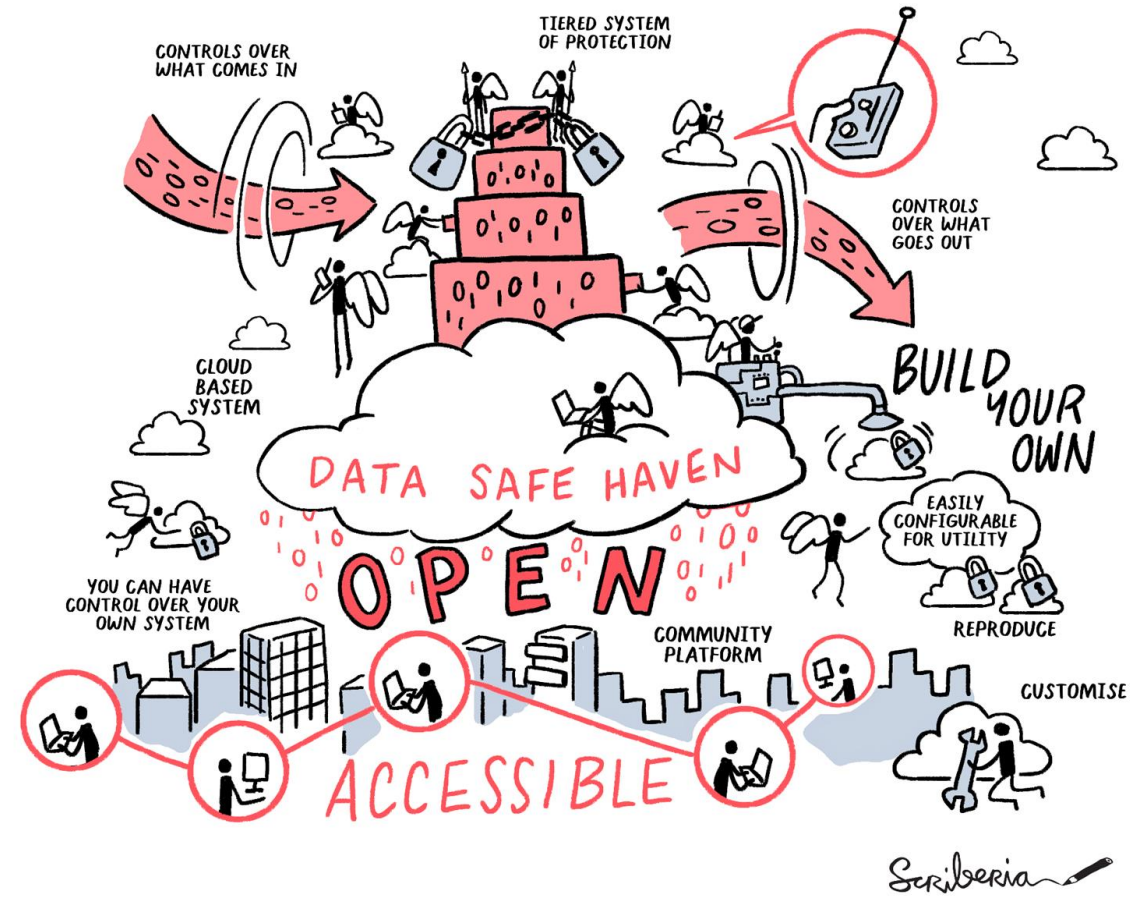


# The Alan Turing Institute

---

## Python and R package management in TREs

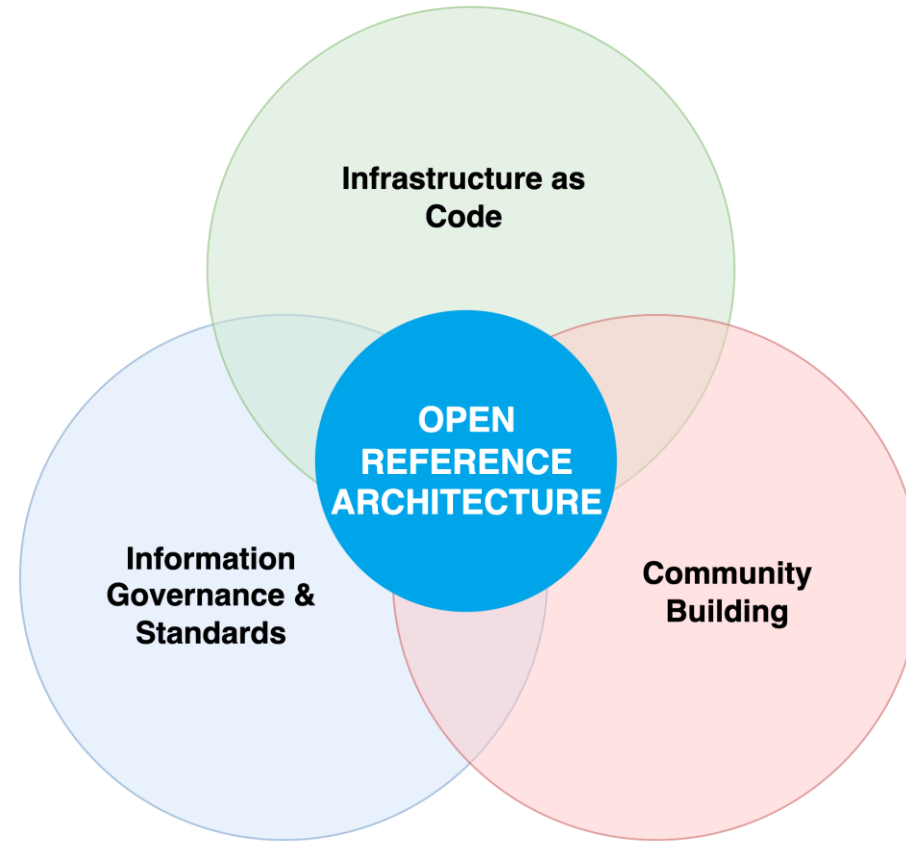


James Robinson, Senior RSE ([jrobinson@turing.ac.uk](mailto:jrobinson@turing.ac.uk))

Data Safe Haven

---

# Data Safe Haven project



---

# Data Safe Haven project

## Infrastructure as code

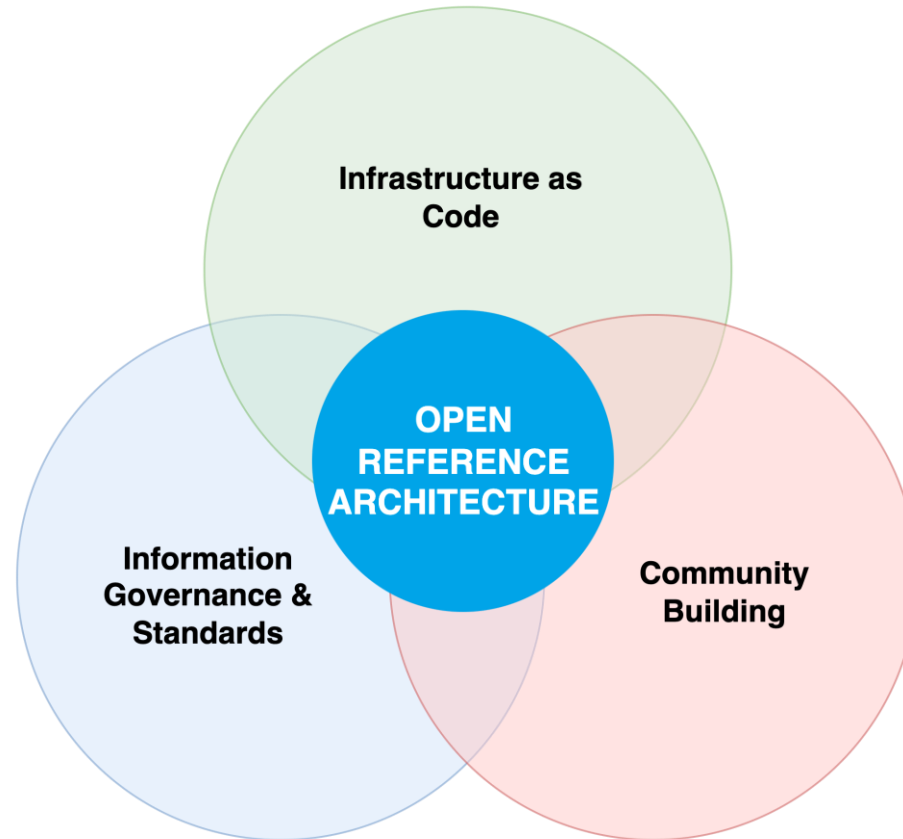
Building a deployable, [reproducible](#) infrastructure

## Information governance & standards

An accompanying [IG framework](#) that is standardised, auditable, accessible and practical

## Community building

An open-first [community](#) that empowers contributions from all stakeholders



---

# Data Safe Haven project

## Infrastructure as code

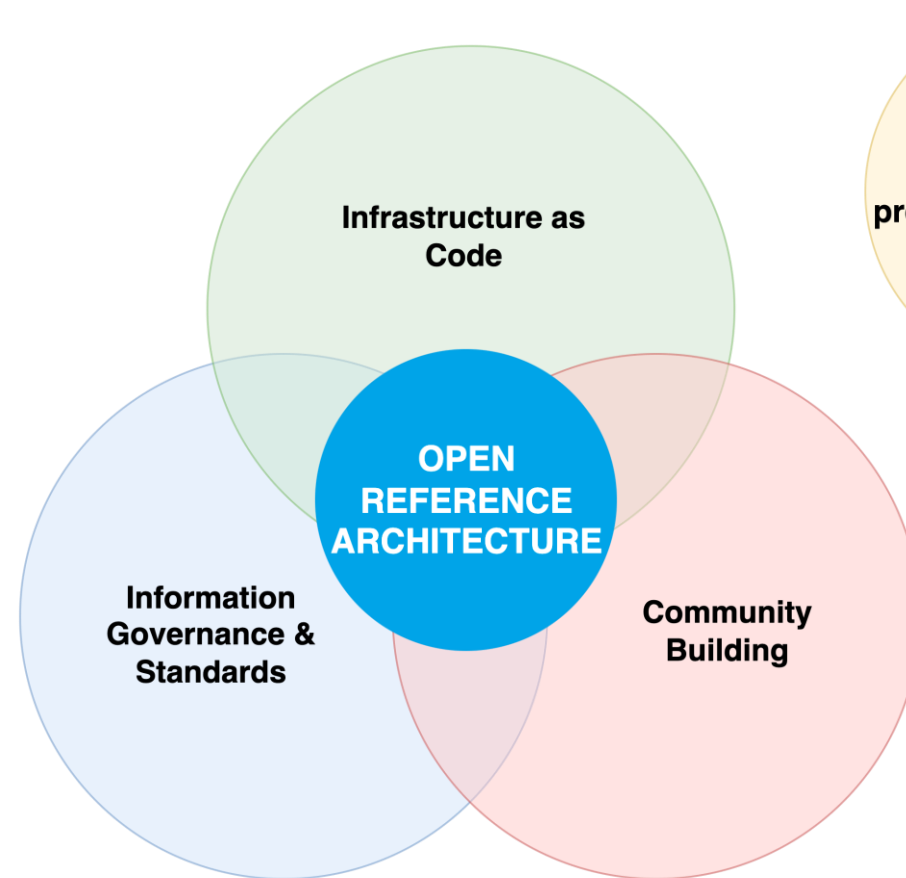
Building a deployable, [reproducible](#) infrastructure

## Information governance & standards

An accompanying [IG framework](#) that is standardised, auditable, accessible and practical

## Community building

An open-first [community](#) that empowers contributions from all stakeholders



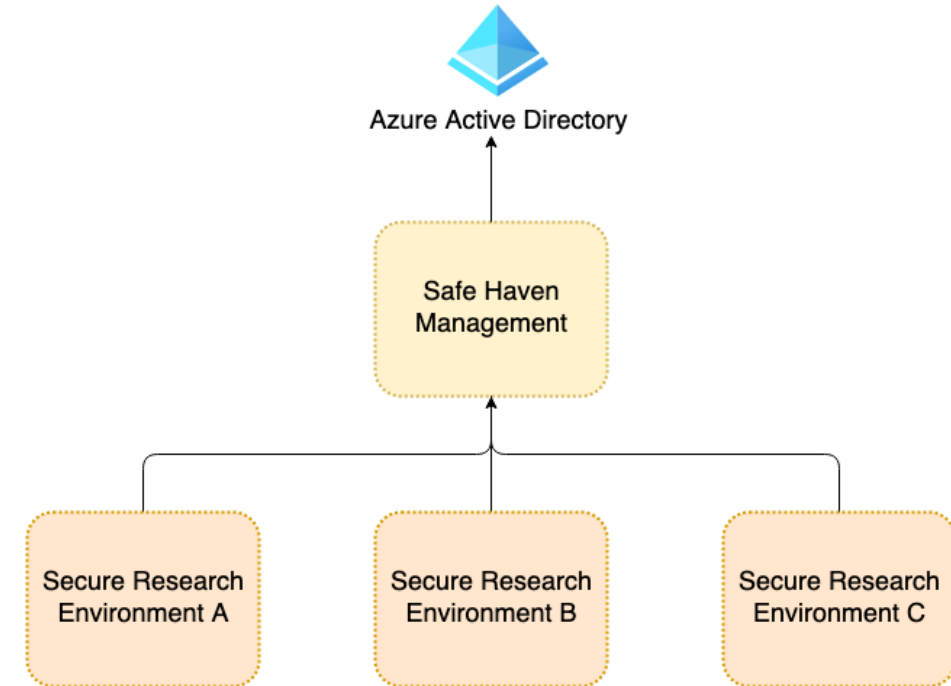
TRESA - a Turing  
production deployment

## A Turing production deployment

A [real-world](#) application and validation of the research workstreams

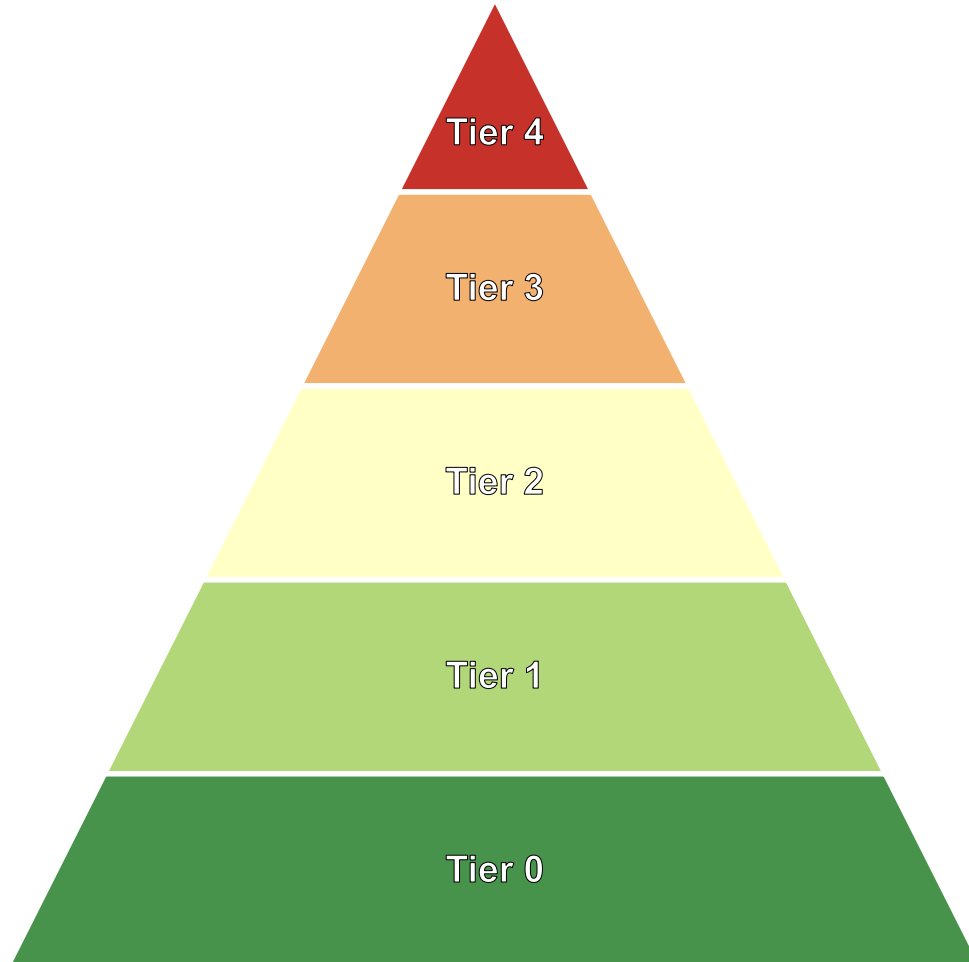
# Data Safe Haven architecture

- **Hub-and-spoke** model
- Central management component
  - Identity service
  - Authentication
  - Package repository access
- **Per-project** research environments
  - Compute
  - Databases
  - Collaborative tools
- **Reproducible** deployments
- **Productive** research environments
- Open to **contributions** rather than pushing 'our' work



---

# Security tiers



Personal data that could threaten safety, security, health  
Data likely to be attacked by state actors

Weakly pseudonymised personal data  
Data likely to be attacked by individuals or small groups

Strongly pseudonymised or synthetic personal data  
Commercial or legal risk

Low risk from disclosure (e.g. having research scooped)

Publicly available, open information

# Software Packages



# Tiered approach

- **Low-risk** projects can install **any package**
  - from upstream repositories
- **Medium-risk** projects can install **any package**
  - through local proxies
- **High-risk** projects can install a **subset**
  - managed allowlist of available packages

# Why not allow all packages?

- **Typo-squatting**
  - scikit-learn vs. scikit-Iearn
- **Malware hijacking**
  - Insertion of malware into widely-used package
- **Ransomware hijacking**
  - Insertion of ransomware into widely-used package
- **Targeted attacks**
  - Dedicated package that targets your TRE in particular

# Which packages to allow?



- Identify packages that are:
  - widely-used
  - well-maintained
- Use this to generate a list of core packages
- Add all dependencies of above packages
- Allow users to install from this allowlist


# Package vetting process

- User makes a **package request** with supporting information:
- Package **details**:
  - Name
  - Audience
  - Number of authors
  - Download statistics
  - List of dependencies
- Explanation of **why** this package is needed
  - Why can't **existing** packages do this?
  - Does this **replace** another package or is it **supplementary**?
- Discussion between maintainers and requester continues:
  - request is either approved or denied


Request for Arrow (R package) to be added to the Tier 3 approve list working in tier 3 SRE #1388

 Closed  3 tasks done DDelbarre opened this issue on Feb 16 · 19 comments · Fixed by #1391

 DDelbarre commented on Feb 16 Member 


 **Checklist**

- I have searched open and closed issues for duplicates.
- This is a request for a new software package to be added to the Data Safe Haven
- The package is still missing in the [latest version](#).

 **Package details**

Provide details about the package you would like to see added:


- Package name: arrow
- Target audience: "core"
- Package version (if different from latest): latest
- Package repository: CRAN
- Number of authors/contributors to the package codebase: 12 authors
- Any existing versions that should not be used (linking to publicly-accessible CVE databases if relevant)
- Download statistics (recent and longer-term, for both current and previous versions): 1.7 million (all time), 80K (last month)
- List of packages that this package depends on:
  - assertthat
  - bit64 (>= 0.9-7)
  - glue
  - methods
  - purrr
  - R6
  - rlang (>= 1.0.0)
  - stats
  - tidyselect (>= 1.0.0)
  - utils
  - vctrs

 **Why is this needed?**

Arrow allows for working with Parquet files, which are a data format that lots of our project data (EDoN) will be provided in. More generally, Parquet is gaining popularity as an efficient column based format for the storage of large data files. Compared to csv files, parquet files can be read/written to up to 50x faster, and be compressed to 10% of the size of a csv file, so it is likely that there will be other projects on the Safe Haven that will use these files over time. Arrow also has functionality to work with large files such as being able to query a file on disk and only import subsets of the data to R, which is useful for big data projects.

None of the currently available R packages for working with Parquet files are on the Tier 3 approved list of packages, so there is no current alternative. There are two other packages that work with Parquet files (sfarrow and parqr), but both of these require arrow as a dependency (and are not so widely used anyway).

I cannot think of any risks that this package would introduce if it was included on the approved list. One of the authors of arrow is Apache, so it also has a reputable author involved with its development. It seems that the dependencies are already approved packages.



# Automation

- Use [Powershell script](#) to generate dependencies
  - Take information from [libraries.io](#) and [rstudio.com](#)
  - Script [freely available](#): BSD licence
- Script generates GitHub **PR**
  - One [commit](#) with changes
  - Project [owners](#) able to approve or deny

## Update PyPI and CRAN allow lists #1601

 Open

github-actions wants to merge 1 commit into [develop](#) from [package-allowlist-updates](#)



Conversation 0

Commits 1

Checks 0

Files changed 2



github-actions bot commented 10 hours ago



### Summary

- Apply package allowlist diff from [af99ee8](#) on 2023-09-13



### Related issues

None



### Tests

Allow-list only



Update PyPI and CRAN allow lists

✓ 614142b

# Typo-squatting

- Pre-defined list of allowed packages
- We provide a **default list**
- Can be **customised** for each project
- Packages can be easily **added** or **removed**

STOPPED BY ALLOWLIST

# Malware hijacking

- Most often via **dependencies**
- Likely to be short-lived for **highly-used** packages
- Attacks typically aim to **exfiltrate** data:
  - Bitcoin wallets
  - SSH keys
  - Passwords
- Attempts should be caught by **standard** protection
  - Restricted **network** connectivity
  - Scanning of **outgoing** connections

## PyTorch Dependency Confusion Attack

In December 2022, **PyTorch** disclosed a **malicious dependency** posing as a legitimate library in their popular machine learning framework. The attack targeted users who installed PyTorch-nightly via Linux pip between December 25, 2022 and December 30, 2022, and worked using a namespace or dependency confusion tactic.

### “cobo-python-api” targeted developers using Mac computers

Another attack in PyPI applied dependency confusion attempting to trick developers into downloading a tainted version of the crypto library **Cobo Custody Restful** in *cobo-python-api*. The package does not have an official distribution through PyPI, so

CAUGHT BY OTHER MEASURES

# Ransomware hijacking

- Importance depends on how system is set-up
- In our case this is **mitigated**:
  - Regular system **backups**
  - Easy for us to teardown and **redploy**

**MITIGATED BY SYSTEM DESIGN**



# Targeted attacks

- Package contains code/data **designed** to attack your system from inside
- Must come from someone who **knows your system** well
  - Safe Users: How much do you **trust** your users/admins?
  - What **sanctions** can you apply in cases of misuse?
- Must be added to the package **allowlist**
  - What is your **approval** process?
- Blocking access to this package will not **stop** a **malicious** user
  - It might make their job harder though

**MITIGATED BY POLICIES**

# Discussion Points

# Eliminating risk

- It is not possible to completely **eliminate risks** from malicious users
- Users with access to a Turing-complete **programming language** can still attack your system

**What can/should be done to mitigate risks?**

# Common approved packages

- Currently no **common set** of approved packages
- We are interested in helping to **develop** one

**Is anyone from this community interested in this?**

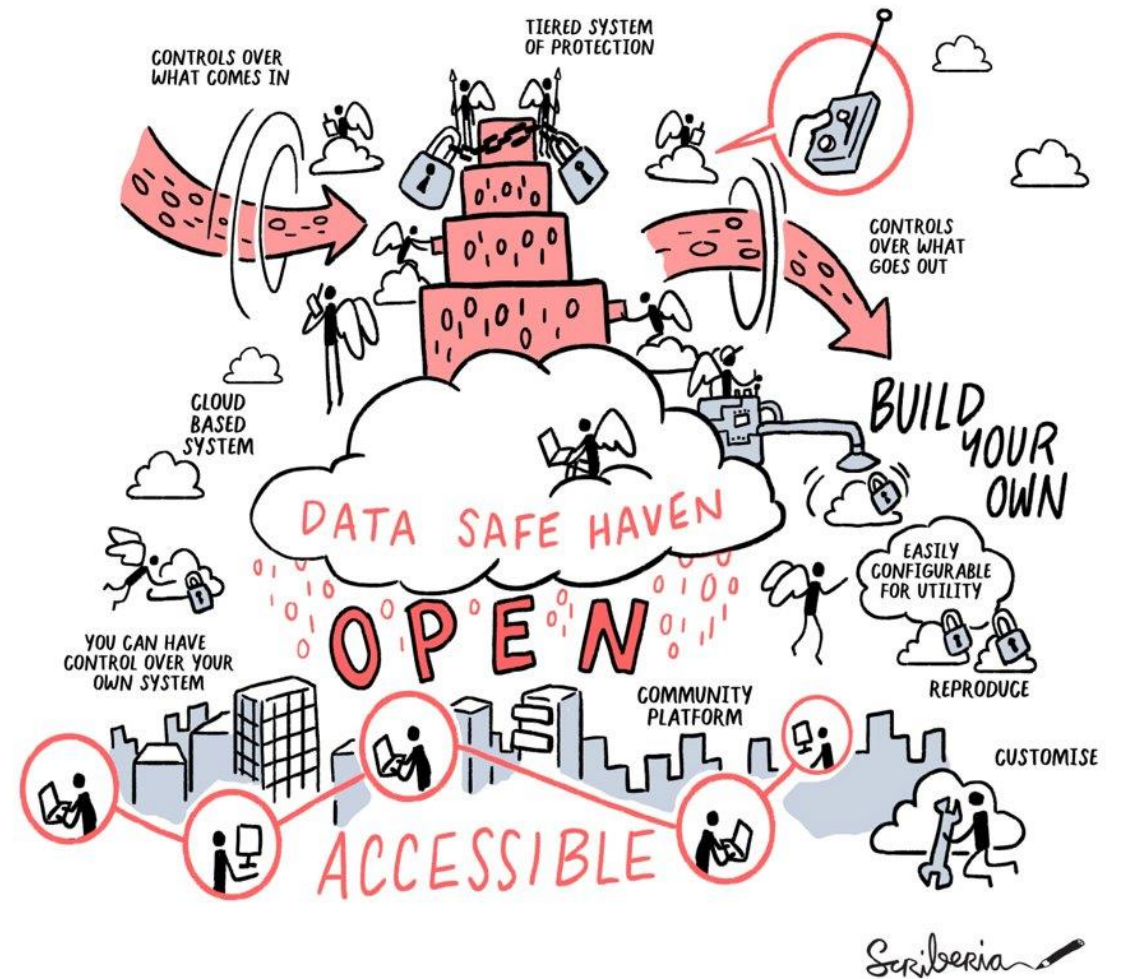
# Summary

- Give different projects different **levels of access** to packages
- Start with a **core package list** and generate dependencies
- Open and easy process for users to **request new** packages
- Interested in working with others to build **common solutions**

Backup

# Our 'North Star'

To remove barriers to working **safely** and **effectively** with sensitive data, by promoting and demonstrating a **culture of open, community-led development** of interoperable foundational **infrastructure** and **governance**.



# Open codebase

The screenshot shows the GitHub repository page for 'data-safe-haven' by 'alan-turing-institute'. The repository is public and has 13 branches, 18 tags, 5,432 commits, 36 stars, and 10 forks. The main content area displays a list of files and folders with their commit dates. The 'About' section on the right provides information about the repository, including a link to the documentation, a list of related repositories, and a list of releases.

File/Folder	Description	Last Commit
.devcontainer	modify location of requirements.txt	2 months ago
.github	Merge pull request #1483 from craddm/docs-...	3 weeks ago
deployment	Update SRD package versions	last month
docs	remove unused snippet	3 weeks ago
environment_configs	Update PyPI and CRAN allow lists	2 months ago
tests	Update sample config files	4 months ago
.PSScriptAnalyzerSett...	Removed residual uses of Write-Host	2 years ago
.PSScriptFormatterSe...	Use <SRE ID> and <SHM ID> across codebase...	2 years ago
.flake8	Nexus option for tier-2 mirrors (#764)	3 years ago
.gitattributes	Added .gitattributes to ensure that .sh files al...	4 years ago
.gitignore	remove custom gitignore	2 months ago
.lychee.toml	Reduce lychee verbosity	2 months ago

The screenshot shows the 'Overview' page for 'data-safe-haven' on The Alan Turing Institute website. The page features a navigation menu with links to 'Overview', 'Design', 'Deployment', 'Processes', and 'Roles'. The main content area is titled 'Section Navigation' and lists several topics: 'What is the Data Safe Haven?', 'Why use the Data Safe Haven?', and 'Sensitivity tiers'. The 'Overview' section is highlighted, and the page content includes a heading 'Background and concepts' and a sub-heading 'What is the Data Safe Haven?' with a brief explanation: 'Basic explanation of what the Data Safe Haven is about.' Other sections include 'Why use the Data Safe Haven?' (Reasons why you might want to consider using the Data Safe Haven.), 'Sensitivity tiers' (Details of the five sensitivity tiers that projects are classified into.), and 'Further resources'.



Code and documentation now freely available



Being used in production at Turing and beyond

(University of Nottingham, East Midlands SNSDE, + adapted by The Health Foundation)

**The  
Alan Turing  
Institute**

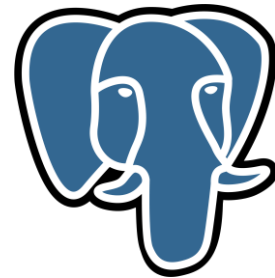


# Enabling high-quality research

- Single **batteries-included** environment
- Developed based on user requirements
- Scalable compute including **GPUs**
- Data in **databases** and/or **cloud** storage
- Access to subset of **PyPI** and **CRAN**
- Servers for **version control**, and **collaborative** document writing



HedgeDoc



LibreOffice  
The Document Foundation



Julia



The Alan Turing Institute

# Getting Involved



[ReadtheDocs](#)



[Open repository](#)



[Slack workspace](#)



[Email contact](#)