OWASP Top 10 2020 Data Analysis Plan

# Goals

To collect the most comprehensive dataset related to identified application vulnerabilities to-date to enable analysis for the Top 10 and other future research as well. This data should come from a variety of sources; security vendors and consultancies, bug bounties, along with company/organizational contributions. Data will be normalized to allow for level comparison between Human assisted Tooling and Tooling assisted Humans.

# Analysis Infrastructure

Plan to leverage the OWASP Azure Cloud Infrastructure to collect, analyze, and store the data contributed.

# Contributions

We plan to support both known and pseudo-anonymous contributions. The preference is for contributions to be known; this immensely helps with the validation/quality/confidence of the data submitted. If the submitter prefers to have their data stored anonymously and even go as far as submitting the data anonymously, then it will have to be classified as “unverified” vs. “verified”.

Verified Data Contribution

Scenario 1: The submitter is known and has agreed to be identified as a contributing party.

Scenario 2: The submitter is known but would rather not be publicly identified.

Scenario 3: The submitter is known but does not want it recorded in the dataset.

Unverified Data Contribution

Scenario 4: The submitter is anonymous.

The analysis of the data will be conducted with a careful distinction when the unverified data is part of the dataset that was analyzed.

## Contribution Process

There are a few ways that data can be contributed:

1. Email a CSV/Excel file with the dataset(s) to [brian.glas@owasp.org](mailto:brian.glas@owasp.org)
2. Upload a CSV/Excel file to a “contribution page”

## Contribution Period

We plan to accept contributions to the Top 10 2020 during March-May of 2020 for data dating from 2017 to current.

# Data Structure

The following data elements are \*required or optional:

## Per DataSet

Contributor Name (org or anon)

Contributor Contact Email

Time period (2019, 2018, 2017)

\*Number of applications tested

\*CWEs w/ number of applications found in

Type of testing (TaH, HaT, Tools)

Primary Language (code)

Geographic Region (Global, North America, EU, Asia, other)

Primary Industry (Multiple, Financial, Industrial, Software, ??)

Whether or not data contains retests or the same applications multiple times (T/F)

If a contributor has two types of datasets, one from HaT and one from TaH sources, then it is recommended to submit them as two separate datasets.

# Process

At a high level, we plan to perform a level of data normalization; however, we will keep a version of the raw data contributed for future analysis. We will analyze the CWE distribution of the datasets and potentially reclassify some CWEs to consolidate them into larger buckets. We will carefully document all normalization actions taken so it is clear what has been done.

[Proposed]

Develops general CWSS scores for the top 20-30 CWEs and include potential impact into the Top 10 weighting.