sesam: Ensuring Privacy for a Interdisciplinary Longitudinal Study*

Boris Glavic , Klaus Dittrich and the sesam Study Team†‡
IFI Department of Informatics, University Zürich, Zürich, Switzerland

Abstract: Most medical, biological and social studies face the problem of storing information about subjects for research purposes without violating the subject's privacy. In most cases it is not possible to remove all information that could be linked to a subject, because some of this information is needed for the research itself. This fact holds especially for longitudinal studies, which collect data about a subject at different times and places. Longitudinal studies need to link different data about a specific subject, collected at different times for research and administration use. In this paper we present the security concept proposed for *sesam*, a longitudinal interdisciplinary study that analyses the social, biological and psychological risk factors for the development of psychological diseases. Our security concept is based on pseudonymisation, encrypted data transfer and an electronic data custodianship. This paper is mainly a case study and some of the security problems emerged in the context of *sesam* may not occur in other studies. Nevertheless we believe that an adopted version of our approach could be used in other application scenarios as well.

1 Introduction

sesam ("Swiss Etiological Study of Adjustment and Mental Health") is a longitudinal and cross-sectional interdisciplinary study aiming at opening the door to a decisive breakthrough in understanding the development of mental health and adjustment to the social, psychological, and biological environments in which we live. Mental health has become a vital issue in Swiss society and probably others as well. The enormous costs of health care and the massive individual impact in terms of suffering and disability make it imperative to understand the pre-disease pathways leading to the development of mental disorders and maladjustment. *sesam* will focus on the complex multi-directional interactions of psycho-

*sesam is funded by the SNF (Swiss National Science Foundation)

†Department of Psychology, University of Basel, Basel, Switzerland

‡Judith Alder (Basel), Johannes Bitzer (Basel), Terry Blumenthal (Wake Forest), Jos-Guy Bodenmann (Fribourg), Dieter Brgin (Basel), Gerhard Dammann (Basel), Nicolas Favez (Fribourg/Lausanne), Alexander Grob (Bern), Paul Grossman (Freiburg), Dirk Hellhammer (Basel/Trier), Ralph Hertwig (Basel), Wolfgang Holzgreve (Basel), Irene Hösli (Basel), Irene Knüsel (Zrich), Roselind Lieb (Basel), Jürgen Margraf (Basel), Gunther Meinlschmidt (Basel), Urs Meyer (Basel), Franz Müller-Spahn (Basel), Klaus Opwis (Basel), Andreas Papassotiropoulos (Zürich), Meinrad Perrez (Fribourg), Christopher Pryce (Zürich), Margarete Rieger (Basel), Hartmut Schächinger (Basel/Trier), Silvia Schneider (Basel), Erich Seifritz (Basel), Johannes Siegrist (Düsseldorf), Werner Stadlmayr (Bern), Hans-Christoph Steinhausen (Zürich), Daniel Surbek (Bern), Michaela Wänke (Basel), Frank H. Wilhelm (Basel), and Dieter Wolke (Zürich)

social and genetic-biological variables across time and between generations, by combining longitudinal, cross-sectional and experimental approaches in a coherent interdisciplinary strategy. Beginning with pregnancy and including the entire risk period for the development of most mental disorders, a large population sample of 3,000 children will be studied together with their parents and grandparents. By adding an experimental manipulation of the nurture component of the assumed etiological factors (i.e., preventive intervention modules in precisely defined high-risk subjects), causal understanding will be enhanced. Among the outstanding conditions that Switzerland offers for *sesam* are an excellent technical and societal infrastructure, low mobility of the population, and the long-term perspective of the Swiss National Science Foundation funding. Using Switzerlands unique infrastructure, *sesam* will yield a national treasure for scientists, public policy developers, and the future generations of Swiss citizens, strengthening the country's position in a strategically important field of key relevance for society and economy.

The *sesam* project is divided into a core study that collects data from the complete cohort of paticipating subjects and several individual projects, which study special research problems with subsamples of the cohort. The *sesam* study will recruit subjects at maternity clinics in different cities in Switzerland. In each city a local *sesam* site will be established for data collection. All data collected at the different local *sesam* sites or clinics will be transferred to the *sesam* central site located in Basel, stored in a central database (*sesamDB*) and reviewed for quality assurance.

The *sesam* study will generate a large amount of data like questionnaires, biological analysis, genetic data, multimedia content and sequence data. Most of this data contains personal information about participating subjects. Longitudinal studies need this type of personal subject information for administrative use. The data collected is of prime importance for the study. Data loss could compromise the progress of the project. Therefore an appropriate backup strategy is needed to reduce the probability of data loss.

The remainder of the paper is organized as follows. In section 2 we discuss existing research in the field of security and privacy. The security requirements of the *sesam* project are presented in section 3. An overview of the *sesam* security concept is given in 4. The following sections cover data collection (secion 5), data storage (section 6) and data export (section 7) in more detail. Finally in section 8 we present considerations and an outline for future work.

2 Related work

A lot of research from various computer science communities focuses on security and privacy. In the database community access control, data encryption and anonymous connections were developed to secure database systems [**?**].

A formal definition of the concepts pseudonym and anonymisation we use in our approach is given in [**?**]. The authors define anonymity as the state of not being identifiable within a set of subjects. In the contrary a pseudonym is an identifier for a subject. The strength of the anonymity a pseudonym can provide depends on the knowledge of certain parties

about links between subjects and pseudonyms.

Another important research field is secure data transfer. One goal of secure data transfer is to protect the content of transferred messages from unauthorised access. This can be achieved by using Cryptographic methods [**?**] to encrypt the message content. Other goals are providing anonymity for communicating subjects and authentication of the communicating subjects (for example by using digital certificates [**?**]).

3 Security requirements for *sesam*

The *sesam* study collects a large amount of "critical" data like genetic data and video observations. From an ethical point of view this data should be completely anonymised to protect the subject's privacy. But a complete anonymisation may result in reduced data quality and loss of information. For example, it is possible to identify a subject in a video observation. However complete anonymisation of video observations cannot be achieved without a certain amount of information loss. Data quality is of high importance for *sesam* and thus a complete anonymisation is not applicable. Protecting privacy is further complicated by the fact that it is essential for a longitudinal study to be able to link a subject and the data collected about this subject. E.g. an individual project may focus exclusively on depressive persons. To find the set of depressive subjects it is necessary to link subject information with scientific data (find the names of subjects with the diagnosis depressive). Because of the need to link subjects and scientific data even anonymisation without quality reduction is not applicable for *sesam*. Considering these constraints, protecting the subject's privacy is limited to pseudonymisation of scientific data and protecting the data and mapping between subjects and data from unauthorised access.

Independent of the problem of privacy protection, the scientific data collected by *sesam* represents a substantial value for the project and should be protected from unauthorised access and manipulation. For example, other scientists who do not have the resources or time to conduct a long-term study will be interested in the collected data. As stated before a backup strategy is needed to prevent data loss. It is important to include the backup strategy in the considerations for a security concept to avoid further security vulnerabilities induced by multiple data versions located at different places.

sesam will collect data and perform analysis at different sites. An security concept should enable secure data transfer between the data collection sites and the central site and protect *sesamDB* from unauthorised access.

Having these facts in mind the main requirements for the development of a security concept for *sesam* are:

- secure data collection
- secure data transfer between data collection sites and the central site
- protect mapping between subjects and scientific data
- protect backups from unauthorised access

4 A proposal for a security and privacy protection concept in *sesam*

The security strategy we develop for *sesam* covers the three parts data collection and transfer, data storage and data export. To guarantee a high security standard the central *sesam* database *sesamDB* will be connected neither directly nor indirectly to the Internet. An Internet connection would reduce the effort for communication between the different *sesam* sites, but would result in a less secure system and higher costs for securing *sesamDB*. Even more important a database without an Internet connection will be more likely accepted by the public. To protect the data transfer between the central site and the data collection sites we use an asymmetrical encryption technique. Note that the encryption technique is only loosely integrated in our approach and can be easily replaced. Thus we can change our encryption technique to the future if it is not considered secure any more.

Before we present the topics of data collection, data storage and data access in detail, we first introduce the roles used in our concept.

- *data collector* (collects data about subjects)
- *quality assurance* (reviews data for quality assurance)
- *sesam scientist* (analyses the scientific data)
- *sample collector* (collects samples from subjects)
- *sample manager* (manages samples at the central site)
- *sesamDB admin* (administrates *sesamDB*)
- *mapDB admin* (administrates *mapDB*)
- *data entry* (stores data in *sesamDB*)
- *sesam messenger* (transports data between the different sites)
- *mail manager* (manages incoming and outgoing mail)

Note that our complete data management concept uses more detailed role descriptions, but in this paper we focus on the roles essential for the security concept.

5 Data collection and transfer

Data collected at different *sesam* sites will be transferred to the central site and stored in *sesamDB*. To identify which data was collected about which subject a measurement identifier will be attached to each collected data item. The measurement identifier or MI consists of the subject's name, the date of measurement, the name of the *data collector* and additional information about the measurement. The role *data collector* enters the information needed via an user interface. An asymmetrical encryption technique is used to encrypt the MI. We refer to the result of the encryption process as transfer code or TC. When the collected data is transferred to the central site, the TC is transferred as well. At the central site the TC is decrypted and replaced with a subject pseudonym (see section 6). While each *data collector* holds the public key used for encryption, the private key is stored exclusive in *sesamDB* and is only accessible via the client application used for data entry.

The collection and transfer processes for the various data types collected by *sesam* differ in detail. A detailed description of all processes is beyond the scope of this paper. Instead we explain these processes by means of the processes for questionnaires.

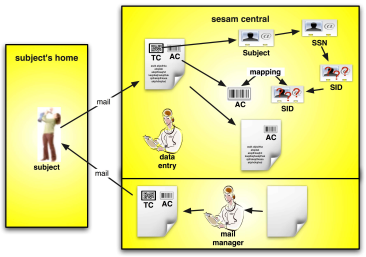

Abbildung 1: data collection for questionnaires

Questionnaires are filled in by the subject's at their homes. The prefabricted questionnaires are prepared and send to the subjects by the role *mail manager*. The *mail manager* uses an application to prepare three labels for the process. The first label contains the mail address of the subject. The second label is a two-dimensional barcode, representing the transfer code TC. The last label is a barcode label representing the archive code AC (a random number including a checksum). The propose of the archive code is discussed in the next section. The *mail manager* labels a prefabricated questionnaire with the TC and AC labels. The questionnaire together with a return envelope is put in an envelope and the envelope is labeled with the address label. This envelope is send to the subject via mail.

6 Data storage

As stated in section 3 it is not possible to remove all identifying information from the collected data and *sesam* needs a facility for connecting subjects and scientific data. Thus it is not feasible to anonymise the collected data and besides the general protection of *sesamDB* we placed a special focus on the protection of the mapping between subjects and scientific data and pseudonymisation. We use pseudonyms called subject identifiers or SIDs to identify the subjects about which scientific data was collected. All personal information like name or address is stored associated with another pseudonym called subject study number or SSN.

The mapping between SID and SSN is not stored in *sesamDB*. We establish an electronic data custodian to control the access to the mapping between the SSNs and the SIDs. The mapping information is stored in a second database located at an external location and administrated by an external organisation. This external database, called *mapDB*, is connected to *sesamDB* via a private connection. *sesam*-employees have no direct access to *mapDB* and can only access the mapping information using a *sesam* client application. These client applications authenticate users and restrict the access to the mapping information to specific use cases. For example, the client application for data entry is used to

decrypt a TC, receive the SSN from *sesamDB*, find the associated SID in *mapDB* and store the scientific data together with the SID in *sesamDB*.

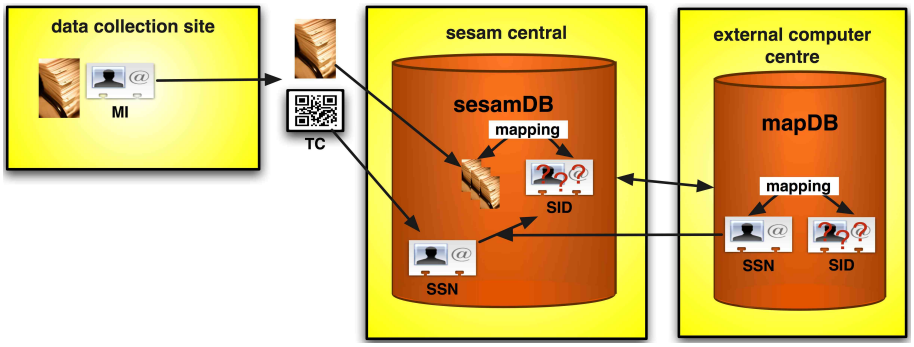

Abbildung 2: overview data collection, transfer and storage

When data is transferred to the central site the measurement information MI is transferred encrypted as the transfer code TC. At the central site the role *data entry* uses a *sesam* client application to enter the data into *sesamDB* and scan the barcode representing the TC. This application decrypts the TC and replaces the subject information in MI with the SSN. The personal information contained in MI is used to find the correct SSN. In a second step the mapping information is received from *mapDB* and the SSN is replaced with the SID. The collected data associated with SID is then stored in *sesamDB*.

Some of the data collected by *sesam*, like biological samples and informed consents, will be archived in physical form. For reference these items are labelled with barcodes each representing a random number with checksum. The mapping between these archive codes or ACs and the SIDs identifying the subjects will be stored in *sesamDB*.

As stated before it is important to include the backup strategy in our security concept. *sesamDB* will be backed up to a second server on a daily basis. This second server is placed in the same location with the *sesamDB* server. In addition tape backups will be performed every week and the tapes will be stored in a secured location outside the central site.

Like in the last section we present a data storage process as an example. We pick up the case of questionnaires, because it includes archive codes. When the subject returns the questionnaire to the central site via mail, the role *data entry* scans the TC and AC labels with a barcode scanner. The questionnaire itself is scanned and interpreted using a text recognition tool. In the next step of the process the TC is replaced by the coresponding SID using the process described before. *data entry* stores the questionnaire data associated with the SID and the SID associated with the AC in *sesamDB*. Before the questionnaire is stored in the archive the TC label is removed.

7 Data access and data export

Access to data stored in the *sesamDB* is restricted to computers located at the *sesam* central site. These computers are connected to *sesamDB* via a local network connection. We require that no computer that is connected to *sesamDB* is connected to the Internet. The access to *sesamDB* is restricted to specialised client applications, which have only access to the data needed for their field of activity. For example the client application used for data export and scientific analysis has access rights for all scientific data, but no access rights for personal subject information and *mapDB*.

The client application used for data export logs all data export queries and stores the log information in *sesamDB*. The log information allows us to monitor the data exports and analyse the exports executed by a specific person. Data is made availible to third parties in aggregated form, without SIDs and with assent of the study direction.

8 Future work and conclusions

We presented a security concept for the *sesam* study. While primarily a case study, some of the considerations and concepts presented in this paper are not specific to *sesam* and could be used in other application scenarios as well.

The *sesam* study is still in the pre-datacollection phase. Our proposed security concept is not implemented yet and will undergo extensive auditing by swiss data security authorities and ethical comissions before implementation.