MALWARE THREATS IN DIGITAL PRESERVATION

Extending the evidence base

Maureen Pennock
The British Library
United Kingdom
maureen.pennock@bl.uk
https://orcid.org/0000-0002-7521-8536

Michael Day
The British Library
United Kingdom
michael.day@bl.uk
https://orcid.org/0000-0002-1443-5334

Evanthia Samaras
University of Technology Sydney, Australia
evanthia.samaras@student.uts.edu.au
https://orcid.org/0000-0001-5683-722X

Abstract – Virus checking is an established process in most pre-ingest digital preservation workflows. It is typically included as part of a general threat model response and there has to date been relatively little research into the virus checking function specifically within a long term context. The British Library recently began a small research project to explore this issue, using data from a legacy digital collection established by the 'Flashback' project and supplementary data provided by the UK Web Archive. Our poster presents this research and findings to date, raising questions about the overhead of virus checking at scale, when organizations should virus-check content, and the legacy capabilities of anti-virus software.

Keywords – digital preservation, malware, Flashback, virus checking

Conference Topics – Exploring New Horizons; The Cutting Edge – Technical Infrastructure & Implementation.

I. INTRODUCTION

Memory organizations typically conduct malware checks on collections as a pre-ingest step in digital preservation workflows, so that 'infected' materials can be identified and dealt with before they would be able to inflict damage on users' computers or the organization's network. Such a step is widely considered to be best practice; however, it is generally employed in response to a general threat model and there is less clarity about the actuality of malware risks over the longer-term, particularly with legacy digital content that is rarely accessed. Many institutions thus incur virus-checking overheads during ingest processing, when it is less than clear exactly how malware risks might manifest. Malware is, effectively, something of an underexplored question in evidence-based digital preservation practices.

Acknowledging this, in late 2018 the British Library initiated a small research project that focused specifically on the virus-checking function within digital preservation and processing workflows, including a three month PhD research placement. This poster introduces that work, provides a high level overview of malware in the Library's legacy 'Flashback' corpus and in the UK Web Archive, and explores some of the research questions that the project seeks to address.

II. WHAT IS MALWARE?

Malware is a catch-all term that refers to various types of 'malicious software,' including viruses, worms, Trojan horses, spyware and ransomware. It is often designed to exploit specific computing environments or applications. Viruses and other types of malware can not only disrupt or damage these environments, they may also provide a means for unauthorized access to computer environments and/or information. According to Farbowitz [1], the "two characteristics common to most malware are that it operates without the consent of the computer user or network administrator, and that much of it self-replicates in order to spread."

Malware and viruses can be identified and removed by anti-virus (AV) software. These use multiple techniques to counter the threat of malware. For example, AV software providers identify and collect 'signatures' that correlate to specific instances of malware. The signature databases for the products grow over time and are regularly updated with new definition files. AV software also uses heuristic analysis to identify malware, examining code for suspicious properties.

III. MALWARE AT THE BRITISH LIBRARY

While the British Library does not actively collect malware as collection items, malware is known to exist as part of other items acquired by the Library for preservation. This research is based on malware

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

DOI: 10.1145/nnnnnnn.nnnnnnn

iPRES 2019

identified in two key datasets: the Flashback corpus and the UK Web Archive.

The Flashback corpus consists of disk images of legacy digital content that was originally acquired by the British Library on handheld media such as floppy disk or CD. The corpus dates from between approximately 1980 to 2010 [2]. The UK Web Archive is a more recent dataset, comprised of millions of UK-relevant websites harvested annually under the UK Non-Print Legal Deposit Regulations (2013). Collectively these two sources provide a sample of both legacy and contemporary malware, delivered and acquired through common malware distribution mechanisms for their respective periods.

The Flashback corpus has been the main focus of the analysis to date. At the time that this analysis was undertaken (February 2019), a total of 22,061 disks had been imaged, roughly half of which were CD-ROM or CD-R, just under a third were 3.5" floppy disks, and the remainder were either 5.25" floppies or Digital Versatile Discs (DVDs). All of the disks were virus-checked prior to being imaged. Of these, 120 of the discs were identified as containing some kind of malware. While this would inevitably include an unquantified number of false-positives, this means that 0.54% of the Flashback disk images have so far been classified as containing malware.

The majority of malware-positive files identified within the Flashback corpus to date have been found on disks previously attached to consumer computer magazines, mainly CD-ROMs and DVD-ROMs published during the 2000s, although some have also been located on non-magazine-based items. Not all of these files, however, represent the same level of risk. For example, three of the magazine cover disks contained joke programs, which seem to have been intentionally distributed with the disks. However, a further item, a disk acquired with PC Gamer No. 58 (July 1998, Disk B), was found to contain three files infected with the Marburg virus, designed to run on Windows 95/98. The visual effect of the Marburg virus is repeated instances of the Windows error icon overlaid on screen content, though it was also found to delete integrity bases of known AV products [3]. In addition, a large number of files contained macro viruses written for MS Word, although there are also many examples of files identified as Trojans and, at the other extreme, a file containing a fragment of a boot-sector virus that was apparently frequently encountered 'in the wild' in the floppy disk era.

Rates of malware in the UK Web Archive are calculated differently from those in the Flashback corpus. The Flashback concept of an 'item' does not transfer to the UK Web Archive, as the boundaries of any given website are fluid, due to the inherent 'linking' system upon which the web is reliant. Malware data for the UK Web Archive is thus currently only available as a percentage of the overall size of the data. We refer to this as the percentage of the stored data flagged as malware-infected, rather than the collection. Up until the end of 2017, malware in the UK Web Archive accounted for less than 0.01% of the stored data.

IV. RESEARCH QUESTIONS

The percentage of stored data from the UK Web Archive is significantly lower than the percentage of malware infected items in the Flashback corpus. That may be explained by the difference in measurement processes. Further work is planned in coming months to address this, as is further identification of the various malware and their function. Nonetheless, the relatively high percentage of malware-flagged items found in the Flashback corpus indicates that virus checking remains necessary. Given the inevitable processing overheads of virus checking at scale, and considering that access rates for content are relatively low, we intend to explore whether it might be more efficient to virus check primarily at the point of access, rather than at ingest. To answer this question we need a better understanding of at least two things: a) the threat model for storing malware in a long-term digital repository; and, b) the performance of virus checking software over time.

Threat-model analysis will consider, for example, the functionality of the malware found in our samples and the likelihood of malware activation within a controlled read-only storage environment. Regarding performance of AV software over time, we expect that the capabilities of AV software to identify malware should improve as it encounters more samples: this would suggest that delaying the virus-scanning process may result in a more comprehensive output (including a reduction in the number of 'false positives'). On the other hand there is a gap in the literature to definitively confirm that legacy signatures relating to obsolete malware are never removed from providers' databases: this may result in a limit as to how long it is viable to delay the virus checking function. Additional evidence is needed either way.

The poster will reflect up-to-date findings for these research questions and will prompt valuable debate and discussion with conference attendees to help drive this currently under-researched area forwards.

REFERENCES

[1] J. Farbowitz, *More than digital dirt: preserving malware in archives, museums, and libraries*. MA Thesis, New York University, 2016, https://archive.org/details/16sThesisFarbowitzFinal

[2] M. Day, et al., "The preservation of disk-based content at the British Library: Lessons from the Flashback project," *Alexandria: The Journal of National and International Library and Information Issues*, vol. 26, no. 3, pp. 216-234, 2016, https://doi.org/10.1177/0955749016669775

[3] F-Secure Knowledge Base: Marburg, https://www.f-secure.com/v-descs/marburg.shtml