STANDARD OPERATING PROCEDURE
Do not Photocopy

Document Information Classification: Unrestricted

Title:	**Importing Content into the TRE**
Effective Date:	**14 Feb 2019**
Reference Number:	**SOP-05-03**
Version Number:	**1.10**
Owner:	**Information Security Manager,**
Review Date:	**29 Oct 2020**

Table of Contents

1. Purpose

This document describes the procedure for the transfer of datasets into, and out of the TRE. It covers the process of importing data into the TRE when a project is initiated.

The procedure for securely deleting datasets from the TRE is described in a separate document: SOP-05-15 Deletion of TRE Datasets. Data exports and corresponding output checking is covered by SOP-07-02 TRE Data Export and Output Checking.

The TRE, and all systems contained within, is involved in handling information that must be managed in a way that ensures its confidentiality, availability and integrity. Implementing security controls throughout the lifecycle of a system can help the TRE achieve its ISMS objectives, regulatory requirements and the needs of its users.

2. Scope

All datasets transferred to, accessed and stored within the TRE.

All derived data created by members of a TRE project that need to be exported out the TRE.

3. Responsibilities

The TRE Operations Manager is responsible for:

- Coordinating the import of data for each project
- Coordinating the export of derived data for each project
- Deciding whether the dataset will be imported directly onto a virtual workstation or a file server belonging to the TRE project
- Ensuring the dataset integrity throughout whilst held within the TRE
- Verifying that data can be restored
-

The TRE System Administrator is responsible for:

- Transferring the dataset from the source (data provider's server) to the TRE storage allocated to the project that has requested that data
- Performing checks of the file digital signatures to guarantee the integrity of the dataset has not been compromised during transfer into TRE, during general use, and during any backups that take place

The project's Principal Investigator is responsible for:

- Providing the necessary details, as described in sections 4.2 to 4.5 of this SOP

4. Procedure

4.1. TRE Project Environment

Each TRE project is allocated a dedicated secure and isolated network that includes a firewall. Within this network at least one data analytics workstation is installed, which is mounted to a networked storage volume.

4.2. Data Import

Once a project has completed all information governance contracts and agreements necessary to obtain the required data, a member of the TRE Operations team will manually transfer the relevant datasets directly onto the file-server allocated to the corresponding project.

The Principal Investigator of the project is responsible for providing the TRE Operations Team with a list of each dataset/file/folder that requires downloading into the TRE. This ensures the TRE Operations team able to register each file within the dataset asset database. Additionally, and where possible, the Data Controller should provide a checksum (digital signature) for each file, as it will then enable the TRE Operations team to verify the data integrity post-import and during its lifecycle (optional extra) within the TRE (see section 4.4).

The Secure File Transfer Protocol (SFTP) is the preferred method for transferring data in and out of the TRE. The TRE Service supports two user-cases for SFTP transfers:

1) The preferred data import method is for the Data Providers to make the datasets available on an SFTP server. A member of the TRE Operations team will then connect to that SFTP server (from the TRE) and download the datasets directly into the TRE. This ensures that the datasets never reside on any other form of storage other than the source file-system (e.g. the Data Provider's SFTP server) and the TRE file-system.

2) For Data Providers who are not able to host their own SFTP server, the TRE contains two 'DropZone' machines; one of the 'main' University network [the 'internet'] and one on the N3/HSCN NHS network. A DropZone is an SFTP server that features elevated security controls beyond the base Unix SFTP service. It doubles up as a quarantine area, allowing data to be imported within the secure boundaries of the TRE without being made available to the destination research project.

For both the above methods, the data provider must supply the public-facing (WAN) IP address of the machine that is hosting the data to be downloaded into the TRE (this is the IP address that the TRE machine will connect to via SFTP).

It is also be necessary to provide the TRE Operations team with an SSH public key associated with an SSH key-pair installed on the SFTP server.

4.3. Secure Data Transfer Channel

Within the context of SFTP use-case method 1) the default configuration of the firewall allocated to each TRE project does not allow any inbound or outbound traffic. Therefore, prior to the data transfer commencing, a temporary firewall rule must be implemented allowing the project's file-server to establish a connection to the data provider's file hosting server, and for the data to be downloaded. This firewall rule must be removed once the datasets have been successfully downloaded. This rule should be a one-to-one type, which requires the following information:

- source (data provider's file hosting server) IP address
- destination (TRE project file-server) IP address
- data transfer protocol and port (e.g. 443 for https)

4.4. Dataset Import Verification

Within the overall process of TRE Data Management, there are three forms of data checking and the first two are covered in this document:

1. Verifying the integrity of datasets that have been imported into the TRE (to make sure no data corruption occurred during transit)
2. Verifying that the imported datasets meet the TRE user's requirements and expectations (to make sure the data contains the exact information content requested, is of the required quality, and is presented in the correct format
3. Verifying that an extract of data generated by the TRE user can be exported from the TRE (to make sure there is a negligible risk of disclosing any personal information to anyone viewing the exported data). Refer to document SOP-07-02 TRE Data Export and Output Checking for further details.

Step 1: If possible, the data provider should supply the digital signatures (checksum) of all files that are to be downloaded into the TRE. These signatures will allow the TRE Operations team to ensure the data integrity has not become compromised during the download. The TRE Operations team must generate a checksum for each file that has been imported into the TRE before they are first accessed by any user. Integrity checks record the 'digital signature' of a file at a bit-level and cannot reveal any human-readable information; therefore there is negligible risk of disclosing personal information when generating a checksum. The TRE Operations team will place a record of these digital signatures within a dedicated directory, residing within the TRE user's home area where the raw datasets were downloaded. A copy of these digital signatures is backed up onto TRE data storage.

Step 2: The checksum 'hash' value must be recorded in two places: The corresponding field within the 'Import' worksheet of the TRE Infrastructure Record, and the 'Properties Measured' field within each Dataset record within the Q-Pulse Asset module. Note, unless otherwise specified by the data providers, it is only necessary to perform an integrity check when the data arrives in the TRE project's storage volume, and not also within a DropZone machine that may have been used as an intermediary transfer location.

Step 3: Following Step 1 the TRE Operations Team will contact the project's Principal Investigator to arrange for further verification of the downloaded datasets. This may be necessary in cases where subject expertise is required to determine the content and quality of the acquired data. It is also possible that software tools will be required to perform this verification. For example, it might only be possible to determine the number of records in a dataset by rendering it in an analytical tool such as STATA. The Principle Investigator is responsible for making the TRE Operations team aware of the requirement to use software tools to perform verification at least 10 working days in advance of performing such checks.

The verification of the data must assess the file according to the following criteria. Referring to other documents like the Project Application Form or data specifications, it is necessary to assess whether they match the data that was requested:

- File format
- Number of files
- Checksum of file size

For files containing variables:

- Number of variables within each file
- Identity and format of each variable

For other files e.g. images:

- Sample some files and use appropriate TRE software to ensure they can be opened and seem to be what they're supposed to be.

If an issue is identified with the imported datasets, the TRE user will liaise with the data provider regarding any issues, and request corrected files if necessary.

Once data import verification is completed satisfactorily, a member of the TRE Operations team should ensure data are organised appropriately for named users on the TRE project to access. This could involve masking or transforming certain variables, creation of new TRE user accounts, and ensuring required software is available to the TRE users to work with the files. Once completed, an acknowledgment should be sent to the data provider and the lead of the TRE project.

4.5. Restrictions on Access to TRE Project Data

Regardless of the data verification requirements, members of the TRE project and named within the corresponding Data Sharing Agreement as authorized users of that data are the only people permitted to view the content of files. This rule becomes effective immediately following import of data into the TRE, and stipulates that no file must be opened, unzipped/decompressed, read, executed or parsed in any way by anyone who isn't authorised to do so.

The exception to this rule is that a member of the TRE Operations team is permitted to conduct an integrity check of the downloaded file to ensure it was downloaded successfully.

Another exception to this rule is when a member of the TRE Operations team is asked to perform data output checking. This is usually conducted on derived data, e.g. tabulated output from a data analysis software application, so it still remains unlikely that someone who isn't a member of the TRE project will be required to view the raw datasets originally imported into the TRE. SOP-07-02 TRE Data Export and Output Checking must be consulted for full details of the export and validation procedure.

5. Cross-referenced ISMS Documents

Number	Type	Title
SOP-05-15	ISMS\SOP\Asset and Supplier Management - SOP	Secure Deletion of TRE Data
SOP-07-02	ISMS\SOP\Information Governance - SOP	TRE Data Export and Output Checking
SOP-03-02	ISMS\SOP\TRE Operations - SOP	TRE User Manual and Agreement

SOP-05-01	ISMS\SOP\Asset and Supplier Management - SOP	Bringing Assets into the TRE
SOP-07-02	ISMS\SOP\Information Governance - SOP	TRE Data Export and Output Checking

6. Appendices

None