AI4CYBER

TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR
CYBERSECURITY REINFORCEMENT AND SYSTEM RESILIENCE

Cybersecurity Datasets – Opportunities and Challenges

Erkuden Rios, Project Coordinator, tecnalia

MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

PRECINCT CONFERENCE, 16th May 2023, Brussels.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

AI4CYBER is a collection of AI services

AI4CYBER is an Ecosystem Framework of Next generation **AI-based services** for critical system **robustness, resilience**, and appropriate **response in the face of advanced and AI-powered cyberattacks**.

11 Key Results that cover **6 cybersecurity areas**

A collection of 11 innovative resilience and autonomous response services that leverage AI models and Big Data, aimed to be encapsulated in cybersecurity tools to ensure a continuum of system protection.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

Objective of the Workshop

- Push **open collaboration and open innovation** with respect to **data requirements in cybersecurity research,** by discussing the difficulties and opportunities of using and sharing cybersecurity related datasets for cybersecurity research and cybersecurity solution development.
- In the workshop we will discuss on available open datasets, sources of datasets, reusable datasets, means and principles for sharing datasets, etc.
- <u>Suggested attendees</u>: Cybersecurity researchers dealing with AI used in threat detection, threat simulation, code testing, incident response, etc.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

Challenges of cybersecurity datasets

- The **accuracy** of AI algorithms and models is directly dependant on the **quality and amount of data** used to train them.
- When using models and algorithms to develop AI-based cybersecurity solutions, the learning requires **big amounts of well-structed and sanitized data** which are sometimes difficult to get.
- Some literature works include **open datasets**, but they are difficult to reuse as a basis of further research.
- The **confidentiality** (and sometimes secrecy) of cybersecurity data makes it difficult to be shared.
- Commercial datasets are **limited**.
- Creating the **datasets synthetically** is often hard because it is not easy to replicate the nature of real attacks and getting realistic results is challenging.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

Opportunities of cybersecurity dataset sharing

- Improved cybersecurity research by learning from previous works.
  - Learn how others get or build their datasets and their models.
  - Learn data synthetization techniques and realistic data (quality data) creation.
  - Learn anonymisation and aggregation techniques.
- Test your models on top of other datasets to benchmark and improve your model.
- Verify the quality and correctness of other researchers' solutions.
- Well-structured and high quality pools of FAIR data.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

FAIR Data

- Findability: open (meta)data shall be easily searchable and locatable -> include identifiers, keywords, version numbers and metadata.
- Accessibility: openness of the (meta)data -> clarify and enable means of access, identify the repository where the data will be deposited, access protocols to and any potential restriction.
- Interoperability: refers to the ability of the (meta)data to integrate or work together with minimal effort with other (meta)data and tools -> clarify vocabularies, standards, formats and methodologies used.
- Reusability: the ability of the (meta)data to be utilized in other contexts -> include attributes, clear and accessible licenses, follow standards and identify provenance.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

From ChatGPT - metrics commonly used to evaluate the quality of datasets in AI

- **Accuracy**: Measures the proportion of correctly labeled instances in a dataset. It indicates how well the dataset represents the ground truth.
- **Completeness**: Measures the extent to which the dataset covers the entire range of possible values and scenarios. It evaluates whether the dataset contains sufficient samples to adequately represent the target population.
- **Consistency**: Assesses the absence of contradictions or conflicts within the dataset. It measures the coherence and uniformity of the data.
- **Reliability**: Measures the degree of trustworthiness and dependability of the dataset. It evaluates factors such as data collection methods, sources, and potential biases.
- **Relevance**: Determines the degree to which the dataset aligns with the objectives and requirements of the AI task. It evaluates the suitability and applicability of the dataset for the specific problem at hand.
- **Precision**: Measures the proportion of true positives out of the total positive instances. It is particularly relevant for datasets used in classification tasks.
- **Recall**: Measures the proportion of true positives identified out of the total actual positive instances. It is also important in classification tasks, especially when it is crucial to minimize false negatives.
- **Bias**: Evaluates the presence of systematic errors or prejudices in the dataset that may lead to unfair or discriminatory outcomes. It is essential to assess and mitigate biases in AI systems.
- **Noise**: Assesses the presence of irrelevant or misleading data within the dataset. It measures the level of interference or distortion that may impact the performance of AI models.
- **Diversity**: Measures the variety and heterogeneity of the dataset in terms of different attributes, categories, or scenarios. A diverse dataset helps ensure the model's generalization capability.
- **Balance**: Evaluates the distribution of instances across different classes or categories. It assesses whether the dataset has an equal representation of different classes or if it is skewed towards certain categories.
- **Annotation Quality**: Applies specifically to labeled datasets. It assesses the accuracy and consistency of the annotations or labels assigned to the instances in the dataset.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450.

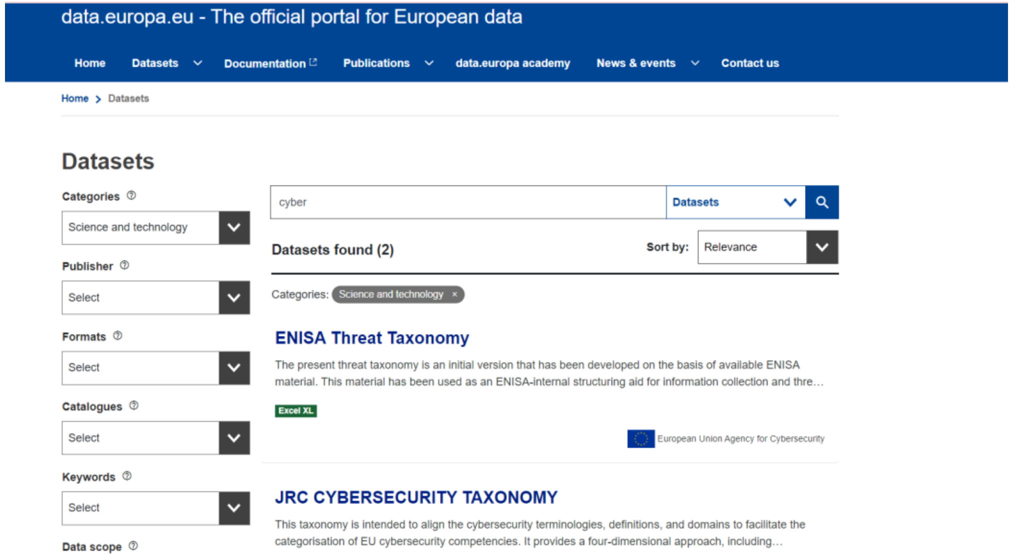
Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

European data pools

- European Data Portal: https://data.europa.eu/en
  - “Science & technology”, “cyber” ->

ENISA Threat taxonomy, JRC Cybersecutity Taxonomy

- Zenodo: https://zenodo.org/
  - “cybersecurity” -> 11 csv, 8 xlsx, some are studies and listings.



Datasets

Categories: Science and technology

Publisher: Select

Formats: Select

Catalogues: Select

Keywords: Select

Data scope

Search: cyber — Datasets

Datasets found (2) — Sort by: Relevance

Categories: Science and technology ×

ENISA Threat Taxonomy

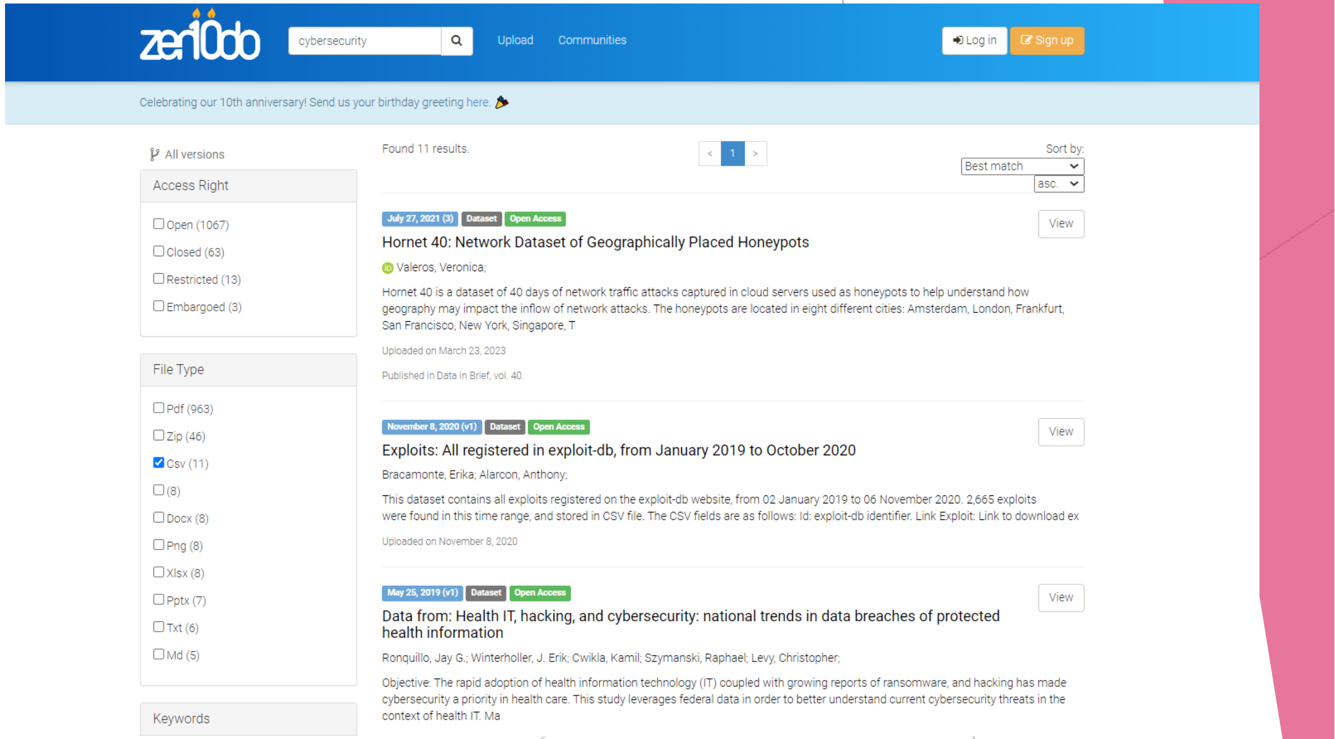
The present threat taxonomy is an initial version that has been developed on the basis of available ENISA material. This material has been used as an ENISA-internal structuring aid for information collection and thre…

Excel XL

European Union Agency for Cybersecurity

JRC CYBERSECURITY TAXONOMY

This taxonomy is intended to align the cybersecurity terminologies, definitions, and domains to facilitate the categorisation of EU cybersecurity competencies. It provides a four-dimensional approach, including…



Celebrating our 10th anniversary! Send us your birthday greeting here.

All versions

Access Right
- □ Open (1067)
- □ Closed (63)
- □ Restricted (13)
- □ Embargoed (3)

File Type
- □ Pdf (963)
- □ Zip (46)
- ☑ Csv (11)
- □ (8)
- □ Docx (8)
- □ Png (8)
- □ Xlsx (8)
- □ Pptx (7)
- □ Txt (6)
- □ Md (5)

Keywords

Found 11 results.

Sort by: Best match, asc.

July 27, 2021 (3) | Dataset | Open Access

Hornet 40: Network Dataset of Geographically Placed Honeypots

Valeros, Veronica;

Hornet 40 is a dataset of 40 days of network traffic attacks captured in cloud servers used as honeypots to help understand how geography may impact the inflow of network attacks. The honeypots are located in eight different cities: Amsterdam, London, Frankfurt, San Francisco, New York, Singapore, T

Uploaded on March 23, 2023

Published in Data in Brief, vol. 40.

View

November 8, 2020 (v1) | Dataset | Open Access

Exploits: All registered in exploit-db, from January 2019 to October 2020

Bracamonte, Erika; Alarcon, Anthony;

This dataset contains all exploits registered on the exploit-db website, from 02 January 2019 to 06 November 2020. 2,665 exploits were found in this time range, and stored in CSV file. The CSV fields are as follows: Id: exploit-db identifier. Link Exploit: Link to download ex

Uploaded on November 8, 2020

View

May 25, 2019 (v1) | Dataset | Open Access

Data from: Health IT, hacking, and cybersecurity: national trends in data breaches of protected health information

Ronquillo, Jay G.; Winterholler, J. Erik; Cwikla, Kamil; Szymanski, Raphael; Levy, Christopher;

Objective: The rapid adoption of health information technology (IT) coupled with growing reports of ransomware, and hacking has made cybersecurity a priority in health care. This study leverages federal data in order to better understand current cybersecurity threats in the context of health IT. Ma

View

Other **data pools** (e.g. IEEE DataPort) require subscription for accessing most datasets.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

Open discussion

- What are the **major barriers** to sharing?
- What possible **solutions** do you see for increasing dataset sharing?
- How to **avoid sharing sensitive information** in the datasets? Anonymisation, pseudonimisation, etc. do always work?
- What **trustworthy open data pools** do you know so as the datasets can be shared? Zenodo (Open Aire), JRC, etc.
- Could an **EU cybersecurity researchers' community** be created with the goal of sharing datasets and information about the usage of the datasets, in which researchers contribute to with datasets and related info?

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

***Disclaimer:** Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*

AI4CYBER

TRUSTWORTHY ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY REINFORCEMENT AND SYSTEM RESILIENCE

https://ai4cyber.eu

https://twitter.com/Ai4Cyber

https://www.linkedin.com/company/ai4cyber/

Erkuden Rios

Project Coordinator

erkuden.ríos@tecnalia.com

Thank you for your attention!

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070450.

Disclaimer: *Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.*