SMART
Internet-Draft

M. McFadden
internet policy advisors ltd
A. Mills
UWE - Bristol

Intended status: Informational
Expires: September 6, 2020

March 6, 2020

Textual Analysis Methodology for Security Considerations Sections
draft-mcfadden-smart-rfc3552-textual-research-01.txt

Status of this Memo

Copyright Notice

Abstract

RFC3552 provides guidance to authors in crafting RFC text on Security Considerations. The RFC is more than fifteen years old. With the threat landscape and security ecosystem significantly changed since the RFC was published, RFC3552 is a candidate for update. This draft proposes that, prior to drafting an update to RFC3552, an examination of recent, published Security Considerations sections be carried out as a baseline for how to improve RFC3552. It suggests a methodology for examining Security Considerations sections in published RFCs and the extraction of both quantitative and qualitative information that could inform a revision of the older guidance. It also reports on a recent experiment on textual analysis of sixteen years of RFC Security Consideration sections.

Table of Contents

1. Introduction

[RFC2223] requires that all RFCs have a Security Consideration section. The motivation of the section is both to encourage RFC authors to consider security in protocol design and to inform readers of relevant security issues. RFC3552 was published in July of 2003 to give guidance to RFC authors on how to write a good Security Considerations section. It is structured in three parts: a tutorial and definitional section, then a series of guidelines, and finally a series of examples.

It is possible to observe that the Internet security landscape has changed significantly since the publication of RFC3552. Rather than an immediate attempt to draft and discuss a revision to the older RFC, it may be prudent to learn from the experience of more than fifteen years of documents published since RFC3552 was approved for publication.

It is possible that an examination of published Security Considerations sections of existing documents could give both quantitative and qualitative insight on how to proceed with a newer version of the Security Considerations guidelines. The motivation is to inform any discussion of a revision with quantitative and qualitative data gleaned from years of published RFCs.

This document proposes a methodology for such research.

This scope of this proposal is for the research itself. Discussion of relevant issues, document organization and revised content for a revision of RFC3552 is out of scope. Instead, the motivation is to guide a piece of research that would later form part of the foundation for a discussion of a revision to RFC3552.

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

In this document, these words will appear with that interpretation only when in ALL CAPS. Lower case uses of these words are not to be interpreted as carrying significance described in RFC 2119.

3. Motivation

Since 1998, all RFCs have been required to have a Security Considerations section. The authors of RFC3552 observed that "historically, such sections have been relatively weak." The motivation for RFC3552 was, in part, to improve the quality of Security Considerations sections.

Today the Internet threat model, the landscape of attacks, and our understanding of how to craft protocols that are more robust and resilient has changed significantly. Experience in both protocol design and implementation has greatly improved our understanding of the security implications of choices made during protocol design.

It is possible that a revision of RFC3552, reflecting the changes to the Internet and our understanding of the evolved security landscape and threat model, is appropriate. The IAB is currently examining and reassessing the Internet's threat model [1].

The IAB has previously discussed a potential revision to RFC3552 in its report from the Strengthening the Internet (STRINT) Workshop. In section 2 of [RFC7687], the editors report that "...the IETF may be in a position to start to develop an update to BCP 72 [RFC3552], most likely as a new RFC enhancing that BCP and dealing with recommendations on how to mitigate PM and how to reflect that in IETF work."

If a revision were to be contemplated, it would be useful to learn from the body of experience of crafting Security Considerations sections in recent years. That body of experience could inform the discussion of what makes up a good Security Considerations section by collecting real-world data from existing RFCs. It would be possible to have a survey of the existing Security Considerations sections in published RFCs. The data collected from that survey could provide one source of information for discussion of how to improve upon RFC3552 in the current environment.

For such a survey to be successful, an outline of some basic goals and a methodology would be required. This document provides those goals and methodology. The intent is that individuals or organizations could then carry out such a survey, publish the results and use that data to inform any discussion of a potential 3552bis.

This draft also documents the results of a recent experiment to conduct an automated survey of words in Security Considerations sections.

3.1. Non-goals and scoping

This document specifically does not make suggestions for changes to RFC3552. It also does not identify changes to the Internet threat model or the general security landscape that has changed since that RFC has been published.

The scope of this document is to provide a basic set of goals for research on existing Security Considerations sections and establish a methodology for conducting that research.

3.2. Research Group

The research work suggested in this document was envisioned and intended to be carried out as a research activity of the proposed Stopping Malware and Researching Threats (SMART) research group in the IRTF. The work could also be conducted independently and submitted as an Independent Submission in the IETF.

4. Goals for Surveying Existing Security Considerations Sections

A cursory examination of recent years' Security Considerations sections shows that authors publish a wide variety of these sections. This is natural since the RFC series has a diverse set of purposes and readership.

However, even a cursory examination shows that published Security Considerations sections have some clear characteristics. Identifying useful characteristics and then surveying the existing base of published RFCs may provide a useful base of information for a later discussion of revising RFC3552.

The goal of surveying existing Security Considerations sections is to provide quantitative and qualitative data, from existing, published RFCs, that can be used to inform a discussion of revising RFC3552.

5. Methodology

5.1. Methodology Overview

The survey of existing Security Considerations sections would examine a subset of RFCs published since the publication of RFC3552. RFCs obsoleted by later publications, RFCs that are reports from IAB activities and IETF, IRTF, and IESG administrative RFC are omitted from consideration.

The survey should select a specific timeframe, across which, all RFCs published in that period are examined.

The examination proceeds in two parts: a quantitative examination of the Security Considerations sections and then a qualitative examination.

As an example, the quantitative examination might survey and collect data on the source of the RFC (e.g. Security Area, Routing Area, Transport Area), whether the RFC extends the Security Considerations section of a previously published document, the wordcount of the section, and the existence of specific keywords.

The qualitative analysis might group Security Considerations sections by particular characteristics - those characteristics being discovered, in part, during an initial examination of the published documents.

5.2. Quantitative Methodology

Once the set of RFCs (where the size of the set is said to be n-set) to be considered is established, the quantitative analysis proceeds as follows for each item in the set:

- recording the date of publication
- recording the source of the original draft
- recording the category of the RFC (e.g. Informational, etc.)
- recording the size of the Security Considerations section in words and paragraphs
- recording whether or not the section updates or extends the Security Considerations section of a previously published document
- record whether or not examples exist in the Security Considerations section
- record whether or not example code appears in the Security Considerations section
- extracting the text and creating a new text removing the 100 most common English words

- against the new text created in the step above, perform text analytics - for instance, create a count of the number of occurrences of expected keywords

The result would be a series of metrics for n-set that establish certain characteristics of the Security Considerations sections of published RFCs. Once the quantitative data was gathered, further analysis of the data could be conducted (for instance, finding relationships between certain features of the RFCs).

5.3. Qualitative Methodology

The documents could also be assigned qualitative characteristics as a result of the survey. For instance, based on characteristics of the document, the Security Considerations could be characterized as "extensive" or "limited."

It is also clear that analysis of the Security Considerations could lead to other groupings. For instance, an analysis of recent RFCs shows that those documents which focus on cipher suites have quite different security considerations sections compared to those that extend and existing protocol. Identification of those characteristics might be possible during an initial survey. In another case, those characteristics might emerge during the survey execution.

5.4. Implications of the Size of n-set

Since part of the execution of the survey has to be done via human intervention, the size of n-set has an effect on whether or not volunteers or organizations take on the effort. While it would be helpful to have as large a sample size as possible for the collection of data to support the analysis. It may be necessary to limit the size of n-set in practice.

One way to do this is to limit the range of dates for the RFCs being analyzed. A cursory, initial examination of Security Considerations sections seems to indicate that, in recent years, a clear set of prototypical security considerations sections has emerged and that there are distinct type of sections. By limiting the RFCs for the set of considered document to a specific, recent timeframe the goal is to focus the analysis on recent practice in crafting Security Considerations sections and moving them through the document approval process.

Another approach to solving the potential problem of the size of n-set is to incorporate a sampling regime for the selection of RFCs to

be examined. This would be a meaningful approach in the event where the timeframe was extended, but where it was still desirable to reduce the size of n-set.

This proposal suggests to use the timeframe limitation but not incorporate sampling.

6. Experimental Activity

One of the authors has conducted an experiment that is consistent with many of the features of the methodology in Section 5 above. This experiment uses a pair of Python scripts to extract the Security Considerations sections from historic RFCs and then parse those sections to get word frequency information from those Security Considerations.

The initial experiment was motivated by a desire to see if one could detect changes in Security Considerations section wording after significant security incidents in the public Internet. In particular, the experiment was designed to detect changes in the frequency of words over time.

6.1. Experiment Methodology

The RFC series was grouped into input files based on the year of publication of the RFC.

Using HTML versions of the RFC series document as input, these were put through an open source parser. The parser then identified the words "Security Consideration" or "Security" in header text. It then output that text to a temporary file in UTF-8 encoding until the parser encountered the next section.

The parser removed non-textual material from the temporary files including hyphens, RFC references, anchor URLs, other sections references, standalone letters and other characters that were not words.

It then built a frequency list for all words not in a designated list of words not to be counted. This list is a variable and could be changed to include, or exclude, words from the designated list.

6.2. Stopword List

The following list of words were used as the designated list of words not to be counted:

. Also

. Could

. Would

. However

. One

. See

. Use

. Therefore

. Discussed

. New

. March

. Type

. Even

. Following

. Without

. Bradner

. Using

. Described

. Might

. Thus

. Two

. Since

. Different

. Number

. Via

. Mechanism

. Used

. Tl

. Header

. Field

. Name

. Sent

6.3. Resulting Characterization

The result of this experiment is a pair of files for each year starting in 2003. The two files for each year are:

. A word frequency file sorted by the number of times a particular word appears in the Security Considerations section of RFCs published in that year; and,

. An RFC Count file that counts how many times each RFC was mentioned within the Security Considerations sections.

The idea behind the second file was to see if there was a trend or change in the RFCs cited and what this might suggest or say in regards to the content of these sections. For example in 2004 the highest referenced RFC was [RFC3410] Applicability Statements for SNMP in 2009 it was [RFC4301] Security Architecture for IP though [RFC3410] was also referenced a high number of times.

As [RFC4301] came out in 2005 we would not expect it to be referenced in 2004, but the reference count in 2009 could indicate that there were a number of RFCs which likely simply referred to the Security Considerations Section of this RFC in a line similar to "this extends the security consideration of <insert RFC here>." This could then be used to help narrow down qualitative focus on this highly referenced RFCs and to also see if in some cases lip service is all that is occurring within other Security Considerations Sections.

Another result, included with the word frequency file, is a list of words similar to the word "security" based on context analysis. This

is another indicator that can be used to look at how the language of the RFC series is changing. For example looking at 2004 the most similar words are:

. Used, ipsec, mode, authentication, implementation, message, may, watcher, method and block.

In 2009:

. Message, attacker, syslog, used, attack, information, transport, gruu, may and case.

Yet another result was a file that provides comparative data for word counts in the Security Considerations and Privacy Considerations sections of published RFCs. The result provides a look at whether the length of those sections might have changed over time.

A final result was a Frequency count over the entire period examined for Internet Standards, BCPs, and Proposed Standards. This result gives indication of whether or not the average length of these sections has changed - either over time, or in response to specific security incidents on the public Internet.

6.4. Indicative Results

This draft is focused on proposing a methodology and not on the experiment being reported on here. However, there are some indicative results that may be of use as a future methodology is considered. It is worth observing that the original motivation for the experiment - to see if Security Considerations sections changed in the face of security-related events on the public Internet - showed that no significant re-wording took place over the timeframe studied.

6.4.1. Top Ten Word Counts in Four Sample Years

Choosing four sample years - 2019 2014 2009 and 2004 as examples, the experiment found the following most frequent words in Security Considerations sections (the lists are in most frequent to tenth most frequent).

. 2019 - security, server, data, message, may, network, attack, information, client, xmpp-grid

. 2014 - security, information, attack, message, may, used, server, data, authentication, network

. 2009 - security, may, message, address, attack, used, packet, protocol, network, information

. 2004 - security, may, key, authentication, object, used, information, message, attack, access

6.4.2. Top Ten Word Counts Without RFC2119 Words in Four Sample Years

Taking the same data and removing the normative words that are defined in RFC2119 leads to slightly different results.

. 2019 - security, server, data, message, network, attack, information, client, xmpp-grid, document

. 2014 - security, information, message, used, server, data, authentication, network, attacker

. 2009 - security, message, address, attack, used, packet, protocol, network, information, object

. 2004 - security, key, authentication, object, used, information, message, attack, access, user

6.4.3. Normative RFC2119 Words in Security Considerations

The word MAY always appears more often than any other RFC2119 word in Security Considerations sections. The word MUST most often appears after MAY and is often in the top 15 words sorted by frequency.

However, the word SHOULD hardly ever appears in the top 100 most frequent words for any year of published RFCs.

Most Frequent Words in Proposed Standards Security Considerations

Over the entire period 2003-2019, the most frequent non-normative words in Security Considerations sections was:

. Security, message, attack, server, information, key, authentication, network, protocol, client

A list of the 75 most commonly, non-normative words is provided in Appendix B.

7. Security Considerations

This document describes goals and a methodology for surveying the existing body of Security Considerations in published RFCs. It does not create, extend or modify any protocols. Its intent is to provide a foundation for a data-driven discussion of the guidelines for writing a Security Considerations section in an RFC.

8. IANA Considerations

Upon publication, this document has no required actions for IANA.

9. References

9.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC2223] Postel J. and Reynolds J., ISI, "Instructions to RFC Authors", RFC2223, October 1997.

[RFC3552] Rescorla E. and Korver B.(Editors), "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC3552, July 2003.

[RFC7687] Farrell S., Wenning R., Bos B., Blanchet M. and Tschofenig H., "Report from the Strengthening the Internet (STRINT) Workshop, RFC 7687, December 2015

9.2. Informative References

[1] Model-t -- Discussions of changes in Internet deployment patterns and their impact on the Internet threat model, https://www.ietf.org/mailman/listinfo/model-t

[2] Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Appendix A. Document History

[[To be removed from the final document]]

-00

Initial Internet Draft

-01

Section 6 and Appendix B are added. Significant editing of Section 3 on Motivation and Section 5 on Methodology. Several typos fixed.

Appendix B. 75 Most Common Words in Security Considerations Sections

Over the entire period 2003-2019, the 75 most frequent words in Security Considerations sections was (in order by frequency):

security, message, attack, data, used, may, authentication, key, access, protocol, information, must, address, transport, process, model, client, server, network, ipfix, tl, user, traffic, packet, object, operation, control, service, ipp, example, document, implementation, measurement, collecting, secure, header, attacker, identity, value, job, need, support, snmp, provide, printer, uri, certificate, authenticated, possible, name, content, source, connection, field, set, system, dtls, cause, sensitive, domain, provides, configuration, router, privacy, protection, peer, nacm, layer, ip, device, exporting, within, request, large, and signature.

Authors' Addresses

Mark McFadden
Internet policy advisors ltd
Madison Wisconsin US

Email: mark@internetpolicyadvisors.com

Alan Mills
University of the West of England, Bristol
Bristol BS16 1QY United Kingdom

Email: Alan2.Mills@live.uwe.ac.uk