential or sensitive to be released via download (such as through the End User Licence or Special Licence routes). These data have not been subject to any suppression, perturbation or other anonymisation techniques.
>
> We provide access via our Secure Lab to a variety of business microdata from the Office for National Statistics (ONS), and social survey data from a range of suppliers. These controlled data collections typically contain the detailed geographies of respondents' locations (including postcodes and grid references), and include variables deemed too sensitive for release. In the future, we will also provide researchers with secure access to administrative and transaction data sources.
>
> Access to these data via the UK Data Service Secure Lab offers Secure Remote Access and Safe Centre facilities. These enable researchers to access and undertake analyses, without the need for downloading the data. This reassures data providers that the data remain confidential but provides researchers with access to a rich source of detailed data.

and access is organised as one of the options in the Secure Lab facility[95]:

[93] GESIS 2018

[94] See UKDS: https://ukdataservice.ac.uk/help/access-policy/types-of-data-access/; [Accessed February 21, 2022e]

[95] See UKDS: https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/; [Accessed February 23, 2022a]

Who can use the Secure Lab?[96]

Access is available to researchers who can fulfil the access requirements for the data required:

- for ONS data: researchers must be an ONS Accredited Researcher, and the forms required to apply for this status will be made available during the application process. Further details are available from: ONS Approved Researcher Scheme[97],
- for non-government data: researchers must be based at a UK academic institution or an ESRC-funded research centre and be an ESRC Accredited Researcher. PhD and research students can request access but will need to apply jointly with their supervisors,
- the satisfactory completion of training,
- an agreement by the user and their institution to a User Agreement.

On Site Access

On site access means that the data can only be accessed within the physical premises of the data provider. The user has to physically enter a room that is under the control of the data provider where access to the data is provided via a terminal with a secure login facility.

This level of access is not provided as an access level classification in CLARIN ERIC

At UKDA this type of sensitive data is termed Controlled, and detailed as Safe Centre Facility[98], and access is organised as one of the options in the Secure Lab facility[99].

GESIS:

GESIS provides on-site data access via its Secure Data Centre (SDC)[100]. Here data can be analysed by appointment and on signing a contract for on-site use at a Safe Room guest workstation in Cologne.

UKDS: https://ukdataservice.ac.uk/help/secure-lab/securelab-faqs/; [Accessed February 23, 2022b]
UKDS: https://ukdataservice.ac.uk/help/secure-lab/am-i-eligible-to-apply-to-access-securelab/; [Accessed February 23, 2022f]

[96] See UKDS: https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/; [Accessed February 23, 2022a]

[97] Office for National Statistics: https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme; [Accessed February 23, 2022]

[98] See UKDS: https://ukdataservice.ac.uk/help/access-policy/types-of-data-access/; [Accessed February 21, 2022e]

[99] See UKDS: https://ukdataservice.ac.uk/find-data/access-conditions/secure-application-requirements/; [Accessed February 23, 2022a]

[100] GESIS: https://www.gesis.org/en/services/processing-and-analyzing-data/research-visits/secure-data-center-sdc; [Accessed March 14, 2022]

Annex 3. Use Case from FORS

Introduction

Many research infrastructures in Europe that maintain data archives are faced with a similar problem - they are expected by funders and universities to preserve and disseminate research data of all kinds but are not adequately equipped to make accessible *sensitive data* for secondary use. In some cases, research data are turned away by an archive because they cannot be fully anonymised, where the risks of harm to respondents from disclosure and identification are too high. In other cases, researchers do not even try to deposit their data, believing that the archive does not have the capacity to safely disseminate their sensitive data. Thus, despite data sharing requirements from research funders and institutions, many such datasets never see the light of day after project completion.

This has been the status quo for digital data archiving in Europe, at least for the social sciences, over the past decades. While initiatives and technical advances have been made in recent years to render sensitive data accessible under highly restricted conditions (e.g., at the UKDA and GESIS), most European data archives lag behind and have not yet put into place systems that would make it possible to share sensitive data. This includes FORS, the Swiss Centre of Expertise in the Social Sciences, national service provider for Switzerland within CESSDA.

Like many other data archives for the social sciences, even those that have been in existence for decades, FORS has yet to address the challenge of archiving and disseminating sensitive data. Its policy has been from its inception in 2008 (and before that with its predecessor SIDOS, beginning in 1993) that only anonymised data can be deposited and distributed on its online platform. Its current platform SWISSUbase allows for a variety of access restrictions according to levels of disclosure risk, but still does not allow for the storage and dissemination of sensitive data.

In recent years, however, it has become evident that the status quo is inadequate for serving its user community, notably due to strengthened data sharing policies from the main research funder in the country - the Swiss National Science Foundation (SNSF). The SNSF now requires that all data from funded projects be made available to the research community after project completion. Since most existing archival options, including at FORS, do not allow for the dissemination of sensitive data, this means that many researchers in Switzerland cannot fulfil the demands of the funder. It also means that FORS cannot fulfil its mandate to facilitate access to all kinds of research data.

In 2021, FORS therefore embarked internally on an exploration of possible solutions that would make it possible to archive and disseminate sensitive data. This reflection was encouraged and inspired as well by the participation of FORS in task 5.4 of the SSHOC project, which aimed to advance the cause of making sensitive data more accessible in Europe by way of remote access systems. The goals of this exploration were twofold. First, intention was to assess what forms such a solution might take, and the corresponding implications from strategic, technical, archival, and legal perspectives. Several criteria guided this work, including the following general questions:

How might sensitive data be archived and disseminated in relation to our current infrastructure and services?

- How might sensitive data be archived and disseminated in line with relevant data protection laws?
- How would such a system address our institutional strategy, direction, and objectives?
- What would be the resources needed to design and implement such a system? And what are the available resources (human and financial) that could be devoted to this work in the coming years?
- What capacities and skills would be needed in-house to run such a system?
- What would be the demand for such data in the social science community?
- What specific models for access to sensitive data correspond best to needs and would best serve our goals?
- How should such a system be put into place, all at once, or incrementally?
- How might such a system be linked in the long run to a larger European network of remote access systems?

Second, it was recognised among SSHOC task 5.4 partners that the internal reflections at FORS might make an important contribution to the field, since there are many institutions in similar situations that have little guidance on how to proceed "from scratch" towards a solution for access to sensitive data. The idea was thus born to employ the experience at FORS as a *case study* that might help guide other institutions embarking on the same path.

The case study presented in what follows describes in detail the internal assessment of FORS regarding the possible adoption of a remote access solution for sensitive data, including key considerations and decision points from strategic, technical, archival, and legal perspectives. Following the description of our deliberations and final decision, practical recommendations are provided regarding how to move systematically in the direction of selecting and implementing an appropriate solution for dissemination of sensitive data. It is hoped that this work will assist other institutions that recognise the need to implement such a solution, but that are not sure how to move in this direction.

Approach

The first step was to fix a framework in which the study could be conducted, including primary objectives and relevant parameters. This was established within the management of FORS, where the exploratory work and goals were approved. This included the rights to time of the head of the FORS data service unit, as well as one member of the IT staff, both participants in SSHOC task 5.4. These two individuals were responsible for organising and carrying out the work. The former is both a member of the FORS management and head of unit, and thus capable of coordinating the project and linking it to the relevant actors within the institution. It was indispensable also to have an IT person on the project team, given the highly technical nature of the work. The study was also granted rights to several hours of time of other relevant staff within FORS, in case of interviews. The head of IT at FORS was also kept informed about the project and its developments.

The overall objective was to determine whether FORS should move concretely in the direction of a remote access system, and if so, how. It was agreed that the work would potentially involve four phases: 1) a needs assessment; 2) evaluation of existing solutions in relation to identified needs; 3) selection of an existing solution, and 4) implementation of the solution. It was not known in advance whether FORS would in fact continue to aim for implementing a solution after the first two phases. That is, it was also a possibility that the idea would be abandoned after concluding that it was not feasible or desirable to put into place a system for access to sensitive data. There was no certainty on the likely time frame - implementation might take more or less time, depending on the nature of the solution and available resources.

A process for carrying out the study was determined by the two project members. First, an IT approach was adopted, with three-week "sprints" that laid out the work for the period, along with responsibilities. During each sprint, project meetings were held once per week. Each meeting involved review of current status in relation to the sprint goals and a larger project planning that was established in the beginning. It also focused on whatever had been assigned after the previous meeting. Last, assignments were agreed upon for the next week. Meeting discussions were documented in notes in a Google doc, including action items. A Slack channel was also used to facilitate communication between the two project members.

Needs assessment

Definitions, scope, and method. The goal of our needs assessment was to identify what is currently lacking technically and, in our services, to assess the importance of what is lacking, and to clarify what would be needed in order to fill the gap. The intention was also to address how such a service and tool for sensitive data might look like in the future at FORS. It was decided to conduct interviews with particular FORS staff members from different perspectives - strategic, technical, archival, and legal, since each of these would have an impact on the shape of any future solution. For this, we would develop four separate interview protocols. Beyond the protocols and the specific questions that would be posed, we thought that it would be useful to bring the discussions to a more concrete level by presenting a brief "concept" for how a remote access system might look at FORS (see below for details).

Towards these ends, some initial questions were posed, and tentative answers are listed:

- How will we define "sensitive data"?
    - We can define sensitive data as any data that cannot be handled by our current system.
- What other definitions are needed to continue?
    - Safe room
    - Remote desktop access
- What is the scope of our solution? (e.g., for what types of data? for what geographical range? for what disciplines?)
    - At least for the social sciences (still need to define social sciences)
    - Other disciplines related to SWISSUbase (e.g., humanities, linguistics, health, and medical sciences)
    - Need to research the question of what types of data can be published
- Who would the solution serve?

- At first this should be for Swiss users, but later maybe can be accessed by people outside of Switzerland (as users, but also as data producers)

At a subsequent meeting, these questions of definition and scope were discussed and agreed. It was decided to define sensitive data for our purposes as follows: "Any data that cannot be handled by our current system. *This could be because of disclosure risk or copyright issues, or because the data producers demand restrictive conditions not available in the current system.*" We also agreed on definitions for "remote desktop access" (or "virtual data room") and "safe room" (or "physical data room") that would be employed for the interviews, aligning these with the definitions adopted in SSHOC task 5.4. In addition, the strengths, and weaknesses of these two main types were documented and summarised (see exhibit 1).

Exhibit 1

- Data room vs remote access: Pros and cons
  - For remote access there is the benefit of flexibility for the researchers, and this widens access for those who might not be able to travel. Also, in most cases there is less work for the data service staff once the project is set up. However, other than a signed contract, it is difficult to prevent the user from taking a picture of the screen, be it with the print screen feature of his machine or with a phone camera.
  - For the physical data room, you can ask the user to leave his/her phone at the door and a print screen will be of no use, since there can be no Internet connection to allow printing. However, a physical data room means that the user must travel to the data room (depending on the country, this can be highly inconvenient) and he/she can only access it on certain hours (working hours).
- Technical remarks
  - In both cases, physical and virtual data rooms, a virtual machine is used to prevent Internet access or usage of USB ports (since it is much easier to restrict access on a virtual machine than on the physical machine).
  - It should be possible to remove the login access to a user at any time, should the contract between the user and provider end.
  - As a more general remark, it can be possible for the user to ask for his analysis output (to have it sent to him/her). The produced data should then be evaluated by the provider to make sure that secure data do not leave the room.

The meetings of the project team involved reflections on what might be best from an institutional point of view. For example, a discussion was documented in the notes that had to do with the issue of relying on "trust" and how a solution might be integrated into our archival platform SWISSUbase:

> "Additional notes: A virtual room will offer advantages to both users and DSUs in SUB, but it will also introduce risks, in comparison with a physical safe room. In this zone with increased risks, we must rely on trust to some extent. The risks in the "trust-zone" can be mitigated with some measures, like user authentication, contracts, training, formal requests, and justification of need for the data, and documentation and logs of users who have used the system. In our interviews, we need to see if the mitigated risks of the virtual room would be acceptable from different perspectives - archival and legal, but also from the point of view of data producers (we might need to interview a few data producers for this...).

> We spoke in today's meeting about how a virtual room or safe room could be integrated into SWISSUbase. We agreed that sensitive datasets should appear in the SWISSUbase catalogue, but where interested users would be redirected to a virtual room or physical safe room to access the data. This would be ideal, since data could be discovered in the catalogue, even if access is through a different system. The technical implications of this are to be explored in the interviews."

Development of a protocol for interviews. Along with establishing definitions and the scope of a possible system for accessing sensitive data at FORS, we posed some additional questions that would be key to assessing needs internally, and that would feed into the development of the four interview protocols for staff:

- What kinds of data cannot be archived currently because of sensitivity?
- Could we archive these data if sufficient safeguards were put into place?
- What would these safeguards need to be? (Technical? legal?)
- What would be the conditions for archiving such data (for example, regarding informed consent)?
- How often do we have to turn away researchers from archiving because of sensitivity issues? And how many researchers themselves do not even try to archive because of sensitivity issues?
- What does the law say about preserving and making available sensitive data?
- If sensitive data could be archived, under what conditions could these be accessed? (Safeguards for access and proper use)
- How willing are users to go to a safe room in Switzerland? Would they prefer a remote desktop solution? Why?
- To what extent should the tool be "integrated" into SWISSUbase? What are the different possibilities regarding integration?
- What are the technical implications regarding the storage and preservation of sensitive data?
- What are the legal implications regarding the storage and preservation of sensitive data?
- Would remote desktop access be feasible in the Swiss legal framework?
- Would a safe room be feasible?
- How might we use such a solution to strengthen collaboration with other data stakeholders (e.g., OFS, journals, SNSF)? (e.g., link hub)
- How would such a solution address requirement of funders and journals?

Based on these initial questions, we elaborated separate interview protocols for each of the four perspectives, presented in exhibit 2. Each was prefaced with a description of the SSHOC project, the purpose of the case study and interview, as well as relevant definitions.

Further, we crafted a "concept" piece for how a system for access to sensitive data might function in practice at FORS. This was developed by the project team, who benefited from work done in SSHOC task 5.4, where various solutions were identified, studied, and compared. The concept aimed to stimulate and bring to a more concrete level the discussions in the interviews. It presupposed a typical desktop remote access solution, although no decision had yet been taken regarding the form of any future system. It is presented in exhibit 3.

Exhibit 2	
Strategic	● From a strategic point of view, which would be preferable, a physical data room or a virtual data room? ● How might we use such a solution to strengthen collaboration with other data stakeholders (e.g., OFS, journals, SNSF)? (e.g., link hub) ● How would such a solution address requirements of funders and journals? ● Is there an institutional interest in joining a European remote access network for sensitive data?
Archive	● What kinds of data cannot be archived in FORSbase currently because of sensitivity issues? Why can't these be archived? ● Could we archive these data if sufficient safeguards were put into place? ● What would these safeguards need to be? (Technical? legal?) ● What would be the conditions for archiving such data? For example, regarding informed consent. ● How often do we have to turn away researchers from archiving because of sensitivity issues? ● If you had to guess, how many researchers themselves do not even try to archive because of sensitivity issues? ● In your view, how important would it be to have a system in SWISSUbase that allowed for the acquisition and dissemination of sensitive data? ● Do you think that this is needed, and why? ● Are there any other things you would like to say regarding the acquisition of sensitive data? ● If sensitive data could be archived at FORS, under what conditions could these be accessed? (Safeguards for access and proper use) ● In your opinion, how willing are users to go to a safe room in Switzerland? ● Would users prefer a virtual data room solution compared to a physical safe room? Why? ● From the archive perspective, which would be the preferred solution for access - a safe room or virtual room solution? ● Are there some data that are so sensitive that a virtual data room solution would not be appropriate? ● From an archival point of view, should the sensitive data be curated and preserved in the same way as the non-sensitive data in SWISSUbase? ● For how long should we preserve the sensitive data? Forever? ● Should sensitive data be reviewed from time to time to see whether they can be reclassified as "normal" data (i.e., non-sensitive data)? ● Should it be possible for linked documentation files to be publicly accessed for sensitive data? Should the virtual room make these files available along with the data? ● What level of resources would be needed to run such an operation in the archive, in general?
Technical	● To what extent should the remote desktop tool be "integrated" into SUB? What are the different possibilities regarding integration? ● Related to integration, would it make sense to allow for datasets in SUB where access to restricted data was limited to a special procedure (e.g., where permission was required)? ● Which is preferable, a system developed in-house or a commercial solution "off-the-shelf"? ● What are the technical implications regarding the storage and preservation of sensitive data? ● Would a SWISSUbase safe room be feasible at FORS from a technical point of view?

	● Are there technical conditions (at FORS) with respect to the choice of a remote access system?
Data protection	● What does the law say about preserving and making available sensitive data? ● Could FORS hold and disseminate sensitive data under certain conditions? What would the conditions need to be? ● What are the legal implications regarding the storage and preservation of sensitive data? ● Would remote desktop access be feasible in the Swiss legal framework? ● Is there anything else we should know from a legal point of view in thinking about these issues?

Exhibit 3: *Concept for a SWISSUbase virtual data room*

This solution would somehow be integrated into the larger SWISSUbase system (details still to be determined). It would be available to researchers working in Switzerland in the social sciences, and possibly other disciplines that work with sensitive data (e.g., humanities, linguistics, health sciences, etc.). It would have the following features:

General: Researchers have access to the requested data in a virtual environment - they have access to the data and tools on the virtual machine from their own work or personal computer, but there is no way to download or copy the data or the results of the analyses.

Storage: All data, secured or "normal" in SWISSUbase, will be stored on Swiss servers.

Catalogue: Datasets with secure data should be discoverable in the SWISSUbase catalogue. They should have an info saying to the user that the data are secure and that they should apply to consult those data.

Data requests: There will be a request system, where users can apply to consult secure data. Only verified/authenticated users should be able to send such requests. (A verified user should be connected to SUB with an institutional email address and should have been verified by a data-curator). On this request form, the user should know what his rights and obligations will be (e.g., no taking pictures of the data). He will also be informed, but not in too much detail for security reasons, on how he will access the data.

A data-curator should examine a request to access secure data. In the request form, the user should explain his motivations and why the data are needed.

Contract: If the request is approved, the user will have to sign a contract where he will be asked not to copy the data in any shape or form (among other things) and when he will be able to access the virtual room. Here, he will have a list of the tools available in the virtual environment.

Access: The day of the appointment, the user will receive credentials to log in to the virtual environment. These credentials can be removed from the user at any time. Then the user can connect to the virtual environment (a VM on a server) and do his analyses.

Outputs: If the user needs the results of his analysis sent back to him, he can ask for it. But everything that will leave the virtual room should be examined by a data-curator (to prevent leak of secure data) and by a technical person (to prevent leak of information about the system that can help to pirate it).

Logs and information management:

The user credentials will be destroyed after he accessed the virtual environment. If the user wants to access any other secure data, he will have to fill in another request and then new credentials will be sent to him.

The project team then considered how to prepare and administer the various interviews. First, we selected the appropriate individuals to speak with. For the "strategy" interviews, we chose two members of the FORS management, one of whom is head of the unit that is responsible for our archival infrastructure SWISSUbase, where a system of access to sensitive data would likely be located. Also included was the head of Data Tools and Services, who manages SWISSUbase. For the technical interview we chose a member of IT who is familiar with issues concerning remote access, having participated in past CESSDA projects related to the development of a European remote access network (ERAN). For the archival perspective, we choose the head of our data archive for the social sciences, as well as an archivist who has been employed at FORS since its inception in 2008. Last, we selected for the "legal'' interview an in-house expert on Swiss data protection law.

We decided to speak informally with each of these staff members to inform them about the project and ask them about their willingness to participate in an interview. After securing their agreement to participate, we sent them in advance our specific questions, the concept text, and some brief instructions on how the interviews would be conducted (e.g., orally, online). In addition, we supplied all interview participants with the following definitions:

> *Sensitive data*: any type of research data that cannot be openly shared and hence needs special safeguards in terms of access restrictions to prevent its unregulated distribution. The sensitivity of the data may be caused by the personal information contained in the data, the nature of its content (e.g., business related data) or the intellectual property rights associated with it.
>
> *Physical safe room*: This offers researchers the opportunity to work onsite at the premises designated by the data producers. Typically, before any results of statistical analysis are physically released to the researchers for further use, they are checked by the data producers or a delegated service to ensure that statistical confidentiality is maintained, and data corruption avoided (output control). Safe rooms can have access facilities that can connect users to sensitive data located at different physical locations or organisations.
>
> *Virtual room (remote desktop access)*: A virtual data room is a cloud solution for securing and sharing confidential information. With the remote desktop procedure, data are stored and processed exclusively on the servers of the data-retaining organisation. The user interface is transferred to the researchers' screen via a secure connection (virtual desktop). The researchers' access device is only used to communicate with the server. The applications and data are physically located exclusively on the server of the data-retaining organisation, whereby viewing and browsing the results and data is possible within a familiar desktop environment.

Interview findings

Strategy. The FORS director expressed openness regarding the desirability of either a physical data room or a virtual one. However, he believed physical data rooms would work well in the context of SWISSUbase, could be federated within Switzerland, and might be ideal for giving access to sensitive administrative data and linked data. In any event, from his perspective it would be important to continue the conceptual work and to begin lightly, on a small scale, without an intensive investment of resources (which are tight at FORS). In this view, he thought that a physical data room might be easier at the beginning. Once a decision has been made, FORS should be pragmatic and cautious as it proceeds, with an eye on how the tool would be used over time. The director cautioned that FORS should look to other existing examples of physical data rooms before proceeding, such as the one at the Swiss Federal Statistical Office.

The head of the FORS unit INDEV (Infrastructure and Development) was in favour of going in the direction of both physical and virtual data rooms, but doing this over time, starting with a single physical room at FORS. We would need to develop a road map for this, taking into consideration technical, legal, and policy perspectives, as well as possible end goals. We would need to consider as well whether to offer such a solution to communities outside of the social sciences (e.g., linguistics, humanities, medical sciences) or to other institutions. There was discussion of the potential of such a system especially for the health and medical sciences, for example, for linking patient data between university hospitals. Until now, this has not been centralised in Switzerland, and systems are not interoperable.

The head of INDEV noted the importance of the topic for the SNSF, since a system for sensitive data does not yet exist within Switzerland and could help to fill the gap with respect to tools for fulfilling the SNSF's open data requirements. In the long- run, we might aim to further develop a national network for access to sensitive data, provided that additional resources are found for this.

The head of the group Data Tools, and Services (DTS) emphasised that analysis of needs is crucial - we should be sure to assess the needs of our target communities before implementing a solution. We should also be careful to consider limits on what might be shared by way of such a system, since certain formats can quickly lead to data volume and cost challenges (e.g., videos). Along these lines, the director agreed that long-term storage of high-volume data would raise cost issues, which might go beyond our institutional mandate. At some point, this might have to become a paid service for certain data.

Archive. According to our colleagues from within the FORS archive, there are sometimes data that cannot be archived because of their sensitivity, usually because the data cannot be sufficiently anonymised. This might be because of small sample sizes, or because of a lack of resources in the project team to do the work of anonymisation. In other cases, there have been copyright reasons that prevented archiving (e.g., data on representations of animals in the media, where the researcher did not have the right to publish a corpus of newspaper articles in database form). Occasionally, there are projects with a mix of sensitive and non-sensitive data, that is, with some data that can be archived and others that cannot. Thus, there are times when we must turn away researchers who would like to archive their data but cannot because of sensitivity issues. It is assumed that there are many more cases where researchers do not even try to deposit their data at FORS, believing that these cannot be shared because of their sensitivity.

It was agreed that with sufficient safeguards such sensitive data could be archived, but not within the current archival system at FORS. Beyond the technical barriers, there may still be practical or legal constraints that prevent archiving of such data (e.g., resource issues for data anonymisation, or lack of informed consent for data sharing). There was also agreement that it would be useful to have a system linked to SWISSUbase that allows for the acquisition and dissemination of sensitive data, even if this would not overcome all obstacles to data sharing. Before committing to a new solution, we must ask ourselves how many researchers would deposit their sensitive data or use sensitive data from such a system? There is no way to predict this currently, and so we should be cautious.

With respect to the acquisition of sensitive data by the archive at FORS, our colleagues warned that we would need to educate researchers regarding concepts and distinctions (e.g., on the notions of sharing and replication), since this domain may be subject to confusion. In keeping such data within our archival system, we would need to ensure sufficient security and controls on outputs by archive staff. This would also have resource implications. As for access, both colleagues believed that users would be willing to go to a physical data room in Switzerland, since it is a small country and distances would probably not prevent travel. Nonetheless, they thought that a virtual data room would be the preference among users, mainly for its convenience.

From the archive's perspective, there is no real preference between a physical and virtual data room solution. But in principle the one that produces the least additional work would be preferable. With respect to archival practice, there are no significant differences between the curation of sensitive and non-sensitive data. The differences are rather with respect to access. As with non-sensitive data, sensitive data should be preserved for the long-term (indefinitely), and over time should become less sensitive, which could lead to changes perhaps in access conditions. Also, documentation associated with data should be made available, unless the contents of the documentation do not permit this for data protection reasons.

The current resources of the data archive at FORS would likely not permit the addition of a full service devoted to sensitive data. Most likely a new system would be costly in resources and time, but this remains to be seen, depending especially on the demand.

Data protection. The discussion with our in-house expert on data protection and legal issues began with the caveat that many of the answers to our questions should be prefaced with "it depends". With respect to what the law says about preserving and making available sensitive data, it should first be noted that there is nothing specific in the law about archiving. Rather, the law addresses the "processing" of data, specifically any kind of process related to personal data. Here, archiving is a form of processing.

The first principle of relevance in the context of personal or sensitive data is that of "legality", meaning that there needs to be a legal basis for processing. This exists in various laws in Switzerland. There are two conditions that must be respected to collect and work with personal data - 1) destruction of data after the goal is achieved, and 2) publication of results where people are not identified. A second principle is that of "finality" - there must always be a specified goal in gathering the data. For archiving the data, this must be explicitly stated as part of the goal of a project.

Anonymised data are not subject to the need for a legal basis and would not need to be disseminated through a physical or virtual safe room. However, it is always difficult to determine whether data are truly anonymised. To determine whether data qualify as anonymised, you need to conduct risk disclosures on a case-by-case basis, assessing possible harm, the magnitude of the risk, and the probability of a breach.

FORS could legally hold and disseminate sensitive data under certain conditions. Notably, it would have to be a system of research data for researchers - the system could not be open outside of the research framework. The legal basis affords the right to handle personal data within the research context only. There is no law that allows dissemination of personal data unless we are in the "research exception" of the law.

There are three conditions for being able to make available sensitive data. First, researchers should explain to participants that there will be a storage and dissemination of their data to researchers. Second, participants should be informed that they have the right to ask about who is using their data. Third, the goal of the research has to be specified, including about long-term storage and use. Importantly for the archive from a legal perspective, it must be demonstrable that the archive has taken measures to ensure the security of the personal data.

A physical or virtual data room would be feasible in the Swiss legal framework, especially since users cannot take the data away with them, and so this form of access may not be considered as “communication of data”. Also, the technical system in itself hinders improper use of sensitive data. Most important is that research participants are informed that their data will be stored at FORS, treated in a secure manner, and used only for research purposes. Such a solution could disseminate any sensitive data, under the conditions described above. However, beyond the legal constraints, both researchers and the data archive should consider the ethical implications of making these data available, that is, whether harm might come to participants. Further, this requires case by case consideration.

Technical. The technical expert from the FORS IT group considered the concept piece to be fairly “classic”, that is, what one would expect from such a solution for remote access. While he was not opposed in principle to such a solution, his strong view was that a physical safe room would be easier to set up and maintain. It would also be less costly. And importantly, it would provide for a more secure environment, with fewer weak points where a breach could occur. The preparation work would be about the same.

Setting up a physical data room could be done with minimal technical hardware and software - this could even be done in-house. We would just need a computer in a room with no possibility for Internet or use of a USB port. We would also need a camera for filming, to prevent people from cheating (e.g., taking a picture of the screen). Someone in IT would have to set up and run a secure server. We would need a room of course, but not even a continuous one permanently devoted to the service, but on an as-need basis.

The physical data room server would be separate from the archival system SWISSUbase.

The platform itself will be separated (virtual servers), with dedicated login and particular software. It would probably be easier for a data-curator to manually download the sensitive data and put it inside the virtual environment (also safer). Encryption would be needed, and we would need to work out how data would be transferred to FORS. The physical data room and virtual data room are not incompatible, and correspond to different sensitivity levels, and so in the long-run FORS could offer both. However, at first it might be better to start light and easy, offering a safe room environment, and then testing and gaining experience, also to see how much demand there would be. We also discussed the possibility of a collaboration with an existing safe room in Switzerland, perhaps at a university or else with SWITCH. This could save us the trouble of setting something up ourselves. There would need to be a contractual agreement for this.

The technical system for a physical safe room could easily allow for choices available for data-depositors with respect to whether their permission is required for each access (with "None" or "With author agreement"). It is possible that an existing solution might be available that corresponds to our needs. However, it is hard to say without any further analysis. First, we do not know the functionalities offered by commercial solutions, we do not know their prices, and it depends probably on the security level we want to achieve. In any case, we would need infrastructure (UNIL vs SWITCH), because we do not expect to have a cloud provider in Switzerland offering such a possibility. We would have to see whether infrastructure has already been built by another University in Switzerland.

Regarding *s*torage and preservation of sensitive data, from a technical point of view this is always based on risk evaluation - 100% risk avoidance is not possible. It is an exponential curve: the more you want to approach 100%, the more complex / costly the solution will be. In the end, there will always be a trade-off. In general, encryption is a good way to go. There are basically 3 levels: 1) local encryption: data only encrypted on storage; 2) “remote" encryption: application runs in one provider, storage is done at another provider and encrypted (silos); and 3) the highest level would be "end-to-end encryption", but it is rather complex to achieve.

Last, there are technical limitations (at FORS) with respect to the choice of a system for access to sensitive data. First, we lack strong expertise in security issues. Second, we currently lack knowledge about secure rooms in general. Third, we lack knowledge about system administration, and lastly, we lack resources to carry out the work. Each of these would have to be addressed to move ahead with establishing and maintaining a solution.

Summary:

The interviews helped to clarify the needs at FORS from different perspectives. Fortunately, the various perspectives converge and lead to some conclusions concerning the likely direction for FORS. From a strategic point of view, there is clearly a need for FORS to offer a solution for access to sensitive data, primarily because our funder’s open data policy requires the sharing of data from funded projects and expects FORS to offer the possibility of preserving and disseminating all types of data. Further, our funder has expressed interest in such a solution to the extent that it could be used more broadly for other scientific disciplines where there is also a need. This aligned well with the further development of SWISSUbase, where a solution for access to sensitive data might be of value for disciplines beyond the social sciences (e.g., linguistics, humanities, medical and health sciences).

From an institutional perspective, it was made clear that building a comprehensive and ideal solution in a short time is not feasible, and that FORS should take a pragmatic approach. This means putting into place a minimally adequate solution, studying the effects, and then extending the service over time if needed, depending on the demand and available resources.

From archival and technical points of view, it should be possible to offer a service for access to sensitive data at FORS, provided that there are sufficient additional resources. Beginning on a small “pilot” scale would help to mitigate the resource challenges associated with setting up and providing a new service. Regarding data protection, there should be no legal barriers that prevent FORS from offering such a service, provided that particular conditions are respected in terms of security, policy, and data workflow.

Given what was discussed and discovered from the interviews, the FORS management therefore took the decision to move ahead in 2022 with a new pilot service for access to sensitive data. At the start, this will involve a physical data room with a simple workflow, run by the FORS data archive group and supported by FORS IT. The characteristics of the room are described in the revised “concept” in exhibit 4.

The work on the physical data room will begin in January 2022, beginning with the design of the room's technical specifications, attribution of a room that can be used, and setting up of a computer and server. On the archive side, we will establish a workflow and related data policies early in 2022. We will also prepare new texts for our FORS website where the service can be presented and described. Once everything is in place, we will announce and promote the service to the social science community.

Exhibit 4 - *Concept 2.0*

There will be a safe room at FORS where researchers come to access sensitive data under controlled conditions.

Technical and physical environment:

The safe room will be in a secure space on the 5th floor of the Géopolis building. This will be a dedicated room that can be used at designated days and times, but that can be used as well for other purposes. The room will include a camera so that visits can be monitored (to prevent cheating).

The room will contain a computer that will be used for the purpose of accessing sensitive data. The computer will not have Internet access (either by cable or wifi) or connection for printing. The computer will have pre-installed certain programs (e.g., SPSS, Excel, Word, etc.) that are commonly used for analyses.

All sensitive data will be archived on a secure, independent, and encrypted server. Sensitive data residing on the secure server will be transferred to the computer as needed when visits arise. The transfer will be done by an IT staff member, who will prepare the environment for the data user.

Workflow:

Datasets with secure data should be findable in the SWISSUbase catalogue, where there will be datasets with no data. For each of these datasets, there should be an info saying that the data are secure and that they should apply to consult those data, with a link to the request system.

The dataset description and file description of secure data should be on the same database as the rest of the data, to prevent complication to display it on the catalogue.

We will need a request system, where users can apply to consult sensitive data. Only verified users should be able to send such requests. (A verified user should be connected to SUB with an institutional email address and should have been verified by a data-curator). On this request form, the user should know what will be his/her rights and obligations (for example, no taking smartphones or cameras into the safe room). He/she will also be informed, not in too much detail for security reasons, on how he/she will access the data. The data requester should explain his/her motivations for accessing the data.

A data-curator should examine the request to access secure data. The data producer should also review and approve the request. If the request is approved, the user will have to sign a contract where he/she will be instructed not to copy the data in any shape or form or try to identify individual respondents (among other things). Here, he/she will have a list of the tools available on the secure computer. If necessary, he/she can ask for more tools (within a reasonable time period and with motivations. This should be evaluated and done by a technical team).

Shortly before the appointment, the archive staff person (i.e., safe room attendant) will install the requested data and documentation in a clearly indicated folder. IT staff will install any additional approved software.

The day of the appointment, the user will present himself/herself to the safe room attendant (data curator) at the scheduled time and place. He/she will receive credentials to log into the secure computer. Then the user can connect to the secure computer and conduct his/her analyses.

If the user would like the results of his/her analyses, he/she can ask for the output in digital form, for example, on a USB stick. All output that will leave the safe room should be examined by the safe-room attendant to ensure that there are no personal data and that no individual respondents could be identified. The technical team will keep a record of every user that requested an access to the safe room and that accessed it.

The user credentials will be destroyed after he/she has accessed the safe room. If the user wants to access any other secure data, new credentials will be sent, and he/she will have to fill another request.

Implementation

This first implementation will act as a prototype for a more elaborate implementation that will come later. We will implement this prototype with our current system, meaning that this will be separate from our current system. Several modifications need to be carried out to have a completely working and secure solution (e.g., encryption, request system, new dataset “type”, how to submit sensitive data).

We had to write protocols on how to retrieve the data from our server, so that it can be placed on the computer in the safe room for the researcher to consult. The safe room computer must be prepared as well, with different software (PDF reader, STATA, R, ...). The computer will not be connected to the Internet, and so the transfer of the data from our servers to the PC must be done by USB key.

Before erasing the data on the safe room’s PC, we need to retrieve the researcher’s output. This will be placed by the researcher in a designated folder. It will be analysed by an archivist to be sure that no sensitive data leave the safe room, and then released to the researcher. We will save on our database each access to the safe room (who consulted what, when, with what output.

Future directions

Once the new service is in operation, the pilot will explore the level of interest and need in the social science community in Switzerland, will test and adjust the workflow and tool, and will re-examine the question of needed resources.

In parallel, FORS will begin to look out for new data that might be candidates for the data room, and we will begin to contact key stakeholders (e.g., the SNSF and the Swiss Federal Office of Statistics, institutional partners) to see how the service might be taken up more broadly. One interesting possibility would be to use the physical data room to store and disseminate linked data tied to the Swiss LinkHub initiative, since currently there is no technical solution nationally that is capable of handling this kind of sensitive data.

In the mid- to long-term, if there is significant uptake of the physical safe room, we can envisage certain developments. First, we could explore the possibility of adding a virtual data room (i.e., remote desktop access), so that users could access the data from their own computers, but still under highly restrictive conditions. The physical data room would then be reserved for the most sensitive data, with the virtual data room used for less sensitive data.

Second, assuming the success of the service, FORS could look to link up its physical data room facility with other institutions nationally that (would like to) offer a similar service. At the international level, ideally, FORS could also consider joining the European Remote Access Network (ERAN), beginning with bilateral collaborations with GESIS and the UKDS. These connections could already be explored during the course of 2022.

Conclusions and recommendations

Many infrastructures do not have solutions for sensitive data. For those that recognise the need and have the interest, many are not sure what options are available to them and how to proceed. FORS underwent this reflection and took exploratory steps in a systematic way. As part of SSHOC 5.4, FORS documented its internal deliberations and have made these available as part of this case study. It is hoped that this might be helpful for other institutions that would like to move down this path.

The lessons learned by FORS are transmitted in the form of the following recommendations:

1) *Consider the strategic importance for your institution.*

Creating and maintaining a system for access to sensitive data brings with it costs and risks that should be justified in terms of potential benefits to one's institution in the long-run. Therefore, there should be a careful assessment of the strategic importance of such a solution for one's institution. Of course, the solution should fit into your institution's mandate and mission and should be offered to the primary public that is served. Further, the solution should fill gaps that are not currently addressed in terms of services in relation to user needs. Ideally, the solution should strengthen or extend your institution's partnerships with key stakeholders.

2) *Conduct a needs assessment.*

A first useful step should be to assess the need for a service for access to sensitive data. Mainly, what is the likelihood of obtaining future sensitive data that cannot otherwise be disseminated? FORS realised this based on the number of datasets that it needed to turn down, or after hearing from researchers that they decided not to share their data because of data protection issues. We therefore anticipate that there will be a significant demand for a solution for access to sensitive data. Also, are there top-down pressures from funders or journals that will incite researchers to share their sensitive data? If so, this will provide an incentive for developing a solution.

3) *Consider the data protection issues.*

Would your institution be capable of handling personal data in compliance with relevant national and/or European legal requirements? To this end, consider whether adequate expertise is available to you regarding data protection and research ethics. Any adopted solution that deals with personal data must be judged by competent experts to be compliant with all relevant legislation.

4) *Consider how a solution would fit into the existing archival workflow*.

In many cases there will be an existing archive workflow and system that handles non-sensitive data. Adapting this to accommodate sensitive data will require careful consideration and possible modifications to the archival workflow, especially in relation to the chosen technical solution. For example, will the sensitive data be kept on separate servers? How will backups be carried out and how will SIPs, AIPs, and DIPs be generated and maintained if the data are not handled within the general archive environment? What special measures will be needed in the workflow to ensure the security of the data, including limited access among archive staff? How will metadata for sensitive data be integrated into a larger catalogue? At FORS we have begun to identify and reflect on these and other important related questions. While we can imagine how things will unfold and already take certain decisions, we expect that much will need to be clarified and adjusted when our chosen solution goes into operation and begins to be piloted.

5) *Consider the technical possibilities and implications*.

We recommend that a first look be given internally to one's institution's existing technical systems and capacities. What is available that might lend itself to addressing the needs regarding the preservation and dissemination of sensitive data? Could existing internal systems be adapted for these purposes? Is an external system needed?

Review existing external solutions in relation to needs. Which correspond most closely? Check those external solutions in depth as some details might not correspond to the needs (example: the data are stored in another country). Is there an external solution provided by a technical partner? What would be the cost to integrate such a solution? And what about building an in-house solution?

In any case, it would be best if the technical team have knowledge on security issues and service administration. It would be important to plan what would be needed for the technical solution first, and then identify what existing internal systems and capacities are available, what gaps there might be, and what (if any) external solutions might be needed.

6) Take into account available resources.

A key factor will be the resources available to invest in a solution for access to sensitive data, both in the short-term and long-term. If resources are limited, then it might be necessary to begin on a small-scale and to adopt a "wait and see" approach, developing the system over time according to demand. Besides moving more slowly, another approach might be to change the objectives or scope of such a service, for example, in terms of who can be served, or which types of data can be accepted. Also, as FORS has realised, a modest safe room can be a pragmatic and light way to begin. Not to be underestimated as well are the IT resources needed to set up and maintain a solution. Sufficient IT support should be assured from the start, regardless of the solution adopted.

7) Set up a good process for decision-making.

We recommend an inclusive process for defining and examining the key issues, including discussing with relevant colleagues and coming up with a plan and solution that fit to the various perspectives in the best possible way. The selected solution should derive from and be compatible with the various strategic, legal, archival, technical, and resource conditions.

Certain techniques can be helpful in defining and clarifying issues in the decision-making process, including in-house interviews, and the design of draft "concepts", which can be tested against the different perspectives.

Last, it can be useful to document your process. This creates a trace that can serve as a resource for reflection and discussion. It also allows for a more conscious and deliberate structuring of the decision-making process.