Title	**Secure Remote Access system for an upgraded CESSDA RI (D10.3)**
Work Package	WP10
Authors	Rob Grim, Pascal Heus, Tim Mulcahy and Jostein Ryssevik
Source	Metadata Technology
Dissemination Level	PU (Public)

Summary/abstract

This report is deliverable 10.3: “Functional Specification for SDC”.

Secure Remote Access system for an upgraded CESSDA RI

Version Sep 2009

Rob Grim
Tilburg University

Pascal Heus
Metadata Technology Ltd.

Tim Mulcahy
NORC at University of Chicago

Jostein Ryssevik
Ideas2evidence

Contact
info@metadatatechnology.com

Abbreviations

CESSDA	Council of European Social Science Data Archives
DDI	Data Documentation Initiative
NSI	National Statistical Institute
PUF	Public Use File
SDC	Statistical Disclosure Control
SDS	Secure Date Services
SRA	Secure Remote Access
SDMX	Statistical Data and Metadata Exchange Standard
SOA	Service-oriented architecture
SUF	Scientific Use File
VPN	Virtual Private Network
XML	eXtensible Markup Language

Credits

We would like to thank the UK Data Archive and the CESSDA teams for their help and support in preparing this document. In particular we thank Hilary Beedham, Melanie Wright, Mus Amshet and Reza Afkhami for their feedback and for kindly hosting us during our visit to the SDS in Colchester.

We also want to acknowledge the technical inputs from the NORC's Data Enclave team and contributions from Kurt Roemer, Jason Southern, Steve Ash and Ali Collins from Citrix Inc.

Executive Summary

The need for improved data access, quality, and sharing has only heightened in recent years. Every day, policymakers and other decision makers depend on access to sensitive data or derivative products to make important, empirically based decisions that affect societies, global economy and living conditions.[1] Providing secure access to sensitive microdata is essential to modern day political, social, and economic well-being. Access alone, however, will not ensure data quality and utility. The microdata must be complemented by high-quality metadata documentation, fostering the replication standard[2] in an environment that promotes collaborative work and knowledge sharing. No less critical is data transparency across the research process to ensure that public policy and the analytical work that underpins it are both generalizable and replicable.

This paper, commissioned by The Council of European Social Science Data Archives[3] (CESSDA), under Work Package 10 of the Preparatory Phase Project (PPP), provides a comprehensive discussion of various secure remote access (SRA) platforms currently in place around the world, as well as advantages and disadvantages to the various models. It also outlines the technical, organizational, operational, statistical and legal challenges associated with operating such facilities. It serves as an initial roadmap for CESSDA as it works toward a solution to create and sustain a truly integrated European data infrastructure in which the European social science researchers have access to the data resources they need to conduct research of the highest quality.

The authors provide an overview of selected data modalities and details on technical requirements, infrastructure, and configurations on models of potential interest, including the: (1) standalone model; (2) shared remote access model; (3) user remote access location; and (4) cross-national configurations. Or particular note, the paper compares and contrasts several remote access facilities, including: Statistics Denmark, Statistics Sweden, Statistics Netherlands, NORC's Data Enclave, and the UK Data Archive's upcoming Secure Data Service (SDS). In addition to emphasizing that new modes of microdata access bring both challenges and opportunities, the authors note the "privacy paradox," and recommend striking an adequate tradeoff between disclosure risk and information loss. The paper draws on advances in the social science and the computer science fields, the sum of which provides CESSDA insight into designing and implementing European wide secure data access platform.

[1] Nonperturbative strategies protect respondent identity by producing partial suppressions or reductions of detail on the original dataset.
[2] http://gking.harvard.edu/projects/repl.shtml
[3] http://www.cessda.org/

The authors point out various challenges involved in developing and implementing SRA facilities. In addition to IT security and other technical, organizational, and operational concerns, SRA facilities must ensure data confidentiality. The paper emphasizes that statistical disclosure control (SDC) in a secure remote access facility – one that provides researchers direct access to sensitive, underlying microdata – means that the focus of the “data treatment” has moved from controlling inputs to controlling outputs.

The paper also provides detail on ways to leverage metadata best practices to increase data quality, as well as effective collaborative and knowledge management. At a minimum, the researcher environment should be interactive, dynamic, one that leverages technology, rich metadata, collaborative spaces, and social networking tools. In terms of SRA organizational requirements, the paper emphasizes the need for a solid general management and organizational structure, and points to the importance of various legal issues, particularly related to data sharing across borders, relevant European legislation for access to confidential data and legal regulations that govern access to confidential data within the European Community. The authors furthermore note the critical importance of developing and faithfully implementing rigorous security and training plans. In addition, some high level cost estimates are provided, assuming varying levels of SRA sophistication. The paper concludes with two case studies, use case scenarios, and practical recommendations and solutions.

High level recommendations include the following:

- CESSDA should consider the relative advantages and disadvantages of the various data access options – both architecturally and organizationally - described in this report with an eye toward developing a hybrid model that best meets the needs of its members, partners and users.

- CESSDA’s technical solution should focus on designing and implementing a secure remote data access that encompass lessons learned from experience gained in developing the NORC Data Enclave, The UK Data Archive’s Secure Data Service, Statistics Denmark, Statistics Sweden, and Statistics Netherlands and others.

- In addition to technical, legal, and organizational issues that need to be addressed, CESSDA must also come to a consensus across its member organizations in terms of how to ensure cross border data confidentiality, recognizing that statistical disclosure control in remote access modalities requires a fundamentally different approach to proscriptive rules-based methods.

- Regardless of the data access solution CESSDA implements, it should follow the portfolio approach to protecting data confidentiality, one that includes adequate technical (e.g., IT, systems, and network); operational/organizational (e.g., management protocols, physical security, cost structure); educational (e.g., data specific training on study design, sampling frame, and correct use of weights); statistical (as per selected disclosure control technique(s)); and legal protection (e.g., contracts, nondisclosure agreements).

Introduction

The Council of European Social Science Data Archives[4] (CESSDA), a network of social science data archives in 20 European countries, is in the early stages of preparing for a major effort to design and implement a truly integrated European data infrastructure. This timely yet challenging effort will provide European social science researchers access to data , within a single European system using a common set of protocols and procedures.

In so doing, CESSDA will evolve from an entity in which each member organisation coordinates with national resources on their own, to one that emphasises its common ground and shared goals, objectives, and overall vision for accomplishing a coordinated pan-European experience.

In this context, one major challenge relates the practical implementation of how exactly secure access will be provided to confidential data. One solution that has received increasing attention is providing secure remote access, in which researchers have direct access to sensitive microdata. Indeed Secure Remote Access (SRA) facilities that are now being implemented in various data organisations constitute an important step forward. This report serves to provide strategic direction to those involved in designing and implementing CESSDA's integrated European data solution.

The report, commissioned by CESSDA under Work Package 10 of the Preparatory Phase Project[5] (PPP), seeks to assess the UK Data Archive Secure Data Service (SDS) and other existing models for access to sensitive data and provides recommendations for an optimal model for CESSDA. Authors of the report include Pascal Heus (Metadata Technology Ltd, UK), Tim Mulcahy (NORC at the University of Chicago, USA), Rob Grim (Tilburg University, Netherlands) and Jostein Ryssevik (Ideas2evidence, Norway) in consultation with a team from the UK Data archive and other domain experts.

[4] http://www.cessda.org/

[5] http://www.cessda.org/project/

Background

"New technology, particularly the development of user-friendly thin client systems, has made the provision of lab facilities increasingly appealing. The result is that demands upon NSIs to improve access to data are increasingly being met by innovative lab solutions. Along with flexible remote job submission systems, the provision of "research environments" (where manipulation of data and the choice of statistical models are both largely unrestricted) is therefore growing strongly. This growth in use of research environments presents a problem for statistical disclosure control" (Felix Ritchie, 2007, UK Office for National Statistics).

****

National statistical institutes (NSIs) collect and disseminate data on all aspects of individuals and businesses. Access to microdata however varies across countries and across data types (Ritchie, 2007). One core function of these institutes is to make available data at the aggregate tabular level. Increasingly, however, NSIs are being encouraged to provide access to the underlying microdata so researchers may pose and answer questions of their own choosing (Abowd and Lane, 2003). This pressure creates both opportunities and challenges. While on the one hand the reputations of statistical agencies may be enhanced as data dissemination may lead to better and more timely policy responses; on the other, new modes of microdata access create new challenges to data security and confidentiality.

Although the mission of NSI's is to collect and disseminate high quality data, a fundamental tension exists in that they must also protect respondent confidentiality (Lane, 2003). Madsen and others have summarized this "privacy paradox" in terms of optimizing the trade-off between disclosure risk and information loss, also referred to as the trade-off dilemma or statistical disclosure control problem. This process of masking data to reduce the probability of re-identifying individuals and enterprises and protecting confidentiality (avoiding disclosure) has become more complex as both technological advances and public perceptions have evolved over time. The choice of different disclosure protection techniques, e.g., the decision to top-code, data swap, or suppress information, affects data quality. Fortunately, statistical disclosure limitation techniques have kept pace with the rapidly growing changes affecting data access and dissemination (Census Bureau, 2002; Lane, 2003; Abowd and Lane, 2003).

Ensuring statistical protection by delivering the appropriate data to the researchers and applying relevant disclosure review processes to outgoing information are typically responsibilities shared by the data producer and the remote access facility agency. The operating environment must therefore be equipped with the relevant tools to facilitate such operations, manage the information flows, and audit the various processes. Involving the users in these activities, training them on disclosure issues, and providing access to reference materials is also an important component.

Access modalities to sensitive data

A number of different approaches currently exist for protecting confidential data. Data producers may choose to lock the data away; anonymize the data and create public use files; legally bind (licenses, contracts) trusted users; or deposit the data in physical data enclaves, research centres, remote execution or other secure remote access facilities. Regardless of one's choice, providing access to sensitive data involves a wide array of complex issues.

Access to sensitive data sets has traditionally been restricted to physical data enclaves or research data centres. Once the only option for researcher data access, these types of facilities have been around for many years and remain a highly secure option. However, they are expensive to operate and, by nature, require users to travel, sometimes long distances, to conduct their research. This precludes access for many potential researchers without the resources or time to travel to the enclaves. Delays also frequently occur during the initial project clearance process, background checking, and long waits on responses to microdata output requests. A well known example of such a facility is the US Census Bureau data centres.

Alleviating the need to be physically located at the data site, remote execution facilities were established that allow users to submit processing scripts to be executed in batch mode, the results of which are subsequently reviewed by statisticians before being released publicly. Users also have access to fake (or synthetic) data files to prepare analytical scripts. This approach however is non-interactive and can be very slow when having to perform numerous, complex analyses. The Luxembourg Income Study[6], the International Service Data Center[7] (JoSuA) at IZA in Germany, and the Remote Access Data Laboratory[8] (RADL) in Australia, are examples of such facilities.

One classic approach for safely disseminating potentially disclosive data involves applying data anonymization techniques. Although these methods facilitate production of public use files that may be made widely accessible on CD-ROM or through the web, it also reduces the usefulness of the statistical information and therefore may fail to meet researchers' needs. Complex modelling techniques can also be used to generate synthetic data; but likewise may not produce statistically meaningful datasets and can be resource intensive. Light anonymization techniques combined with contractual agreements can likewise be applied to deliver more meaningful datasets to the researchers on encrypted CD-ROM or over secure connections. Recent experience (American Statistical Association, 2008; Kennickell and Lane, 2006) however, points to limitations to this approach.

[6] http://www.lisproject.org/

[7] http://metadata.iza.org/josua_home.php

[8] http://www.abs.gov.au/websitedbs/D3310114.nsf/home/CURF:+Remote+Access+Data+Laboratory+(RADL)

At the same time, technological progress in the last decade has opened the door to new secure access mechanisms. High speed private or encrypted networks now offer the option of linking data access facilities to one another, thereby connecting geographically distributed researchers and data. This is, for example, the approach currently being taken by the Canadian Research Data Centre Network[9] whose access facilities and branches have recently been connected through a high speed encrypted network, greatly facilitating the management and exchange of the data/metadata and fostering collaboration across sites.

The option that has gained significant prominence in recent years, and is the focus of this document, is providing secure remote access to sensitive data through virtual data centres over a public network such as the Internet. Technologies such as Citrix have demonstrated the security of such an approach by delivering solutions for the financial, health, manufacturing, government and other sectors – the specifications of which could easily be adapted to social science data. Operating such facilities shares some of the challenges of the physical enclaves and requires specific technological expertise, but also includes numerous benefits, e.g., user friendliness and knowledge sharing. Examples are provided in annex (p.9) and additional information may be found on the Citrix web site[10].

It is important to note that each data access "solution" inevitably will have both advantages and disadvantages. No single solution however will ever reduce the risk of disclosure to zero. There will always be a human, technological, or statistical factor that may come into play and open the door for disclosure concerns. Therefore we recommend combining statistical, operational, and technological protection (i.e. a portfolio protection approach). Making sure the data are accessed for statistical research purposes only, binding the users through legal agreement, educating the researchers, and providing access to trusted researchers, are all organizational techniques that increase the likelihood of maintaining sensitive data securely. The vast majority of researchers take their oath of confidentiality seriously and would not intentionally breach confidentiality. It is nonetheless a reality that one single breach could significantly damage the trust of respondents, lower respondent rates, and threaten the reliability and validity of vital statistical data.

Although the contents of this report focus mostly on issues related to designing and implementing secure remote access (SRA) platforms for disclosive data, the information provided herein may also be useful to those interested in providing access to less sensitive data, such as scientific use files (SUFs) or public use files (PUFs). Indeed it is often desirable to import these types of data into the closed environment for the purpose of merging with disclosive data for analytical purposes. It is important to note, however, that SRAs should not be seen as replacing other existing data dissemination modalities, or as a reason to limit the production of SUFs

[9] http://www.statcan.gc.ca/rdc-cdr/network-reseau-eng.htm

[10] http://www.citrix.com/lang/English/ps2/segments/index.asp

or PUFs, as such data satisfy the needs of many users. This report also focuses on official statistics and government data whose accreditation and security requirements are expected to meet the needs for other data depositors (non-NSIs). The solutions are therefore also applicable to other data providers such as academics or non-governmental institutions.

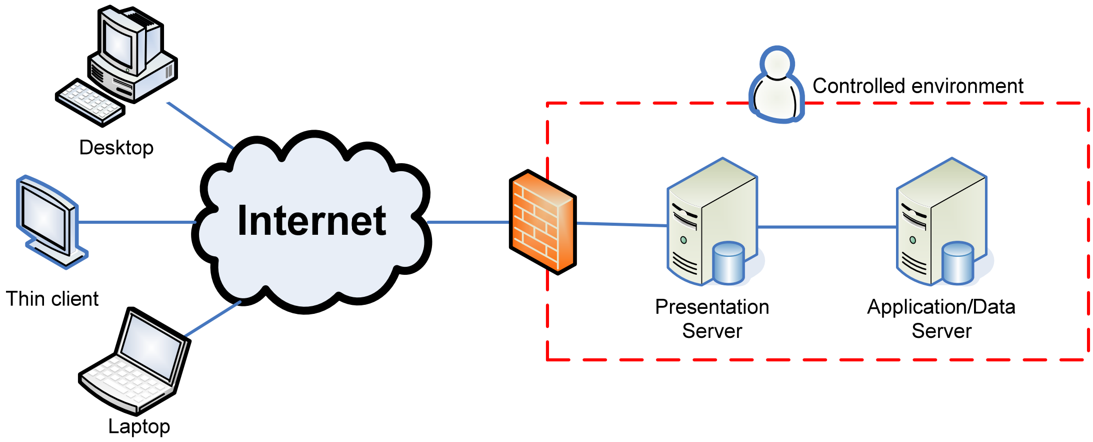
Overview of Secure Remote access

Introduction

What is secure remote access and how does it work?

Providing secure remote access to disclosive data is made possible by implementing a combination of modern technology and organizational principles.

From the technical perspective, the typical architecture is a modern interpretation of the legacy mainframe computer paradigm, whereby all the data reside on the central system upon which processing applications are executed. Users interact with the central system using a remote terminal; however, no output is exported from the closed and secure environment until cleared for public release, i.e., after having undergone a formal review process for potential unauthorized disclosure issues.



Today's architecture is much more sophisticated. Terminals come in various shapes (desktop, laptops, thin clients, even phones). Elaborate security layers surround the system (firewall, encryption, strong authentication, etc.). Servers and desktops operate in a virtual environment, and communication can takes place over a high speed public network (Internet).

The main features that distinguish this system include the following:

1. No information may move in or out of the centrally protected environment without the permission of the system managers

2. Only authorized users can access the resource

3. Access to the data resources and applications are closely controlled. The only external access to the data is what is displayed on researchers' computer screens, i.e., only screen updates, mouse clicks and keystrokes transit over the network over a secure encrypted channel.

Implementation Challenges

Technology alone however is not sufficient to ensure data confidentiality. As noted previously, providing secure remote access to sensitive data presents challenges that are not only technological in nature. While a state of the art IT architecture plays a central role in the feasibility of such an approach, other aspects must be taken into consideration when envisioning the most appropriate solution. This report addresses these issues in a stepwise manner.

Finding creative ways to address the fundamental tension between data dissemination and the protection of respondent confidentiality goes to the core of each statistical institution's mission (Abowd and Lane, 2003). One innovative way to minimize disclosure risk and protect data is to distribute the risk through a portfolio approach, one that combines technical, organizational, operational, statistical, educational and legal protections. This approach builds on Markowitz's (1959) optimal portfolio theory, that "*any two data protection methods are correlated in their risk of disclosure of confidential information, but not perfectly. Combining the two methods can, then, produce greater data utility for any given level of disclosure risk in exactly the same way that an investor can achieve greater expected return for any given level of investment risk by combining the risky assets into a portfolio*."

Although contracts and licensing agreements (legal protection) may be effective means for addressing breaches ex post facto, alone they will not ensure data confidentiality. Neither will any particular statistical protection technique used to de-identify microdata. Nor will any other specific technical or operational component. Indeed, no single component will adequately protect data confidentiality. However, by combining the key components in a synergistic, portfolio perspective, the risk of disclosure is minimized. The ideal secure remote data access system therefore will include technical and operational security, including rigid IT security protocols and statistical protection of the data, as well as statistical disclosure control processes of researchers' output prior to publication. It also will include a clear set of legal protocols to ensure that only authorized researchers, from trusted institutions, be permitted access, and that the research is conducted for statistical purposes only (Lane, Heus, and Mulcahy, 2008).

Technical challenges

Remote access platforms that leverage Virtual Private Network[11] (VPN) technology protect data by controlling the environment in which the research is conducted. Data are not distributed to researchers; rather, researcher access to data is distributed and

[11] http://en.wikipedia.org/wiki/Virtual_private_network

controlled, preventing outsiders from reading the information transmitted between the researcher's computer and the host network. Files may not be downloaded. Users cannot use the "cut and paste" feature or save or print data on a local computer. Data producers may also choose to implement physical restraints on the researchers, e.g., webcams, biometrics, RSA cards, secure rooms, and electronic card entry. Statistical applications and data (read only) are provided through the host network.

In addition, researchers are also provided with friendly tools that support online data and metadata searching, browsing and analysis. To enable high-quality use of the data, all data are delivered in tandem with the related metadata, providing the researchers with the necessary context for data analysis.

Security is central to designing and implementing the overarching architecture. Sufficient measures must be in place to ensure that users are properly identified and the data are fully protected, including authentication, authorization, encryption, monitoring, and backups. These challenges are not specific to social science and can be addressed by adopting industry standard solutions. Products however must be customized to meet the specific need of a statistical data management environment.

The environment must also be producer- and researcher-friendly, which implies a scalable infrastructure that can meet the need of the data providers and scale up to meet users' demand. This necessitates technical specifications for the computer hardware, software products, and network configuration that fit the various scenarios. This is particularly important given that, in a remote access environment, researchers concurrently perform data analysis on the remote server, which requires considerable processing power and system memory. In addition, while disk space is expensive, hosting tens or hundreds of users in a shared environment requires a sophisticated and safe storage system.[12]

Operational challenges

Operational aspects are similarly important. The data hosting institution is advised to establish a set of operational procedures to ensure appropriate access. Regardless of the type of data being disseminated, the data custodian is typically interested in ensuring that only trusted, approved, and authorized researchers have access to the data. This reduces the risk of malicious disclosure and reduces the risk that the study respondents will have negative perceptions associated with data access.

A related issue has to do with the economic costs associated with providing secure access to data. Although the cost of providing access depends on the modality (e.g., public use microdata, licensing, remote access sites, and research data centres) the

[12] Depending on the data provider, various remote access options must also be taken into consideration, from typical computers operating over the Internet to more secure access station (thin clients) or monitored remote access facilities. These various options will be discussed on "Client hardware and configurations" (p.40).

feasibility and need to develop cost effective solutions for making high quality public use data sets is ever increasing. In addition, the opportunity costs of accessing research data centres are substantial, e.g. costs related to staffing and technology (hardware, software). There are also potential reputation costs and the costs associated with identification of the sample entities and the potential disclosure of confidential attributes (Abowd and Lane, 2003; Kennickell and Lane, 2006)

Because providing access also results in a support burden, appropriate operational incentives should be put in place that reflect the cost of support. Examples of such operational incentives could include charging the marginal cost of statistical disclosure review, charging for excessive storage costs, and charging a small weekly access fee to ensure that there is an incentive for projects to be completed in a timely fashion. Similarly, since most data custodians want to provide access to data in order to promote data analysis, operational incentives should be put in place to promote analysis. This could include highlighting the work of particular researchers, instituting a working paper series, or, as discussed in a subsequent section, actively promoting the development of a virtual organization around a particular dataset (Lane, Heus, Mulcahy, 2008).

Training and knowledge management challenges

Researcher training is another critical component of the portfolio protection approach. The training should focus on the importance of having safe projects (approved projects); safe people (i.e. authorized researchers); safe settings (i.e. remote access); and safe conduct (care with handling and releasing data). The goal is to instil a “culture of confidentiality” among all authorized researchers. Fundamental to fostering this shared trust, users must have a clear understanding of their researcher responsibilities. Along with the shared benefit of gaining access to sensitive data comes a shared burden of ensuring data confidentiality.

Before obtaining access to data, researchers should be trained on the legal background of each data source, various disclosure definitions, as well as the principles and practicalities of disclosure control. Researchers also should be trained on the nuances of the datasets, which increases the likelihood that they will generate valid and reliable analysis. For example, data producers could consider providing training on the data themselves, including information about the study design, sampling frame and the correct use of weights. One critical part is to train and encourage researchers to conduct a preliminary disclosure review of their research prior to requesting formal release. This enables the researcher to understand what type of information is needed before release is permitted. It also sensitizes researchers to the time and resources required to conduct a thorough disclosure review, thereby reducing the number and volume of requests.

Researchers also should be trained in best practices in working within a shared environment, microdata and metadata documentation best practices, etc. Users benefit greatly from the comprehensive metadata that should surround all datasets. Researchers also are encouraged to collaborate and view statistical data analysis as both a knowledge capture and sharing experience. A data enclave, as a closed environment, presents an ideal opportunity to manage such information by taking advantage of the appropriate metadata specifications and collaborative tools. Training sessions should emphasize that documenting data and capturing knowledge can be achieved by combining relevant metadata specifications and related best practices with knowledge sharing and collaboration tools. An effective secure remote data access facility must therefore take into account issues such as:

- Leveraging metadata specifications such the Data Documentation Initiative (DDI) to provide high quality documentation
- Facilitating the capture of the researcher processes and knowledge
- Providing a solid understanding of how the data are being used
- Facilitating the archive and dissemination of researcher outputs
- Providing users with collaborative spaces that foster knowledge exchange

In the particular context of a network of access facilities such as CESSDA, we must also examine how the metadata and the community knowledge, typically less sensitive in nature than disclosive data, can be exchanged and/or shared between different facilities, therefore establishing a broader and more active knowledge space.

Statistical protection challenges

Protecting the data from misuse and potential disclosure, intentional or not, is a fundamental aspect of any data access facility. This can be addressed in several ways:

- Minimizing the risk when the data are provided to the users (while maintaining maximum usefulness)
- Monitoring the information that leaves the enclave environment and
- Educating the users

Historically, NSIs have protected data confidentiality by constructing a set of unique identifiers that substitute for variables including explicit personal/organizational identifiers, such as name, address, phone number, Social Security Number and Taxpayer Identification Number. NSIs also have limited researchers’ access to the data they need for their specific research questions if necessary. To accomplish this, NSIs created custom analytic data files that contained a subset of the columns (and even rows) from the master data set. With the advent of researchers working directly with microdata in secure remote data access platforms, no longer is the focus on ensuring the non-disclosiveness of aggregates or generating non-disclosive (“public use” files) dataset.

Legal challenges

Another set of issues includes legal and ethical protections. While many laws govern data dissemination, often there is less clarity about the prevailing legal framework. Indeed there is a fundamental uncertainty about data ownership – whether data constitute private or public property. Who owns the data? The respondent him or herself? The data collection organization? Researchers who analyze or otherwise add value to the information? What about researchers who purchase interest in the data, etc (Lane, 2003)?

As a starting point, the host should develop a set of business rules that clearly defines the roles and responsibilities of data custodian, data producers, and researchers. For example, the host could develop a Memorandum of Understanding with the data producer that codifies the general parameters for providing researcher access and the prevailing legal framework (Ritchie, 2009). The host might require that researchers enter into a formal contract within which access policies and penalties are made clear. Host institutions might also require that an official with signature authority from the researcher's institution also sign the agreement, binding both the researcher and the institution to the legal agreement. Secure host institutions might also choose to have researchers sign nondisclosure agreements specific to each data set. Nondisclosure agreements also should be signed by all external contractors (e.g., system service and maintenance). It may also be useful to develop researcher profiles and to reserve the right to conduct background investigations on researchers.

A pledge of confidentiality assumes that publicly available data will be anonymized or otherwise masked to ensure that they cannot be used to identify a specific person, household, or organization, either directly or indirectly by statistical inference (National Research Council, 2005). Despite taking every reasonable means possible to protect data confidentiality, breaches however may still occur. This is why it is important not only to have a contract in place that binds both the researcher and his/her institution, but also to have in place fines and penalties. The researcher and the institution must be clear that if there is a breach, inadvertent or otherwise, penalties will arise. The risk of facing civil and/or criminal sanctions both individually and institutionally is not insignificant. Although the risk of potential fines and imprisonment serves as a deterrent to nefarious behaviour, it may also be prudent to include penalties to researchers and institutions in the form of restricting privileges to seeking grant funding opportunities. In this manner, research institutions bear a significant portion of the risk and thus may also be more apt to more closely monitor researcher contractor compliance.

Safe people, projects, settings, and outputs

Another way to envision the portfolio protection approach to risk management is to suggest that safe use of data is really the combination of prior decisions made. Although safe people are the key component of ensuring confidentiality, it takes three criteria: safe people, safe projects, and safe settings to increase the likelihood of safe outputs.

Criterion	Meaning
Safe projects	The project has been reviewed to ensure that it has a valid research aim
Safe	The researchers can be trusted not to misuse their access

people	
Safe data	The data have been treated to limit disclosure risk
Safe settings	Technical solutions limit the options for misuse of data
Safe outputs	Checking of outputs produced by researchers to reduce the likelihood of identifying respondents in statistical outputs

- *Safe people* – authorized researchers must demonstrate that they are trustworthy. Data access may be contingent on the potential users achieving some sort of researcher accreditation (similar to the UK Data Archive SDS requiring researchers to achieve Approved Researcher (AR) status). Researchers should be affiliated with reputable academic institutions, and these institutions must share some of the risk by signing a formal contract that bonds not only the researcher but also the research institution to a contract. Researcher training is also crucial in emphasizing a culture of confidentiality (this is discussed more fully in "Training" p.70).

- *Safe projects* – researchers must clearly demonstrate (through a formal proposal process or otherwise) that the proposed research project is for a specific statistical or other research purpose (not law enforcement, marketing, etc.), and that available public use data are insufficient to answer the research questions posed.

- *Safe settings* – a SRA facility must be technically secure. When researchers access data through secure remote access facilities, all of the technical work is carried out on the internal server. Researchers' machines are essentially made into dumb terminals. Users only see screen shots of the output (i.e., data may only be accessed in read only mode). Researchers can read/write on the work area, and they have own private work area where no one else can access. Researchers cannot take anything in or out of the virtual facility. Since researchers do not have Internet access, all import and export requests must go through the facility managers. If researchers want to take anything into the environment (program or data), they must submit through a secure file transport protocol site where lab staff access and copy to the appropriate destination. Similarly, after completing work, output that researchers would like exported from the SRA must be placed in an output review folder where SRA statisticians review and make recommendations on release. Researchers cannot change anything on the input or output drive, cannot cut and paste, print, or map network drives.

- *Safe outputs* – All results/output must be reviewed to make certain it is safe to release, i.e. it is non disclosive. Researchers should adhere to a set of formal guidelines (or checklist) of items that must be attended to before submitting a table for disclosure review. All derivate files must be included in the output review request as well as a summary of the research methods. A statistician or small group of statisticians must also have a set of guidelines to apply. Imposing strict rules on statistical output alone however will not suffice, as most microdata output requests by nature are context specific and thus must be manually reviewed.

Example of existing remote access facilities

Over the last few years, a handful of remote access facilities have been established throughout the world by national statistical offices as well as by independent data producers and data publishers. The facilities vary along a variety of dimensions, mostly reflecting the differences in requirements deriving from national data protection laws. Below we highlight and describe select facilities of particular relevance to this effort.

Statistics Denmark

Statistics Denmark was the first European NSI to establish a remote access facility for confidential microdata and the Danish approach has served as a model for several other statistical offices worldwide. The facility was established as an alternative to an existing on-site data enclave (research centre) and for several years the two services operated in parallel. The on-site data enclave was closed in 2009, leaving remote access as the only alternative for researchers requiring access to microdata.

The remote access facility provides access to survey data as well as data from statistical registries. Currently more than 600 researchers are registered as users; and on an average day, as many as 50 users may log on.

Originally the facility was based on a small number of high-end Unix-servers (1 with 16 CPUs and 2 with 8) and a Citrix-based remote desktop solution. This platform is currently being replaced by additional Windows servers and Windows Terminal Services. The current storage capacity on dedicated file-servers is 7 TB (for master files as well as user accounts). The system is configured to handle 50 simultaneous users.

All master files are stored as SAS-files. Conversion to other file formats as well as data preparation (subsetting, linking etc) is handled by the internal staff and users only have access to data specially prepared specifically for each research project. This policy is rooted in a strict interpretation of the need-to-know principle where individual researchers are authorized to only see data that they explicitly need to meet their research requirements.

The remote desktop provides access to standard statistical software such as SAS, SPSS and Stata and a limited number of specialized packages. Users who need access to unsupported software may bring a dedicated server with the software installed to the remote access facility. The server will be integrated in the server park and hosted by the facility. All costs related to hosted servers are covered by the users. More than 50 external servers (with a combined storage capacity of 10 TB) are currently hosted by the facility. In addition to providing access to specialized software the solution is used by research organizations that require privileged access to storage and computing capacity.

Authorization for data access is based on a two-stage process:

1. General authorization of the research organization
2. Authorization of individual research projects within the organization to obtain access to narrowly specified data.

The security arrangements are relatively weak. Access is provided from the researchers' own computers, and authentication is based on a combination of user-id, project-id and a password generated by a pin-code device (similar to the ones used by on-line banking services). User-ids are checked against the individual users' IP-address (or the IP-range of the organization).

The remote desktops have standard settings that prevent users from downloading data or printing outputs from the statistical packages. The only way to retrieve information from the facility is by mail. Researchers place tables and other output in a disclosure review outbox, after which the cleared output is emailed to researchers. There is no time-delay, and only 1 out of 10 outputs are randomly checked. There is also an upper limit of 2 MB on individual outputs.

As this control is random and-after-the-fact, it cannot claim to ensure data protection. In this sense disclosure cannot be prevented, only sanctioned. Any breach of the general disclosure rules, however, results in denial of access for the entire research organization. According to Statistics Denmark this is a strong enough incentive to establish the necessary discipline amongst the external researchers. No severe compromise of the disclosure rule has been detected to date, although a few close calls have been reported.

The argument for using random and after-the-fact output control is mainly that the users would not accept the delays related to a system based on a more complete and rigorous control system. Statistics Denmark is also concerned about the amount of time spent on output reviews.

A staff of14 manages the remote access facility.

For further information on this service, see: http://www.dst.dk/forskning

Statistics Sweden

The remote access facility at Statistics Sweden (Microdata Online Access, MONA) was established in 2005 and is to a large extent modelled on the Danish system. Statistics Sweden has no on-site research centre, so MONA is providing the only access route to confidential microdata for Swedish researchers. There are however a few major research organizations in Sweden that have been authorized to host microdata on their own local servers. It is the intention of the statistical office to migrate these users to MONA as soon as their current agreement runs out.

The service provides access to survey data as well as data from statistical registries and currently includes 350 registered users.

The system is based on 4 application servers and 1 fileserver, all Windows-based. The file server has a storage capacity of 5 TB and is used for master files as well as user accounts. The system is configured as a standard remote desktop solution based on Windows Terminal Services.

As in Denmark, data preparation and conversion is handled by the unit, and only specially prepared and project specific datasets are made accessible to the users. Currently the following software systems are supported on the remote desktop: SAS, SPSS, STATA, GAUSS, Super Cross, SQL Query Analyzer, Excel, Word and RAPS. Unsupported software can be included provided that the user covers the license costs.

Authorization, authentication and output control is more or less identical to the Danish system. This is largely due to the comparatively liberal data protection legislation in the Scandinavian countries, as well as a culture based on trust.

The unit managing the remote access facility currently has a staff of 6 full-time employees. Preparation of data and specialized data consultancy services are however provided by the various productions units within Statistics Sweden.

For further information on this service, see:
http://www.scb.se/Pages/List____257147.aspx

Statistics Netherlands

Statistics Netherlands established a remote access facility in 2006 as a parallel offering to their on-site research centre. The remote access facility currently has approximately 200 users, compared to 130 users of its on-site service.

The remote access facility hosts about 400 well documented datasets. The majority of data comes from sample surveys, only 10 percent of which derives from statistical registries.

The solution is Citrix based and runs on 16 identical Windows servers, each configured to handle 8 simultaneous users. A load balance mechanism allocates users across the available hardware resources. Statistics Netherlands holds that many small servers provide a more robust and scalable solution than a system based on fewer large servers (as in Denmark). The facility also has a file-server that houses all the master files and user account information. The following software packages are available for all users: SPSS, Stata, StatTransfer, MS Office, Blaise or WinEdit. SAS, Ox and Gauss are available as a chargeable service.

In contrast to the practice in Denmark and Sweden all data conversion and data preparation work in Statistics Netherlands is handled by the users. The users obtain access to complete copies of the master files. The need-to-know principle applies less in the Netherlands compared to the Scandinavian countries.

The authorization, authentication and output control mechanisms are however significantly stronger than in Denmark and Sweden.

The facility can only be accessed through dedicated computers (typically one) at each local research organization. These computers as well as their location are checked and authorized by a Statistics Netherlands staff member. Typically these computers are located in a locked room with no other activities or functions. About 40 computers across the Netherlands are currently authorized for remote access.

Authentication is handled by a fingerprint reader installed by Statistics Netherlands on all authorized remote access computers. The user must re-authenticate every 30 minute to keep the connection open. The cost of the fingerprint reader and its installation is covered by the research institute.

In contrast to Denmark and Sweden, Statistics Netherlands is controlling all outputs submitted for outbox clearance by users. When output is submitted, it is removed from the outbox within 5 minutes and sent to the remote access unit for control. The disclosure review is conducted by Statistics Netherlands a staff member with specialized knowledge and experience working with the datasets (approximately 40 different staff members). The control is thorough and requires one man-hour per output on average.

To reduce the number of outputs to a manageable number, only the first 4 outputs of each research project is free. All additional outputs are subject to fees. The users also are encouraged to only submit tables and results that will be used in their research publications. Currently only 600-700 outputs are reviewed per year adding up to approximately half a man-year of control activities. The average duration from output request to delivery is 5 working days.

The unit managing the remote access facility has currently a staff of 16 full-time employees. This includes 6 people focused specifically on data preparation and documentation.

For more information on this service, see:
http://www.cbs.nl/en-GB/menu/informatie/onderzoekers/microdatabestanden/algemeen/default.htm

NORC at the University of Chicago (USA)

Although public use data can be disseminated through a number of established commercial and academic archives, to date there is a more limited range of options available to other entities seeking to disseminate sensitive microdata that have not been fully de-identified for public use. While some of the largest federal statistical agencies (e.g. U.S. Census Bureau and Bureau of Labor Statistics) have sufficient economies of scale to develop advanced in-house solutions that serve the needs of external researchers, smaller data producers often lack the resources to archive, curate, and disseminate the datasets that they have collected.

NORC at the University of Chicago created a secure remote data access facility in 2006 to respond directly to this need. In conjunction with the National Institute of Standards and Technology (NIST) Technology Innovation Program (TIP), this initiative represents one of the first attempts in the U.S. to provide secure, remote access to confidential microdata collected and managed by federal statistical agencies and other data producers. It combines elements from the computing and social sciences that not only provide technical security, but also create an environment whereby researchers can conduct high-quality research. In particular, this facilitates cutting edge metadata documentation best practices and research dissemination, and helps demonstrate the benefits of researcher access to the producer in addition to helping meet the replication standard.

The NORC Model espouses a portfolio protection approach to data access that includes statistical protection (mainly deleting obvious identifiers), screening of researchers, training researchers in legal and ethical confidentiality requirements, and both secure onsite and remote access. The protocols include high level technical security which has certified by NIST, as well as by other federal agencies for which NORC collects data such as BLS, NSF, IRS and HHS. Other protocols include: a review process and legal agreements to ensure that only authorized researchers from approved institutions access data; audit logs and audit trails to monitor research behaviour during data access; and full disclosure review of statistical results before they are permitted to leave the secure environment.

One of the key, distinguishing features of the enclave is that it offers an e-collaborative environment within which researchers can share knowledge (code, scripts, macros) about the data and hence provide information to fellow researchers as well as back to the data producers that can be used in archiving and curating the data. The research environment, or collaboratory space, features discussion forums, wikis and blogs and instant messaging (IM). The enclave also engages the research community by developing a knowledge infrastructure around each available dataset through its research collaboratory, which enables geographically dispersed researchers to collaborate and share information. Not only does this promote high quality research, but it also increases meaningful interaction among producers, researchers, and data custodians. The creation of a research community around important datasets is the first step in developing a core body of knowledge about the data thereby fostering the replication of research and ensuring that research based on the data are acceptable to academic journals and policy makers. Indeed, creating a research community involves creating a collaborative environment within which ideas, code and results can be exchanged. In addition to providing collaborative tools, Data Enclave staff produces detailed, Data Documentation Initiative (DDI) compliant, metadata documentation specific to each dataset. Another important guiding principle of the NORC data enclave is its emphasis on obtaining feedback and input from researchers. Datasets should not be static in nature, but should rather be set up to adapt and change in response to evolving research questions and needs. In order to facilitate this, Data Enclave staff work with the data custodian to promote the interaction between producers and researchers that creates the healthy survey lifecycle.

The research community is further developed by means of a rigorous, remote researcher training program. This serves several purposes. First, researchers are trained in confidentiality protection, so individual respondents will not be re-identified. Second, they learn about the sampling frame, questionnaire design, weighting and other dataset specifics from the data producer. Third, they meet each other in a group environment, which has been shown to be critical to building trust and collaboration. Finally, as a condition of access, the researchers are required to demonstrate that they are serving data producers' missions. They can do this by enhancing the database infrastructure by providing detailed metadata documentation, adding information or data to the survey, or providing their code to the agency and subsequent researchers. Researchers also are required to provide their research output for dissemination by the agency, as well as evaluation and feedback of the survey.

Currently, there are six data producers in the Data Enclave, three federal statistical agencies including the U.S. Departments of Commerce (NIST), Energy, and Agriculture and three foundations including the Ewing Marion Kauffman Foundation, Annie E. Casey Foundation, and the National Science Foundation. Approximately 100 approved users conduct research in the enclave, conducting analyses on issues such as U.S. businesses and communities, children, youth and family issues, agricultural economics, educational trends and policy, and energy consumption. Examples of researcher topics include entrepreneurship, joint ventures and strategic alliances, innovation process, start-ups and new business characteristics, founding partnerships, capital structure, minority and woman owned firms, agricultural economics, location effects, firm performance outcomes, and intellectual property. Researchers accessing educational microdata focus on issues related to tracking the employment history and research productivity of members of the science, engineering and health doctoral labour force as they move through their careers in research and practice as well as educational histories, funding sources, and postdoctoral plans from all recipients of research doctorates earned from U.S. institutions.

The NORC Data Enclave currently operates on 8 Dell PowerEdge 1855 or 1955 blade servers configured with 8-32Gb of memory, 4 virtual servers (using VMWare ESXi) and a NetApp StoreVault S500 network attached storage. All servers are connected over a gigabit network. A diagram of the architecture is available in Annex 3: NORC SRA Diagram (p.109). Users in the environment have access to statistical packages such as SAS, Stata, R, LimDep/NLogit, LISREL, and Matlab, along with Microsoft Office 2007 suite.

For further information, see http://www.norc.org/DataEnclave

UK Data Archive Secure Data Service (SDS)

The UKDA Secure Data Service is a new service funded by the ESRC to allow controlled restricted access procedures for making more detailed microdata files available to some users (Approved Researchers), subject to conditions of eligibility, purpose of use, security procedures, and other features associated with access to the SDS data.

Building on the success of other secure data enclaves worldwide, and employing security technologies used by the military and banking sectors, the SDS will allow trained researchers to remotely access data which is held securely on central SDS servers at the UK Data Archive. The aim of the service is to provide approved academics unprecedented access to valuable data for research from their home institutions, with all of the necessary safeguards to ensure that data are held, accessed and handled securely.

The SDS follows a model which suggests that the safe use of data should cover the elements of safe project, safe people, safe setting and safe output (Ritchie, 2006. see Figure 1). In order to achieve this goal, data security depends on several factors, including technical, legal, contractual, and educational.

The technical model that has emerged is one which shares many similarities with both the ONS VML[13] and the NORC Secure Data Enclave[14]. It is based around a Citrix infrastructure which turns the end user's computer into a 'remote terminal' giving access to data, statistical software, and collaboratory spaces on a central secure server held within the UK Data Archive. The system is flexible in that, depending upon the wishes of the data custodian, access can be restricted to particular users (safe people) and/or particular locations (safe rooms/machines). It maintains security in that all data manipulation occurs on the server, which is maintained to very strict security protocols.

Beyond the general security policy, the secure server itself will be subject to additional security measures and controls. Approved researchers will access the

[13] http://www.ons.gov.uk/about/who-we-are/our-services/unpublished-data/business-data/vml/index.html

[14] http://www.norc.org/DataEnclave

proposed SDS by using VPN (Virtual Private Network/thin-client) technology, which encrypts the data transmitted between the researcher's computer and the host network. Other components of the VPN technology allow control to be established over which network resources the external researcher can access on the host network. The service will employ a Citrix XenApp server farm, which participates on two networks.

The Approved Researcher logs onto the SDS system remotely via a web secure (HTTPS) browser. All data processing is carried out on a central secure server, which processes all requests centrally and returns information about the results. With this technology, although all applications (SPSS, STATA, etc) and data run on a central server at the UKDA/SDS, the Approved Researcher still interacts with a full Windows graphical user interface. This means that the researcher never has to install any complex applications on his/her remote computer – the only application required by the Approved Researcher is a web browser.

Users of the SDS will be required to be either "ONS Approved Researchers"[15] or "ESRC Accredited Researchers". The first of these is defined by the Statistics and Registration Services Act 2007 as "an individual to whom the Board has granted access, for the purposes of statistical research, to personal information held by it."[16]
An "ESRC Accredited Researcher", will have a similar status to an ONS Approved Researcher, i.e. a person who has been granted access for the purposes of statistical research to personal information which has been licensed to the ESDS/UK Data Archive[17] University of Essex for dissemination on behalf of a government department or some other data provider. Neither of these two types of users will be able to use the SDS without appropriate training. Mandatory training will allow the UKDA to ensure that end-users are fully aware of any penalties which they might incur if they cause a breach.

The SDS disclosure staff will divide the outputs from SDS into three main categories:

- Safe: No risk / very low risk of disclosure – output will be released promptly

- Uncertain outputs: Low or medium risk of disclosure – output will be considered carefully, with some dialogue with the researcher as necessary, perhaps to collapse categories, remove one or more variables or suppress some cells

- Unsafe: High risk of disclosure – output will be blocked in its current form and won't be released. This is the responsibility of the researcher to produce safe outputs and demonstrate that they are free from the disclosure risks.

The SDS will benefit users in:

[15] http://www.data-archive.ac.uk/orderingData/agreements/ARFormsandNotes.doc

[16] Statistics and Registration Services Act 2007 § 39 (5).

[17] http://www.esds.ac.uk/aandp/access/licence.asp

- Ability to work in their own private work areas or in shared areas with other approved researchers
- Access to enhanced, highly sensitive available data storage in tandem with the related metadata through increased capacity and environmental protection
- Possibility of data linkage exercise with using existing data in the UKDA or other administrative data
- Collaborative functionality including survey and document library, SPSS/ STATA code library, knowledge repository, disclosure review and technical assistance
- Flexibility, access can be restricted to particular users (safe people) and/or particular locations (safe rooms/machines)
- A self-contained secure 'home away from home' service with familiar analytical environments
- Capability for growth and expansion
- A consolidated environment built from the ground up with security and data protection in mind
- Server management processes including auditing, change control, monitoring and alarm notification

This service is to operate fully in autumn 2009. For more information please contact securedata@ukda.ac.uk

Additional reviews

A 2003 study on remote access facilities[18] was conducted by Sandra Rowland on behalf of the Committee on National Statistics of the National Academy of Sciences (NAS). The paper was presented at the Access to Research Data: Assessing Risks and Opportunities NAS workshop October 16-17, 2003. It covers some of the examples presented above along with other systems such as the Luxembourg Income Study, Australian Bureau of Statistics, monitored remote access facilities in US Federal Agencies, and research projects in the United States. While now slightly outdated, the paper provides valuable insight on different approaches and architectures.

Recommended Options

While all the above models use a remote access technology, they illustrate various architectural and organizational approaches.

An important decision host institutions must weigh when deciding whether or not or how to provide access to disclosive data is the degree to which users will have access to the data (i.e., full or partial). Whenever possible, the authors recommend the open model whereby the researcher has access to the full dataset and has the freedom to subset variables and observations. This model also carries with it several advantages:

[18] 1. S. Rowland, “An examination of monitored, remote microdata access systems,” in NAS Workshop on Access to Research Data: Assessing Risks and Opportunities, October, 2003.

- It provides a flexible research environment where users can explore the data and shape them to meet their requirements. This is often an exploratory process and needs often change during the analytical processes.
- It greatly reduces the burden on the SRA facility staff as customizing dataset for each research project is resource intensive
- It is essential to provide an effective collaborative environment. Restricting access to data subsets also restricts the collaborative space as users can only communicate with others entitled to see the same information.
- Rather then investing in restricting the inputs, the Netherlands and NORC models demonstrate that controlling the outputs is more cost effective.

The models we describe above also illustrate the importance of a flexible server side architecture that can scale to the users or institutional demand and adapt to customer needs. The recent emergence of platform virtualization technologies effectively meets these needs. Solutions that show particular promise include facilities like the NORC Data Enclave and the UK Data Archive SDS that take advantage of such technologies and follow IT industry standard practices. We strongly recommend that future architectures likewise leverage knowledge that comprises these approaches, through a well balanced mix of physical and virtual servers.

Alternative Approaches

As previously noted, secure remote access is not the only available option to provide virtual access to sensitive data. For example, web based analysis engines or remote execution solutions are possible. Such tools however do not provide the level of interactivity necessary for in-depth research and considerably restrict the potential for collaboration.

These types of systems may however be appropriate for audiences such as the junior researchers, or more casual users or individuals that may not qualify as accredited researchers.[19] The various approaches are not mutually exclusive; indeed, complementing a SRA with hybrid solutions may also be a good option.

Web based analysis or batch execution facilities should be fairly easy to deploy on the same infrastructure as the SRA. The publication of the data through such tools should also be greatly facilitated by the SRA ingest and documentation processes. Some potential products include:

[19] Although, the prime philosophy of the SRA is to facilitate access to disclosive data through the user's own desktop, there are situations where an approved researcher or member of the research team may not be able to meet the logistic/physical security conditions which are necessary to have access to the disclosive data (e.g., lockable office, shared offices, hot desks, etc) or the data owner would not permit access to their data through the user's desktop. Under these circumstances an SRA safe room at local establishment would bridge the gap between protecting data confidentiality and maximizing data access, where there are limitations on the user's side or restrictions imposed by the data owners.

- *Space Time Research*[20]: This Australian company provides a suite of products (SuperView, SuperWeb, SuperCross, and SuperStar) that delivers high performance dynamic data processing and visualization tools on the web or the desktop. It is particularly well suited for microdata and facilitates ad hoc anonymization procedures through the integration of the ESSnet sponsored tau-argus tools[21] or custom programming.

- *Nesstar*[22]: As a product that is already in use amongst the CESSDA community, Nesstar is a software system for data publishing and online analysis. The product compliance with the DDI specification makes it a particularly attractive option.

- *Berkeley SDA*[23]: Survey Documentation and Analysis (SDA) is a set of programs for the documentation and Web-based analysis of survey data. The platform is developed and maintained by the Computer-assisted Survey Methods Program (CSM) at the University of California, Berkeley (the group that also developed the CASES software package). The latest release of the software will include some disclosure control capabilities[24]. The product is fairly inexpensive and can read DDI using a conversion tool to the internal Data Description Language (DDL).

- Exanda: An open source web based tabulation and visualization engine (for up to three variables) under development at GESIS[25] in Germany. This tool is based on the DDI 3 metadata specification. Further information should be become available later this year.

- *Josua*[26]: Developed by the International Data Service Center at IZA, Germany, JoSua is an instrument for controlled remote data processing (Job Submission Application). Originally designed to provide international researchers access to German labour market data JoSuA, it has matured into a flexible data analysis tool offering a considerable degree of automation designed to meet the individual needs and specifications of each individual data provider. It is available as service or for deployment at data providers' sites.

- *Webtab*[27]*:* An on-line tabulation service offered by Luxembourg Income Study that supplements an existing job submission and remote execution service. Webtab was released in august 2009 and include crude disclosure control by simply suppressing table-cells with less than 15 observations.

[20] http://www.str.com.au/

[21] http://neon.vb.cbs.nl/casc/

[22] http://www.nesstar.com

[23] http://sda.berkeley.edu/

[24] http://sda.berkeley.edu/man34h/disclosure.htm

[25] For further information, contact Joachim Wackerow <wackerow@zuma-mannheim.de>

[26] http://idsc.iza.org/index.php?page=4

[27] http://www.lisproject.org/web-tabulator/web-tabulator.htm

Technological requirements

Overview of Infrastructure

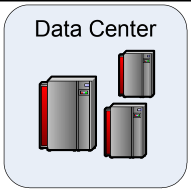
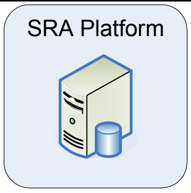
The overall technical infrastructure to support SRA facilities includes the following components:

- *Remote access technology*: to provide secure access to the facility (based on the Citrix platform or similar technologies)
- *Servers*: to deliver and support the terminal, data and collaborative services
- *Clients hardware*: for the users to run the terminal services and access SRA facilities
- *Network*: to ensure the user connectivity and information exchanges between facilities or agencies
- *Storage*: for master files, user space, backups, etc.
- *Software*: for the operating system, office operations, statistical analysis, data/metadata management, collaboration/knowledge management
- *Security components*: for authentication, access control, encryption, backup and disaster recovery, room security, etc.

Configurations

SRA configurations include three parts:

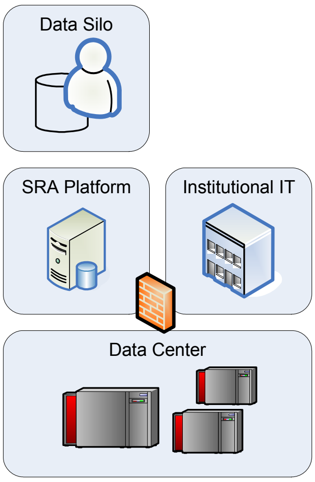
- the data centre providing the core IT platform
- the remote access platform
- the virtual data silo providing access to surveys and registries from a data source or provider

	The data centre is the traditional IT infrastructure providing the core hardware, system software, network connectivity and security. This can be a dedicated environment but is often used for other institutional purposes.
	The remote access platform is essentially the hardware, software, and other components dedicated to the SRA facility. This may include servers, firewalls, encryption products, virtualization platforms (like Citrix XenApp), and customized software licensed for users and administrators, etc.
	The data silo is essentially the space visible to a user when logged into the system. This may include a data collection from a specific provider or multiple survey collections. Accessible to authorized users, this is the area in which researchers work with the underlying microdata and communicate with each other, essentially a closed community around the data in the silo.

Of particular relevance to CESSDA, any of these three components can be setup and managed independently, and may be configured to realize various cost efficiencies, regardless of implementation model, i.e.

- data centres can be shared
- multiple remote access platforms can be hosted in single data centre and
- many data silos can be deployed in a single remote access platform

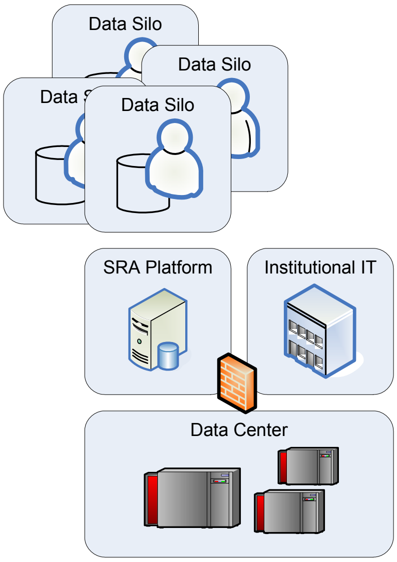
The stand alone model



The simplest version of a SRA facility is the stand alone model. A single institution hosts the entire solution providing access to its own data.

In this case, the data centre infrastructure is typically also used for other institutional IT needs. For security reasons, security layers need to be in place to fully isolate the secure remote access facility from the institutional users or applications. This configuration is a typical out-of-the box installation.

The shared remote access platform model

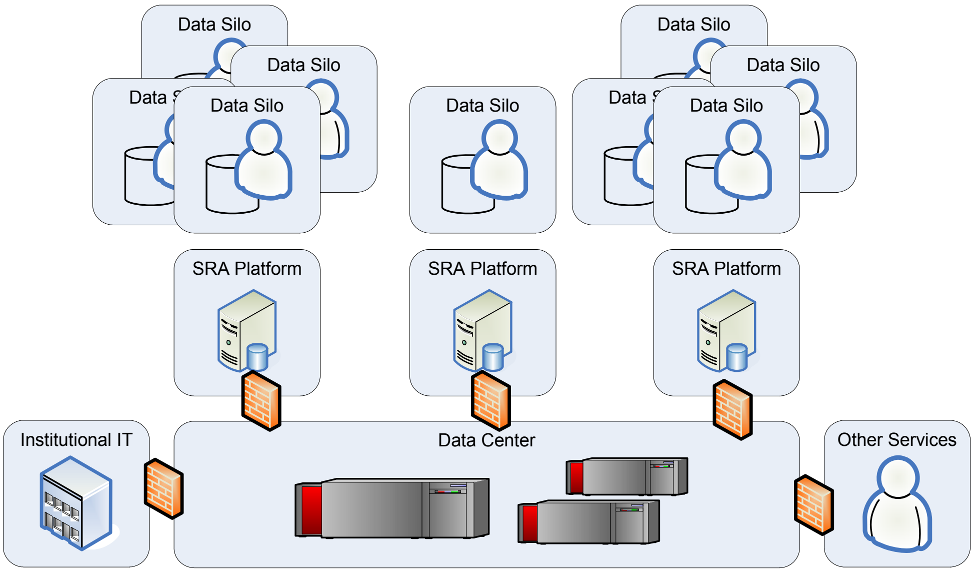


The shared remote access platform is a natural extension of the stand alone facility. The host institution essentially takes advantage of its existing IT infrastructure and expertise to provide secure remote access solutions to other data providers (and to itself if applicable).

This model provides several advantages as it alleviates the need for a data provider the solution. It also can greatly reduce overall costs by sharing the data centre, SRA platform, ICT expertise, data and metadata management services, user training & support, etc.

This is the model used by the UK Data Archive SDS and NORC and may work well as a configuration option for CESSDA.

The shared data centre model



The shared data centre model is an attempt to maximize cost effectiveness by having several remote access platforms, potentially managed by different organizations, hosted on a common IT infrastructure. The IT expertise and related core services essentially are managed at a single facility.

This infrastructure also can be used for institutional purposes and to provide access to other general public or private services (web services, registries, repositories, etc.). This may be particularly attractive for CESSDA to consider in that it could provide data centres to several agencies interested in hosting an SRA, possibly spanning multiple countries, while at the same time delivering other secure services.

User Remote Access Location

Technically, anyone who is considered an authorized user (approved researcher, system administrators, managers, data producers, etc.) may connect to the SRA facility from anywhere in the world, but this is generally less desirable when working with potentially disclosive data. Both the environment and the geographic locations play an important role in determining accessibility and user rights. Technical and legal aspects are discussed in greater detail later in this document

Readers should note that the term “node” or “terminal” below refers to any computer (desktop, laptop, or thin client) equipped with the remote access software and optional security hardware.

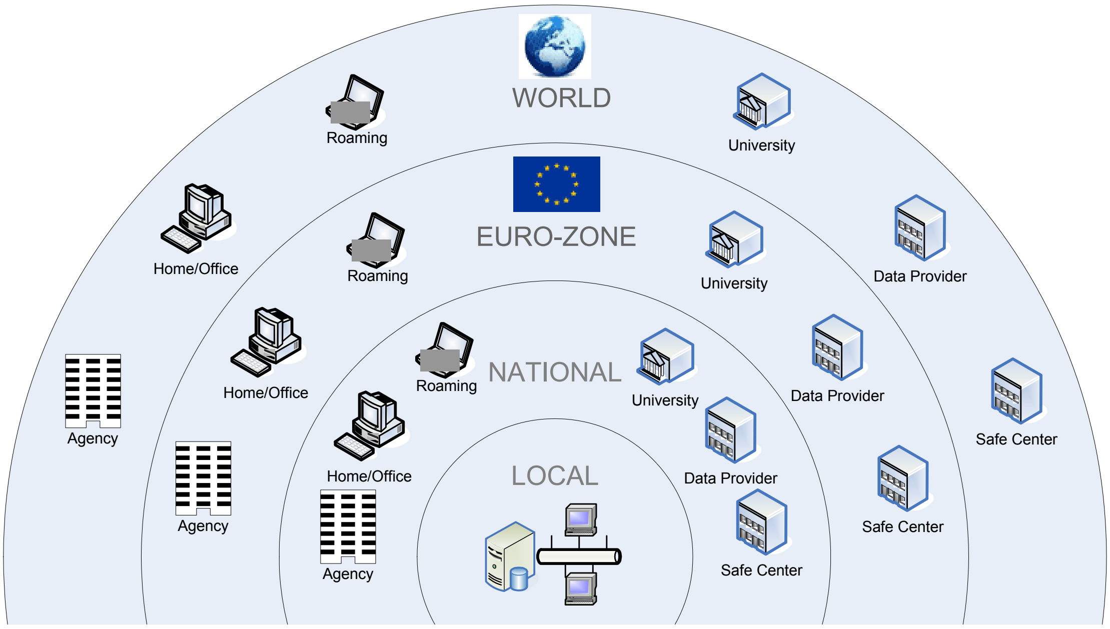
The environment within which the user accesses the SRA includes the following components:

- The *SRA facility:* nodes are located at the data centre of the hosting institution. System administrators and SRA managers typically access from this environment, but it is not as safe as other environments such as access controlled rooms or onsite research data centres. Note that local institutional users are typically considered regular researchers.

- The *data producer* site: the SRA will often provide special access rights and conditions to the data producer. These users are considered safe and responsible but the security of the environment may vary based on the institutional capacity.
- A *safe institution* that provides access to potentially disclosive data or has made special data sharing arrangements to provide access to the SRA facility. This for example may include a research data centre or an agency with a safe access room. Such an environment is usually considered to be highly secure and is staffed with responsible individuals trained in legal and confidentiality issues.
- A *regular institution* such as a university, governmental agency, or research institute. The access point can be located in a dedicated room or in a regular office. Such an environment provides limited control, which may require the host to add security features into the remote terminal or the safe room. Local IT staff may be available to respond to technical questions generally are not proficient in legal or data confidentiality issues
- The *personal office or home* is the stand alone researcher accessing from personal office space or from home. Limited or no technical support is available locally, and this is not considered a safe environment.
- The *roaming user* is the researcher who is working remotely from a laptop and does not work from a specific location. This type of access is particularly difficult to control and should only be offered to well trained and trusted users.

The geographic location where the data access node resides is generally considered the centre point of the legal framework under which users are bound. In the case of CESSDA, we can break this down into:

- *National*: the location/user is the country of origin of the data available in the SRA silo.
- *Euro zone*: the location/user is in a country that falls into the European zone (and regulations)
- *Non-Euro zone*: the location/user is in a country outside the European zone (and regulations). This could be further broken down based on the existence of multi-lateral or bi-lateral agreements.
- *Anywhere*: this is the case of the roaming user that can access from any location. The remote access facility should in this case use various techniques to determine the appropriate user category. If it can’t be determined with a high degree of certainty, access can simply be denied (or significantly restricted).



When establishing a SRA facility, it is important to weigh the advantages and disadvantages, at the data silo level, of the various access options. Ideally, such policies should be harmonized across data providers and SRA.

Data silo configuration example:

	National	Euro-Zone	World
SRA Staff	Authorized from SRA office computers or admin laptops.	Authorized from admin laptops	Not authorized
Data provider	Authorized from selected office computers.	n/a	n/a
Safe Institution	Authorized from local workstations	Authorized from SRA approved workstations	Authorized with special agreement and using SRA approved workstations. Requires secure room.
Regular Institution	Authorized from selected office computers or SRA approved workstations	Authorized from SRA approved workstations. Requires safe room.	Not authorized
Home or personal office	Authorized on personal desktop or SRA approved workstation. Personal desktop requires installation of SRA security software/hardware.	Authorized from SRA approved workstations	Not authorized
Roaming user	May be authorized on a case by case basis. Laptop requires installation of SRA security software/hardware.	Not authorized	Not authorized

Cross-national configurations

It is important for CESSDA to note any of the models presented above can technically be implemented across national boundaries:

- A data silo could contain data from different countries
- An SRA facility could host data silos from other countries
- A data centre could be shared by multiple countries

The first option, using a single facility for accessing data from multiple countries, is highly attractive for researchers in that it facilitates cross country analyses. The other two options would be highly beneficial to countries or agencies that do not have the capacity to host their own facilities or sustain the necessary IT infrastructure.

While these options provide advantages such as reducing operational costs, facilitating management, harmonization services, and others, they also include a number of challenges such as issues related the storage of confidential data in foreign countries, complexities in data disclosure procedures, or system ownership and management. These issues are discussed in greater detail later in the report (see “Sharing data across borders / legal aspects” on p. 78)

Despite these challenges, CESSDA is remarkably well positioned to lead the effort to identify common ground, and, on a case by case basis, provide cross-national cost effective solutions for remote access to sensitive data. A solution might include providing remote access to European surveys or harmonized cross country datasets, much to the benefit of CESSDA’s member organizations, data providers and researchers.

Remote Access Technology: Citrix XenApp

Selecting a platform

Selecting the most appropriate platform to provide remote access services is critical. Decision criteria include: performance, security, scalability, cost, industry recognition, long term sustainability, technical support, community support, proven technology, etc.

The commercial market space provides several options like Citrix XenApp[28], Microsoft Terminal Services[29], Quest Software vWorkspace[30], Ericom Software PowerTerm WebConnect[31]. Citrix XenApp (previously knows as Citrix Presentation Server) however has for several years been the market leader and is often considered a de facto industry standard. XenApp has faced limited competition in the thin client computing space. Though Microsoft’s new Terminal Services for Windows Server 2008 is emerging as a potential competitor to XenApp, this is limited to the small and medium business market space and therefore is not yet considered a solid alternative.

XenApp provides a cost effective, robust and scalable platform. It is aligned on several security standards and best practices. For example, it is EAL 2 certified[32] and

[28] http://www.citrix.com/english/ps2/products/product.asp?contentid=186

[29] http://www.microsoft.com/windowsserver2008/en/us/rds-product-home.aspx

[30] http://www.vworkspace.com/

[31] http://www.ericom.com/ptj.asp

[32] http://en.wikipedia.org/wiki/Evaluation_Assurance_Level

a Common Criteria Report[33] was prepared by the UK Certification Body, CESG, using the UK IT Security Evaluation and Certification Scheme.

Some academic institutions use XenApp to provide students and faculty remote access to applications running on campus servers. Similarly, because of the higher degree of security built into Citrix and other remote access platforms, various military branches extensively use XenApp to offer data access to sensitive military systems from a remote location for deployed and remote personnel. The same security argument is a driving factor behind Citrix's success in healthcare, financial, government and other services which are regulated industries in which customer data security and protection is paramount. Several illustrative use cases are presented in Annex (p. 102) with more available on the Citrix web site[34]. As discussed in the previous section, Citrix XenApp has also been proven a successful platform in implementing several SRA facilities (See "Example of existing remote access facilities", p. 21).

We therefore strongly recommend and repeatedly refer to Citrix XenApp as a critical component of any solution as CESSDA work toward a European wide remote data access system. .

What is Citrix XenApp

Citrix XenApp (formerly Citrix MetaFrame Server and Citrix Presentation Server) is an application virtualization/application delivery product that allows users to connect to their corporate applications. XenApp can either host applications on central servers and allow users to interact with them remotely, or stream and deliver them to user devices for local execution.

How XenApp works

Utilizing integrated application virtualization technology, XenApp isolates applications from the underlying operating system and from other applications to increase compatibility and manageability. Applications are streamed from a centralized location into an isolation environment on the target device where they execute. The target device can be a user PC or a server in the data centre.

Citrix XenApp is unique in that it is a complete application delivery system, offering both online and offline application access through a combination of application hosting and application streaming directly to user devices. For the implementation of a SRA service, we are only interested in the hosted application as information should not be permitted to be exported from the secure server.

Hosted application delivery uses application streaming to deliver applications to hosting servers in the data centre. XenApp then connects the user to the server to

[33] http://www.citrix.com/English/SS/supportThird.asp?slID=162512&tlID=162515
[34] http://www.citrix.com/lang/English/ps2/segments/index.asp

which the application has been delivered. *The application executes entirely on this server*. The user interacts with the application remotely by sending mouse-clicks and keystrokes to the server. The server then responds by sending screen updates back to the user's device.

User interaction with the application is seamless. Printers, drives, peripherals and even the clipboard work in the exact same manner as if the application were installed locally. Hosted application delivery via XenApp allows any user on any operating system to access any application. XenApp enables Windows, Mac, Linux, UNIX, Thin clients, iPhone, Windows Mobile devices, and even Symbian and Java-enabled devices to run any Windows or UNIX-based applications using hosted application delivery. To the user, the application would appear as if it was installed and running on their computer (seamless desktop integration), whereas in reality, the application is running on a server in a corporate environment.

Hosted application delivery like that available in Citrix XenApp and Microsoft Terminal Services are reminiscent of the mainframe-terminal system, where a central powerful computer does most of the processing work and smaller, much less powerful machines provide the user interface.

Citrix XenApp Products

For the purpose of implementing secure remote access capabilities, the following Citrix products are relevant:

- XenApp Platinum Edition
- Access Gateway Enterprise Edition[35]

These include features such as SmartAccess[36], SmartAuditor[37] or Password Manager[38], and others.

In addition, several other partner products can be used for increased security or integration purposes. For example:

- Ping Identify as a SAML bridge to Shibboleth
- Thin clients from IGEL[39], WYSE[40], VLX[41] or other think clients
- Imprivata OneSign platform[42]
- Token keys like RSA SecurID[43] or other one time password generator

See the Citrix Ready[44] web site for further information and options

[35] http://www.citrix.com/English/ps2/products/feature.asp?contentID=26144
[36] http://www.citrix.com/English/ps2/products/subfeature.asp?contentID=163990
[37] http://www.citrix.com/English/ps2/products/subfeature.asp?contentID=682169
[38] http://www.citrix.com/english/ps2/products/product.asp?contentID=7181
[39] http://www.igel.de/
[40] http://www.wyse.com/products/hardware/thinclients/index.asp
[41] http://www.vxl.net/citrix/vxl_&_citrix_essential.html
[42] http://www.imprivata.com/onesign_platform
[43] http://www.rsa.com/node.aspx?id=1156

SRA Specific Environment Configuration

Some specific configuration steps must also be taken to customize the environment to the SRA specific needs. This is necessary to ensure that the system is properly isolated from the outside world as well as restricting the user actions that can be performed within the environment. To configure appropriately:

- Close all connectivity to the outside world (like Internet)
- Tighten up computer and user security policies to prevent certain user operations (lock redirections, prevent changes to taskbar, disable active desktop, lock in the start menu and remove admin functions, etc.)
- Use custom login and logout scripts (i.e. to map network drives, create shortcuts, register software licenses, configure software specific features, disable auto-update features, customize the user shell, etc.)
- Rather than providing access to all software through a full desktop, it might be useful in some case to deliver selected packages outside the desktop as individually hosted applications
- Force the user to logout of the environment to regain control of the local machine
- Disable all email functionalities
- Establish secure mechanisms/procedures to move file in and out of the system
- Establish procedures for offline software updates
- Configure disk space for user, team, producer, SRA archive

Requirements summary

☑ **Citrix product configuration and licensing**

☑ **Citrix partner product configuration and licensing**

☑ **Citrix Environment custom configuration**

Server Hardware and configuration

Platform Virtualization

Given the need for a flexible and scalable architecture, we highly recommend operating both physical and virtual servers.

Platform virtualization[45] is performed on a given hardware platform by host software (a control program), which creates a simulated computer environment, a virtual machine, for its guest software. In case of server consolidation, many small physical servers are replaced by one larger physical server, to increase the utilization of costly hardware resources such as CPUs. Although hardware is consolidated, typically operating systems (OSs) are not. Instead, each OS running on a physical server becomes converted to a distinct OS running inside a virtual machine. The large server can "host" many such "guest" virtual machines.

[44] http://www.citrix.com/ready

[45] http://en.wikipedia.org/wiki/Platform_virtualization

A virtual machine can be more easily controlled and audited from the outside than a physical one, and is more flexibly configured. A new virtual machine can be provisioned as needed without the need for an up-front hardware purchase. Also, a virtual machine can easily be relocated from one physical machine to another as needed. Because of the ease of relocating, virtual machines may also be useful in disaster recovery scenarios.

There are several available solutions for platform virtualization. We particularly recommend:

- *VMWare*[46]: an industry leader in virtualization technologies. This is for example the solution that has been chosen by the NORC Data Enclave and the UK Data Archive Secure Data Services.
- *Citrix XenServer*[47]: a solid choice for virtualization provided by the same company that produces the remote access platform (single vendor).
- *Microsoft Hyper-V*[48]*: a* virtualization technology that now comes standard in the Enterprise or Data Center version of Microsoft Windows Server 2008:

The authors of this report highly recommend implementing a perfectly harmonized virtualization solution across the entire CESSDA network.

Server Hardware

The servers' characteristics can vary considerably based on the number of concurrent users to support for the service. A virtualization based approach will often call for high end servers used to host the virtual servers. In general, system characteristics will need to be discussed with the platform vendor to determine the optimal configuration based on the number of concurrent users, the application to be hosted and the size and complexity of the datasets.

Scalability Issues

The server configuration is an essential factor that impacts the scalability of the architecture. The virtualization approach provides the necessary flexibility to allow the system to grow or shrink on demand to meet users' needs. This is however limited somewhat by the physical capacity of the underlying hardware that must be carefully planned to ensure that it can support the demand. Growing the core platform however likewise has little impact on the platform availability as virtualization hides this from the end user.

For the Citrix environment, a XenApp 5 Scalability Analysis[49] is available on the Citrix web site. This study outlines that a standard server with 8-16 GB of memory should be able to sustain 100+ concurrent users performing office operations. SRA users however do not fit that profile as statistical analysis put a heavier load on the memory

[46] http://www.vmware.com/

[47] www.xensource.com

[48] www.microsoft.com/windowsserver2008/en/us/hyperv.aspx

[49] http://support.citrix.com/article/ctx119108

and processors. The NORC Data Enclave experience, based on SAS usage, suggests that a similar setup can concurrently support about 30 - 40 light users (short jobs, mostly non concurrent) or 8 heavy users (long jobs, lots of variables). This however may need to be adjusted based on the behaviour of users and the size of the datasets.

CESSDA also should note that there are several options available to support power users, for example reserving the system during off-peak hours for special operations or having dedicated virtual environments instantiated on a case by case basis to avoid disrupting other users.

Requirements summary

- ☑ **Platform virtualization solution**
- ☑ **Server hardware specifications**
- ☑ **Server configuration**
- ☑ **System scalability plan**

Client hardware and configurations

To access the facility, users will need to run the terminal client software on a remote computer. The machine can take various shapes and forms (desktop, laptop, or thin client), can be managed by different parties (the user itself, a corporate IT unit, the secure data service provider, or a third party), and can operate from various geographical location or facilities (worldwide, institution, in a specific office, in a secure room, etc.). The most appropriate model to use will depend on the level of trust or in the user, legal agreements, and the data provider requirements.

Client security issues and protection options

While no hard data or information is delivered by the secure server to the terminal (only the screen “image”), security remains a primary concern and the machine is not impregnable. Various methods can be used to try to capture these images, attempt to steal the user identity or tamper with the system. While these actions would be in violation of the user agreement (and are unlikely to happen), they can potentially occur without the researcher’s awareness.

It is important to emphasize here that the risk of information leaking out of the environment remains minimal. Verifying researcher identity and making sure no unauthorized users obtain access is a more relevant concern. The simple username/password model should be reinforced with additional levels of authentication and environment / behaviour monitoring methods. The initial level of control is the researcher clearance process, legal bindings and training, but this may be insufficient. Other potential issues can then be alleviated by retaining control over the client computer operating the terminal software or by monitoring the environment.

To strengthen the solution, various layers of protection can surround the client terminal machines such as:

- *Client hardware model*: the type of computer used to host the client terminal (desktop, laptop or thin client) can make a significant difference in terms of protection.
- *System ownership*: the computer itself can be owned and managed by the user, an institution, the SRA facility, or possibly a third party that provides different level of controls over the local capabilities and security features.
- *Physical protection*: keeping the machine in a secure box, adding various theft-protection systems, or screen filters can help deter tampering attempts.
- *Monitoring and control options*: deploying self-diagnostic utilities, monitoring the room from where the client operates (locally or remotely), monitoring the user behaviour, using remote control tools and providing technical support can significantly add to the overall security of the system.
- *Machine identity*: Authenticating the remote computer using IP address, MAC address, CPU serial number and other machine signature mechanisms strengthens the integrity of the remote client.
- *Network access control*: controlling the systems that the remote computer can connect to using static routes, locking routers and permitting no DNS services, prevents the user from downloading software, accessing other web site, or sharing their screen with other users.
- *Biometric authentication*: adding user authentication mechanisms such as fingerprint, iris reads, or facial recognition further control who can access the facility.
- *User re-authentication*: requiring the user to re-identify at regular or random intervals can be an effective to prevent "session-piracy" (the user logs in and gives controls to someone else). This however needs to remain user friendly and minimize the burden on the researcher.
- Location / Proximity detection: making sure the machine is physically located where it is supposed to be using IP geolocation, GPS, WiFi, proximity keys or proximity devices are useful to reduce risk of piracy.
- Keyboard encryption: making sure keyboard keystrokes cannot be captured reduces the risk of account and password theft.
- Operating system and environment control: making sure the local user only has access to relevant applications and operates in a controlled environment reduces risk of system tampering.
- Secure / dedicated room: having the terminal hosted in a dedicated room potentially equipped with monitoring and access control mechanism is a very effective security mechanism.
- Virtual security, monitoring and support centre: taking advantage of technology to virtualize various monitoring and support resources can significantly reduce the cost and burden of maintaining the infrastructure.

These are discussed in further details in "Annex 1: Remote Client Protection" (p.95).

The option of enclosing the access node in a secure access room of course remains an attractive and, in some case, a necessary choice. The combination of a dedicated room with a highly secure station actually provides the highest level of security. What is important in this case is to consider all technological options for the monitoring and control of the room as a smart station that can provide a gateway to the facility for remote or virtual monitoring which can in turn significantly lower costs.

Sample client configurations

It is recommended that remote access facilities provide multiple clients configuration options in order to meet various conditions. The choice of which one to use or which safety features to configure will typically be based on who is using it, where it is located, configuration and maintenance costs, and other factors.

Level 0: Open Terminal	Any regular computer under the control of the user that requires users to install the Citrix software. Such a machine can technically access the service from anywhere though this can be controlled through server side policies (e.g., location, machine identity, hours of operations, behavioural control, etc.). This model is appropriate for highly trusted and responsible users.
Level 1: Corporate Terminal	A desktop or thin client computer under the control of a system administrator from the institution where the user works. The user is actually not allowed to configure the machine or install applications. Various security options may be required to be configured by the hosting agency.
Level 2: SRA Terminal	A standard desktop or thin client computer under the control of the remote access facility or contracted third party. The system is fully pre-configured and cannot be altered by the user. It could be used as a local machine, may be equipped with office applications and may have access to the Internet. Remote access allows administrator to control the system if needed. This system may be equipped with light security features but can typically be shipped and deployed by the user / local admin.
Level 3: Secure SRA terminal	A thin client computer that can only be used to access the data facility and is under the full control of the remote access facility or a contracted third party. It can only be used as a terminal and no local operations are allowed (see “Annex 1: Remote Client Protection”, p.95). Depending on the options selected, it can be shipped and deployed by the user or existing local IT administrator or may require on site installation by a certified SRA administrator (an individual with the clearance and training to do so). These systems can typically be centrally managed and monitored.
Level 4: Terminal in dedicated / secure room	Any of the level 1-3 configurations installed in a dedicated or secured room. Room monitoring and access control may be remotely or/and locally managed.

Note that the National Opinion Research Center (NORC) and Metadata Technology are currently investigating the issue of implementing secure terminals. Sample configurations are expected to become available in 2010.

Requirements summary

☑ **Specification for client configurations**
☑ **Harmonized or data silo client requirements**

Data Centre

The data centre hosting the remote access facility is a critical component to ensure the availability of the services and security of the hosted data. A data centre hosts the system servers and generally includes redundant or backup power supplies, redundant data communications connections, environmental controls (e.g., air conditioning, fire suppression) and security devices.

The centre design should follow industry standard practices to ensure continuous operations. The Telecommunication and Industry Association TIA-942[50] Data Centre Standards, for example, describes the requirements for the data centre infrastructure. The simplest is a Tier 1 data centre, which is basically a computer room, following basic guidelines for the installation of computer systems. The most stringent level is a Tier 4 data centre, which is designed to host mission critical computer systems, with fully redundant subsystems and compartmentalized security zones controlled by biometric access controls methods.

While it is unlikely that a data centre will be established for the sole purpose of deploying a secure remote access facility, it is crucial for institutions hosting such solutions to ensure that their existing centre meets the operational and security requirements.

It is important to note that the infrastructure costs and expertise required to operate a data centre along with their associated costs are significant. Sharing facilities at the institutional, national or cross-national level is therefore highly beneficial, provided that the centre meets the necessary operational and security requirements.

Data centre security

Data centre security plans should typically include the following components:

- Access control: to the data centre facility
- Environment control and monitoring: temperature, humidity, air flow, wetness/flood, etc.
- Fire suppression system
- Local and off site backups (see also Storage, p.52)
- Disaster recovery plan:

[50] See http://www.adc.com/us/en/Library/Literature/102264AE.pdf and http://www.tiaonline.org/standards/catalog/search.cfm?standards_criteria=TIA-942

Requirements summary

☑ **Specifications for data centre infrastructure**
☑ **Data Centre Security Plan**

Network

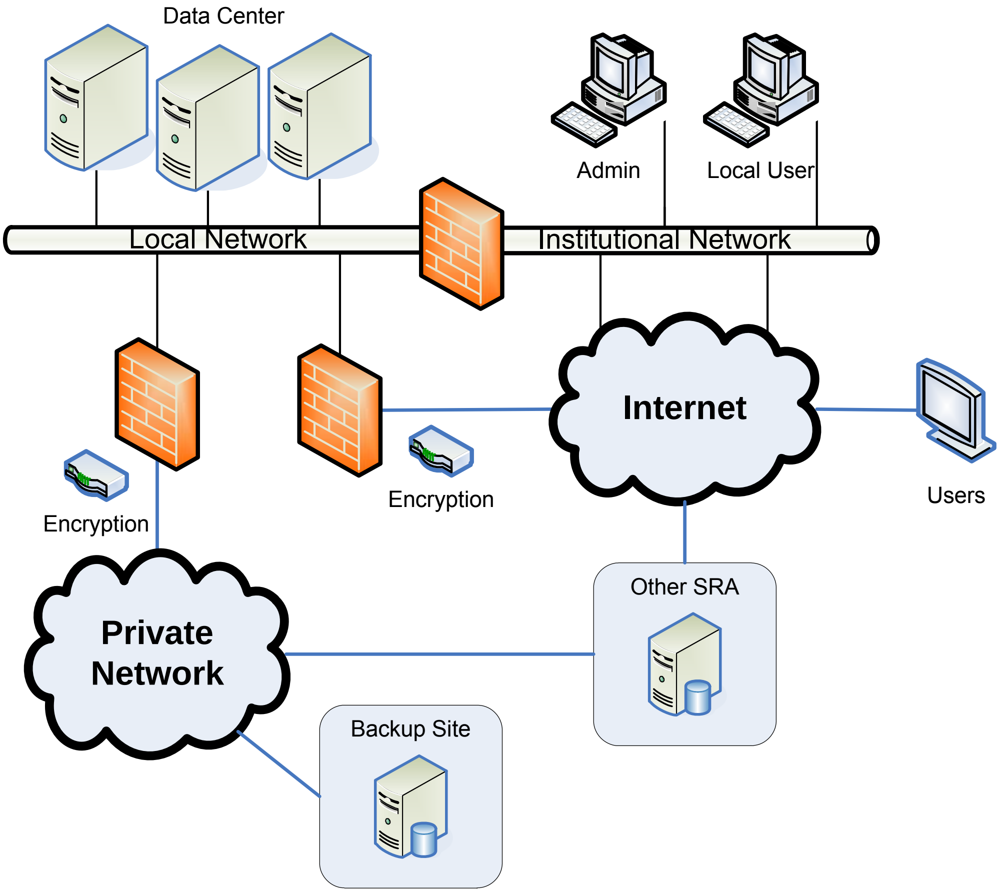
Various communication channels will be necessary for the servers to communicate with each other as well as the SRA terminals and other facilities. These can be broken down into the following categories:

- *Local network*: for communication between physical servers, network attached storage and other peripherals within the data centre. These should be implemented on a very high speed network with minimum megabit capability. In a secure data centre, this network does not require encryption.
- *Institutional network*: the data centre will typically be connected to an institutional network. This will be used by system and data/metadata administrators to access and manage the facility. Local users may also obtain access to the system for research or other technical purposes. This typically includes a secure bridge between the two local networks.
- *Private networks*: can be used to provide connections between data centres, secure remote access facilities, external backup centres and user access points (for example a research data centre). They provide an increased level of security compared to public networks such as the internet.
    - *Leased*: private lines can be purchase or rented to provide a permanent connection between locations; for high availability, dual connections can be established for redundancy.
    - *Over public networks*: running Virtual Private Networks[51] (VPN) over public networks such as the Internet, typically using encrypted channels is also a commonly used approach. This is somewhat less secure and may not provide as much control on the available bandwidth but is less expensive.

Public network: most users will connect to the SRA facility over the Internet. While Citrix does not require a large bandwidth to deliver the terminal connectivity, a stable and broadband connection is necessary on the data centre site to deliver quality services to the end user.

The local, private, institutional and public network topologies will therefore need to be well specified and documented. Relevant service level agreements must be established with network providers to ensure service quality and availability.

[51] http://en.wikipedia.org/wiki/Virtual_private_network



Network security

Protection of the communication channels and of the information exchange between the connected components is essential. The network security plan will have to include products such as firewalls, crypto capable routers or other network encryption devices, Transport Layer Security[52] (TLS) or Secure Sockets Layer (SSL) certificates for communication over the Internet. In addition, relevant network management solutions[53] and monitoring tools[54] will need to be put in place to ensure proper operation, administration, maintenance, provisioning and security of the telecommunication environment.

Requirements summary

☑ **Specifications for local area and institutional network**
☑ **Specifications for private networks**
☑ **Specifications for public networks and connectivity**
☑ **Network Security Plan**

[52] http://en.wikipedia.org/wiki/Transport_Layer_Security
[53] http://en.wikipedia.org/wiki/Network_management
[54] http://en.wikipedia.org/wiki/Network_monitoring

Software

Several software components are necessary to operate the facility, manage the environment, and deliver services to the users. This section provides a list of required or recommended software to be licensed and deployed in the environment.

Infrastructure software

- *Operating system(s)*: this is typically Microsoft Windows to ensure support for a variety of software packages but could in some cases be a Linux/Unix based solution. A virtual environment provides the flexibility to run multiple operating systems on the same physical server.
- *Virtualization platform*: such as VMWare, XenServer, or Microsoft Hypervisor
- *Web server*: Microsoft Windows server comes with Internet Information Services[55] pre-installed. The Apache HTTP server[56] however remains the most popular choice and is available for multiple platforms. This might therefore be a better choice in a harmonized environment.
- *Database server*: to support the infrastructure and application, one or more database servers will need to be deployed. The natural choices are likely Microsoft SQL server or the open source MySql server. Others such as IBM DB2 or Oracle are also excellent choices (but might be comparatively more expensive)
- *Security software*: backup / restore packages (Symantec, CA ArcServer), anti-virus and other protection software (Norton, MacAfee, CA), encryption tools (PGP), etc.
- *Management software*: hardware, network and other system management and monitoring tools (often equipment specific)

Note that most of these components typically need to be licensed per sever, whether physical or virtual.

Citrix

Citrix XenApp will need to be installed on the data centre physical servers. The product is available in four editions: XenApp Fundamentals, XenApp Advanced, XenApp Enterprise, and XenApp Platinum. Feature comparison information is available on the Citrix web site[57]. The Platinum Edition is the recommended choice. For North America, suggested retail pricing is per concurrent user (CCU) and includes one year of Subscription Advantage:

- Advanced Edition – US $350
- Enterprise Edition – US $450
- Platinum Edition – US $600

Based on selected security mechanisms and other features, various Citrix partner products also will need to be acquired.

[55] http://www.iis.net/

[56] http://httpd.apache.org/

[57] http://www.citrix.com/English/ps2/products/feature.asp?contentID=1686588

Office productivity software

All users accessing the environment will likely need a standard productivity suite to manage documents, spreadsheets, presentations and possibly small databases.

Two options include: (1) the free open source OpenOffice suite and (2) Microsoft Office for Windows. The decision will likely be driven by Microsoft licensing issues. Note that OpenOffice can always be deployed alongside MS Office as well and it is a free product.

The Adobe Acrobat reader is also an essential tool to deploy in the environment.

Statistical analysis packages

This is likely one of the most challenging aspect of a remote access facility. When selecting analytical packages, striking an appropriate balance between users' requirements, management and support options, cost, and licensing issues can be a complicated process.

Ideally the SAS[58] / SPSS[59] / Stata[60] triad should be available in the environment. These are among the most broadly used researcher packages. They however all have slightly different licensing models which may raise issues in some cases. A long list of other statistical, econometric or mathematical analysis packages may also be considered, such as "R"[61], S and S-PLUS[62], GAUSS[63], LimDep[64], Matlab[65], Mathematica[66] and others. Decisions will be affected by:

- user demand / popularity
- licensing model and costs (further discussed below, see p.50)
- management complexity
- platform compatibility and
- data compatibility

We recommend engaging the research community to gauge researcher's demands. An interesting exercise CESSDA could conduct would be to survey potential users to determine their favourite packages and how/when they are used.

Management complexity is linked to the package deployment procedure, the upgrade mechanisms, and the issue that researchers often like to use their favourite extensions or plug-ins, requiring that they be installed, upgraded and maintained.

[58] http://www.sas.com/
[59] http://spss.com/
[60] http://stata.com/
[61] http://www.r-project.org/
[62] http://www.insightful.com/products/splus/default.asp
[63] http://www.aptech.com/
[64] http://www.limdep.com/
[65] www.mathworks.com
[66] www.wolfram.com

Technical support is also an issue but we suggest in this case taking a community driven approach where experienced users can provide support to one another. The SRA staff may not have the expertise in all statistical packages.

Platform compatibility is a measure of how well the package fits into the operating environment. Not all products are available for all operating systems. Some may not fully comply with security requirements or properly operate on a 64-bit platform. And copy protection like a dongle key may make it incompatible with a virtual environment.

Data compatibility is a measure of which data formats a package can use as input. In order to support a wide variety of statistical packages, data administrators need to maintain the master data in several formats. We typically recommend having the data available in the core formats (SAS, SPSS, and Stata) along with an ASCII version for import in other packages. If a product requires a different proprietary format, it should usually not be installed as it would oblige the data manager to produce and maintain even more files. Note that an emerging technique to address this problem is also to keep, along with the ASCII data, metadata in a XML format such as DDI. Using a simple XML transformation technique, it is possibly to generate import scripts (setup files) on demand for many packages. This transformation need only be written once per package. Such solutions have been prototyped by the UK Data Archive, Open Data Foundation and the International Household Survey Network and are likely to become more widely available by the time the SRA facilities are implemented.

The reality however is that technical and licensing constraints will limit the standard configuration choices and will directly impact the utility and availability of tools available to researchers. If necessary, on demand custom environments can be deployed for specific users or research groups. A list of packages and weighted selection criteria should be established to determine whether or not to integrate into the SRA environment.

Note that by determining popularity scores SRAs could statistically determine the average cost per user. For example, in an environment hosting 100 users, a package that cost €10K/year and is needed by 100% of the users is equivalent to €100/user/year. A popularity score of 50% would double that number and 10% would be €1K/user/year. This could be further adjusted by the average number of concurrent users and other factors such as management costs to produce a figure representing the overall score.

SRA Management tools

In addition to the end user packages, the SRA system and facility administrators require a collection of tools to manage data, metadata and environment, requiring licenses based on the number of potential users and includes:

- Data conversion utilities, such as StatTransfer[67] (note that DBMSCopy has been discontinued by DataFlux/SAS in 2008)

[67] http://www.stattransfer.com/

- Multimedia software to manage internal web sites or produce training materials such as the Adobe[68] Creative Suite, Adobe Capture or Camtasia Studio[69], etc.
- File compression software such as WinZip[70] or 7Zip[71]
- Metadata management tools to maintain survey documentation. This includes:
  - DDI editors, e.g., the Nesstar Publisher[72], the IHSN Management Toolkit and other emerging DDI solutions.
  - XML editors such as XMLSpy[73], Oxygen[74], Stylus Studio[75] or Editix[76]
  - Adobe Acrobat[77] for the conversion of document into a system independent format (this software may be included in multimedia solution or package)

Note that these applications should be available only to administrators and can be deployed in a separate environment such as a virtual Manager Desktop restricted to authorized users.

Collaboration platform

The SRA environment is more than just a place to share data and documentation files. It should provide the user with dynamic environment that fosters collaboration and knowledge sharing. This is discussed in further detail in the Metadata, Collaborative and Knowledge Management section (p.62).

Relevant software will need to be deployed in order to support the user community which may include wiki, blogs, discussion groups, events and news, instant messaging and other social networking tools.

One emerging product that may be particularly attractive for a collaborative environment is the recently announced Google Wave platform[78]. Google Wave is "a personal communication and collaboration tool" announced by Google at the Google I/O conference on May 27, 2009. It is a web based service, computing platform, and communications protocol designed to merge e-mail, instant messaging, wiki, and social networking. While currently in its early developmental stages, this tool could potentially revolutionize the way users communicate. Given that it will be made available as open source software, it can potentially be deployed in the closed SRA environment.

[68] http://www.adobe.com/
[69] http://www.techsmith.com/camtasia.asp
[70] http://www.winzip.com
[71] http://www.7-zip.org/
[72] http://www.nesstar.com/software/publisher.html
[73] http://www.xmlspy.com
[74] http://www.oxygenxml.com/
[75] http://www.stylusstudio.com/
[76] http://www.editix.com/
[77] http://www.adobe.com/products/acrobatpro/
[78] http://wave.google.com/ and http://en.wikipedia.org/wiki/Google_Wave

One important aspect to consider when selecting collaborative tools is their ability to integrate into a single sign-on system to alleviate the need for the user to authenticate multiple times.

Conferencing platform

When working with remote users, it is important to be able to communicate effectively. This may take place outside the SRA environment, in particular for training purposes as it is usually a requirement for obtaining access to the facility. Requiring users to travel to a training centre can be challenging and is not particularly cost effective. We therefore recommend having a web based conferencing[79] platform available to be able to virtually deliver training to geographically distributed users. Popular solutions include CISCO WebEx[80] and Adobe Connect[81]. The UK Data Archive and CESSDA also report having a positive experience with Marratech.[82]

Software licensing issues

Software licensing presents considerable challenges for a virtual remote access environment, in particular for researchers' targeted packages, as the number of users can quickly grow. This can have a major impact on the cost of deploying a particular package in the SRA environment. While limited options are available when it comes to required software (similar to the core infrastructure or remote access), this can be a key selection criteria for end user software such as analytical packages.

Licensing models include:

- *Per server*: One license fee per server (physical or virtual), independent of the number of users. This model is very attractive for the SRA as it comes at a fixed cost. It is however no longer very common. A variation is the per-CPU model which takes the number of processors into account.
- *Enterprise license*: The package can be used by any number of users belonging to a particular organization. This typically involves a high fee as it assumes a large number of users but is very attractive for a SRA as it comes at a fixed cost. The challenge here is that not all external researchers fall in the category of an institutional user. One option however is to temporarily make the user an "employee" of the organization, for example by "hiring" them as subcontractors for the duration of the research project. While this may have some administrative and legal implications, it can provide significant savings for the SRA facility.
- *Per concurrent user*: One license is required for every user using the software at the same time. Citrix XenApp and Stata network license, for example, fall into this category. This is a good model for a SRA, though estimating the peak number of simultaneous users can be challenging.
- *Per user or per seat*: One license is required for each user. This is a very common model that can end up being very expensive for SRA.

[79] http://en.wikipedia.org/wiki/Web_conferencing

[80] http://www.webex.com/

[81] http://www.adobe.com/products/acrobatconnectpro/

[82] http://www.marratech.com/

- *Client access license* (CALs)[83]: One license per user connecting to a server side product. This model is very commonly used by Microsoft and is similar to the per user model.

  Several other variations exist and some packages have fairly complex models (Space Time Research is an example).

The *duration* of the license is also very important. Many packages use a perpetual licensing model essentially meaning a one time fee for unlimited usage and no expiry date. These products typically come with the option to purchase a yearly maintenance fee that provides free upgrades and technical support. Other packages require renewal of the license on a regular basis (like yearly) which can significantly increase the operational costs (SAS for example falls into this category).

The legal status of the licensing institution can also allow for special discounts. Academic institutions or non-profit organizations often have access to deep discounts and governmental agencies typically receive preferred pricing.

Users groups in the SRA context can be broken down into system administrators (operating the data centre), staff (managing the SRA facility), and end users (researchers, data providers and others accessing data silos). The licensing plan for a SRA facility should document the size of each group and packages in use.

In some case, operating custom virtual desktops or restricting access to some applications to specific individuals can help address the licensing issues.

Regardless, it is important to carefully examine the licensing model for all packages that will potentially be deployed in the SRA environment. Properly costing software makes a significant difference in the SRA operational budget. Note that this further emphasizes the advantages in operating shared or common facilities as this can considerably reduce overall licensing costs.

Requirements summary

☑ **Specifications core infrastructure software**
☑ **Specifications for Citrix XenApp platform and partner products**
☑ **Specifications for office productivity and other user software**
☑ **Specifications for statistical analysis packages**
☑ **Specifications for SRA management tools**
☑ **Specifications for collaboration tools**
☑ **Specifications for web conferencing solutions**
☑ **Software Licensing Plan**

[83] http://en.wikipedia.org/wiki/Client_Access_License

Storage

A large amount of information is expected to be stored in the remote access facility servers. Sufficient disk space must be available to address the system, data archive, and users' needs. When planning a remote access facility and designing specifications, the following should be taken into consideration:

- A clean separation should be maintained between the various storage areas such as the operating system / environment, software / applications, master datasets and documentation (local data archive)
- Network attached storage[84] (NAS) should generally be used to maintain the storage areas independent from the servers
- For security purposes, we recommend that the data holding storage area (at least) be encrypted
- To reduce the risk of information loss, disks drives should operate in a redundant configuration[85] (RAID) in case of drive failure
- Users' space should be controlled by quota, and disk space management and shared etiquette should be explained during user training
- Backup/restore: based on backup strategies, relevant storage devices will need to be available in the data centre to support both short term and long term preservation of information.

Requirements Summary

☑ **Storage Hardware specifications**
☑ **Primary storage configuration**
☑ **Secondary / backup storage configuration**
☑ **Storage Security Plan**

Other Security Issues

CESSDA user management and SSO

Earlier in this report we emphasized the importance of and various mechanisms for properly authenticating SRA users. Doing so in a user friendly and limiting burden on users should be paramount, In fact, once a user has been properly authenticated, he should not be prompted again (unless required by security policies).

This single sign-on[86] (SSO) philosophy should first apply within the SRA environment where all applications used at the facility should in general support an SSO based system. In the particular case of CESSDA, however, this could also extend across facilities and an external user should not be required to register multiple accounts or remember passwords to obtain access to different facilities. This issue is not specific

[84] http://en.wikipedia.org/wiki/Network-attached_storage
[85] http://en.wikipedia.org/wiki/RAID
[86] http://en.wikipedia.org/wiki/Single_sign-on

to the SRA, and CESSDA is considering adopting Shibboleth[87] as a general SSO solution for its users. As an SAML[88] based system, integration with the Citrix platform may be achieved by using Password manager or through Ping Identity's[89] flagship product PingFederate. The Shibboleth platform however not may be considered sufficient for an SRA facility; but it could be combined with additional SRA specific authentication mechanisms, enforced locally by the Citrix XenApp server (such as biometric, proximity keys, token generators, etc.).

Another useful feature that could be implemented in the particular context of CESSDA is the maintenance of a shared researcher registry (database). While not technically an IT issue, it would require the relevant architecture to support it. Such registry would essentially allow all SRA facilities to check and/or document the profile of a user to see prior researcher activities. This could greatly facilitate and speed up the researcher clearance process by validating credentials as well as helping to identify potentially unsafe users. Combined with general user demographics, potentially available through the SSO system, this information would also be highly valuable for statistical purpose to understand user characteristics. This essentially parallels the idea of a CESSDA and European accreditation system outlined by Roxane Silberman[90].

Encryption

An SRA environment requires various encryption solutions to ensure that the data and other sensitive information are properly protected from unauthorized access or use. These aspects should be summarized and well documented in the SRA security plan, including encryption:

- *On the network*: any information exported from the data centre over the network should be properly encrypted. This is built into the Citrix technology but other channels are often used to exchange sensitive information. Both software (such as TLS / SSL certificates[91]) and hardware solutions (such as encrypted routers or network cards) should be deployed as needed across the environment. Networks, whether wireless or cabled, are particularly prone to eavesdropping and all data transmission should be properly protected.

- *On the servers and storage devices*: while the data centre security system should provide a significant layer of protection, servers, disk drives and other storage

[87] http://shibboleth.internet2.edu/ and http://en.wikipedia.org/wiki/Shibboleth_(Internet2)

[88] http://www.oasis-open.org/committees/security/ and http://en.wikipedia.org/wiki/Security_Assertion_Markup_Language

[89] http://www.pingidentity.com/

[90] "CESSDA and European accreditation", Roxane Silberman, CCDSHS/Réseau Quetelet, MICRODATA ACCESS – NEW DEVELOPMENTS AND A WAY FORWARD, Luxembourg, December 2008 http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/3_3_CESSDA.pdf

[91] http://en.wikipedia.org/wiki/Transport_Layer_Security

media can always be stolen, lost or potentially accessed by third party when eventually discarded. Sensitive data and information should therefore always be properly encrypted when digitally preserved. Depending on performance requirements, this can be enforced by software (like at the operating system level) or specialized hardware components. This particularly applies to files stored off site for backup and disaster recovery purposes.

- *During external data exchange*: when exchanging data with providers (for example during ingest or for disclosure review purposes) or possibly other SRA or agencies, it may be necessary to use portable media such as CD, DVD, USB keys, and others. It is crucial in these situations that all sensitive data be well protected in case the device gets lost. Many software products are available to encrypt files such as the licensed version of WinZip[92] or pGp[93]. A recently released USB product may be particularly well-suited for this purpose is the IronKey[94] providing AES 256-bit encryption, compliance with Security Level 3 of FIPS 140-2[95], optional self destruct mechanisms and other security features. The Kanguru Micro Drive with AES encryption[96] is also a good option.

Backup / Restore / Disaster recovery

As with any information system, an SRA facility must have solid backup/restore and disaster recovery strategy. Indeed these are often standard requirements for certification of infrastructures. The details of such strategies will not be discussed but they must at the minimum include:

- On site backup
- Off site backup
- Disaster recovery plan

It is also important that these procedures be tested on a regular basis to ensure their effectiveness and to make refinements as necessary.

A critical aspect to be taken into consideration for SRA facilities is the storage of data outside the hosting institution. Legal issues may often prevent such data from being stored by third parties or outside national boundaries. In a network of facilities, cross-SRA backup storage & recovery services might be an effective way to address this problem.

Requirements summary

☑ **User management and SSO strategy**
☑ **Encryption requirements and tools**
☑ **Backup/Restore strategy**
☑ **Disaster Recovery Plan**

[92] http://www.winzip.com
[93] http://www.pgp.com/
[94] https://www.ironkey.com/
[95] http://en.wikipedia.org/wiki/FIPS_140-2
[96] http://www.kanguru.com/aesmicrodrive.html

Reference Architecture

The following reference architectures are available from the Citrix web site and provide valuable insight for planning purposes.

- Provisioning Services for XenApp - Reference Architecture[97]
- Simplifying Application Delivery to the Virtual Desktop - Reference Architecture[98]
- XenApp 5 Scalability Analysis[99]
- XenApp and XenServer - Reference Architecture[100]
- Simplifying the Migration to XenApp 5 with XenServer - Reference Architecture[101]
- Citrix Access Gateway Enterprise Edition Integration Guide for Citrix XenApp and Citrix XenDesktop[102]

Statistical requirements / Risk management

Introduction

NSI's, survey organizations, academic researchers, and business establishments collect and disseminate data on people, businesses, or other entities under specific pledges of confidentiality. Typically, some type of statistical protection is applied to protect data confidentiality before results are made public. Yet, in order to protect respondent confidentiality, data quality is routinely compromised (Zayatz, 2005) as sensitive information, such as income, is typically rounded or top coded. Accurately assessing statistical risk is therefore of paramount interest. If data are released too liberally, the risk of statistical disclosure becomes significant; if risk assessments are too conservative, released data will be of suboptimal quality and utility (Eliot, in Doyle et al).[103]

[97] http://support.citrix.com/article/ctx120512

[98] http://support.citrix.com/article/ctx120516

[99] http://support.citrix.com/article/ctx119108

[100] http://support.citrix.com/article/ctx117922

[101] http://support.citrix.com/article/CTX119495

[102] http://support.citrix.com/article/CTX119426

[103] A good example of the resultant difficulties is illustrated in a paper by Stuart Soroka and Chris Wlezien entitled 'How Measures Matter' (2002). The authors ran the same model on three different quality UK budget datasets: the unadjusted data (i.e. what is reported by the UK Government to OECD); data adjusted by simply treating public corporations consistently, and the full adjustments for backward compatibility. The first model yielded insignificant results in the wrong direction. The second yielded insignificant results in the right direction. The third confirmed the model. It is worth noting that, despite the potential consequences, few, if any, statistical agencies inform researchers about the potential consequences of disclosure protection techniques and edits on the quality of their analysis (Kennickell & Lane, 2006).

Strategies to prevent unauthorized and inappropriate disclosure of identifiable information inevitably involve some degree of data content modification and/or data access restrictions. At a minimum, all direct identifiers such as names and addresses must be suppressed before releasing researcher output to the public, despite the fact that it clearly limits the quality of the potential analyses. The end goal should always be to minimize data loss with an eye toward improving the precision of results and thus empowering evidenced-based decision making.

As discussed earlier in the document, managing risk is essential to ensure the integrity of the SRA facility and the trust of producers and users in the system. This section examines one of the most sensitive issues: respondent identity protection. While the risk of re-identifying respondents cannot be reduced to zero, data providers and SRAs are advised to implement appropriate risk management practices customized to each dataset. In addition these facilities should have adequate IT security systems in place to protect the source data and effective output disclosure control processes.

This can be achieved in three ways:

- Controlling the input by assessing the risk assess and ensuring authorized access to the data and documentation
- Responsible analysis by training users on relevant issues and by establishing legal agreements and
- Controlling the output through effective statistical disclosure control

Although this report focuses mostly on output disclosure control, it also discusses input disclosure control. Training and legal control issues are discussed in other sections of this document.

Risk assessment and disclosure control

What is statistical disclosure?

Statistical disclosure occurs when information on an individual, household, or business is disclosed through the release of a dataset that allows an individual's identity to become known even though direct identifiers have been removed. This kind of identification disclosure happens when identification information held by an unauthorized user is able link to data, held in a file that has been cleaned of direct identifiers, through key (or common) variables that allow the intruder to derive information about particular individuals that the intruder does not already know.[104]

Most disclosure control is context-specific. It either involves primary or secondary disclosure. Primary disclosure concerns are addressed by looking at the individual cells and checking for class disclosure. Secondary disclosure concerns are by nature much more difficult to check and are derived by combining data from different tables and sources, using non-suppressed information (Mulcahy, Lane 2007).

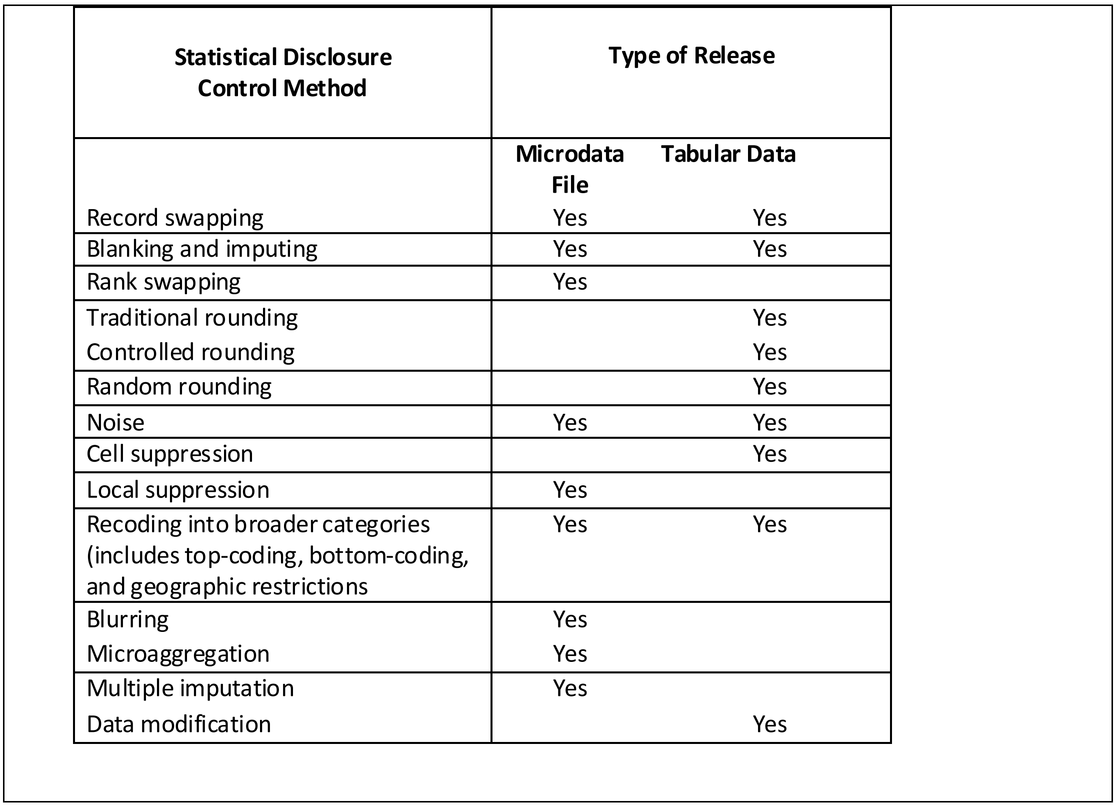
Essentially, there are three types of disclosure; identity, attribute and residual. Identity disclosure occurs when an individual can be identified from the released output, leading to information being provided. Attribute disclosure occurs when confidential information is revealed and can be attributed to an individual. It is not necessary for a specific individual to be identified or for a specific value to be given for attribute disclosure to occur. For example, publishing a narrow range for the salary of persons exercising a particular profession in one region may constitute a disclosure. Residual disclosure can occur when released information can be combined to obtain confidential data. Care must be taken to examine all output to be released. While a table alone may not disclose confidential information, disclosure can occur by combining information from several sources, including external ones, e.g., suppressed data in one table can be derived from other tables (Statistics Canada, 2005).

Fienberg (2003) summarized the technical goals of disclosure limitation techniques as follows: (i) inferences should be the same as if we had original complete data; (ii) researchers should have the ability to reverse disclosure protection mechanism, not for individual identification, but for inferences about parameters in statistical models; (iii) there should be sufficient variables to allow for proper multivariate analyses and (iv) researchers should not only have the ability to assess goodness of fit of models but also be provided with most summary information, such as residuals (to identify outliers). The core guiding principle, however, should be to generate released data that are as close to the frontier as possible (Abowd and Lane, 2004).

[104] Measuring the chance that respondents may be re-identified is based on methods such as k-anonymyty, l-diversity, l-completeness, SUDA, Poisson model or record linkages. These are not typically computationally intensive and do not require a deep expertise in statistics.

Methods for applying disclosure control methods generally fall into one of two categories: perturbative and non-perturbative (Willenborg and De Waal, 2001). Perturbative methods involve distorting the microdata before being provided to researchers, and include: additive noise, data swapping, micro aggregation, and post randomization (PRAM), data distortion by probability distribution, re-sampling, Lossy compression, multiple imputation, camouflage, rank swapping, and rounding. This method typically follows a 3-step process that involves (1) measuring the risk; (2) reducing the risk; and (3) assessing the information loss.

By contrast, non-perturbative methods, such as global recoding, local suppression, and sampling, do not alter the data. Rather, they produce partial suppressions or reductions of detail on the original dataset. Proper use of disclosure control methods, particularly perturbative, requires a significant amount of expertise and can also be computationally and resource intensive. The table below provides some guidance on which SDC methods apply to various types of data releases (e.g. microdata files or tabular data)



Statistical Disclosure Control Method	**Type of Release**	
	Microdata File	**Tabular Data**
Record swapping	Yes	Yes
Blanking and imputing	Yes	Yes
Rank swapping	Yes	
Traditional rounding		Yes
Controlled rounding		Yes
Random rounding		Yes
Noise	Yes	Yes
Cell suppression		Yes
Local suppression	Yes	
Recoding into broader categories (includes top-coding, bottom-coding, and geographic restrictions	Yes	Yes
Blurring	Yes	
Microaggregation	Yes	
Multiple imputation	Yes	
Data modification		Yes

These methods have been widely discussed and documented and we refer the reader to standard literature for more information.

Disclosure control in SRA

Applying disclosure control on a dataset (or statistical microdata output) is a delicate operation that demands not only a solid understanding of the techniques - and therefore statistical and mathematical expertise - but also familiarity with the underlying dataset. Performing such operations in an SRA environment should therefore be fully controlled, take place in coordination with the data producer, and be guided by mutually agreed principles and procedures by the data producer and archivist.

Typically, providing researchers access to the same set of files, ideally unperturbed microdata, is the recommended approach. By contrast, performing dataset customization per research project requires significant resources and greatly reduces the potential for collaboration. Data producers should prepare a set of master files suitable for efficient analysis in a controlled SRA environment. Facilities also should have in place appropriate statistical staff capable of conducting ongoing, and sometimes extremely sophisticated, disclosure control reviews. These issues are even more complex for CESSDA, for example in cases of cross-country dataset output release requests or in case the data providers or legislation require project level customization.

At a minimum, the authors recommend that SRA staff be familiar with simple risk assessment methods as these are necessary for quality control operations from data ingest to output clearance review.

New challenges to statistical disclosure control

A number of NSIs currently provide access to disclosive microdata through physically controlled laboratories. Challenges arise in the need to develop disclosure limitation techniques that are flexible enough to be used in a wide variety of situations. Considerable effort has gone into developing disclosure limitation methods for tabular data that effectively lower disclosure risk and provide products with high utility to legitimate users (Duncan, 2001; Duncan et al, 1993; Willenborg an d de Waal 1996, 200). These techniques include cell suppression, local suppression, global recoding, rounding, and various forms of perturbation (Federal Committee 1994). Under cell suppression, for example, the values of table cells that pose confidentiality problems are determined and suppressed (as primary suppressions) as are values of additional cells that can be inferred from released table margins (as secondary suppressions) (Cox, 1980). Perturbation is used through controlled rounding (Cox, 1987), versions of post-randomization response (Gouweleeuw et al. 1998), and Markov perturbation approaches, which have been proposed in various forms by Duncan and Feinberg (1999), Fienberg et al. (1998), and Feinberg et al. (2001).

Ritchie, Abowd Lane and others argue that disclosure control in remote access modalities requires a fundamentally different approach to proscriptive rules-based methods – the “principles-example” approach. This explicitly underscores the limitations of trying to specify exact rules, and places the focus on an understanding of principles to which rules can be more flexibly tied. Ritchie (2007) goes so far as to call for new discussion of what constitutes effective SDC in a remote microdata access environment, cautioning that the range of analysis carried out in virtual centres goes far beyond the traditional models used for designing SDC rules. “While

SDC for aggregation and anonymization is regularly tested and developed, the lack of discussion about rules for research outputs means that there is little independent scrutiny of the internal rules the research centre managers have developed; nor is there much sharing of 'best practice', Ritchie notes.

What's more, with the advent of researchers working directly with microdata in secure remote data access platforms, no longer is the focus on ensuring the non-disclosiveness of aggregates or generating non-disclosive ("public use" files) dataset (Ritchie, 2007). The focus seems to be moved from controlling inputs to controlling outputs. While on the one hand providing access to microdata provides researchers the autonomy to explore analyses above and beyond simple linear aggregation, the range of research outputs and inherent risk expands considerably in moving away from linear aggregates (e.g., linear and non-linear estimation, simulation, probabilistic modelling, Bayesian analysis, factor analysis, dynamic modelling, transition data, etc) (Ritchie, 2007).

Despite significant progress in providing authorized users access to sensitive microdata, the literature on disclosure control has failed to keep pace. Indeed the international community seems transfixed on antiquated concerns about the physical aspects of safe settings, or on preparing safe files for distribution (see, for example, UN (2003, 2006); Domingo-Ferrer and Torra (2004), Domingo-Ferrer and Franconi (2006). Of particular note, none of the projects listed under the Eurostat methodological programme address the issue of controlling research outputs from disclosive microdata. Apart from Reznek (2004), Corscadden et al (2006) Steel and Reznek (2006), and Ritchie (2006a, 2006b), which all discuss the release of analytical outputs, there appears to be little analysis of some of the general problems that arise when researchers are given free rein over data (Ritchie, 2007). The table below, used by ONS's statistical review team provides a helpful guide that distinguishes safe from unsafe output.

Tools

A number of disclosure control techniques can be applied using standard statistical packages such as SAS, Stata or SPSS. The CASC/CENEX/ESS web site[105] on statistical disclosure control provides access to the micro-Argus package as well as excellent methodological references and resources. The International Household Survey Network[106] is also working on microdata anonymization tools[107] that should become available later this year. Space Time Research software also has integrates statistical disclosure control algorithms that leverage the micro-Argus package.

General guidelines for developing a disclosure review process

[105] http://neon.vb.cbs.nl/casc/

[106] http://www.ihsn.org

[107] http://www.fsd.uta.fi/iassist2009/presentations/F2_Dupriez.ppt

When researchers are ready to have their results reviewed for public release, a good rule of practice is to require that researchers review an SDC checklist or guidelines document that clearly demonstrates what needs to be completed before the disclosure review process may begin. For example, researchers should provide a brief description of their project. They also should identify which file(s) they are requesting to be reviewed and identify the exact location of the file(s). Researchers also should specify the dataset(s) and variables from which the output derives; and identify the underlying cell sizes for each variable, including regression coefficients based on discrete variables.

Disclosure control in a research environment is no simple exercise. All results are context specific, and hence no absolute rules can be defined. Researchers therefore must be trained on the basic threshold and dominance rules, and how these are applied in practice. Researchers furthermore should review the SDC checklist to ensure that they have obeyed rules where relevant, e.g., no cells with less than 10 units (individuals or enterprises) and local unit analysis (threshold rule) must show enterprise count (even when there is no information associated with each cell). Researchers should be careful when tabulating raw data (threshold rule); using "lumpy variables, such as "investment"; and when researching small geographical areas (dominance rule)**.**

Users should also be reminded that graphs are simply tables in another form (i.e., they display frequencies) and that they should treat quantiles as tables (and remember to display frequencies). As a general rule, researchers should avoid reporting minimum, maximum, and median values. Regression results generally do not present disclosure concerns unless on dummy variables in a table; on public explanatory variables; and in potentially disclosive situations when differencing hiding coefficients makes linear and non linear estimation completely non-disclosive (note: panel models are inherently safe). Tables such as the one below provide general guidelines for researchers to keep in mind.

Output	**Classification**
Frequency and magnitude tables, including means	Unsafe
Percentiles (inc max and min, and median)	Unsafe
Mode	Safe
Higher moments of distributions	Safe
Graphs/pictorial representations (actual data)	Unsafe
Graphs (fitted values)	Safe
Estimation residuals	Unsafe
Linear regression coefficients	Safe
Non-linear regression coefficients	Safe
Summary and test statistics from estimates (R^2, χ^2 etc)	Safe
Cross-product matrices	Unsafe
Variance-covariance matrices	Safe

Correlation coefficients	Unsafe
Herfindahl/Ellison-Glaeser indexes	Safe
Gini coefficients	Safe
Oaxaca and other decompositions	Safe
Index numbers in general	Unsafe
Concentration ratios	Unsafe

I terms of presenting the output to the review team, tables and regression results should be easy to read. In particular, researchers should try to make it easy for reviewers to locate the cell count. Researchers should err on the side of requesting a smaller number of large tables to be proofed, rather than lots of smaller tables. For example, if a researcher wanted to tabulate the mean sales by sic in STATA, the "by sic: sum sales" command gives a lot of messy tables, which also include min and max values (which we prefer to suppress). Instead, one could use "collapse (mean) sales (count) entref (count) wow, by(sic)" and then "list" to obtain a very neat table that's much easier to read and only includes the relevant information. In addition, frequencies supporting each cell should be clearly displayed and these frequencies should be unweighted.

With respect to the volume of output submitted to the review team, a large amount of output might be a reason for rejecting output so, for example, it is easier to check repeated regressions if the syntax is written in loops. Researchers should not submit log files from a long interactive session. Neither should they submit results that are peripheral to their analysis. In addition, researchers should avoid unnecessary detail, e.g., "sum, detail" in STATA. If researchers require a large number of regressions to be cleared, they should also submit the program. Where possible, researchers should write programs in loops to make them easier for the review team to read.

Practical Issues

In addition, there are a number of practical issues related to developing and implementing a statistical disclosure control process that need to be addressed. For example, how much and what types of human capital and other resources need to be dedicated to the effort? Appropriate staffing levels depend largely on the complexity of the surveys in question, the degree to which there are disclosive variables, the number and complexity of researcher output requests, etc. To the extent possible, it is optimal to limit the number of statisticians reviewing output for public release. For small to medium size facilities, one or two seem optimal. In this manner, the SDC reviewers become familiar with the datasets, problematic variables, and increase their knowledge about what output has been released in the past, what methods were used to protect respondent identity, etc. Although residual disclosure concerns will likely remain, identity and attribute disclosure may be well-controlled. By contrast, if one prefers to staff a disclosure control team with a larger number of reviewers, these issues may be exacerbated and also may increase the risk of breach.

Regardless of whether an SRA decides to staff its SDC team with one or two fully committed statisticians or a larger group of part time statisticians, they should operate in a consistent manner. There should be a formal protocol on how the process works, both from the reviewer and researcher perspective. Clear guidelines must be in place, so all reviewers consistently apply agreed upon measures. Implementing risk scales is one effective approach, e.g.:

- Safe: No risk / very low risk of disclosure – output will be released promptly
- Unsure outputs: Low or medium risk of disclosure – output will be considered carefully, with some dialogue with the researcher as necessary, perhaps to collapse categories, remove one or more variables or suppress some cells.
- Unsafe: High risk of disclosure – output will be blocked in its current form and won't be released. This is the responsibility of researcher to produce safe outputs and demonstrate that they are free from the disclosure risks.

However, most output review will involve at least some degree of manual review. In these cases, SDC reviewers should maintain detailed notes of each review, problems identified, and resolution strategies. Maintaining a comprehensive but dynamic document codifying all business rules and clarifying why decisions were made will prove beneficial for training purposes, historical review and continuous improvement, and as a record for investigating the root causes of breaches.

Preferably all statistical output should be reviewed by the SDC review team before being made public, although entities that must respond to a heavy volume of requests might choose to randomly sample among the total number of output requests. Selecting output at random, however, in no uncertain terms means that the data producer accepts that there is a greater risk of undesired data disclosure. Therefore, to the extent that sampling is used, the output should be derived from variables that have a demonstrably lower risk for confidentiality breach.

Perhaps the most important aspect of the building and sustaining an adequate SDC process is to inculcate the researchers and institutions in a "culture of confidentiality" – one in which the research, institution, data producer, and archivist all share some of the risk. Guidelines must be clear. Researchers must clearly understand their responsibilities. And all must be made aware that the benefit of data access will go away the moment that one of the crucial players in the process fails to recognize this crucial, shared responsibility. In so doing, researchers may better appreciate the need for the process and may be more inclined to do their due diligence in preparing output that has met all the requisite parameters for final SDC review. What's more, researchers may gain a better sense of all that is involved in the SDC process and thus may appreciate the time, resources, and effort involved.

Metadata, Collaborative and Knowledge Management

Overview

A virtual remote access facility is more than a place to access data and perform analysis. It should provide the users with high quality, well documented data as well as a dynamic environment that promotes effective research, collaboration and knowledge sharing. It is also an opportunity for data producers to interact virtually with its user community by sustaining a productive dialogue that increases an understanding of how the data are being used. This can be accomplished by leveraging on technology, rich metadata, collaborative spaces, and social networking tools.

Metadata

There is no question that high quality data must be surrounded by comprehensive documentation. This should be more than a collection of electronic documents and a SRA should leverage on social science metadata specifications and community best practices. The most relevant specification in this context is the Data Documentation Initiative[108] (DDI) but other such as the Dublin Core[109] (DC), the Statistical Data and Metadata Exchange Standards[110], (SDMX), Metadata Encoding and Transmission Standard[111] (METS), ISO 11179[112], and other are also important.

The metadata issues have been widely documented by the community and are well known within the CESSDA space so it will not be further discussed here. We refer the reader to standard literature and other CESSDA PPP work packages[113]. What is important is that a metadata strategy must be integrant part of the SRA facility.

Archive Metadata

Just like a data archive or a research data centre, an SRA facility must have clear procedures in place for the ingest of data and documentation into the environment. An SRA will have to maintain its own internal archive where the master data will be prepared and packaged for delivery to the end users. During ingest, the quality of the data, documentation and metadata will need to be validated and gaps filled as necessary. This will typically involve working closely with the data provider.

As it is not unusual for data to come with limited documentation, improving the quality of the datasets through the capture of standards based metadata can be an attractive

[108] http://www.ddialliance.org
[109] http://www.dublincore.org
[110] http://www.sdmx.org
[111] http://www.loc.gov/standards/mets/
[112] http://metadata-standards.org/11179/
[113] http://www.cessda.org/project/wps.html

add-on service or even a major role for a SRA facility. This makes the environment not only an attractive solution to provide access to data but also an opportunity to improve of the overall quality of information. CESSDA, as a sponsor of best practices, could be ideally positioning to offer such services to producers and other depositors.

Furthermore, bringing together data from different sources or even possibly from different countries can be a challenging process that can not only be documented using standards but also greatly facilitated by the availability of good metadata. The SRA facility could therefore also provide an environment to support such activities.

Archived datasets also do not always come in a researcher friendly format. Easy to manage large rectangular files are often preferred by archivists for preservation purposes but a hierarchical version is often much more appropriate for research purposes. Reshaping incoming datasets may therefore also be an add-on activity of the SRA ingest process.

As a general rule, an SRA should adopt a standards based data and metadata management framework following community best practices but with a focus on delivering information to the researchers.

Researcher metadata

A less familiar aspect of social science metadata is researcher metadata. Just like in data centres, the closed nature of a SRA facility makes it an ideal environment for the capture and exploration of such information.

Example of researcher metadata includes project description, research topics, feedback on data quality, citations and references, survey or variable ratings, dataset sub settings or variable selections, research paper peer review, system usage, etc. Such information provides several potential benefits including, as examples:

- Understanding of data usage
- Facilitating the research process by automating the production of documentation, code, or citations
- Support peer review
- Manage shared libraries (documents, scripts, etc.)
- Collect user quality feedback or perception of the data
- Ensure the security of the environment
- Manage research projects

As this is an explorative area, research metadata are not always based on standards and often takes various shapes. Some can be captured in XML specifications such as DDI while other may be stored in other XML or proprietary formats.

Several capture mechanisms can be used in the SRA environment:

- *Automated*: using system logs (operating system, Citrix), harvesting utilities (collecting scripts, information on file systems)

- *Semi-automated*: using forms, citation management systems, user feedback mechanisms
- *Manual*: using specialized tools/editors, by capturing at the source code level
- *Web based*: using social networking or other collaborative tools

Individually, these various metadata components already provide valuable information but their power is typically unleashed once combined across multiple sources (for example using structure DDI XML with variable usage and research topics). Such systems are being extensively explored in the NORC secure data enclave.

An SRA facility should leverage on its environment to capture rich researcher metadata.

Tools

The SRA facility will need to be equipped with the right set of tools to support the ingest, archival and research related processes. This includes some of the packages mentioned in the SRA Management tools section (p. 50), relevant configuration of the environment logging mechanisms, and potentially the development of proprietary solutions.

Benefits of metadata

A metadata driven environment is essential to the success of a SRA facility to ensure :

- Quality of the data
- Support for better research
- Understanding of data usage and research activities
- Deployment of collaborative platforms
- Dialog between the producer and the users
- Reporting to the data providers

It also enables the information exchange mechanisms with other SRAs and the publication of metadata on an intranet or public web sites.

Requirements summary

☑ **Metadata strategy**
☑ **Archival metadata, ingest processes and quality assurance procedures**
☑ **Research metadata capture plan**
☑ **Metadata management tools**

Collaboration & Knowledge capture

The ability to provide a virtual environment that fosters collaborative behaviour and captures knowledge is a major benefit of SRA facilities. It not only allows for geographically distributed users to communicate with each other but also leverages information that would typically have remained undocumented. It also increases the awareness of what others are doing with the data, encourages collaboration, promotes the user-producer dialog, and reduces duplication of effort.

While social science researchers in the past worked mostly in isolation, a virtual space literally has the potential to change the way in which research is conducted. Awareness of this type of environment and a broader knowledge of how to take advantage of it has increased due in part to the emergence of social networking and information sharing sites on the Internet.

Privacy concerns, however, remain an important issue. The researcher must be assured that his or her personal information or private research data remain protected. Three spaces should therefore be created in the SRA environment:

- The *individual space* where the user stores his or her personal information.
- A *team space* where individual working on the same research project can exchange information, share files, manage tasks, etc., and
- A *data silo space* where researchers working on the same data sources can dialog with each other, the data producers and the SRA staff.

Note that the data silo is only possible when the master data are made available to all researchers in that silo. If the researchers only obtain access to a subset of the master data for their specific project (a SRA administered under a “need-to-know principle), this essentially precludes all communication with other groups as the likelihood of disclosure is increased dramatically.

The collaborative and knowledge capture toolbox can include classic tools such as wiki, discussion groups, blogs, calendar of events, announcements, file and document libraries, and instant messaging, as well as potentially custom metadata driven tools. Note that metadata can also play a considerable role in seeding or feeding content to the users (for example, using DDI to seed or publish initial content in a Wiki). Likewise, as mentioned in the previous section on metadata, web 2.0 technologies also can be used to collect valuable metadata. Overall content management should generally be open to everyone to foster contributions and openness. A technical support tool is also an essential component. Users should be able to reach out for help within and outside the environment.

Note that the effectiveness of these tools greatly depends on the size of the community, the presence of active users willing to contribute or drive the content and participation from data producers and SRA staff. For example, instant messaging may work well for small communities but a wiki may not succeed given the limited amount of contributors. As mentioned before, the Google Wave[114] platform pre-released this summer may be a particularly well suited tool for the SRA environment.

Requirements summary

☑ **SRA Communication Strategy**

☑ **SRA Collaborative tools**

☑ **SRA Knowledge Management Tools**

[114] http://wave.google.com

Communication outside the SRA environment

While most of the core activities take place inside the secure and closed SRA facility, communication with users and the public outside the environment is an important aspect. A public portal for an SRA facility should be available to provide information to prospective researchers and data depositors, showcase the researcher's results, advocate data availability, describe access procedures, provide access to metadata and possibly public use or synthetic data file, etc. Such a web site should be dynamic and contain information for various audiences:

- the general public
- potential researchers
- potential data providers and
- partners and other research networks

In the case of CESSDA, where several facilities are likely to have to co-exist and where researchers may come from many different places, it could also be an attractive option to establish a virtual "SRA Intranet" accessible to accredited researchers, data providers and SRA staff. The CESSDA single sign-on system could be used to authenticate users and authorise access to such resources based on their profile. Similar to the SRA, such a private network could host various collaborative and knowledge sharing tools, the main difference being that no confidential information could be disclosed in such an environment. Some of these issues are being examined by other CESSDA PPP work packages.

The authors recommend that CESSDA consider maintaining an external facility for users to communicate with each other especially in isolated SRAs. Individuals have a natural tendency to talk to each other and exchange information outside the protected environment. Providing tools to do so would ensure that the exchanged knowledge is properly captured and that mechanisms are in place to monitor content for quality improvement and potential disclosive communications.

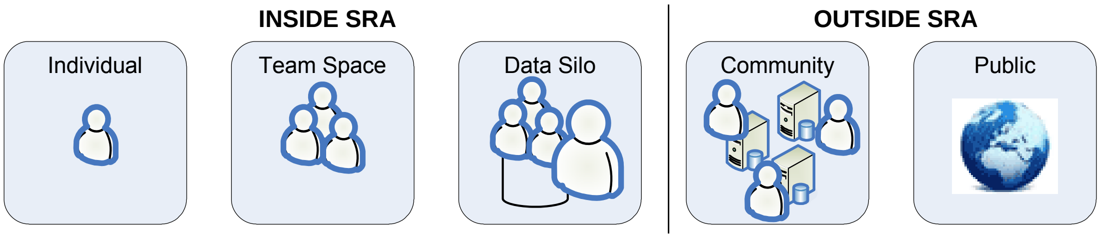
Requirements summary

☑ **SRA Public Portal**

☑ **SRA Intranet (optional)**

☑ **Integration in CESSDA portal**

Summary of communication zones



The individual user space. Essential for privacy and personal research	The research team area to facilitate group collaboration and activity coordination	The collaborative and knowledge surrounding a particular data producer or collection.	Communication between SRA users outside the environment. Information exchanged here must be non-disclosive but can span across SRAs or borders.	The public perspective of the SRA. Critical to communicate with stakeholders, potential users, advocate services and data availability, etc.

Organizational requirements

Advisory Board

One crucial component behind the success of an SRA is an advisory board comprised of internationally recognized experts providing multiple perspectives on various issues of importance to the initiative. Board members could provide expertise from multiple perspectives, from science and innovation to feedback on the business model and effective management strategies. To the external world, high level advisors provide a measure of credibility to the program and also serve to increase the visibility (and arguably importance) of the initiative. Internally, particularly in the start-up phase, panels of outside experts offer a convenient opportunity to test various modalities and functionality, and obtain insight both from the researcher and data producer perspectives.

To the extent possible, when conceiving of and developing the composition of an advisory group for newly emerging SRAs, one should try to maintain a representative balance of experts across academia, government and non-government entities (or better yet: to create a panel comprised of individuals with experience across multiple sectors and disciplines. Advisory group members also should be selected according to specific expertise they bring to the group; and that expertise should be closely aligned with the particular needs of the SRA in question. Lastly, in addition to being helpful in terms outreach and dissemination, expert advisory groups provide diverse, cross disciplinary perspectives, honest feedback, comprehensive review of policies, practices, and guidelines, and recommendations on how to make refinements to the SDAF model.

Requirements summary

☑ **SRA Advisory Board Members**

Legal Issues

Managing an SRA facility includes addressing a range of legal issues such as:

- Service level agreements with the data provider
- Service level agreements with the user and their institution

- Non-disclosure agreements
- Data access agreements
- General access policies
- Service level agreements with IT service providers (connectivity, backup, equipment vendors, etc.)
- Software licensing
- Etc.

National and European specific regulations will further complicate these matters, particularly for SRA hosting cross-national datasets (this issue is further discussed in the next section). For example, using cameras to monitor users may be a requirement for some data providers but is illegal in Germany.

These aspects must be carefully examined and fully documented and standard or harmonized templates should be drafted to facilitate the SRA management processes.

Requirements summary

☑ **Reference on international and national legal issues**

☑ **Templates**

Certification & Security Standards

Important characteristics of a SRA facility include its reliability and security. While the hosting agency can extensively document the technical features and advocate best practices, meeting industry standards or third party certification are essential to further strength the level of trust data producers and users have in the system. Meeting such requirements may also often be a requirement in order to host potentially disclosive or sensitive datasets.

Numerous international and national information security standards that can be considered[115] such as:

- International: ISO/IEC 27001: - 27002, ISO/IEC 20000, ISO/WD 31000.
- Europe: BS 25999, BS 7799-3, KongTraG, Basell II, DPA, EUDP, IAS, Companies Act, BDSG, LOP, Reg 357, Article 46, King II Report, Banking Act.
- North America: Bill 3494/2000, Bill 3221/2004, Bill 198, COBIT, COSO, SAS 70, Sarbanes-Oxley, Homeland Security, CMMI.
- South America: NBR 17799/27001, NTP 17799, NCH 2777, SB Regulations, Decree 83, Specific Local Requirements.
- Asia: Japan Privacy, Japanese SOX, Basell II & FICS.
- Australia and New Zealand: AS/NZS 4360, CLERP 9, PA&PAA.

A detailed security plan should be prepared that extensively document the various mechanisms used to protect the facility and the data as well as certification and

[115] Source: Data Center, Security & Outsourcing Newsletter, April 2009 and May 2009 issues (see http://www.globalcrossing.com/news/dc_security_out/dc_security_out_news.aspx)

standards with which the SRA complies. This plan should be revised on a regular basis (at least annually).

The NORC Data Enclave for example is fully compliant with DOC IT Security Program Policy, Section 6.5.2, the Federal Information Security Management Act, provisions of mandatory Federal Information Processing Standards (FIPS) and all other applicable NIST Data IT system and physical security requirements. An overview of the data protection plan is available on the NORC DE web site[116].

CESSDA will need to investigate further which standard should apply and consider how, organisationally, it will implement the certification process. Providing international and country specific guidelines could then be used as a requirement for CESSDA certified SRA.

Requirements summary

☑ **Certification procedures and Standards**

☑ **SRA Security Plan**

Training Plan

As we have seen, training is one of the pillars of the portfolio approach. This is an essential activity for all that are involved in SRAs. The training plan should provide detail on programs and delivery methods for the researcher, the SRA staff and the data providers. Some suggested modules are listed below:

Researcher	- How to access and use the SRA facility - Taking advantage of the collaborative tools - Data disclosure issues and output review - Data collection specific training
SRA Management Staff	- Operation and maintenance of the environment - Data and metadata management - Technical support - Delivery of trainings - Data collection specific training - Data disclosure and output review - General administration
SRA IT Staff	- Standard IT training for management of data centre - Citrix XenApp - IT Security - Role and responsibilities of SRA facility and importance of data protection - Training on statistical packages (configuration,

[116] http://norc.org/DataEnclave/Data+Security/IT+Security+Compliance/Data+Protection+Plan/Data+Protection+Plan.htm

	extensions)
Data Provider	- How to access and use the SRA facility - Taking advantage of the collaborative environment and dialog with users - Data and metadata standards, best practices and tools

A harmonized training program could lead to user certification valid across multiple CESSDA SRA facilities. The preparation and maintenance of the training materials could also be shared to reduce the overall costs and ensure consistency. As noted previously, dataset specific trainings should optimally be delivered by the data producers.

Training Content

This section provides information on recommended components of SRA training modules.

Using the SRA environment

- Connecting to the SRA: login/authentication mechanisms, protecting your user identity, installation of Citrix software (if applicable).
- Overview of the environment: interacting with the desktop, locked down functionalities, available software, file areas (personal, team, data and documentation, etc.), list of available datasets.
- Overview of collaborative: web based tools, sharing knowledge.
- Metadata and documentation: introduction, importance, using metadata, contributing metadata.
- SRA procedures: how to move information in, out to move information out, overview of disclosure review, technical support, etc.
- Miscellaneous issues: system performance, backup/restore procedures.

Data Disclosure Review

- Mission of the SRA
- Portfolio approach to data security and utility
- Data disclosure principles
- Protecting inputs vs. protecting the outputs
- Primary (Identify, attribute) and secondary (residual) disclosure
- Practicalities, safe vs. unsafe methods
- Disclosure output review processes: packaging materials for output requests check list, intermediate vs. final clearance, processing time, etc.
- Recommended readings, references, tools

Data specific training

- Background on the survey collection
- Description of datasets and data files
- Available documentation / metadata

- Sampling and weighting procedures
- Difference between SRA and other versions (SUF, PUF)
- Merging and comparability across time/geography/topic
- Availability of specific scripts/libraries/tools to facilitate/support analysis
- Strength and weaknesses / know issues
- Public / non-public datasets available to complement / merge
- Ongoing and future releases
- Data specific technical support

Training delivery

Various training delivery methods should also be considered and where possible, linked into the proposed Virtual Centre of Competence. Methods to consider include:

- Traditional in classroom training: this is the preferred approach as it also provides an opportunity for the producer, SRA managers and researchers to meet and get to know each other. Given the geographic distribution of remote users however this is not always practical and can be expensive.
- Virtual training using facilities such as WebEx, Adobe Connect or Marratech. This approach is very effective with remote users but is less interactive and can present technological challenges (connectivity, client software, etc.).
- Printed references: interactive training methods should be combined with references materials and documentations. These can be made available either in the SRA environment or on the public web site.
- Multimedia: when possible, we recommend recording training sessions for replay by participants or new users. This is often a feature of virtual training facilities. Producing short training or educational videos on disclosure issues, specific data collections, using the SRA environment, and other topics is also a good way to support the users and reduce training costs, particularly when reusable by multiple SRAs. Materials can be made available to researchers on web sites (public or in SRA), DVD or other media.

Requirements summary

☑ **Training Plan**
☑ **Training Methods**

Multilingual / multicultural environment

The multi-national and multi-cultural CESSDA SRA environment will present particular challenges as the data, metadata, collaborative, software and other components will need to support and operate in multiple languages. While other PPP work packages are looking into some of these issue (e.g., the WP4 on controlled vocabularies or WP9 on harmonization and conversion), it will also impact the SRA infrastructure in various ways. CESSDA SRA is well advised therefore to include an Internationalization Plan that documents and addresses these particular issues.

Requirements summary

☑ **Internationalization Plan**

Costs

Implementation and maintenance

Given the large number of options available when establishing and operating an SRA facility, it is challenging to come up with accurate figures for costing such infrastructures.

It is generally assumed that the SRA will be established on top of an existing IT infrastructure and will therefore only require adding the necessary components. If this is not the case, a considerable upfront investment may be necessary to establish the data centre and hire staff.

Choices that will significantly impact the cost of a SRA include:

- the availability of an existing data centre (impact staff and initial costs)
- the number of concurrent users (impacts staff, infrastructure, software licensing)
- the supported statistical packages (impacts software licensing)
- the number of data silos (impacts staff and infrastructure, usually increase # of users)
- the need of per project disclosure control or dataset customization (impacts staff)

We could tentatively classify facilities as follows:

- small: 1-3 data silos with less than 50 users registered researchers and/or 5-10 concurrent users
- medium: 3-10 data silos with 50-300 registered researchers and/or up to 30 concurrent users
- large: over 10 data silos or hosting multiple SRAs, over 300 registered researchers or 30 concurrent users

A small entry level SRA with base software can likely be established for as little as €50K, with medium facilities ranging between €100K and €500K, while large infrastructure will require between €500K into millions. A 20% annual maintenance fee should be factored in, as per routine annual maintenance, IT infrastructure upgrades, etc., while staff time is recurrent.

One important fact to take into consideration is that duplicating an existing model and sharing lessons learned may potentially reduce overall costs significantly. As an example, the initial setup of the NORC Data Enclave, a medium size facility, is estimated to have cost around $750K-$1M (€525K-700K). The same model is now being replicated at the UK Data Archive with an initial budget of about £200 (€225K) and University of Pennsylvania has established small internal secure data facility for around €75K. Having a set of various size vs. utility reference architectures harmonized across the European framework would greatly decrease the investment of establishing a network of SRA. It might also allow such networks to negotiate preferred rates with equipment and software vendors as well as build common technical expertise amongst staff.

Staff

In terms of staffing, the following positions need to be filled:

- SRA manager(s): to coordinate activities, report to stakeholders, advocate the facility, and dialog with data providers.
- IT administrators: to configure, maintain and operate the data centre and the SRA.
- Data/Metadata administrator(s): to manage the data/metadata in the internal archive and distribute information to researchers.
- Knowledge manager(s): to maintain and moderate the collaborative environment and feed the public/internal web site.
- Training / Technical support team: to deliver training, help user accessing and using the environment, and answer on dataset specific questions.
- Statistician(s): for disclosure control and output review.
- Support staff: for registration, project processing, accounting, general administration, contracting, legal advisor, etc.

A small SRA facility can likely operate with a team of 2-4 individuals dedicated part time to the facility. In medium size SRAs, a team of 4-8 full time and part time individuals is likely necessary. Large SRAs require a full team of dedicated staff. It is also important to remember that each individual should have at least one backup person in case of absence. This is particularly critical for IT expert, technical support staff and output reviewers. Note that it is common and expected in small or medium facilities for a single individual to assume multiple responsibilities and therefore cover several positions.

Regarding technical support, the SRA staff should provide full training and assistance on how to operate the environment. For dataset specific questions, a first line of support can optionally be provided by the SRA staff but we strongly recommend involving the data producer in this activity. For example, all the data providers to the NORC DE assign a dedicated contact per research project. Engaging the researchers themselves in assisting each other is also recommended, particularly of statistical software specific questions.

The need for full time or part time statistician will greatly vary based on the number of users, their level of productivity, services provided; and disclosure review policies.

Requirements summary

☑ **Costing Plan**

☑ **Staffing Plan**

Sharing data across borders / legal aspects

Overview

Providing access to disclosive data raises numerous legal and organizational challenges such as determining if an individual is an accredited researcher, which legal framework the user falls under when accessing the data, contractual agreements with institutions and researchers, ways to legally prosecute users in case of breaches of agreements, statistical disclosure control and privacy, and many others. We assume the reader is familiar with these issues, as many have been extensively discussed and documented. The situation gets more complicated when operating in a remote access environment as users may connect from foreign countries or obtain access to data falling under different jurisdictions. Data silos may also host data from external or foreign sources or multi-country datasets. These will need to be addressed at the national, European and international level by relevant regulatory bodies.

The two sections below cover (1) the legal aspects that are in place for access to confidential microdata in Europe and (2) present-day arrangements for access to microdata that are - and which in the future may be - held and controlled by NSI's and/or CESSDA members. First we focus on currently available arrangements within Europe and the CESSDA member countries for remote access, remote execution and microdata access in safe centres. Thereafter scenarios and practical recommendations are discussed that could be of use to CESSDA's aim to enhance the conditions for access to microdata for research purposes.

WP10 Audit Report

The CESSDA WP10 "Audit report on access mechanisms and availability of official statistics across the European Research Infrastructure", provides an excellent overview of the range and complexities of existing arrangements for access to microdata in Europe [19]. The report also describes the current role of CESSDA in providing access to government microdata and other microdata services to the research community. Furthermore, the report sums up a number of recommendations for CESSDA to enhance its intermediary role in the European research infrastructure. For example, according to the report additional effort should be put into fostering and preparing (additional) cross-national agreements for microdata access with NSIs and other ESS statistical bodies. Also a more prominent role of the research community is foreseen in the governance structure of CESSDA, as well as the need for a pan-European long term cooperation between CESSDA and authorities involved in the ESS.

Looking more closely at the recommendations made in the WP10 report we signify a number of areas that need further clarification if the aim of CESSDA is to take up a more prominent role in providing access to confidential microdata. The issues are spelled out below.

If CESSDA aims at a role in the storage, access-management and the provision of microdata research data services, the following issues should be negotiated with statistical authorities:

- The release of confidential data from statistical authorities to safe data centres and data laboratories;
- A distributed storage of data sets in CESSDA-member countries vs. a centralized storage of datasets. A central storage option is preferred when datasets from different countries can be embedded in an international research program. The LIS study is a best practice for this situation. A decentralized approach to storing data and managing access is on the other hand more feasible if legal complications at a national level are foreseen. However, it should also be noted here that European Regulations in principle precede the National law and therefore can act as a precautionary measure to prevent custom legal complications.

Legislation

European legislation for access to confidential data

Besides the use of microdata for statistical information provision on the characteristics of a population, microdata are an important source for socio-economic research and policy development and evaluation. The collection of microdata is very costly, and intelligent approaches are consequently needed to ensure that the data collected are used efficiently. The use of statistical microdata is embedded in a national and European legal framework that safeguards the privacy concerns of individuals and other information providers. Part of this legal framework is the recognition that access to microdata for scientific research purposes and policy evaluation is important. Microdata are collected by a wide range of organizations i.e. NSI's, governmental organizations (i.e. central banks), international organizations (i.e. Eurostat, OECD), research institutes and privately held companies. The increased demand for access to microdata clearly has implications for the research infrastructure both at a national and at the pan-European level.

CESSDA is now recognized as an important European Research Infrastructure [1], representing the social science data archives of 20 countries across Europe. CESSDA has also more than 30 years of experience of collaborating with NSI's, research institutes which have collected datasets of national importance and research funding organizations. This sub-section first highlights the legal regulations that are enforced within the European Community (EC) and the CESSDA member countries for access to confidential microdata. The second part of this sub-section will sketch the contours of the scenarios that CESSDA may follow to strengthen its intermediary and proactive role to foster an enhanced infrastructure for access to microdata.

Over the last few years several documents have been published which focus on the need to improve arrangements and the infrastructure for access to confidential microdata [2-5]. Museux and Bujnowska [6] presented earlier this year the contours of a European infrastructure for microdata access encompassing the following (not mutually exclusive) options: 1. the set up of a number of decentralized centres for access to confidential data, 2. the set up an integrated network of safe centres with remote access facilities for confidential data, 3. a more intensive collaboration with social science data archives in which all prerequisites for access to confidential data are addressed and 4. an increased effort of all stakeholders involved, directed to the implementation of the new legal framework.

The establishment of the cessda-ERIC, under Council Regulation 723/2009, will strengthen existing, and build new links between data providers and the network of European data archives. The cessda-ERIC will co-ordinate activity between its membership and the owners and distributors of data. It will enable knowledge transfer and exchange and will work closely with organisations with similar goals to build a sustainable, collaborative and innovative system for the benefit of researchers across the European Research Area.

Legal regulations that govern access to confidential data within the European Community (EC)

Access to confidential data in Europe is governed by the framework regulation for European Statistics[117] and the European Statistical System (ESS) [8-10]. Access to confidential data for scientific purposes was first acknowledged by the European Community (EC) in 2002 with the acceptance of Commission Regulation 831/2002, thereby completing the legal instruments on confidentiality at the European level. In 2009 Regulation No 223/2009 on European Statistics has been adopted [11, 12], allowing more flexibility and further cooperation on the exchange of confidential data and access to such data for research purposes. The following quote stems from the new regulation that is in place since April 2009:

> "*The research community should enjoy wider access to confidential data used for the development, production and dissemination of European statistics, for analysis in the interest of scientific progress in Europe. Access to confidential data by researchers for scientific purposes should therefore be improved without compromising the high level of protection that confidential statistical data require.*"

Following the adoption of the new regulation, and incorporating the notion that cumbersome procedures to grant microdata access for researchers were recently perceived as one of the weaknesses of the ESS, Eurostat will thereto [13].

- increase the efforts to meet the expectations of users.

[117] European statistics are defined in the Council Regulation (EC) No 322/97 on Community statistics. The Council Regulation provides a legal directive for statistics produced and disseminated by national statistical authorities and Eurostat as the Community's statistical authority. Statistical authorities are the NSI's (NSIs), other statistical bodies in charge of producing and disseminating European statistics and Eurostat (Community level).

- increase efforts to smooth procedures for researchers having access to data covered by regulation 831/2002 and its modifications [12].
- ensure the support and the dissemination of results to the ESSnet safe centres/remote access.
- explore the feasibility of simplifying the procedures for supplying data to researchers via data archives.

The new framework regulation clearly offers a starting-point to improve microdata access for research purposes within the Member States and a base to extend the network of authorities that could liaise with the partners involved in the ESS. The EC regulation provides an adequate legal base to release and provide access to micro-data for scientific purposes. It should also be noted however that in parallel to the Community legislation, each Member State has its own rules and procedures governing the confidentiality of microdata and the provisions for access to microdata for research purposes [14]. European countries therefore also differ on the degree of conformance to the European Statistics Code of Practice.

The next section provides a summary of the EC regulations that are relevant for the release and access to confidential statistical data. An overview of the Community Legislation in force in the field of statistics is available at the website of the European Statistical System [15].[118] The subsequent section then briefly comments on the European Statistics Code of Practice.

Overview
Within the European Community the following regulations are enforced for access to confidential data:
Council Regulation (EC) No 322/97 on Community statistics, Chapter V Statistical confidentiality [16].
Council Regulation (EEC, Euratom) No 1588/90 on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities [17].

Both EC regulations deal with the safeguarding of the privacy of the information provider.

1. Commission Regulation (EC) No 831/2002 implementing Council regulation 322/97 on Community statistics, concerning access to confidential data for scientific purposes (with its further amendments) [18].

The implementing regulation (EC) No 831/2002 foresees two possible ways of access to confidential data for scientific purposes:
- Via the Eurostat safe-centre, which provides access to confidential data that are obtained from national authorities
- By distribution of CD/DVDs with anonymised microdata, obtained by modifying the confidential personal records to minimize the risk of disclosure. Both the EU-

[118] Note that the latest regulation is not yet included in this document.

Labour Force Survey and the EU-Statistics on Income and Living Conditions (EU-SILC) make use of this second option [19].

Risk monitoring takes place through: 1. safe settings 2. applying disclosure risk control for microdata sets, according to EC best practices and 3. by applying Eurostat safe centre rules of procedure. Safe people in the EC regulation framework are researchers in EU research bodies / universities under legal/license contract and bodies commissioned by the Commission. There is an admissibility procedure in place for others.

4. Commission Regulation (EC) No 223/2009 on the transmission of data subject to statistical confidentiality [11]. This regulation is in force since April 2009. Of particular relevance is Article 23, Access to confidential data for scientific purposes:

> *"Access to confidential data which only allow for indirect identification of the statistical units may be granted to researchers carrying out statistical analyses for scientific purposes by the Commission (Eurostat) or by the NSIs or other national authorities, within their respective spheres of competence. If the data have been transmitted to the Commission (Eurostat) the approval of the NSI or other national authority which provided the data is required. The modalities, rules and conditions for access at Community level shall be established by the Commission. Those measures, designed to amend non-essential elements of this Regulation by supplementing it, shall be adopted in accordance with the regulatory procedure with scrutiny referred to in Article 27(3)."*

European Statistics Code of Practice, Principles 1 and 5

The latest EC framework regulation (No 223/2009) builds on the 15 principles that comprise the European Statistics Code of Practice which is in place since 2005 within the European Statistical System[119] (ESS) [11]. The Code of practice elaborates two statistical principles for access to microdata for research purposes: the principle of statistical confidentiality (Principle 5) and the principle of accessibility and clarity (Principle 15).

Eurostat recently published a summary report of good practices on the implementation of the European Statistics Code of Practice [14]. In this report good practices are highlighted for a number of countries against the statistical principles that underpin the legal framework of the European Statistical System (ESS). Two of the 15 principles that comprise the Code of Practice are especially relevant in the context of this project: Principle 1, which covers the legislative underpinnings for the production and the release of statistics by national statistical authorities (NSA's) and

[119] The European statistical system (ESS) refers to the partnership comprising Eurostat, NSI's and other national statistical bodies responsible in each Member State for producing and disseminating European statistics.

other statistical bodies, and Principle 5 which covers statistical confidentiality and access to microdata for research purposes.

The first principle focuses on the professional independence of the statistical authorities from other policy, regulatory or administrative departments and bodies, as well as from private sector operators. The Eurostat report [14] signifies the legislative framework of Ireland as a model which could be used by other countries. The report also refers to the Czech Republic, Ireland, Lithuania, Liechtenstein, The Netherlands and Slovenia as further examples of good practice countries in this area.

The Statistical confidentiality principle addresses the privacy of data providers (households, enterprises, administrations and other respondents), the confidentiality of the information they provide and the requirement that the data gathered are used only for statistical purposes. The Statistical confidentiality principle also encompasses the access to microdata for research purposes. Indicators for conformance to the principle of Statistical confidentiality are [20]:

– Statistical confidentiality is guaranteed in law.

– Statistical authority staff signs legal confidentiality commitments on appointment.

– Substantial penalties are prescribed for any willful breaches of statistical confidentiality.

– Instructions and guidelines are provided on the protection of statistical confidentiality in the production and dissemination processes. These guidelines are spelled out in writing and made known to the public.

– Physical and technological provisions are in place to protect the security and integrity of statistical databases.

– Strict protocols apply to external users accessing statistical microdata for research purposes.

Denmark, Finland, Germany, Iceland, Italy, Norway, Slovenia and Spain are mentioned as examples of countries with good-practices for microdata access for research purposes [14].

Each country and authority in the ESS has committed itself to work towards the implementation of the European Statistics Code of Practice during the coming years following a self-regulatory approach.

Case Study: Statistics Netherlands (Centraal Bureau voor de Statistiek)

At Statistics Netherlands (SN) access to microdata is provided via 1. licensed microdata files that are released for scientific research purposes, 2.a remote access facility and 3. on-site access to microdata at SN. The microdata files are disseminated usually on CD-ROM, to interested researchers that are qualified according to the law or to the CCS. Note that the microdata released on CD-ROM are disseminated at a fairly high aggregation level to prevent disclosure. Since mid-2005 a *remote access* facility has been developed, making it possible for researchers to analyse microdata present at SN through a secure connection from workstations in their own institute. The costs involved in setting up a remote access facility to SN data can be traded off with the possibility to analyze the microdata locally at SN. Noteworthy is that researchers who make use of the remote access facility can access the microdata at the same level of detail as researchers having on-site access to the microdata.[120]The interested reader is also referred to the case studies (Annex 1.6 en 1.13) provided by the UNECE [21].[121]

The microdata services are available only to researchers of trustworthy institutes as specified by Dutch law, or of institutes that have special permission to access microdata, subject to approval of the Central Commission for Statistics (CCS) of Statistics Netherlands [22]. Individual researchers using the services are further obliged to sign a confidentiality statement. This confidentiality statement is also co-signed by their institute. Microdata are made accessible under a contract or license to legitimate researchers only. Section 41 of the law cites the researchers that are qualified. These include the universities and other research institutes with a legal foundation, but also Eurostat and the EU NSIs. An appendix to the contract is a confidentiality statement to be signed by each individual researcher with access to the data. The CCS has no principal objection against admitting non-EU universities, for example, but a commercial bank or a journalist would not be eligible.[122]

The legal context for providing access to microdata for academic purposes is provided in the Statistics Netherlands Act (SNA) which was adopted in 2003.[123] Relevant excerpts from the SNA for this project stem from section 41 (part 3):

- Contrary to the provisions of Section 37 the director general may, on request, provide or grant access to a set of data to a department, organisation or institution as referred to in the second subsection for the purposes of statistical or academic research where appropriate measures have been taken to prevent identification of individual persons, households, companies or institutions from those data [subsection 1].
- ‘A set of data as referred to in the first subsection may be provided to or made accessible to:
  a. university, within the meaning of the Higher Education and Research Act;
  b. an organisation or institution for academic research established by law;
  c. planning offices established by or by virtue of the law;
  d. the Community statistical agency and national statistical agencies of the

[120] The authentication procedure however differs between the two procedures: a fingerprint identification is used with remote access.

[121] Note however that some of the information provided in this document is no longer up-to-date.

[122] Statistical Commission, 2007, Annex 13, p. 61 (UNECE Task Force on Microdata Access).

123The Statistics Netherlands Act: is available from: http://www.cbs.nl/NR/rdonlyres/BBD8113D-7EE5-4BE4-8879-685253B31882/0/statlawen.pdf

member states of the European Union;
e. research departments of ministries and other departments, organisations and institutions, in so far as the CCS has given its consent. [subsection 2]

SN does not hold datasets from other countries. Access is granted to both non-profit and profit organisations and is only granted for research that results in publications that are accessible in the public domain. CBS Netherlands controls whether this is the case. The origin of a person is not a criterion for granting or withholding access, as a person does not represent a legal entity.. Furthermore organisations cannot simply apply for access to any dataset they are interested in; organisations are required to motivate the aim of their research before access to special interest files is granted (i.e. health care data, crime data). Before access is granted a research proposal is evaluated on 'scientific soundness' and whether the data is suitable for this research proposal. Initiatives to improve the conditions for access to microdata are also on the agenda of Statistics Netherlands. Mol et al. [22] for example point to future developments that are foreseen at Statistics Netherlands:

"Statistics Netherlands would like to develop its microdata facilities into a central node in the microdata research landscape. To do this, we are exploring possibilities to host data from other parties in our safe environment. In this way, researchers can access data from different organisations in the same environment, thus opening up new possibilities for linking datasets. At present, a first such pilot project with the Dutch National Central Bank is underway. Considering possibilities of setting up European networks to make cross-border access to microdata possible is on the horizon. Statistics Netherlands is currently participating in several European projects aimed towards this goal."

Case Study: Luxembourg Income Study

The Luxembourg Income Study, known as LIS, is a non-profit microdata archive and research institute. LIS, located in Luxembourg since 1983, serves a global community of researchers, educators and policy makers. LIS acquires datasets with income, wealth, employment, and demographic data from a large number of countries, harmonizes them to enable cross-national comparisons, and makes them available for public use by providing registered users with remote access. The LIS archive includes two primary databases, the *Luxembourg Income Study Database*, which focuses on income data, and the newer, smaller *Luxembourg Wealth Study Database*, which focuses on wealth data. LIS enforces a set of fee structure rules to ensure the future financial stability of the project. Access to the LIS is granted under the following restrictions[124]:

- The user must be a researcher working for an academic, government or non-profit organization.

[124] See also: http://www.lisproject.org/data-access/data-access.html for detailed information on LIS' conditions for access to the microdata.

- The use of the microdata is restricted to Social Science research purposes only. No private or commercial use is permitted.

Accredited researchers can use a remote execution service to the microdata that are hosted at LIS. Researchers can submit their SAS, Stata or SPSS program files to LIS where they are then executed under control of LIS staff. Full access to the microdata is only granted on-site, and only for some microdata files and only for some countries. Noteworthy is that LIS does not internally use the distinction between disclosive and non-disclosive data.

Access to the microdata is limited to on-site use only. Users work on a secure computer without access to Internet, email, or any electronic storage devices. Therefore, at the end of the visit, only the aggregated results produced will be available. An authorized LIS staff member checks these prior to their delivery.

Use Cases

To illustrate the various situations that can emerge in an international SRA environment like CESSDA, we can define a set of use cases that take into account the legal framework under which the user operates, the type of data being accessed and the location the user is accessing from.

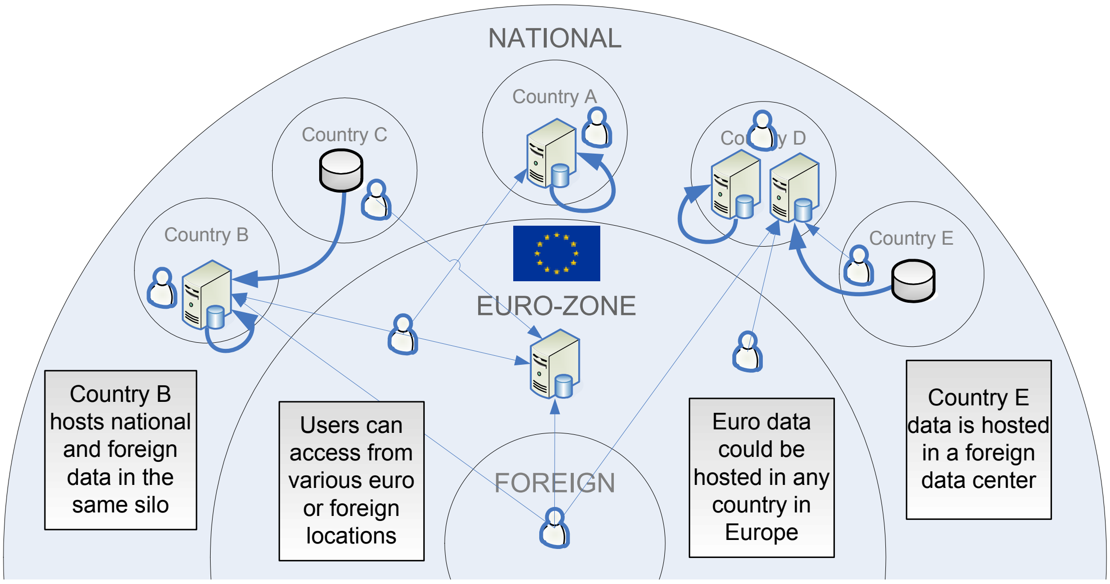
The national legal framework under which the user falls is critical to establish if the access agreement can be properly enforced. As a user getting granted access to disclosive data should always do so under an institutional umbrella (direct personal access is not considered a safe practice), this is often determined by the agency hosting the researcher. We break this down into three categories:

- same country as the data: national laws can apply
- foreign but within the Euro-zone: national laws difficult to enforce but European regulations are applicable
- foreign outside Euro-zone: difficult to enforce unless bilateral agreement is in place. This can also be the used case of a roaming user on the Internet.

While typically the access point from which the user connects to the SRA facility will be in the same as the hosting institution, this may not always be the case. The same categorization as above should therefore apply.

Finally, the type of dataset being made available in the data silo can introduce various levels of complexity. Data sources can be organized as follows:

- National data only: all the dataset being made available belong to the same country and therefore fall under the same legislation. This is the simple use case.
- European data: harmonized European surveys or data collected across countries that can be made accessible under European regulations,.
- Foreign data: a silo that contains dataset from a foreign country
- Multi-national data: a silo that contains datasets from multiple countries


NATIONAL
Country A
Country C
Country D
Country B
Country E
EURO-ZONE
FOREIGN
Country B hosts national and foreign data in the same silo
Users can access from various euro or foreign locations
Euro data could be hosted in any country in Europe
Country E data is hosted in a foreign data center

An important aspect to keep in mind as well is that there are essentially two levels of agreements that are being established:

- the first between the SRA facility and the institution hosting the researcher, with the assumption that the individual falls under the institutional legal framework (sanctions can be taken against both the agency and the users)
- the second between the hosting institution and the individual to ensure that the user is bound to the legal framework or can be prosecuted if needed

The latter can be tricky and is often beyond the control of the SRA.

Data silo with National datasets only

This is the classic case where national datasets are made available through an SRA based in the same country.

		ACCESS POINT		
		National	Euro-zone	Foreign
. LEGAL FRAMEWORK	National	Example: a UK researcher accessing UK data from the UK This should present no particular problem as national legislation applies. Note; EU regulation precedes National law, but NSO' are free to choose how they will distribute and provide access to microdata .	Example: a UK researcher accessing UK data from France Acceptable as long as accreditation and procedures for granting access are in line with European regulations and European Statistics Code of Practice. Contracts need to be in place both in the UK and France (at ESS authority level) Note: Local law might exclude access to foreigners	Example: a UK researcher accessing UK data from Canada Acceptable as long as accreditation and procedures for granting access are in line with European regulations and European Statistics Code of Practice. Contracts need to be in place both in the UK and Canada. (at ESS authority level) Ambiguity of legal arrangements to enforce in case of breaching makes this a special case Note: Local law might exclude access for foreigner
	Euro-zone	Example: a Norwegian researcher accessing UK data from the UK Acceptable as long as accreditation and procedures for granting access are in line with European regulations and European Statistics Code of Practice. Accreditation contracts need to be in place in the UK (at ESS authority level) Note: Local law might exclude access to foreigner	Example: a Norwegian researcher accessing UK data from Norway or France Acceptable as long as accreditation and procedures for granting access are in line with European regulations and European Statistics Code of Practice. Accreditation contracts need to be in place in Norway and France (at ESS authority level) Note: Local law might exclude access to foreigner	Example: a Norwegian researcher accessing UK data from the USA or Australia. Complicated accreditation. A use case would be data gathered on a specific topic for an international research program (i.e. LIS). Agreements need to be in place with all data providers as are accreditation procedures for approved research organizations and researchers Note: Local law might exclude access to foreigner
	Foreign	Example: a USA researcher accessing UK data from the UK Acceptable as long as accreditation and	Example: a USA researcher accessing UK data from the Netherlands Complicated	Example: a USA researcher accessing UK data from the USA Could be granted on a case by case basis

		procedures for granting access are in line with European regulations and European Statistics Code of Practice. Accreditation contracts need to be in place in the UK (at ESS authority level) Note: Local law might exclude access to foreigner	accreditation and hosting procedures depending on the National law of the data providing and hosting country. A possible use case again would be data gathered on a specific topic for an international research program (i.e. LIS) or a situation where National Statistics hosts UK data. Note: Local law might exclude access to foreigner	based on bilateral agreements ((i.e. LIS research centre In New York)

Data silo with foreign datasets

The situation can arise when (1) a European partner does not have the national capacity to host a SRA or (2) a foreign country accepts to deposit data in the SRA facility for local researchers. The most significant barrier to this use case is that the data must be allowed to leave national ground. Beyond this, it becomes similar to the previous use case regarding national data.

There is no complete picture for this case. Germany, Hungary and Norway for example do not allow data to be disseminated outside of the country. The Netherlands prohibits the dissemination abroad of datasets on the full Dutch population. More research is therefore recommended.

Data silo with European or cross-country harmonized datasets

Some of the issue that need to be examined here regards which country can host the data or can data be freely transferred between SRA facilities in different countries.

		ACCESS POINT		
		National	Euro-zone	Foreign
LEGAL FRAMEWORK	National	N/A	This should present no particular problem as European regulations directly apply. Note: Local law might exclude access to foreigners and prohibit data dissemination to other countries	N/A
	Euro-zone	This should present no particular problem as European regulations directly apply	Acceptable as long as accreditation and procedures for granting access are in line with	Should be possible as long as hosting institution can legally bind the user at the foreign access

		Note: Local law might exclude access to foreigners and prohibit data dissemination to other countries	European regulations and European Statistics Code of Practice. Accreditation contracts need to be in place at ESS authority level	point, possibly through a local partner. Should only be authorized for special cases.
	Foreign	N/A	The foreign researchers should instead seek a hosting institution in the Euro-zone	Could be granted on a case by case basis based on bilateral agreements

Data silo with multi-country datasets

Multi-country data silos can lead to very complex situations if one attempts to simultaneously apply national legislations. The recommended way to manage this situation is to simply agree that the data falls under the European regulations which essentially transform it into European data use case. This could be a requirement for CESSDA partners or interested data depositors. The same could be required in case this combines data from countries outside the Euro-zone.

International access

The statistical commission of the UN(ECE) discusses access to microdata from an international perspective and lists a number of options for researchers that want access to (confidential) datasets from other countries. The options include:

1. Request access to licensed anonymised microdata files, where countries are able to do this;
2. Make use of remote access facilities with appropriate safeguards;
3. Collaborate with researchers based in the NSI or the NSI's country, who have access to the microdata.

Summary notes

1. A discussion should be held on the current position of CESSDA in the statistical value chain and the role that CESSDA could play here.
2. The methods that CESSDA could use to support the research community should be clarified further. More specifically the role of CESSDA in the dissemination stream for microdata and the envisaged services need to be articulated [5].
3. The practical implications of a more self-regulatory approach are far from clear yet.
4. The added value for researchers of services delivered by CESSDA compared to having a direct connection with a NSI should be further detailed.

Practical recommendations and solutions[125]

From a high level viewpoint we foresee two scenarios that might be of practical relevance to CESSDA. The first scenario strongly dwells on the current role of CESSDA as a networked intermediary; the second scenario elaborates a more proactive and extensive role for CESSDA in the European research infrastructure. Both are summarized below. By adapting the second scenario CESSDA could extend its role as an 'intermediary organisation' with a role as a 'producer'.

Enhance the current intermediary role of CESSDA through:

- A focus on the harmonization of accreditation procedures and licenses for access to microdata in Europe.
- Fostering liaisons with organizations both at a national level (i. e. NSI, Central Bank) and at an international level (Eurostat), to improve the conditions for access to microdata for scientific research.

This scenario is largely bottom-up. The focus here is on extending and improving ad-hoc solutions, i.e. developing (more) remote access points and safe centres in countries where these are not available and helping researchers/institutes with the accreditation and application forms for access to microdata.

Establish a central role for CESSDA in the research infrastructure through:

Extending the intermediary role of CESSDA as we know it, to an active role as a producer and provider of additional research data services (discovery, comparison, providing detailed information on data quality, metadata services, etc.). This role will strengthen at the same time the connections with NSIs, governmental bodies and the research community. CESSDA can play for instance a crucial role in building a catalogue which provides, for each member country, information on the most relevant datasets that are available for research purposes. Such an initiative for a one-stop-shop with information on the availability of datasets and detailed information on the data items would we welcomed very much by the research community. More specifically, CESSDA could play a key role in developing metadata production systems, metadata repositories and metadata registries to articulate the similarities and differences across countries of the member organisations [23]. CESSDA can also contribute significantly in facilitating search and discovery at a detailed level for data elements (variables), population characteristics and the comparison of attributes that might be eligible for comparative analysis and extensive statistical modelling. The success of LIS is due, at least partly, to the additional research services that are provided.

Take up a lead role in the harmonization of accreditation procedures across Europe for access to microdata.

[125] See also: Description of work for WP10: "*The report needs to describe the potential advantages of extending the SDS system to the whole of CESSDA, as well as the difficulties that might arise. Focus should be both on technical issues and on the data needs of the social science research community. The report should also provide guidance on whether and how any difficulties can be realistically overcome, how long it may take to do so, and at what approximate cost.*"

a. Harmonize all microdata access accreditations and user licenses across countries and harmonize all other procedures (i. e. breaching penalties).

b. With respect to the harmonisation of accreditation procedures CESSDA could probably adapt the following step wise approach:
    - accrediting safe organisations
    - accrediting safe projects.

CESSDA could work out and negotiate concrete proposals with constituencies that focus on safe organizations and safe places (safe centres, remote access points). The accreditation of safe projects (i.e. research proposals) will probably stay in the hands of domain specialists.

Legal section references

1. ESFRI, *European roadmap on research infrastructures*. Update 2008.
2. Elias, P. *Towards an improved research infrastructure for the social sciences: future demands and needs for action. Providing data on the European level*. Working papers 2008 [cited; Available from: http://www.ratswd.de/download/workingpapers2008/46_08.pdf.
3. Lane, J., *Improvements and future challenges for the research infrastructure: administrative transaction data.* 2009, RatSWD
4. Rowland, S., *An examination of monitored, remote microdata access systems* 2003.
5. UNECE, *Principles and guidelines for managing statistical confidentiality and microdata access*, U.T.F.o.M. Access, Editor. 2007.
6. Museux, J. and A. Bujnowska, *Access to confidential data for research: State of the art for EU datasets.* 2009, Eurostat Unit B5 - Methodology and Research.
7. EC, *On the Community legal framework for a European Research Infrastructure Consortium (ERIC).* Official Journal of the European Union, 2009(L206): p. 1-8.
8. Ottosson, H. *The new regulatory frameworks on confidentiality in the European Statistical System* [PDF] 2009 [cited; Available from: http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/2_1_Eurostat.pdf.
9. EC, *Recommendation of the commission on the independence, integrity and accountability of the national and Community statistical authorities*. 2005: Brussels.
10. Museux, J.-M., M. Peeters, and M.-J.S. Santos, *Legal, Political and Methodological Issues in Confidentiality in the European Statistical System*, in *Proceedings of the UNESCO Chair in data privacy international conference on Privacy in Statistical Databases*. 2008, Springer-Verlag: Istanbul, Turkey. p. 324-334.
11. EC, *Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics* Official Journal of the European Union, 2009. 5**2**(L 87).
12. UNECE. *Statistical confidentiality and disclosure protection (Eurostat).Theme 6.11 Data security and statistical confidentiality* 2009 [cited; Available from: http://www1.unece.org/stat/platform/display/DISA/4.6+Statistical+confidentiality+and+disclosure+protection+%28Eurostat%29.
13. Eurostat, *Eurostat self-assessment against the principles and indicators of the European statistics code of practice.* 2006.

14. Eurostat, *Summary of good practices identified during the European Statistics Code of Practice peer reviews carried out during 2006-2008*. 2008.
15. ESS. 2009 [cited; Available from: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_insite/pge_estat/tab_management.
16. EC, *Council Regulation (EC) No 322/97 on Community statistics, Chapter V Statistical confidentiality*.
17. EC, *Council Regulation (EEC, Euratom) No 1588/90*. 1990.
18. EC, *Commission Regulation (EC) No 831/2002 implementing Council regulation 322/97 on Community statistics, concerning access to confidential data for scientific purposes (with its further amendments).* . 2002.
19. Espelage, F. and L. Wahrig, *EU-LFS and EU-SILC: legal, processing and dissemination aspects*, in *EU-LFS and EU-SILC: 1st European User Conference*. 2009, Eurostat.
20. Eurostat, *European Statistics Code of Practice for the national and community statistical authorities*. 2005.
21. UNECE, *Managing Statistical Confidentiality & Microdata Access PRINCIPLES AND GUIDELINES OF GOOD PRACTICE* in *UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE*, UNECE, Editor. 2007, UNITED NATIONS: .
22. Mol, J., F. Hoeve, and O. Ten Bosch, *Controlled and secure access to microdata*. 2009.
23. Zeng, M.L. and J. Qin, *Metadata*. 2008, London: Facet. xvii, 364 p.

Conclusions

This report provides a comprehensive discussion of various secure remote access (SRA) platforms currently in place around the world, and highlights advantages and disadvantages to the various models. It also outlines the technical, organizational, operational, statistical and legal challenges associated with operating SRA facilities. Although one's decision as to which model (or models) to implement largely depends on the needs of the SRA architect and one's ultimate goals and objectives, this document provides a roadmap that is generic enough to guide decision making and to provide valuable context to inform the process along the way.

Among other things, the report emphasizes that a complete solution must combine technology with the appropriate levels of management and protection techniques (notably a portfolio approach). Moreover, it notes through technical specifications and illustrated use cases in social science and across the private sector, that the technologies and methods are now available to design and implement such solutions. Of particular note to CESSDA, the report explores the situation beyond the common use case of the "stand-alone facility" extending it in the context of a network of national and international SRA facilities. It also highlights some of the technical benefits and cost effective strategies involved with shared infrastructure and harmonizing methods.

The document also provides detail on a number of challenges presented in this endeavour and their potential solutions. CESSDA, both by its mandate and international breadth and depth, seems ideally positioned to lead and navigate through the barriers involved in supporting a SRA infrastructure for its consortium members and other stakeholders. Activities that might help develop a solid infrastructure toward the development of a network of SRAs might include:

- Providing infrastructure at the national or cross-national level
- Operating SRA facilities and provide related add-on services
- Providing blue print for technical configurations
- Playing a major role in harmonizing best practices in data/metadata management, disclosure output review, and other areas
- Supporting cross-country data and metadata harmonization efforts
- Coordinating network of researchers.
- Operating portals and collaboration networks
- Furthering research on legal issues related to national, trans-European and international remote access to disclosive data

Lastly, the report makes clear that providing improved and additional access to disclosive data through SRA is highly beneficial to the producers, archives, the researchers, and the social science community at large.

Annex 1: Remote Client Protection

This section examines in greater detail various methods for increasing the platform security by providing protection options for the computer system hosting the remote client terminal software.

Clients Hardware Model

The remote client can be operated on a desktop, laptop or thin client[126]. The preferred solution is the thin client as such machine typically has limited processing and data storage capabilities which considerably reduce the risk of someone attempting to compromise system integrity or steal information. Desktop is the next choice as they are unlikely to move around and are not supposed to operate from different locations. Laptops are in general not recommended as they can easily be moved around and are more challenging to secure.

System Ownership

An important characteristic of the client terminal is ownership: who is responsible for configuring and maintaining the machine itself. This plays a significant role in the level of security and trust of the terminal.

Ownership	Description	Use
User	The system is fully under the control of the user who can install software, component and alter the configuration.	System owned by users present the highest risk as basically no control is possible. An individual with sufficient technical skills can easily compromise the integrity of the terminal. This is therefore only recommended for highly trusted and responsible users and should be combined with extensive monitoring mechanisms.
Institution	The system is locked and controlled by a local system administrator. Users do not have permission to install software or alter the system configuration.	This significantly reduces the risk of system tampering. This is an acceptable option for trusted users. Some of the responsibility is actually transferred to the administrating agency and additional agreement may be established.
SRA facility	The system is configured and managed the secure remote	This offer full control over the station and, depending on

[126] http://en.wikipedia.org/wiki/Thin_client

	access facility. The machine can be shipped to the access point location and installed by local user/administrator or deployed by SRA staff as well.	configuration options (see below), provides highly secure access points and monitoring solutions. It does however require IT capacity and resources to configure deploy and maintain the clients.
Third party	The system is configured and managed by third party in charge of maintaining the remote clients operational and possibly monitor activity. This can be an external contractor or, in the case of CESSDA, a centralized/shared service centre provided by one of the member.	As above, this provides a high level of control and flexibility over the system and alleviates the need for the SRA facility to provide the IT resources.

Physical Protection

Physically protecting the client machine can significantly reduce the risk of tampering with the hardware or the removal of the station. Several companies provide system enclosures, entrapment lock down, screen filters and others.
A few examples are listed below:

- http://www.iboxtech.ca/
- http://www.computersecurity.com
- http://www.securityware.com/
- http://www.secure-it.com

Monitoring and control

Activity monitoring is an important aspect of the system security. The Citrix product line provides several options to facilitate such process on the server side but additional monitoring features can be implemented client side to enhance the overall security.

Self-diagnostic / monitoring utility	Custom utility software can be developed and installed on the client to ensure that the system integrity has not been compromised and that all the security features are operational. Such utility could report any issue or send copies of local logs to a security/monitoring centre or prevent the station from connecting to the remote access facility if it detects any issue.
Room monitoring	The client, if equipped with one of more webcam (build-in, USB or wireless), can be used to actually monitor the room. This could be relayed or even controlled by a virtual security centre. Other security devices could also be attached to the machines to essentially provide additional security measures to the room (i.e.

	motion detectors, door sensors, etc.).
User monitoring	Likewise, a local web cam could be used to monitor the user or take pictures when login in or at regular interval.
User support	Instant messaging software can be used to communicate with the user to provide technical support.
Remote control	Most operating systems allow administrator to take control over a machine remotely. This can be used for maintenance, diagnostic, support purposes.

Machine Identity

Every manufacture computer comes with various pieces of hardware that can be used to uniquely identify the machine. For example, the network card MAC address[127] or the CPU serial number can be used for such purpose. This can be used to define a unique signature to ensure that a user is using a specific system.
This feature is supported by Citrix SmartAccess.

Network Access Control

Considerable control over the system can be gained by limiting the system connectivity to know hosts such as the remote access facility or a monitoring and essentially preventing access to the Internet.

The router connecting the computer to the Internet should now allow for DNS resolution and only permit for network connection to authorized SRA facilities. This will prevent users from connecting to other web sites and potentially download files or application to the computer (this feature should likewise be locked in the browser and the hard disk read only for the local user account).

- Limited to remote access facility (by machine, router, etc.)
- No internet access (DNS, routes)

Biometric Authentication

Equipping the client terminal with biometric[128] hardware / software as an additional level of authentication significantly improves validating the identity of the user.
While these technologies have made huge progress in the past few years, a few drawbacks remain:

- The user registration process often need to be performed by a system administrator familiar with the system and requires the physical presence f the user
- Multiple login are usually required to increase the system accuracy
- Accuracy is not 100% which means we need an alternate login mechanism. This is typically a user-id/password combination which in essence somewhat defeats some of the advantages of biometrics. For such case, an option would be to require the presence, physically or virtually, of a system administrator to confirm

[127] http://en.wikipedia.org/wiki/MAC_address
[128] http://en.wikipedia.org/wiki/Biometric

the user identify if the biometric layers fail. This however may require customization of off the shelf solution.

Fingerprint	This feature is becoming widely available on laptop models and is common on thin clients. It provides a fairly high level of accuracy. The typical implementation of this feature is however to require authentication at login time. This might be insufficient in our case as the user can walk away and give control to someone else. An option would therefore to require the user to re-identify after a given time has elapsed (like every 15-30 min). This should not be a major inconvenience as it would only require a touch of the fingerprint scanner. Statistics Netherlands is using this solution for their SRA but based on a fingerprint reader device installed by a Statistics Netherlands staff member visiting the external research organization.
Iris Scan	Iris scanning is a high accuracy authentication system. This is not typically available off the shelf and limited options seem to be currently available. The UK based company Eyenetwatch offers a range of products[129] like the Panasonic DT120 Authenticam. The product however seems slightly outdated; we suggest contacting the company for more information. TNS Gallup is also using this approach to identify TV-watchers in their TV-monitoring survey but we have not at this time found more information on the technology.
Facial Recognition	SensibleVision FastAccess is a commercial software package that provides enterprise solution for facial recognition. The product provides *continuous authentication*: the system logs the user in when s/he approaches the machine and logs out within a few seconds as s/he walks away. Extensive information is available on the web site[130] as well as a video demo[131]. A "consumer version" of FastAccess is available on selected Dell computers. Some Lenovo and Toshiba[132] laptops have similar features available: - Lenovo has an option called Veriface available on selected IdeaPad models - Toshiba Satellite U400, Satellite M300, Satellite A300 and Satellite P300 are equipped with integrated Webcams. During a short setup process, users tilt their heads up and down and side to side so their notebooks can take and store several photos.
Other options	Other potential mechanisms include voice, signature or gesture recognition.

[129] http://www.eyenetwatch.com/iris/scanners.htm

[130] http://www.sensiblevision.com/products/fastaccess.htm

[131] http://www.sensiblevision.com/products/FAvideo.html

[132] http://explore.toshiba.com/innovation-lab/face-recognition

Citrix partners offer biometric authentication options that integrate directly in the XenApp platform. See the Citrix Ready[133] web site for a catalogue of products.

Location / Proximity Detection

Making sure the terminal operates from a pre-allocated location is an important aspect. While in some cases users might be authorized to move around, most of the time they should be expected to work from one or more authorized locations. There are various ways to ensure or monitor where a terminal is connecting from and takes actions if necessary.

Network Location	Every computer accessing the system over the Internet (or network) is assigned an Internet Protocol or IP address[134]. This is a set of 4 or 6 numbers that can be used to restrict access from specific locations (like an organization, university or even home) using IP ranges or IP geo-location. For example, this feature is directly integrated in the Citrix SmartAccess product and can also be used by a third party to monitor terminal connections.
GPS Location	Small USB Global Position System can potentially be installed or connected to the terminal to determine its precise location. This information can then be used by a monitoring utility to report to a security centre or prevent the machine to operate outside authorized zones like a specific room or building. This means that the terminal will not connect be authorized to connect if removed from the premises. This could also be used to self-report theft in case the computer is stolen. One difficulty of this approach is that most off the shelf thin clients or laptops do not come with a build in GPS which will then likely be an external device. This makes it susceptible to theft. While a self-diagnostic utility (see below) would detect the absence of the GPS device and shut down the machine, it might be a costly option. A significant drawback is that the GPS device usually requires a direct view of the satellites which is not always easy to provide, particularly if the station is in an enclosed room.
Proximity keys	A proximity key is in interesting feature that requires the user to carry a specific object to be authorized to log in the system. One of the weaknesses of this system is that the device can be passed along to another user who can then impersonate the researcher. One way to discourage this is to detect the proximity of more personal objects that

[133] http://www.citrix.com/ready

[134] http://en.wikipedia.org/wiki/IP_address

	an individual is less likely to give away. Imprivata [135] a Citrix partner company providing various proximity solutions and experimenting with using national ID card or passports as proximity devices. Cell phone proximity could also be an interesting option to explore as users would be unlikely to part with such item (most can be detected using Bluetooth pairing).
Proximity devices	Besides carry-on devices, another option to ensure that the terminal is operating from a known location is to detect other devices available n the neighbourhood. A good example of this is WiFi devices such as wireless access points that have a specific hardware signature. The SafeFrontier[136] products for example use such approach to verify mobile device locations.

Encrypted keyboard

Keystroke logging [137] is a commonly used method to attempt to steal accounts, passwords, and other information without the user knowledge. While this situation is unlikely to arise, there are several ways the risk can be alleviated:

- multiple level of authentication (beyond password)l
- one time password (using for example token keys)
- encrypted keyboards (see for example http://www.wireless-computing.com/)
- Anti-keylogging software (like http://www.spyreveal.com)
- system enclosure

Operating system and environment

One decision that needs to be made when specifying a terminal configuration is the operating system and local environment. Several choices are possible such the usual Windows environment (XP, Vista, and Windows 7), Windows Embedded, Linux, or even the upcoming Chrome OS announce by Google. This choice will actually often be driven by the security features that need to be supported but note that many of the products described in this section require a full windows environment.

Given that the users essentially only require a web browser and the Citrix client to connect to the remote access facility, restricting the installed applications to a bare minimum is also a good way to reduce the risks. In a windows environment, further control can also be gained by controlling the user “winlogon” process[138] and boot the user directly in a browser or a custom shell.

[135] http://www.imprivata.com/
[136] http://safefrontier.com/laptoptracking
[137] http://en.wikipedia.org/wiki/Keystroke_logging
[138] http://en.wikipedia.org/wiki/Winlogon

Secure / dedicated room

When a remote access node is deployed at a fixed location, the hosting institution has the option (or may be required) to provide a room dedicated to the access point terminal. This is particularly true when more than one user will be accessing the facility from within the organization (often the case for junior researcher in universities). A dedicated room improves the overall security and allows for the deployment of extra security measures. As an important aspect of establishing a virtual remote access environment is also to reduce the overall operating costs, it is important to note that many of the room security or monitoring features can be operated virtually/remotely, thereby alleviating the need for local staff to be present.

Virtual Security, Monitoring and Support Centre

Configuring, installing, monitoring and administering the terminal units can be a complex process that requires significant IT expertise. Such resources may not be available at the hosting institution or the remote access facility. Transferring these tasks to a third party or establishing a service shared by multiple access facilities is an attractive option to consider as it could reduce operational cost. Such facility could be operated by CESSDA for all the SRA facilities with responsibilities shared amongst partners. The centre roles will include monitoring terminals, remote access room and other security related activities. It could also be used to provide remote technical support to users, perform system diagnostics, or validate a user identify in case a local authentication mechanisms fails (like biometric).

Annex 2: Citrix Case Studies

The following Citrix solutions have been selected from the Citrix web site to illustrate some application of the technologies across various sectors. Additional examples can be found at: http://www.citrix.com/lang/English/ps2/segments/index.asp

A couple of military applications are also documented at:

- http://www.defencetalk.com/citrix-to-boost-us-armys-satellite-communications-capabilities-19996/
- www.accelerasolutions.com/downloads/PACAF.pdf

Addenbrooke's Trust Improves Care and Staff Work/Life Balance with Remote Access

http://www.citrix.com/English/aboutCitrix/caseStudies/caseStudy.asp?storyID=162549

Background

Addenbrooke's Trust, part of Cambridge University Hospitals National Health Service (NHS) Foundation Trust, serves approximately half a million people who live in Cambridge, United Kingdom (UK). With about 6,500 staff, Addenbrooke's hospital provides a wide range of clinical and non-clinical services. The trust is the teaching hospital for the University of Cambridge, a provider of specialist services and a centre for international research

Challenge

Flexible Access to Critical Patient Information to Improve Care
Addenbrooke's 400 consultants, specialists and doctors need to make potentially life-saving decisions, no matter the time of day or night. Under the previous system, doctors could receive phone calls during the night, requiring them to travel many miles to the hospital to consult test results and make a diagnosis. To accelerate delivery of patient care while easing the burden on clinicians, the trust wanted to give doctors real-time access to patient records and test results — including X-rays — anytime, anywhere.

Furthermore, all healthcare providers in the UK must meet best practice guidelines that are stipulated by National Health Service (NHS) Connecting for Health when providing access to patient records through electronic means. These guidelines demand two-factor strong authentication — such as domain authentication and smart cards — to confirm the identity of users who access the network externally.

Key Benefits

On-demand access to clinical information to improve healthcare decision making

Strong security for confidential patient information and compliance with NHS Connecting for Health regulations
Home access to improve work/life balance for staff
Centralised management for greater control and efficiency

Applications Delivered

Meditech pathology results software
HISS patient administration system for demographics and appointment
Web OCS clinical tests scheduling application
PACS digital medical software, such as X-rays
OIT UK's EMR for creation and retrieval of digital medical records
Microsoft® Exchange Server 2003

Networking Environment

Citrix Presentation Server™ running on two HP DL360 Dual Processor servers
Citrix Access Gateway™
Microsoft® Windows Server® 2003 and Windows® 2000 Server
Secure Computing SafeWord for Citrix and RSA SecurID tokens
Any connection type for home users, from ADSL to ISDN dial up
Home PCs and desktops from vendors using a variety of operating systems

Dutch Ministry of Defence: Creates dynamic application and desktop delivery model

http://www.citrix.com/English/aboutCitrix/caseStudies/caseStudy.asp?storyID=1688362

Background

The Ministry of Defence of the Netherlands consists of four operational commands: Royal Navy, Royal Army, Royal Air Force and Royal Marechaussee — the Defence Material Organization and the Support Command. The central staff is located at the Plein in The Hague. The ministry employs almost 70,000 civilian and military personnel. IT support for the Ministry of Defence and governmental chain partners is internally handled by IVENT, a service organization with 2,700 employees.

Challenge:

Continuously adapting IT to a dynamic organization
Around the year 2000, the Ministry of Defence implemented some large and complex client/server applications, including an electronic patient information application that often needed to be updated and maintained. Traditional solutions for software distribution were not satisfactory because it was impossible to reach many thousands of desktops in one weekend, for example, after conversion of a database in the back office.

“Any organization with tens of thousands of desktops and thousands of applications struggles with standardization and maintenance, which is a very complex process,” said Danny de Vries, application delivery specialist at IVENT. “At the Ministry of Defence — one of the largest employers in the Netherlands — this is no different. We looked for a solution with enough flexibility to continuously adapt to our dynamic organization.”

During the following years, numerous application delivery challenges were solved with Citrix XenApp™, including the demand for delivery of virtualized desktops. There was definite demand from a governmental chain partner to adapt the infrastructure to roll out future applications very quickly but make a head start combining the user experience of fat clients with a centralized and controlled environment.

Applications Delivered

Numerous applications, including:

- In-house-developed Medical Information System
- Microsoft® Internet Explorer®
- Published desktop with XenApp
- Virtualized Windows® XP fat client environment with XenDesktop

Networking Environment

- Citrix XenApp™, Enterprise Edition running on 250 HP servers
- Citrix® XenDesktop™, Platinum Edition running on 10 HP servers
- Microsoft ® Windows Server® 2003
- HP EVA storage area network
- RSA tokens

Key Advantages

- Sand-boxed published browsers create very secure Internet connections
- Secure, remote desktop access using any connection
- Adaptability to changing application and desktop needs
- Fat client look and feel in a centralized and controlled environment
- Ability to renew IT without burdening end-users
- Saves on expensive bandwidth upgrades

Barrett Steel Reinforces Customer Service with Citrix Access Gateway

http://www.citrix.com/English/aboutCitrix/caseStudies/caseStudy.asp?storyID=39352

Background

Operating through more than 40 subsidiaries, Barrett Steel is the largest independent general steel logistics company in the UK. Barrett has more than £36 million worth of steel in its warehouses, representing approximately 80,000 tonnes of general steel and special products. The company has sites across England and a large fleet of vehicles to distribute steel where needed. Barrett Steel has tripled in size over the last five years, and plans to continue growing by acquiring new sites.

The Challenge

Providing Affordable, Controlled Access to Users on the Move

In the fast-paced logistics business, responsiveness is crucial. As Barrett Steel continues to grow through acquisition, its centralised IT system, based on Citrix Presentation Server™, has been a key facilitator to the company's expansion strategy. By providing on-demand access to virtualised applications, Citrix Presentation Server enables the steel company to integrate new sites quickly and easily. The Citrix software also allows Barrett Steel to keep IT support costs to a minimum by utilising powerful centralised management and delivery capabilities.

To further improve responsiveness to customer needs, Barrett wanted to give its sales representatives access to real-time customer and stock information, such as the latest steel prices, while travelling or working from home. The goal was to enable them to interact with their office-based systems while at a customer site. Tony Smith, group IT director, explained: "We have a fleet of 50 representatives who spend most of their time on the road, visiting customers. They had a growing need to be able to access systems and push information back in a timely fashion, rather than waiting until they were back in the office a week or two later. Such delays could lead to outdated information in the system and the potential for lost sales or poor customer service."

Key Benefits

Flexible access to real-time data for field sales force to drive sales
Centralised administration to lower IT costs and strengthen security
Granular control of access rights and data encryption for improved security
Faster response to customers

Applications Delivered

A custom-developed sales system, delivered via IBM Client Access
Lotus Domino/Lotus Notes
Microsoft® Office Suite
Company intranet

Networking Environment

Citrix Presentation Server™ running on 10 IBM Netfinity servers
Citrix Access Gateway™ Advanced Edition, two appliances
Microsoft® Windows Server® 2003
250 Boscom thin clients
Inter-site links provided by 1Mbit ADSL connections

Bedell Group Improves Service to Legal and Fiduciary Clients with Citrix

Background

Bedell Group is based in Jersey, a British Crown dependency off the coast of France and one of the world's leading international offshore finance centres. Bedell Group comprises Bedell Cristin, a leading law firm specialising in providing legal advice to the offshore banking and finance industry, and Bedell Trust, one of Jersey's leading independent trust companies which, together with its subsidiaries, provides corporate management and related services.

The Challenge

Support Ambitious Growth whilst Ensuring High-quality Service

Bedell Group has pursued impressive plans for rapid growth in recent years, expanding into a number of new jurisdictions including London, Dublin, Geneva and Guernsey. The company has more than 200 employees and continues to expand at a rate of 20 percent annually.

Central to this expansion strategy is an ongoing drive to increase productivity. Remote access to a broad range of applications was needed to provide flexible working practice, enabling fee earners to respond to the client demands from wherever they might be.

Bedell Group also wanted a technology infrastructure that would support the firm's expansion plans for new offices and provide on-demand access to feature-rich applications from any location or device. Security of data, reliability of application access and scalability to align with growth plans were the three critical requirements.

Key Benefits

Supported business expansion through flexible access and streamlined branch openings

Offset £50,000 in IT hardware replacement costs

Ensured continued protection of sensitive data

Eased delivery and management of applications

Enhanced client service

Increased productivity by offering flexible working practice

Applications Delivered

More than 12 applications, including:

Interwoven MailSite 8.x document/case management solution

Microsoft® Office XP

Best Carpe Diem time tracking and recording software

Interface Interaction contact management solution

WinScribe Internet Author/Internet Typist digital dictation

Networking Environment

Citrix Presentation Server™ running on HP ProLiant DL380 servers
Microsoft® Windows Server® 2003
Frontier Authenticator
Dell Optiplex desktops and Latitude notebooks

CommunityBanks Invests in Secure, Managed Network Access

http://www.citrix.com/English/aboutCitrix/caseStudies/caseStudy.asp?storyID=31780

Background

CommunityBanks, a subsidiary of Community Banks, Inc. (Nasdaq: CMTY), is a financial services company that operates an extensive network of banking offices and ATMs throughout central and northeastern Pennsylvania and northern Maryland. In July 2005, the company merged with Blue Ball National Bank, whose 18 branches serve the southeastern portion of Pennsylvania. Blue Ball is now a division of CommunityBanks. Currently, the combined organization operates 70 branches.

Challenge

High Cost and Complexity of Application Access for Branches

Before Blue Ball National Bank merged with CommunityBanks, it faced the challenge of delivering and maintaining client/server applications on local servers and desktops at its 18 branch locations. Not only was it time-consuming to have IT staff travel to the branches each time a server or application upgrade was needed, but this approach also required shutting down the server for several hours, affecting the office's productivity. Tape backups had to be done on each server as well.

"To improve manageability and reduce the amount of time spent on routine maintenance and support, we wanted to consolidate servers and applications into our operations centre," said Jeff Lyons, IT Department system administrator. "We also planned to move to the Windows server platform and felt a centralized infrastructure would simplify that process."

In addition to reducing costs through server consolidation, the IT team hoped to save money by extending the existing three-year replacement cycle for its 300 desktops with thin-client devices. Finally, like other financial institutions, Blue Ball was concerned about data security and compliance with the Sarbanes-Oxley Act and other regulations. "We were looking for a way to lock down the desktop to prevent introduction of threats and viruses, and also to simplify deployment of security patches, antivirus software and other preventive measures," said Lyons.

Key Benefits

Enabled server consolidation for easier IT administration
Reduced desktop management costs
Supported upgrade to cost-effective thin-client devices
Provided secure, remote access to mobile and home-based users
Provided foundation for regulatory compliance and business continuity

Applications Delivered

Microsoft® Office
Microsoft® Internet Explorer
Harland Encore! (bank account administration suite)
Silverlake (core account processing)
Laser Pro (loan documentation preparation)
Bondpro (bond redemption)
Calyx and Loan Handler (loan tracking)

Networking Environment

Citrix Presentation Server™ running on 10 HP DL360 servers
Citrix Access Gateway™
Citrix Password Manager™
Microsoft® Windows Server® 2003
WAN
HP T5700 thin clients and legacy PCs

Lovells Mitigates Security Risk with Cost-effective Remote Access

http://www.citrix.com/English/aboutCitrix/caseStudies/caseStudy.asp?storyID=164399

Background

Operating from 26 offices across the financial centres of Europe, Asia and the United States, Lovells is one of the world's largest business law firms with more than 3,500 employees. The firm advises many of the world's largest corporations, financial institutions and government organisations, and regularly acts on complex, multi-jurisdictional transactions and commercial disputes.

The Challenge

Flexible Access to Promote Mobility and Business Continuity

As an international law firm, Lovells fields a mobile workforce — including lawyers and consultants — who require secure access to core business systems to provide quality service to their clients around the world. Clients require regular updates on cases, and this confidential information must be stored securely, while also being easily accessible. However, the high security levels within the firm prevented access to all information from outside the physical perimeters of the office.

Recognising the need for lawyers to have efficient and secure access to documents and information whilst away from the office, and facing changing business risks, bird flu being one example, Lovells sought a solution that would provide business continuity in the event of a disaster by enabling its senior lawyers to quickly and easily access core applications remotely. The solution needed to allow users to log on from any location and access their virtual desktop. It also needed to be highly secure, easy to use and manage, and cost-effective.

Key Benefits

Enhanced client service with global access to desktop applications
Enabled employees to work from home in the event of a disruption
End-to-end security protects confidential client information
Remote access reduces need for costly hardware at disaster recovery sites

Applications Delivered

Microsoft® Outlook®
Microsoft® Office Suite
OpenText Document Management software
Sage Carpe Diem time recording software
Lexis InterAction CRM software

Networking Environment

Citrix Presentation Server™ running on 30 HP DL380s servers
Citrix Access Gateway™ Enterprise Edition
Microsoft® Windows Server® 2003
Dell PCs and laptops

United Cerebral Palsy of NYC Supports Mission of Caring with Flexible Access

http://www.citrix.com/English/aboutCitrix/caseStudies/caseStudy.asp?storyID=25013

Background

United Cerebral Palsy of New York City, Inc. (UCP/NYC) is a leading non-profit agency providing direct services, technology and advocacy to children and adults of all ages with cerebral palsy and related disabilities. UCP/NYC provides a wide range of services to more than 10,000 New York City residents with disabilities and their families. The organization, which bills $85 million annually, operates 20 offices, four full-service clinics and 75 residences for cerebral palsy clients, and employs about 1,500 managers, counselors, physicians, nurses and administrative staff.

Challenge

Converting IT from Inhibitor to Enabler of Organizational Growth
Although UCP/NYC is a large enterprise with many locations spread across New York City, a complex array of services and facilities, and millions in revenues, its IT infrastructure had not kept pace with the organization's needs. Specifically, many functions were paper-based and business applications ran locally in the remote offices, forcing staff to stay at work if they needed to put in extra hours and requiring mobile employees to load software on their laptops. "As far as being able to communicate across the sites, it was inefficient," recalled Jim Brown, CIO. "In some respects, every location was an island unto itself."

Further, the organization's headquarters near the 9/11 terrorist attacks on the World Trade Center prompted concerns about how UCP/NYC would cope if the facility had to shut down. "We needed a solution that would allow us to give people remote access to a second datacentre in case something happened in Manhattan," said Brown. In addition, the IT department was investigating the use of thin clients instead of PCs to save money and administration, and wanted a solution that would support them.

As a healthcare organization, UCP/NYC needed to ensure compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations, including managing and monitoring the use of application passwords and safeguarding the privacy of patient information.

Finally, Brown and his team faced an immediate problem delivering FundWare, an "industrial-strength" non-profit accounting package with a very large client component that delivered inadequate performance in a LAN or WAN client/server environment.

Key Benefits

Improved organization-wide communication and efficiency
Improved staff satisfaction and client service with secure remote access
Enabled compliance with HIPAA regulations
Provided a solution to support business continuity strategy
Reduced IT hardware and administrative costs

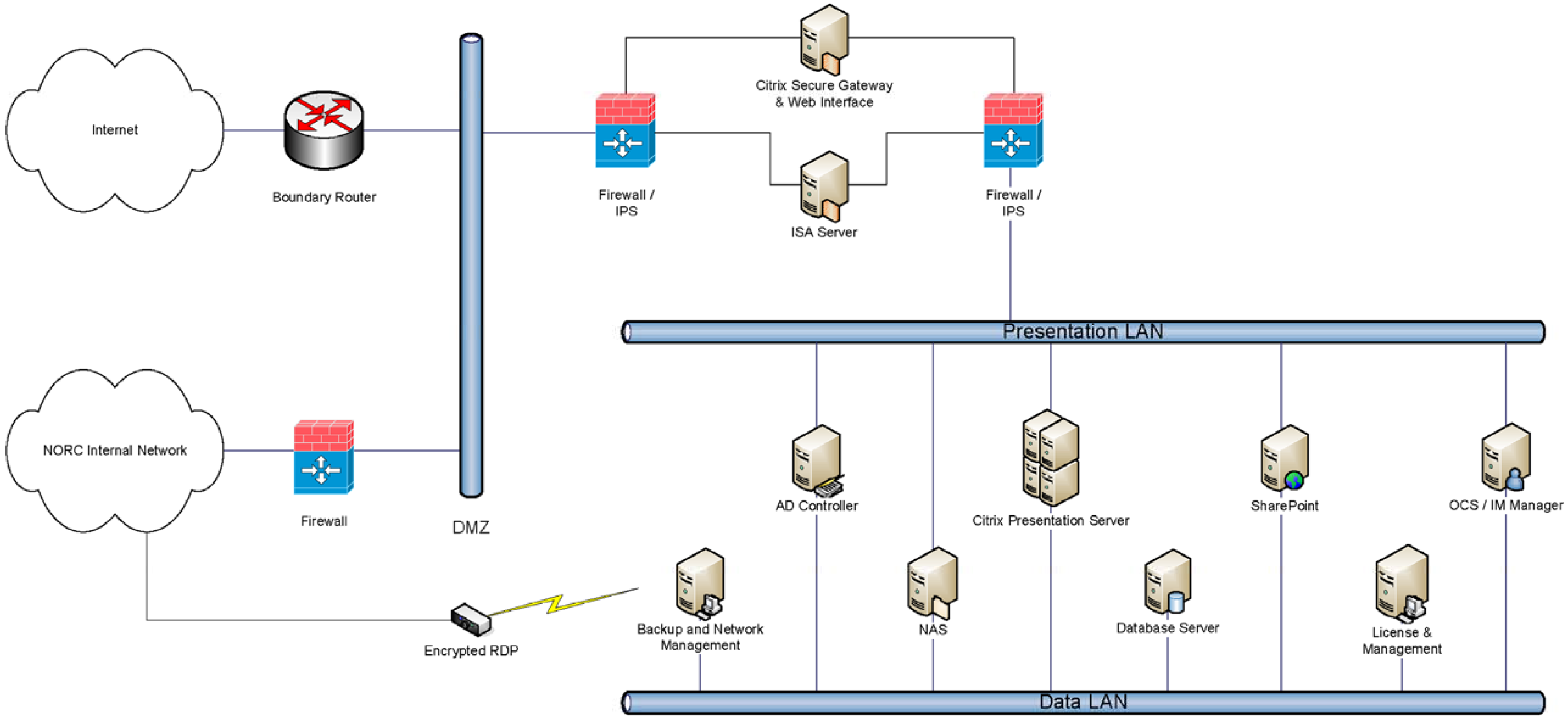
Applications Delivered

10 applications, including:
FundWare nonprofit accounting software
ADP (HR and payroll system)
Medical Manager (medical billing system)
Microsoft productivity applications
Corporate intranet

Networking Environment

Citrix Presentation Server™ running on 25 Dell PowerEdge servers
Citrix Access Gateway™
Citrix Password Manager™
Microsoft® Windows Server® 2003 and Windows® 2000 Server
DSL, dial-up and wireless connections
500 Dell OptiPlex PCs, 50 Wyse thin-client devices and 50 Dell laptops

Annex 3: NORC SRA Diagram


Internet
Boundary Router
DMZ
Firewall / IPS
Citrix Secure Gateway & Web Interface
ISA Server
Firewall / IPS
Presentation LAN
NORC Internal Network
Firewall
AD Controller
Citrix Presentation Server
SharePoint
OCS / IM Manager
Encrypted RDP
Backup and Network Management
NAS
Database Server
License & Management
Data LAN