University of Essex Online

"… But you left the door open!". Security guides in an insecure world.

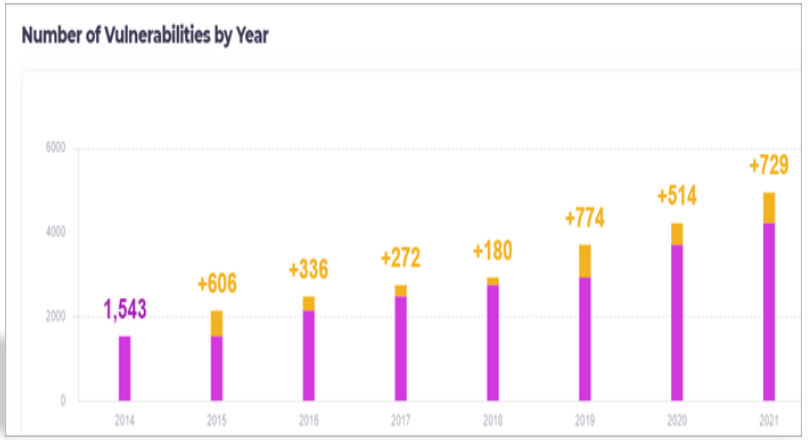
Research Proposal, 2022

Michael Justus, UoEO, Research Methods and Professional Practice, 2022

Background

What is the issue?

- Resolving security needs support and informed answers based on standards.
- Software development is a social activity (Stefik and Hanenberg, 2014).
- Stack Overflow (SO) provides well-documented, correctly tagged questions and answers (Moutidis and Williams, 2021).

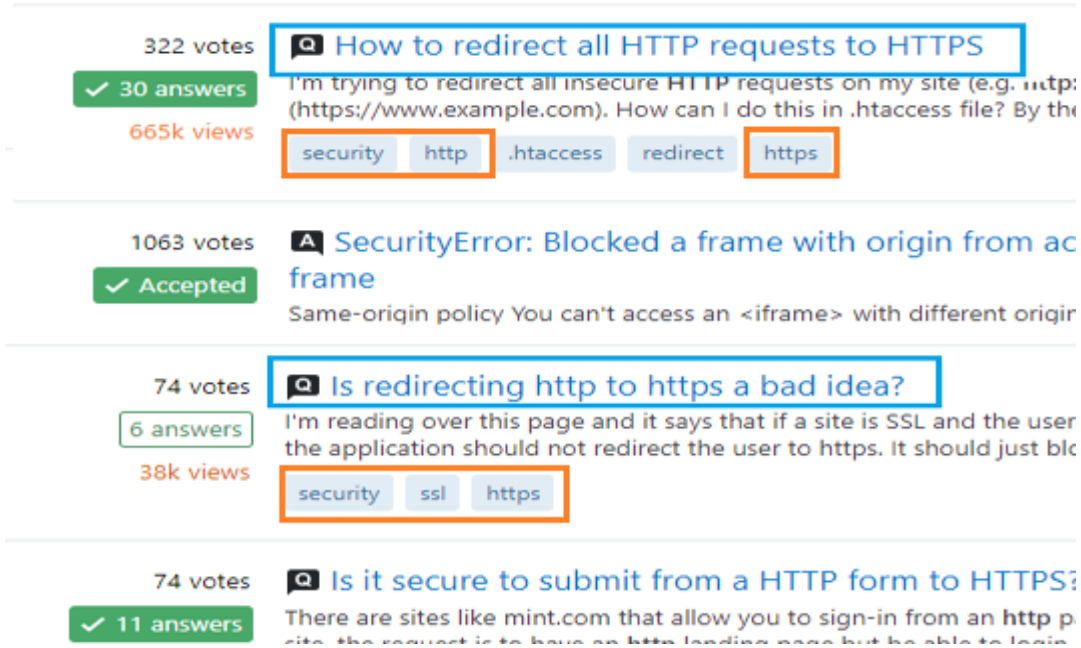


WPScan Statistics, 2022

Why Topic Modelling?

Automatically find topics inside a corpus of text; allows quick exploration of text (Asmussen and Møller, 2019). Can uncover recurring issues in development texts (Johri and Bansal, 2018).

Significance



Sample of SO questions and tags

Developing secure information systems is difficult with similar questions asked repeatedly with no answers referencing industry standards related to the topic.

Researchers have not yet correlated current security guidance with questions/answers on developer support sites such as Stack Overflow

- Linking current industry guidance to developer questions/answers provides:
  - Awareness of standards' influence
  - Guidance for (security) organisations/bodies
  - Support for security compendium

Related Work

- **Q/A**. Barua et al. (2014) recommend analysis on questions/answers for deeper insights.
- **Languages.** Unclear if user participation for new languages is security-related (Chakraborty et al., 2021).
- **Communities.** Security concepts might not permeate developer communities (Moutidis and Williams, 2021).
- **Vulnerabilities.** Not all answers are trustworthy. ~66% Java code outdated, ~6.6% considered harmful. Code may contain vulnerabilities (Ragkhitwetsagul et al., 2019) despite supporting security standards. Neuhaus and Zimmermann (2010) used topic modelling on CVE DB, not SO data.
- **Social.** Lopez et al. (2018) found developers concerned about personal values, responsibility, trust, and fear. Showing they are aware of social impact of security.
- **Privacy.** Tahaei et al. (2020) looked at drivers of implementing privacy policies but did not link to privacy standards.

Others classified SO data by question, type, and code (Allamanis and Sutton, 2013) but did not link to security guidance.

Research Gap

Use topic modelling to determine the strength of correlation between security topics in answers and questions and to public security guidance.

Research Question(s)

RQ1: What is the perceived importance of current security guidance on SO users?

RQ2: What percentage of accepted security answers contain references to the latest security documentation or guidance?

RQ3: How has security guidance impacted developers over time?

Goal: To show the extent to which developers consider and reference OWASP Top 10:2021.

Objective 1: Obtain a ranked list of topics from a subset of SO question/answers relating to security within the scope's time period.

Objective 2: Generate a list of topics found in the latest corpus of security guidance text.

Objective 3: Analyse the relationships between datasets and document results.

Research Methodology

Item	Description	Justification
Approach	Quantitative	Analysis of relations, statistics, counts
Framework	(1) Prepare → (2) Source and Obtain → (3) Pre-process → (4) Topic Models → (5) Analyse ←→ Adjust → (6) Present	Require structured approach to help plan project deliverables
Collection	Literature from Google Scholar, Taylor and Francis, IEEE Explore, University of Essex library, Scopus, ScienceDirect, and computing journals.	Insights into previous methods.
	"Posts" data sets from Stack Overflow (StackExchange Data Dump)	SO provides public data sets
Analysis	Topic modelling library for Java (Mallet) or Python (Gensim).	No need to develop from scratch
	Data analysis via Excel	Familiar with tool
Scope	Posts referring to security, has "security" or words semantically similar	The data supports research gap
	Data limited to two years	Delays, classification processing time and project duration
Limitations	Literature pay-walls;	N/A

Table 1: elements of research methodology

Ethical Considerations

- References

  Acknowledgement will be ascribed to researchers' works used.

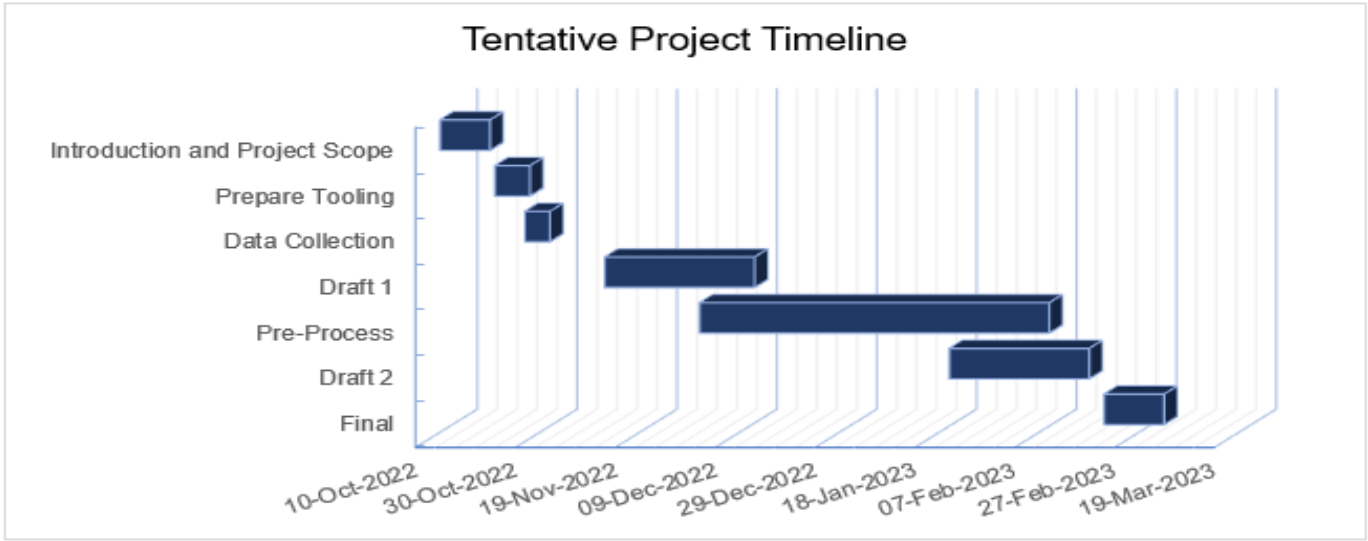
- Data

  All datasets obtained from pubic or open-access sources. All personally identifiable information in any dataset will be discarded.

- Software

  All software used will either have a valid software license or be open-source.

Project Timeline (Tentative)



Name	Start Date	Duration
Final	20/02/2023	12.00
Draft 2	20/01/2023	28.00
Pre-Process	01/12/2022	70.00
Draft 1	12/11/2022	30.00
Data Collection	27/10/2022	5.00
Prepare Tooling	21/10/2022	7.00
Scope	10/10/2022	10.00
	TOTAL DAYS	**162**

Figure 1: tentative timeline showing total days estimated.

Risk Assessment

Risk	Why?	Mitigation
Time constraints	Full-time work	• Re-evaluate objectives or questions • Raise with supervisor
Costs/Hardware	Might require more powerful tools/platforms	• Reduce dataset size and scope • Raise with supervisor • Free cloud-based service
Familiarity	Lack of experience developing a topic model	• Seek guidance/samples from development sites • Seek fellow students with experience
Insufficient literature	Existing literature may be outdated	Search for recent literature
Limited datasets	API or access rights issues	Consider alternate sources
Recency	SO datasets contain outdated responses	None

Table 3: evaluation of risks to project.

Budget

Budget Item	“Free”* available?	Paid-for Cost
Microsoft Excel	No	Yes. I have a personal license.
Visual Studio Code	Yes.	No.
Laptop	No	Yes. I have my own hardware.

Table 4: evaluation of budgetary requirements

* “Free” has limited feature sets.

Artefacts

- Tables.
  - List of security topics (SO, OWASP Top 10:2021)
  - Nominal data
- Graphs.
  - Trends over time
  - Links between topics
- Dataset analysis results.
  - Interpretations of the data
- (Optional) Software algorithms and configurations

Thank you.

References (1/2)

- Allamanis, M. & Sutton, C. (2013). Why, when, and what: analyzing stack overflow questions by topic, type, and code. '*2013 10th Working conference on mining software repositories (MSR)'.* 53-56. IEEE.
- Asmussen, C.B. & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data*, 6(1):1-18.
- Barua, A., Thomas, S.W. & Hassan, A.E. (2014). What are developers talking about? an analysis of topics and trends in stack overflow. *Empirical Software Engineering*, 19(3):619-654.
- Chakraborty, P., Shahriyar, R., Iqbal, A. & Uddin, G. (2021). How do developers discuss and support new programming languages in technical Q&A site? An empirical study of Go, Swift, and Rust in Stack Overflow. Information and Software Technology, 137:106603.
- Johri, V. & Bansal, S. (2018). Identifying trends in technologies and programming languages using Topic Modeling. '*2018 IEEE 12th International Conference on Semantic Computing (ICSC)'.* 391-396. IEEE.

References (2/2)

- Lopez, T., Tun, T.T., Bandara, A., Levine, M., Nuseibeh, B. & Sharp, H. (2018). An investigation of security conversations in stack overflow: Perceptions of security and community involvement. '*Proceedings of the 1st international workshop on security awareness from design to deployment'.* 26-32.

- Moutidis, I. & Williams, H.T. (2021). Community evolution on stack overflow. Plos one, 16(6), p.e0253010.

- Neuhaus, S. & Zimmermann, T. (2010). Security trend analysis with cve topic models. '*2010 IEEE 21st International Symposium on Software Reliability Engineering'.* 111-120. IEEE.

- Ragkhitwetsagul, C., Krinke, J., Paixao, M., Bianco, G. & Oliveto, R., (2019). Toxic code snippets on stack overflow. *IEEE Transactions on Software Engineering*, 47(3):560-581.

- Stefik, A. & Hanenberg, S. (2014). The programming language wars: Questions and responsibilities for the programming language community. '*Proceedings of the 2014 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming & Software*'. 283-299.

- Tahaei, M., Vaniea, K. & Saphra, N. (2020). Understanding privacy-related questions on stack overflow. '*Proceedings of the 2020 CHI conference on human factors in computing systems'.* 1-14