PDI-SSH Call for digital infrastructure 2021

Secure Analysis Environment (SANE)

Main applicant: Tom Emery (EUR-ESSB/ODISSEI)
Team members: Lucas van der Meer (EUR-ESSB/ODISSEI), Roeland Ordelman (NISV/Clariah), Martijn Kleppe (KB/Clariah), Freek Dijkstra, Ivar Janmaat, Annette Langedijk (SURF)
Partners: EUR-ESSB/ODISSEI, NISV/Clariah, KB/Clariah, SURF
Discipline: Social sciences

1 Summary

Privacy, copyright, and competition barriers limit the sharing of sensitive data for scientific purposes. We propose the **Secure Analysis Environment (SANE)**: a virtual container in which the researcher can analyse sensitive data, and yet leaves the data provider in complete control. By following the Five Safes principles, SANE will enable researchers to conduct research on data that up until now are hardly available to them.

SANE comes in two variants. Tinker SANE allows the researcher to see, manipulate and play with the data. In Blind SANE, the researcher submits an algorithm without being able to see the data and the data provider approves the algorithm and output.

SANE uses concepts from the CBS Remote Access Environment, ODISSEI Secure Supercomputer and SURF Data Exchange, to build a generic off-the-shelf solution to be used by any sensitive data provider and researcher. SANE can be used by researchers in any discipline, as illustrated by the involvement of consortia in both the social sciences (ODISSEI) as well as humanities (Clariah).

Potential sensitive data providers include the Dutch Chamber of Commerce (KvK), Funda, National Library of the Netherlands (KB) and Netherlands Institute for Sound and Vision (NISV).

2 Purpose of the initiative

2.1 Sharing sensitive data

While more and more interesting datasets are available through non-academic data providers, there currently is no infrastructure available allowing researchers to analyse sensitive data in a way the data

providers remain in control. As a result, most potential data-providers (such as governments, heritage institutes or commercial parties) are reluctant to share their datasets and they thus remain unused, even though academic breakthroughs could be made once these datasets are available.

2.2 Why data sharing is difficult

Even when there is good will to share data, it is not always easy to do so [Van Atteveldt et al. (2020)].

Privacy is an important barrier to sharing personal data. According to Article 89 of the GDPR, the processing of personal data for scientific research is permitted only if there are sufficient technical and organisational safeguards for the rights and freedoms of the data subject.

Also **copyright** issues impose an important barrier to sharing data. Books, websites, television programs, social media outlets: most are likely copyright protected and thus cannot freely be shared. Strict agreements and technical measures could however allow for data access.

Barriers to **competition** often play a role. Organisations often want to prevent their data from ending up at competitors, or becoming freely available which would damage their business model which is based on the sale of data.

2.3 Proposed solution

We propose the **Secure Analysis Environment (SANE)**: a virtual computer containing pre-approved analysis software (R and Jupyter notebooks, including packages) and access to the sensitive data (Figure 1). It follows the Five safes principles, leaves the data provider in complete control and still allows the researcher to study the data in a convenient way.

The data provider grants access to particular researchers (Safe People). The SANE has a critical safeguard: it is only possible to output any sensitive data to the researchers' own computer, after verification by the data provider (Safe Output). This means that for the researcher, besides its connection to the environment, SANE is locked off from the internet. The data provider can even prevent the researcher from seeing the data (Safe Settings). All actions by the researchers are monitored (Safe Projects). Uploading data can also be prevented, since combining more data may result in de-anonymisation (Safe Data). The full researcher's journey is listed in Appendix A.

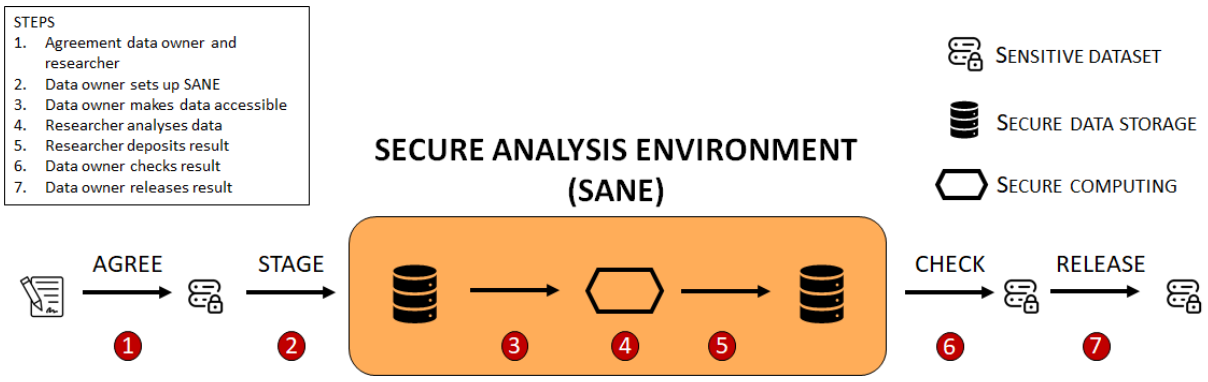

Figure 1 – The Secure Analysis Environment (SANE)

We will build two variants of SANE: Tinker SANE and Blind SANE. Which variant to choose depends on the required level of interaction with the data the researcher needs, and the data provider's confidentiality requirements (see Appendix B for all available customisations).

In Tinker SANE, the researcher gets to see, experiment with, and manipulate the data. The "tinker" variant is most appropriate when the researcher combines several different data sources and where specific characteristics of the combined data determine consequent analytical steps.

On the other hand, in Blind SANE the researcher submits an algorithm, and the data provider prevents the researcher from seeing the data. This is typical in situations with copyright barriers. The "blind" variant can be used for large datasets of which the data structure is known to the researcher, such as historical newspapers at KB or historical TV broadcasts at NISV.

3 Current status

SANE builds on earlier efforts by the project partners (also see section 4). These earlier efforts are either limited in scope (i.e. tied to a specific dataset) or limited in technology readiness.

SURF earlier experimented with algorithm-to-data solutions such as the SURF Data Exchange (DX) prototype. SURF as a trusted third party retrieves data from a data provider and executes the researchers' analysis in a secure (Docker) container, ensuring that no data are leaked to the outside world. DX is designed for situations where the researcher cannot see the data during analysis. The DX is currently at technology readiness level (TRL) 6, and shown with a couple of potential use cases. It is believed that this is beneficial to more use cases, but before it can readily be used, it needs to reach TRL8. The DX prototype will be extended towards the fully functioning and scalable SANE.

Existing tools will be used as much as possible in further development. Appendix C lists the technologies that can already be reused, and the technologies that need to be developed. For example, for the verification of user affiliations, existing SURFconext or SURF Research Access Management (SRAM) infrastructure can be reused, and for incidental storage of confidential data (where an integration to another data repository is too cumbersome), SURF Research Drive can be reused. The SANE approach will heavily rely on the SURF Research Cloud. We will not only consider tools developed by the project partners, but also open source tools developed by other parties, like BBMRI Podium (developed by Lygature) and Personal Health Train Vantage6 (developed by Maastro)[1].

4 Relationship with existing infrastructures

We differentiate SANE from existing infrastructures along two dimensions (Figure 2): data confidentiality and the level of data interaction the researcher needs.

[1] Even if these tools are not used, we do strive for interoperability with those tools, for example by using a compatible programmatic interface. Similarly, we plan to publish as many developed tools as possible with an open source licence, so that others may later build upon it.

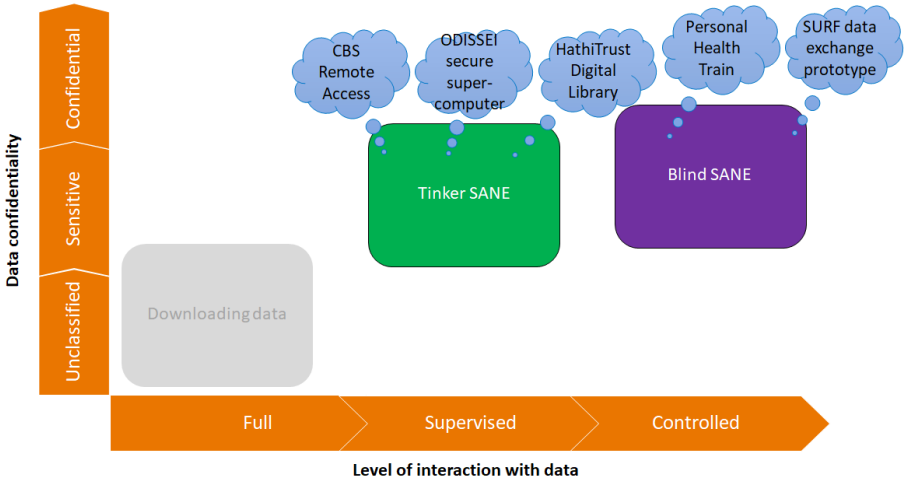

Figure 2 – SANE compared to five existing infrastructures

A similar concept as SANE is being applied actively by CBS. CBS has very strict requirements for allowing access to sensitive data, for instance for pseudonymised microdata on individuals, companies or addresses. It provides a secure analysis environment called CBS Remote Access (RA) environment for about 400 research projects a year. The RA environment was later cloned by ODISSEI, CBS, and SURF for supercomputer purposes, into the ODISSEI Secure Supercomputer (OSSC). We want to build an environment that takes the best of the RA and OSSC environments, but can be used for non-CBS data and without CBS involvement. Additional customisations relevant to the SSH will be added (see Appendix B).

Some data providers, such as the HathiTrust Digital Library, allow 'data capsules' for researchers to analyse the data. However, these solutions are typically only available for the specific data provider or with limited analysis tools, while we will build SANE so that it can be applied generically.

Paradigms where the researcher cannot see the data are also applied in distributed analytics solutions, such as the Personal Health Train. Distributed analytics solutions are designed for the algorithm to visit multiple data stations on different locations, compared to just one data station for SANE. Setting up a distributed analytics solution typically takes months, whereas SANE can be fully set up in just a day.

Furthermore, the participation of Clariah and ODISSEI in the development of SANE ensures that a solution is made that caters to real use cases, and is general enough to cater for possibly many more.

5 Added value

The added value of SANE is that (1) it enables access to currently unavailable datasets by researchers, (2) data providers can allow access to more detailed versions of existing data, (3) any data provider can use it.

Enabling access to new data

At the moment, the majority of confidential datasets are either not available for scientific research, or are available for download for specific researchers after vetting the identity and purpose of the researcher. This project aims to make more confidential datasets available for researchers by giving potential data providers the tools to reduce the risk of confidentiality being compromised.

Extending existing datasets

Data providers pseudonymise the data or remove certain sensitive elements, such as geospatial references, before sharing the data. However, these limit the functionality of the data. SANE provides extra safety for data providers by preventing data leakage and preventing disclosure, meaning that they may not need these limiting measures when using SANE. This makes richer data available to the researcher.

Generic solution

SANE is a generic solution while the CBS Remote Access, ODISSEI Secure Supercomputer and HathiTrust are custom-built for one data provider. This makes SANE a valuable, future-proof solution that is independent of the current institutes involved.

Because it is a generic solution, SANE creates standard specifications for software developers at ODISSEI and Clariah. The developers can use these specifications to produce analytical tools working on data from a wide range of providers, rather than having to adjust them for every individual data provider's environment.

6 Intended users

We verified the proposed solution at a number of potential data providers and researchers across various disciplines within the social sciences, humanities and even health sciences. The interest has been overwhelming. The Dutch Chamber of Commerce (KvK) considers sharing a full export of the ‘handelsregister’, as long as the recipients (Firmbackbone, a PDI-SSH project of the first round) can guarantee that it will only be used for research purposes (privacy and competition barriers). The Deans of Economics/Business Administration validated that hundreds of researchers would use these KvK data.

Data centers such as the National Library of the Netherlands (KB) and Netherlands Institute for Sound and Vision (NISV) have huge data archives that for copyright and privacy barriers cannot be made accessible for large scale analysis. The only way to use the data for research currently is via a broad search and browse facility[2]. However, Humanities researchers want to do more fine-grained and/or large-scale analysis of data. The need for specific analysis tools is rapidly growing as reflected in the increasing requests for data NISV and KB receive in the last few years that can only suboptimally be catered. SANE would bring data-driven Digital Humanities research to the next level.

[2] e.g. mediasuite.clariah.nl or kb.nl/en/resources-research-guides/data-services-apis

Other parties that showed significant interest include the YOUth cohort, Funda, Consultatiebureaus, WODC, NVM and the National Archives. A full list is included in Appendix D.

Both data providers and researchers will be actively involved in the project via design sessions and research pilots.

9 KPIs

We aim for a production-ready system of both Tinker and Blind SANE by the end of the project. If needed, additional features can be developed after 2024, using financing from the project partners.

The data provider will have to create a SANE through a simple process at SURF which includes analysis software and packages requested. No assistance or clearance from technical support is required. The data provider can add further restrictions, such as disabling data uploads, restricting the type of analysis that can be performed, or preventing the researcher from being able to see the data.

After the project, the SANE production system will be available through SURF **structurally**. Partners will have a sustainability model in place by the end of the project. Researchers can request the regular SURF-NWO research support grants to use SANE.

The infrastructure will be considered successful if (1) it motivates data providers to share their data which is currently unavailable for research (five data providers cases during the project, five in the first year after the project), (2) motivates data providers to share richer data (one data provider after the project), and (3) provides usability for researchers that is sufficient to answer their research question (five researchers during the project, five in the first year after the project, all scoring the usability with at least a 7/10[5]).

[5] The exact measurement is yet to be determined.

11 Appendix A: Researcher's journey

The data provider and the researcher follow a predefined process:

1. The data provider agrees that the researcher may use a (pseudonymised) dataset for a specific research question.
2. The data provider creates a SANE through a simple process at SURF which includes analysis software and packages requested (R or Jupyter Notebooks). No assistance or clearance from technical support is required. The data provider can add further restrictions, such as disabling data uploads, restricting the type of analysis that can be performed, or even preventing the researcher from being able to see the data.
3. The data provider makes the dataset available (either a local copy or via remote access to the dataset at the source repository of the data provider), and possibly a dataset of the researcher. The data provider grants access to the researcher.
4. The researcher carries out research on the SANE by remote login to the environment. All actions on the dataset can be monitored by the data provider via logging.
5. The researcher indicates which results should be exported.
6. The data provider checks if the results do not contain any sensitive data from the dataset.
7. After approval, the data provider releases the results.
8. After finishing the research, the data provider can decide to close and store SANE, e.g., for re-opening by the same researchers or for verification purposes. Withdrawing an earlier consent is possible at any stage.

12 Appendix B: Available customisations

In addition, the data provider may choose to add the following functionalities to its SANE:

- The identity of the researcher is verified, e.g. using *SURFconext* or *SURF Research Access Management* (SRAM).
- The purpose of the research is vetted by the data provider, similar to *BBMRI Podium*.
- The data provider and the researcher sign a confidentiality and project agreement.
- The data is temporarily deposited at a data store that allows it to be mounted in a SANE, but not downloaded by researchers. E.g. *SURF Research Drive*.
- The data is accessed directly from the data repository of the data provider, instead of being temporarily deposited. This requires a secure connection between the environment and data source.
- For transport of the data to SANE, a VPN or lightpath is used.
- The data provider can monitor all operations on the confidential data.
- Each SANE is destroyed after use.

Blind SANE and Tinker SANE are by no means the only possible variants. In particular when a researcher wants to run an analysis on multiple confidential datasets, more advanced environments are needed, like a trusted third party that provides a linking table between the datasets, or multiple linked

environments that perform secure multiparty computation (MPC). In this project, only development of the Blink SANE and Tinker SANE are envisioned.

13 Appendix C: Existing and new components

SURF will build SANE partly based on existing components. The table gives an overview of existing components versus components that require development:

Existing components:	**New components:**
• VPN / lightpath • Verification environment (e.g. SURFconext, SRAM) • Data sharing (e.g. Research Drive) • Research Cloud • iRODS • Security Audit • Log file monitoring • Destruction of SANE after usage	• Data staging, access rights • Secure access to storage for data provider, researcher • Pre-processing services • Secure data analysis environment • Availability and access to analysis software • Method for result check and secure release of result • Logging of activity

14 Appendix D: Use cases

Tinker SANE use cases

1. Data from Funda, like the houses database and the search queries. Currently being made available on an ad-hoc basis but for it to be scalable, the access solution would also need to be scalable.
2. G4 municipalities have data on work reintegration that is currently not available for research. They would still pseudonymise the data, as the researcher will not be interested in individuals.
3. Pseudonymised WODC data on court judgements. Data are currently not available at scale.
4. High School admission data from Amsterdam to analyse the effects of sorting and selection on outcomes.
5. YOUth cohort, containing data on individual development of children. Currently shared with researchers on an ad-hoc basis only because of privacy restrictions.
6. Commercial Data from NVM that are not generally available to researchers.
7. Public Housing Lottery data from Amsterdam to examine natural experiments on outcomes.
8. KvK may be willing to share a full export of the handelsregister, as long as the recipients (a project called Firmbackbone) can guarantee that it will only be used for research purposes. Firmbackbone would therefore want to set up such an environment
9. A university teacher has a research data set that students may analyse, but not download.

Blind SANE use cases

1. Audiovisual collections (e.g. Radio/Television data) from Netherlands Institute for Sound and Vision, specifically the metadata for these collections, optionally also extracted features of the sources and viewing/play-out.
2. Access to all full text and metadata of digitised textual data from KB, national Library of the Netherlands such as historical newspapers, magazines and books is currently available through keyword search at Delpher and DBNL, including those still protected by copyright.
3. Archival data from the National Archive
4. Born digital data from the KB such as websites.
5. In general, data from Clariah centers that have copyright or privacy concerns such as Oral History collections from DANS, speech databases from Meertens or Max Planck Institute.
6. Large text corpora from publishers for linguistic research.
7. Spoken data collections provided by or co-created by private organisations for speech research that have privacy concerns or commercial value.