Plan Overview

A Data Management Plan created using DMPonline

Title: THESEUS: Making patching happen

Creator: Michel van Eeten

Principal Investigator: Michel van Eeten, Simon Parkin, Herbert Bos, Fabio Massacci, Lokke Moerel

Data Manager: Michel van Eeten, Simon Parkin, Herbert Bos, Fabio Massacci

Affiliation: Delft University of Technology

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan NWO (September 2020)

ORCID iD: 0000-0002-1091-8486

Project abstract:

A core assumption underlying organizational security practices is that defenders are able to remediate known vulnerabilities in their systems in a timely fashion. Otherwise, attackers can just follow the breadcrumbs laid out by security advisories and exploit known weaknesses. This is indeed what happens in many large breaches. While progress has been made at the level of consumers, with automatic updates and default patching settings, this does not translate to enterprises. They face a painful dilemma: patch too soon and incur potential downtime and failures; patch too late and get compromised by attacks. As a result, organizations take a long time to patch even critical security vulnerabilities. The central objective of THESEUS is to empower organizations to patch much faster. It aims to achieve this by radically changing the risk governance of patching. Changing the risk of patching for enterprises means to develop interdisciplinary breakthroughs at three interdependent levels: -- Systems: reducing risk of patching via new techniques in automatic vulnerability and patch triaging, as well as automatic patch generation with live update for cases where critical patches pose unacceptable availability risks. -- Enterprises: better quantifying risk of patching by assessing and aggregating the results of the patch triaging, as a way to estimate exploit likelihood in a coherent picture that accounts for different attacker models and functional impact. -- Governance: more effectively managing risks of patching by introducing incentive mechanisms via notifications and information sharing, sector-wide benchmarks of patching speed, and potentially legal instruments. THESEUS sets out to (1) bring advances from the lab to real-world settings by working with a large consortium of partners from healthcare and transportation who contribute people, data, and pilots; and (2) replace the status quo, as well as counterproductive solutions like mandatory patching, with a richer set of governance interventions across different levels.

ID: 86060

Start date: 15-10-2021

End date: 15-10-2027

Last modified: 03-11-2021

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

THESEUS: Making patching happen

General Information

Name applicant and project number

Michel van Eeten
NWA.1215.18.006

Name of data management support staff consulted during the preparation of this plan and date of consultation.

My faculty data steward, Nicolas Dintzner, has reviewed this DMP on 15 October 2021 .

1. What data will be collected or produced, and what existing data will be re-used?

1.1 Will you re-use existing data for this research?

If yes: explain which existing data you will re-use and under which terms of use.

- Yes

We will start our work on patching practices by collecting information on current patching activities of partner organizations (healthcare and transport). This may include technical documentation, internal process descriptions and policy documents. Such documents will be considered and handled as confidential by all parties involved in the project.
We plan to re-use existing data from open source tools for vulnerability assessment, network moniotring and exploit generation. Such datatset are routinely used in industry for network scan and vulnerability management.
Such datasets only contain specific technical information that is general (e.g. software version x is vulnerable), it cannot be used for re-identification of data. Examples are NVD (National Vulnerability Database), NMap Database of vulnerability signatures etc.

1.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

During this project, we will create two main types of artefacts:
1. Network measurements regarding system's patch level on the network. This will be gathered using existing open source tools.
We plan to perform external measurements on the public internet. If company network measurements will turn out to be necessary, these will be processed on the company premises, IP addressed will be anonymized, the raw data will processed to extract technically relevant data, eliminate PII and other confidential information and then exfiltrated to the university repository. It is envisaged that the raw data will never leave the company premises.
2. Stakeholder perceptions on patching techniques, both from a technical and organizational perspective.
To avoid accidental re-identification, the identity of the interviewed people will be kept on a separate file for the only purpose of research integrity verification. The information processed for analysis will be keep in a separate file where only anonymus ID will be kept.

Overall these data collection efforts will results in quantitative information regarding network's current status (tabular data – e.g., CSV/JSON), and qualitative information on the network management techniques (audio recordings of interviews, transcripts, summaries, codebooks). Should existing tools not be sufficient for our measurements, additional development may be required (software code). Audio recordings will be destroyed as soon as the interviews are transcribed, but we will store them securely in the interim on our servers with only project member access.
Whenever possible, we will use file formats suitable for long-term preservation and re-use of research data. In our choices, we will adhere to the guidance provided by 4TU.ResearchData. Table below provides an overview of the types of data which will be collected and the associated file formats.

Type of data	Format
Quantitative results	.csv format
Qualitative interviews (*anonymised)*	.txt format
Metadata	.txt format

1.3. How much data storage will your project require in total?

- >1000 GB

We expect to require at peak moments 10 TB of storage, mostly because of large-scale internet network measurements .
For company level measurements, those will never leave the company premises so for the purpose of long term storage a significant less amount is envisaged.
To minimize potential data leak, data transfer between partners will be limited to what is strictly necessary. Each partner will use their own storage solution for their work package. The selected data storage will have to meet the following requirements: approved by their institutions for storage of confidential data, provide the necessary backup system, and ensure secured access through authentication mechanisms. Each partner institution will provide a dedicated contact point for data related concerns throughout the project.
Should partners not be able to obtain such storage in the their institutions, TUD may be able to provide them with such storage for the duration of the project.

2. What metadata and documentation will accompany the data?

2.1 Indicate what documentation will accompany the data.

All public datasets will be accompanied by README files providing documentation necessary for data re-use data, as well as point to relevant publications containing the relevant descriptions of collection methodologies. Guidance provided by 4TU.ResearchData will be followed when preparing the README files.

2.2 Indicate which metadata will be provided to help others identify and discover the data.

Data supporting publications will be made openly available through 4TU.ResearchData, unless: (1) It concerns data from industry partners. For those sets, making data available will depend on the agreements with the partner; (2) it contains sensitive information, such as PII or identifiers (e.g., IP addresses, domains) of vulnerable devices.
The published datasets will be accompanied by rich metadata (adhering to DataCite metadata standard) to ensure that they are findable. In addition, to further aid their discoverability, keywords describing the datasets will be added. 4TU.ResearchData is also using schema.org metadata, meaning that all datasets are indexed in Google Dataset Search. Every dataset will be also assigned a Digital Object Identifier (DOI), to make them citable and persistently available.

3. How will data and metadata be stored and backed up during the research?

3.1 Describe where the data and metadata will be stored and backed up during the project.

- Institution networked research storage

During the course of the research project, all data will be stored on local servers maintained and automatically backed up by the TPM cybersecurity team.
All code will be maintained in a dedicated GitLab version control system provided by TU Delft, which is backed up and maintained by TU Delft ICT.

3.2 How will data security and protection of sensitive data be taken care of during the research?

- Additional security measures (please specify)

During this project we will be handling some personal research data. We have made specific arrangements to ensure the safety and security of these datasets, following guidance of our faculty Data Steward. In brief:

- As also stated on the informed consent form, only team members will have access to the designated server. The storage security is ensured by TU Delft ICT services and the sysadmin of the cybersecurity team.
- The laptops used by project members will be encrypted.

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?

4.1 Will you process and/or store personal data during your project?

If yes, how will compliance with legislation and (institutional) regulation on personal data be ensured?

- Yes

During this project we will be handling some personal research data. We have made specific arrangements to ensure the safety and security of these datasets, following guidance of our faculty Data Steward. In addition, we adhere to the dedicated procedure for ensuring compliance with GDPR legislation established by TU Delft.
In brief:

- As also stated on the informed consent form, only team members have access to the designated server. The storage security is ensured by TU Delft ICT services.
- The laptops used by project members will be encrypted

4.2 How will ownership of the data and intellectual property rights to the data be managed?

The datasets underlying the published papers will be publicly released following NWO's policies, barring the exceptions mentioned under 2.2. During the active phase of research, the PI of each participating institution (TU Delft, VU, Tilburg university) will oversee the access rights to data (and other outputs), as well as any requests for access from external parties for the work led by that institution (as indicated by the affiliation of the lead author of the work.

5. How and when will data be shared and preserved for the long term?

5.1 How will data be selected for long-term preservation?

- All data resulting from the project will be preserved for at least 10 years

All data supporting publications will be made openly available through 4TU.ResearchData, barring the exceptions noted under 2.2. 4TU.ResearchData is a trusted and certified research data repository (it has a Data Seal of Approval certification), and ensures that research data will be preserved for at least 15 years.

5.2 Are there any (legal, IP, privacy related, security related) reasons to restrict access to the data once made publicly available, to limit which data will be made publicly available, or to not make part of the data publicly available?

If yes, please explain.

- Yes

Yes, there are reasons to not share datasets that include, or consist of, data shared with us by industry partners. As per the consortium agreement, restrictions on making data publicly available will be decided on a case-by-case basis.

5.3 What data will be made available for re-use?

- All data resulting from the project will be made available

All raw data, except for audio recordings, will be retained for at least ten years on TU Delft servers for the purposes of validation. The datasets underlying the figures and conclusions in academic papers will be made publicly available through 4TU.ResearchData, in line with the TU Delft Research Data Framework Policy, barring the exceptions noted under 2.2.
For raw data that includes PII, data will be retained for at least ten years on TU Delft servers for the purposes of validation. Personal data will be anonymized and anonymized datasets (the processed data) underlying the figures and conclusions in academic papers will be made publicly available through the 4TU Centre for Research Data, in line with the TU Delft Research Data Framework Policy. The release of interview transcripts as open data will not be feasible, if the transcript risks de-anonymizing the interviewee.

5.4 When will the data be available for re-use, and for how long will the data be available?

- Data available as soon as article is published

We aim to make research data underpinning research papers publicly available by depositing at 4TU.ResearchData at the time of the publication of the corresponding research article, barring the exceptions mentioned under 2.2.

5.5 In which repository will the data be archived and made available for re-use, and under which license?

Barring the exceptions mentioned under 2.2., the datasets underlying the published papers will be published at 4TU.ResearchData, which is a trusted and certified research data repository (Data Seal of Approval certification). All datasets will be licensed under a CC-BY licence which requires attribution/credit for the original creation, while at the same time ensures broadest possible re-use. All datasets will be accompanied by rich and descriptive metadata, compliant with DataCite metadata schema, to ensure that all datasets are findable and accessible online. https://data.4tu.nl/info/en/

5.6 Describe your strategy for publishing the analysis software that will be generated in this project.

The developed software and codes presented in academic papers will be shared on GitHub and those GitHub repositories will be published via 4TU.ResearchData. This way, they will be publicly available to anyone for re-use under an open licence. They will be also assigned a Digital Object Identifier (DOI), to make them citable and persistently available.

6. Data management costs

6.1 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

4TU.ResearchData is able to archive 1TB of data per researcher per year free of charge for all TU Delft researchers. We do not expect to exceed this and therefore there are no additional costs of long term preservation.
The dedicated data manager hired in the project (see the project proposal and staff allocation) will be responsible for data management in the project.