

The Alan Turing Institute

Data Safe Havens

Shared model for classifying data sets and work packages into common sensitivity tiers, with recommended security measures for each tier

A reference implementation on Azure to provide a secure cloud-based platform for remote analysis of sensitive datasets Independent, isolated secure research environments deployed for each project

Shared identity, authorization and access management across project environments

Maximizing researcher productivity while maintaining security appropriate to the tier

Classification – How should you handle your data?

Tier	Classification	Access
Tier 4	Very sensitive personal, commercial or government data	Access only from known dedicated secure rooms Stricter package whitelist
Tier 3	Personal data with weak or no pseudonymisation, or more sensitive commercial or government data	Access only from known physical spaces Access only via managed devices Whitelisted packages
Tier 2	Most commercially sensitive data Strongly pseudonymised personal data	Access only from known networks Remote desktop only No outbound internet No copy/paste Full package mirrors
Tier 1	Data with very low consequences for disclosure	No safe haven required Outbound internet ok Access from internet ok Still require good standard security practices)
Tier 0	Open data	No safe haven required Outbound internet ok Access from internet ok Still require good standard security practices

Using Data Safe Havens

Our model proposes instantiating separate secure environments for each new research project. This is made possible by software-defined infrastructure and will be commonly supported by the public cloud

Step by step

- New Project
- Classify the data and tools
- Deploy a Secure Research Environment
- Ingress the data
- Carry out the research remotely
- Classify the outgoing data as work packages
- Egress the data
- Shut down the Secure Research Environment

Each work package has a dedicated secure research environment for each tier of work package within them, with access controls defined by the work package Classification Tier

Work Package (Egress – Project report)

Once the project is complete, data must be classified before it can be egressed from the secure research environment, to ensure that it is handled appropriately

Classify the work packages

Out of secure storage

Work Package (Egress – Project data)

If egress data is classified as a higher tier, it must be stored appropriately or moved to another safe haven

- Investigator (PI)** - The research project lead, this individual is responsible for ensuring that project staff comply with the Environment's security policies.
- Dataset Provider (DPR)** - A representative of the organisation who provided the dataset under analysis. The Dataset Provider will designate a single representative contact to liaise with the Investigator, authorised to certify sharing of datasets with the researchers.
- Referee** - A Referee volunteers to review code or derived data (data which is computed from the original dataset), providing evidence to the Investigator and Dataset Provider Representative that the researchers are complying with data handling practices.
- Work package** - Datasets should be organized into work packages, which outline the intentions for a phase of work, and contain the datasets which will be required.