Lab 8

Meta Data and theHarvestor

Group 2

, Abdul-Ramon Adeyemi Folarin, &

, , , &

School of Information Technology Administration & Security, Seneca Polytechnic

DPI914NIA: Social Engineering

Professors Chris Robertson & Suranjit Paul

July 22, 2026

Student Integrity Statement: I/we declare that the attached assignment is my/our own work in accordance with the Seneca Academic Policy. No part of this assignment has been copied manually or electronically from any other source (including web sites) or distributed to other students

Table of Contents

Introduction....4
Objectives....5
Meta Data Tools: Online....5
Word Document....6
Spreadsheet Data File....7
JPEG Image....7
Meta Data Tools: Desktop....9
Word Document....9
Spreadsheet Data File....9
JPEG Image....10
Meta Data Tools: Meta Data Collector....10
Running Metagoofil....11
Sourcing the Files Manually....12
Extracting the Metadata....12
theHarvestor....13
Hackertarget Source....14
Urlscan Source....14
Yahoo Source....15
Conclusion....16
References....17
Group Member Contributions....18

Table of Figures

Figure 1 metadata2go extraction of the Word document, showing file and author properties 6
Figure 2 metadata2go extraction of the Word document, showing revision history and software details 6
Figure 3 metadata2go extraction of the spreadsheet data file 7
Figure 4 metadata2go extraction of the JPEG image, showing the colour profile and encoder comment 7
Figure 5 metadata2go extraction of the JPEG image, showing image dimensions and remaining profile data 8
Figure 6 ExifTool extraction of the Word document from the command line 9
Figure 7 ExifTool extraction of the spreadsheet data file 9
Figure 8 ExifTool extraction of the JPEG image 10
Figure 9 Metagoofil run against Seneca Polytechnic across seven document types 11
Figure 10 Manual search-engine query for PDF files on senecapolytechnic.ca 11
Figure 11 Manual download of three public Seneca Polytechnic PDF documents 12
Figure 12 ExifTool extraction of the three downloaded Seneca PDF documents 12
Figure 13 theHarvester hackertarget search returning 80 hosts for Seneca Polytechnic 14
Figure 14 theHarvester urlscan search returning hosts, IP ranges, and indexed URLs 14
Figure 15 theHarvester yahoo search returning 26 email addresses and 30 hosts 15

Table of Tables

Table 1 Comparison of the metadata exposed by each file type using the online tool.....8
Table 2 The same files through the online tool versus the desktop tool.....10
Table 3 Metadata recovered from the three downloaded Seneca PDFs 13
Table 4 theHarvester sources and what each search returned 15

Introduction

Metadata is data about data. It is the information a file quietly stores about itself, such as who created it, what software produced it, and when it was last changed, and it usually travels with the file without the owner paying it any attention. For a social engineer, that hidden layer is valuable, because it can reveal real names, internal usernames, software versions, and sometimes even a physical location, all of which help during the information gathering stage of an attack.

In this lab we looked at three different ways of pulling metadata out of files, and then at a tool that automates the collection of files, and therefore their metadata, from public sources. We worked through an online extractor at metadata2go.com, the ExifTool desktop utility, the Metagoofil document collector, and theHarvester, which gathers emails, hosts, and subdomains from public databases. For each tool we recorded what it found, either against sample files or against Seneca Polytechnic, and we explain why that information would matter to an attacker. All work was carried out in our own Kali environment against publicly available data, as the lab directed.

Objectives

The aims of this lab were to:

- Understand why the metadata held inside documents is useful.
- Extract metadata using a number of different tools.
- Understand what information harvesters do and why they are valuable.
- Demonstrate the use of the tool theHarvester.

Meta Data Tools: Online

This task asked us to run an online metadata extractor against three different file types and pick out any metadata that could be useful for social engineering. We used metadata2go.com [2] with a Word document from an earlier lab, a spreadsheet data file, and a JPEG photo, avoiding anything sensitive since these tools upload the file to a third-party server. The figures below show what each file exposed.

Word Document

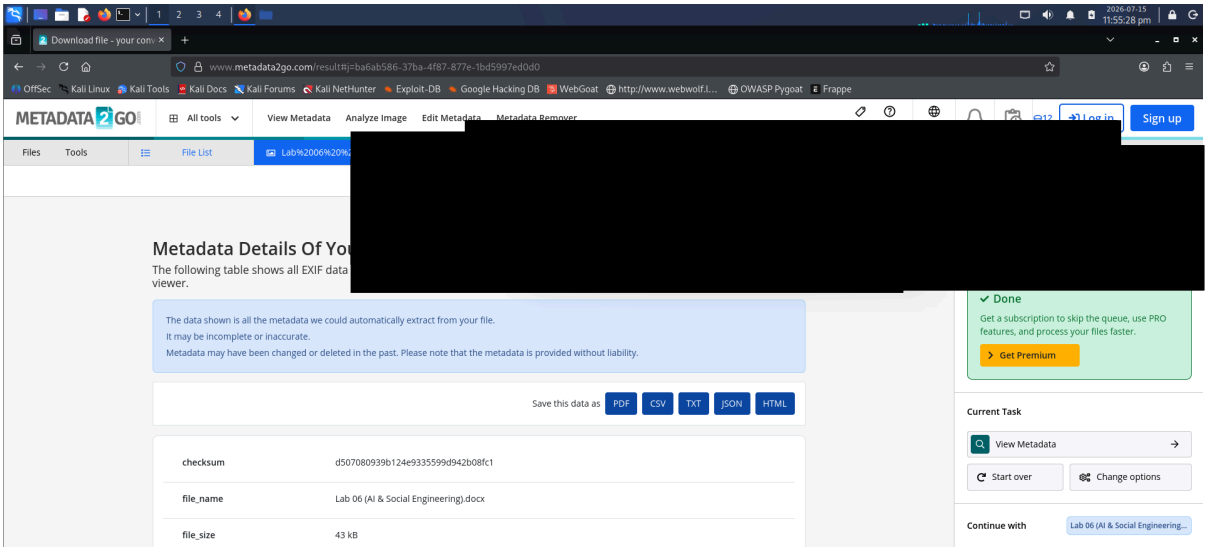


Figure 1 metadata2go extraction of the Word document, showing file and author properties

Figure 1 shows metadata2go reading the Lab 06 DOCX and returning its identity fields. The Creator and Last Modified By values both hold the author's real name, the file is a 43 kB Word document, and it was originally created on 9 February 2023. A real author name ( ) attached to a document is exactly what a social engineer needs to begin a target profile.

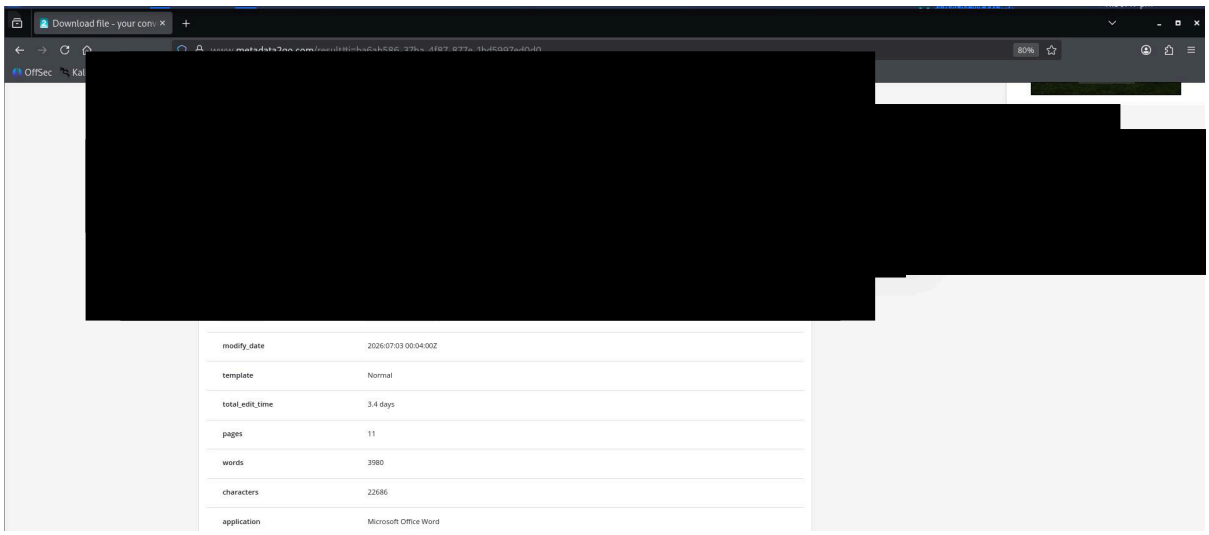


Figure 2 metadata2go extraction of the Word document, showing revision history and software details

Figure 2 the extended properties show the file was saved 209 times, took around 3.4 days of total editing, and was last written in Microsoft Office Word. A high revision count signals an actively used document, and the exact application version tells an attacker which Office build to aim a malicious macro or tailored lure at. The second part of the same extraction shows the document's history: a Revision Number of 209, a Total Edit Time of 3.4 days, a last modified date of 3 July 2026, and Application set to Microsoft Office Word (version 16). The high revision count

marks this as an actively worked file, and the exact Office version tells an attacker which build to target with a malicious macro or a tailore lure.

Spreadsheet Data File

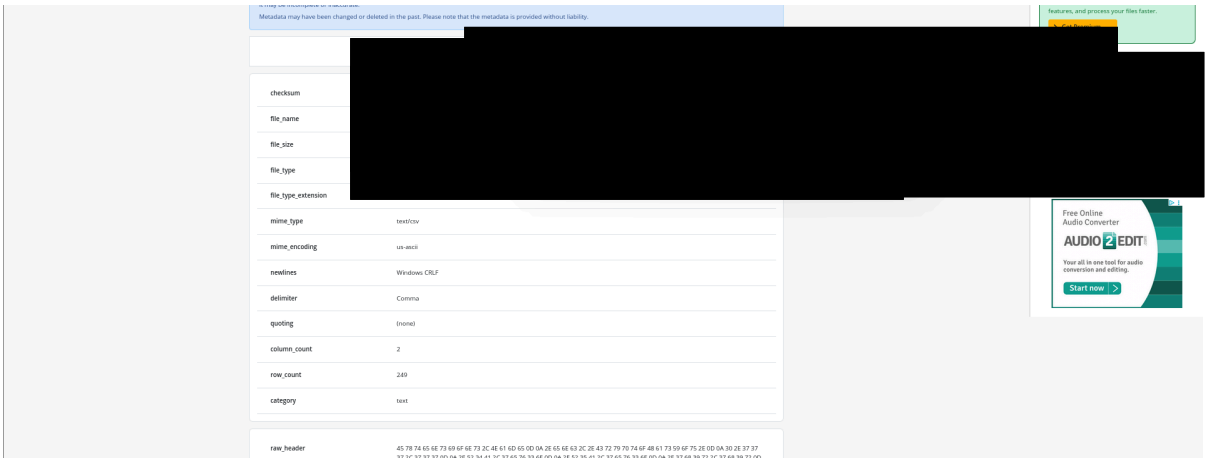


Figure 3 metadata2go extraction of the spreadsheet data file

Figure 3 shows metadata2go reading the ransomware_extensions file, which resolved as a plain CSV rather than a true spreadsheet. The only metadata returned is structural: US-ASCII encoding, Windows CRLF line endings, a comma delimiter, and a count of 2 columns by 249 rows. There are no author, software, or date fields, which shows that flat text files give a social engineer almost nothing to work with.

JPEG Image

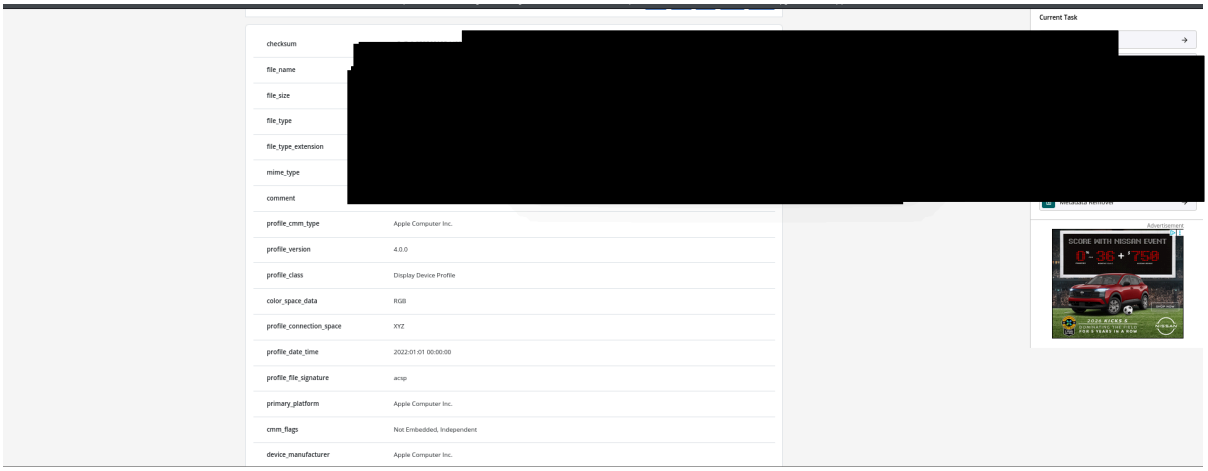


Figure 4 metadata2go extraction of the JPEG image, showing the colour profile and encoder comment

Figure 4 shows metadata2go reading IMG_2332.jpg. The Comment field records that the image was re-encoded by Discord using the jpegli encoder, and the colour profile is an Apple Display P3 profile from 2022. Because the file passed through Discord, its original camera EXIF has been stripped, which is why there is no device model or GPS to be seen.

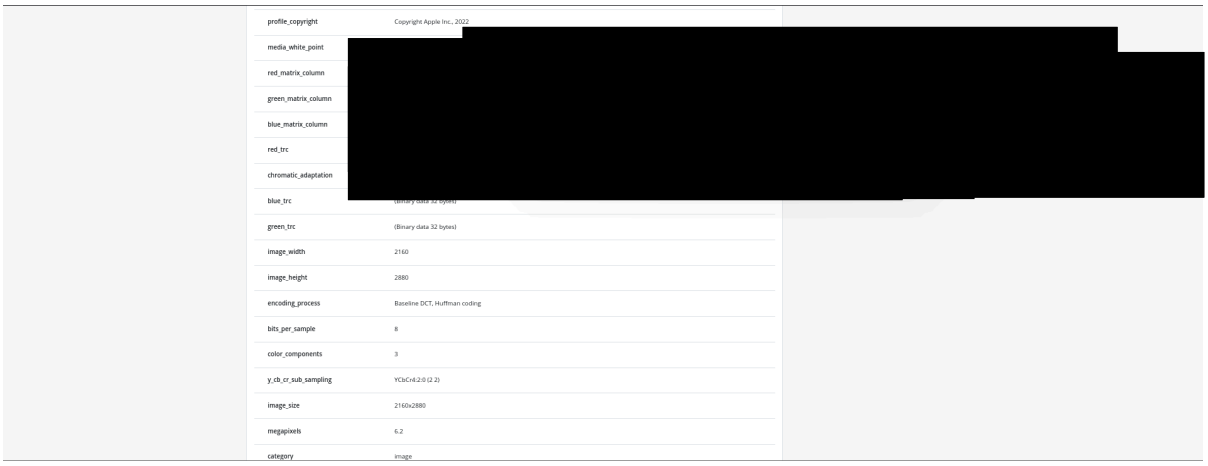


Figure 5 metadata2go extraction of the JPEG image, showing image dimensions and remaining profile data

Figure 5 shows the second part of the JPEG extraction confirms what is left after the platform stripped it: the ICC colour profile details and an image size of 2160 by 2880 pixels (6.2 megapixels), with no GPS coordinates, capture time, or camera model. It demonstrates that although a photo can carry sensitive location data, a version shared through a chat app usually does not.

Across the three files, the Word document leaked the most: a real author name, a heavy revision history, and the exact Office version. The spreadsheet, being a flat CSV, gave up almost nothing, and the JPEG had already had its camera data stripped by Discord. Table 1 sets the three side by side, and the pattern for a social engineer is clear: Office documents are the rich target, while plain data files and platform-shared images are usually not worth the effort.

Table 1 Comparison of the metadata exposed by each file type using the online tool

File (type)	**Key metadata returned**	**Value to a social engineer**
Lab 06 (DOCX)	Author name (Creator and Last Modified By), Revision 209, Total Edit Time 3.4 days, Microsoft Office Word 16, created 2023, modified 2026	Real author name for a target list; an active document; exact Office version for a tailored macro or lure
ransomware_extensions (CSV)	File type CSV, US-ASCII, Windows CRLF, comma delimiter, 2 columns, 249 rows	No author or software data; flat files leak little of use
IMG_2332 (JPEG)	Discord re-encode comment (jpegli), Apple Display P3 profile, 2160x2880, no GPS, no device model	Camera EXIF stripped by the platform; shared images often carry little

Meta Data Tools: Desktop

This task asked us to install ExifTool [1] on Kali, gather the same three files into a folder called Evidence, and extract their metadata from the command line, with the command visible in each capture. ExifTool reads files locally, so nothing is uploaded anywhere. We ran it against each file in turn.

Word Document

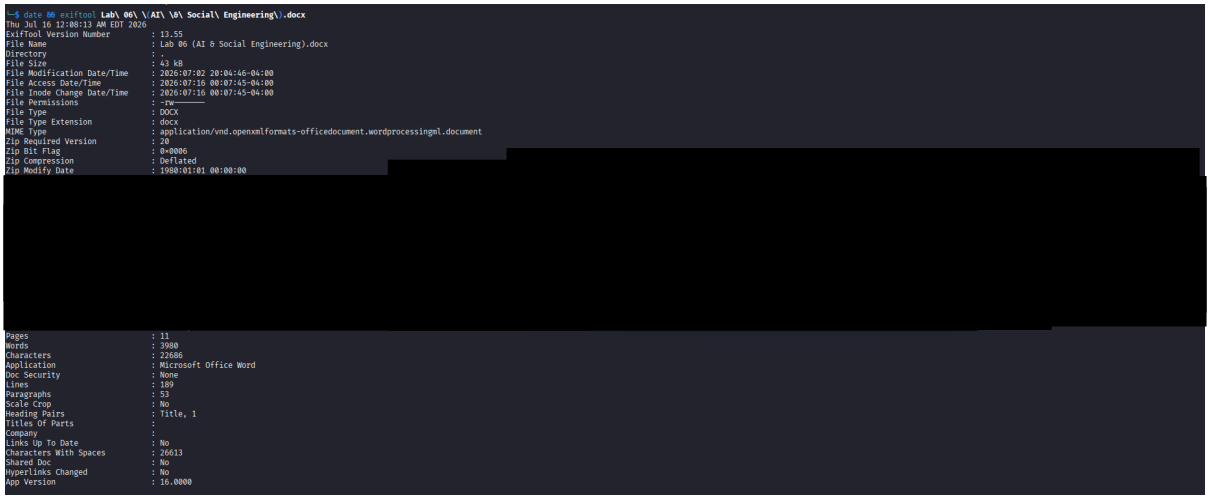


Figure 6 ExifTool extraction of the Word document from the command line

Figure 6 shows ExifTool (version 13.55) run from the Kali command line against the same Lab 06 DOCX. It confirms the Creator and Last Modified By author name, Revision Number 209, Total Edit Time 3.4 days, and Application Microsoft Office Word, matching the online result. The point for an attacker is that all of this personal and software detail is available locally, without ever uploading the file to a third-party site.

Spreadsheet Data File



Figure 7 ExifTool extraction of the spreadsheet data file

Figure 7 shows ExifTool reading the CSV. As with the online tool, it returns only File Type CSV, US-ASCII encoding, a comma delimiter, and 249 rows across 2 columns, with no owner or software fields. It reinforces that plain data files leak nothing useful about who created them.

JPEG Image



Figure 8 ExifTool extraction of the JPEG image

Figure 8 shows ExifTool reading the JPEG. It surfaces the Discord re-encode comment, the Apple Display P3 profile, and the 2160 by 2880 dimensions, but the Device Model field is empty and there is no GPS block. This confirms the identifying camera data was removed before the image reached us.

ExifTool returned the same core information as the online tool and, in places, more raw detail such as the full colour profile block. Table 2 lines the two tools up on the same three files. The value for an attacker is getting the same author and software metadata without ever handing the file to an outside service, which matters when the document is confidential.

Table 2 The same files through the online tool versus the desktop tool

File	metadata2go (online)	ExifTool (desktop)
Lab 06 (DOCX)	Author, revision, edit time, Office version	Same fields, plus more raw detail; runs locally with no upload
ransomware_extensions (CSV)	Structure only (delimiter, rows)	Structure only, identical result
IMG_2332 (JPEG)	Colour profile, dimensions, Discord comment	Same, plus the full ICC profile block

Meta Data Tools: Meta Data Collector

This task asked us to run Metagoofil against Seneca Polytechnic, report what we found, and note anything of interest to a social engineer. Metagoofil is meant to search a domain, download the documents it finds, and expose their metadata in bulk. The version in Kali no longer reads metadata itself [3]; it only downloads files, so ExifTool still does the reading. We also tested the maintained community fork [4] to be thorough, and it returned the same empty result.

Running Metagoofil

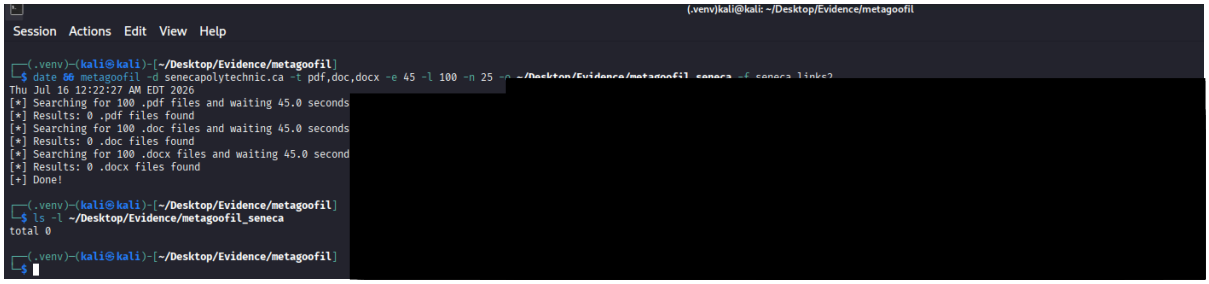


Figure 9 Metagoofil run against Seneca Polytechnic across seven document types

Figure 9 shows Metagoofil running against senecapolytechnic.ca for seven file types (pdf, doc, docx, xls, xlsx, ppt, pptx), and every search returns 0 files found. The tool relies on scraping Google, which now blocks that automated activity, so the failure is in Metagoofil's discovery stage and not evidence that Seneca lacks public documents.

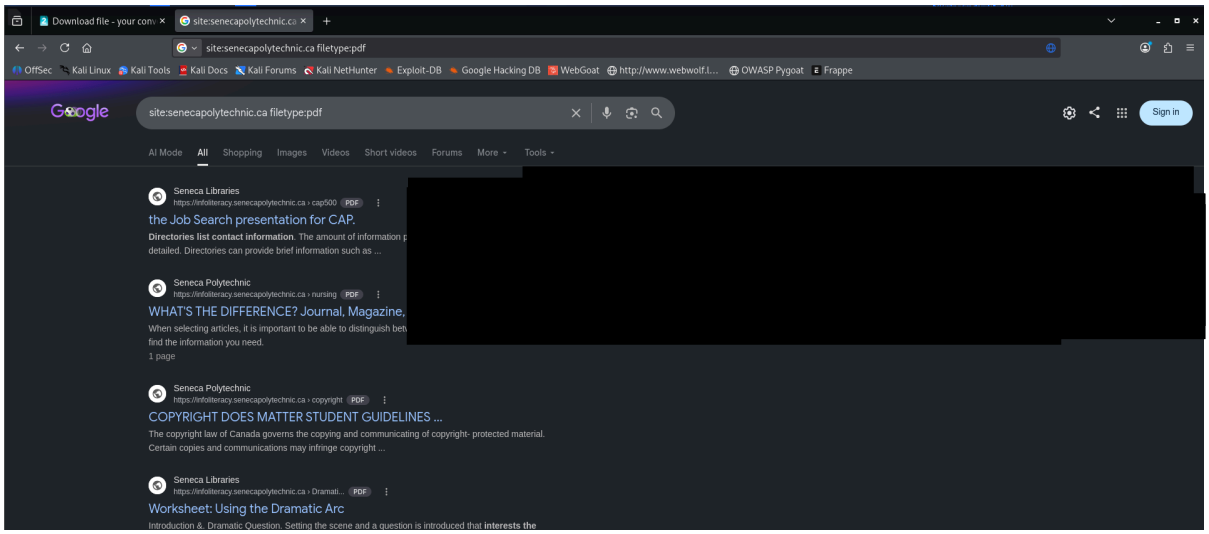


Figure 10 Manual search-engine query for PDF files on senecapolytechnic.ca

Figure 10 shows a normal browser search using the operator site:senecapolytechnic.ca filetype:pdf, which returns numerous public PDFs. It proves the documents exist and are indexed, confirming that Metagoofil's empty result was a limitation of the tool rather than a lack of available files.

Sourcing the Files Manually

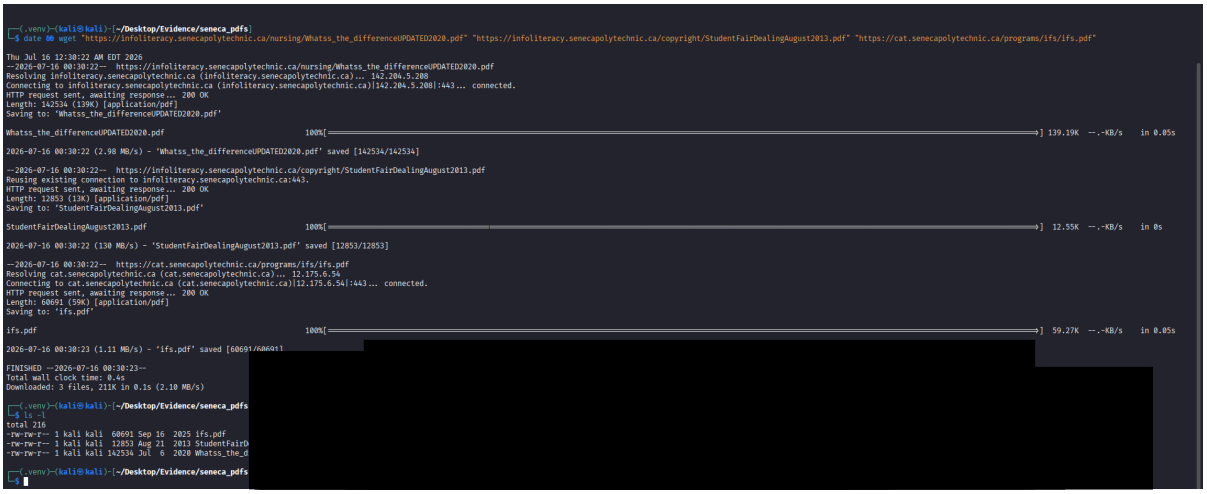


Figure 11 Manual download of three public Seneca Polytechnic PDF documents

Figure 11 shows three public Seneca PDFs being downloaded directly with wget after they were located manually, each returning HTTP 200 OK. Since the automated collector failed, this manual step recreates what Metagoofil was supposed to do, which is the same fallback a social engineer would use when their tooling is blocked.

Extracting the Metadata



Figure 12 ExifTool extraction of the three downloaded Seneca PDF documents

Figure 12 shows ExifTool reading the three downloaded PDFs, and this is where the useful metadata appears. ifs.pdf lists Author and Creator as CourseLeaf with Producer Apache FOP and the title Honours Bachelor of Information Technology Cybersecurity; StudentFairDealingAugust2013.pdf lists Author Wanda, built with Acrobat PDFMaker for Word; and Whatss_the_difference lists Author win7user, produced by Microsoft Word for Office 365. A real first name, a Windows username, and the exact software stack all feed a target list and a picture of the institution's tooling.

So, what did we find? The tool itself found nothing, because Google now blocks the automated searching that both the Kali package and the community fork rely on. That is a useful result on its own: bulk document harvesting against this target through Metagoofil is currently unreliable, and an attacker would have to fall back to manual collection or a different search backend. Once we sourced three public PDFs by hand, though, the metadata was genuinely valuable, as Table 3 summarises. It exposed a real author first name, a Windows account username, and the software behind each document, all of which feed a target list and build a picture of the institution's tooling for a social engineer.

Table 3 Metadata recovered from the three downloaded Seneca PDFs

PDF	Author / Creator	Producing software and date	Value to a social engineer
ifs.pdf	CourseLeaf	Apache FOP 2.1, CourseLeaf, 2025	Reveals the CourseLeaf curriculum system; the title exposes a program name
StudentFairDealing (2013)	Wanda	Acrobat PDFMaker 10.1 for Word, Adobe PDF Library 10.0, 2013	A real staff first name; document built in Word
Whatss_the_difference (2020)	win7user	Microsoft Word for Office 365, 2020	A Windows account username; confirms Office 365 in use

theHarvestor

This task asked us to run three different searches against Seneca Polytechnic with theHarvester [5], report what we found, and highlight anything of interest to a social engineer. theHarvester gathers emails, subdomains, hosts, and related data from public sources, and we deliberately chose three different sources so that each search exposed a different kind of data.

Hackertarget Source

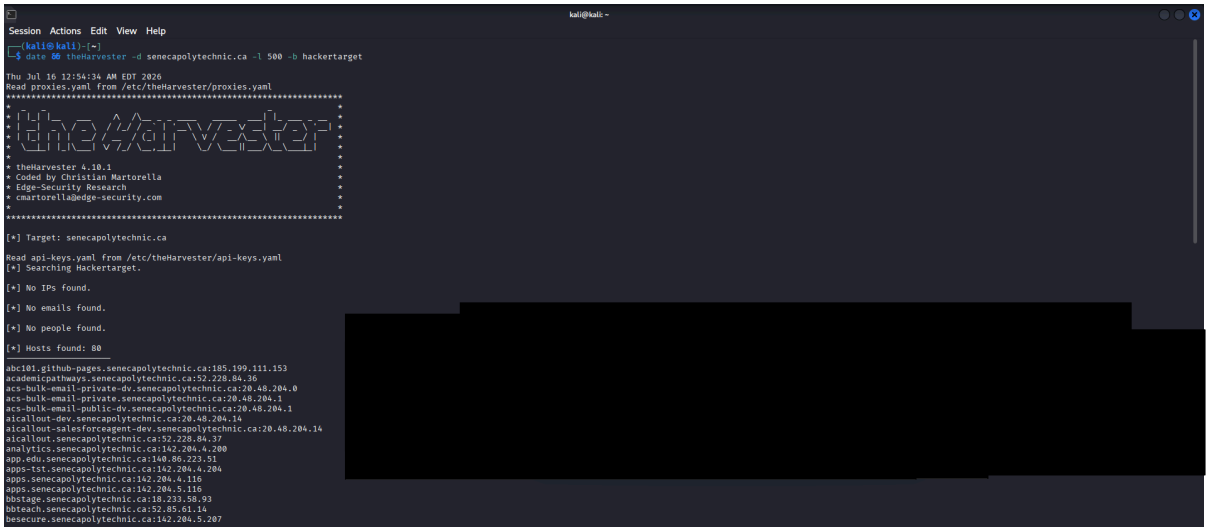


Figure 13 theHarvester hackertarget search returning 80 hosts for Seneca Polytechnic

Figure 13 shows theHarvester using the hackertarget source return 80 hosts with their IP addresses. The list names internal systems directly, including hrerp and financeerp, employeeservicehub, formeremployeesportal, the VPN gateways senecavpnnhegw1 and studentvpnnhegw1, and the mail servers mail1-nh and mail1-sy. Named systems like these let a social engineer build a pretext around a specific service, such as a fake HR ERP password reset.

Urlscan Source

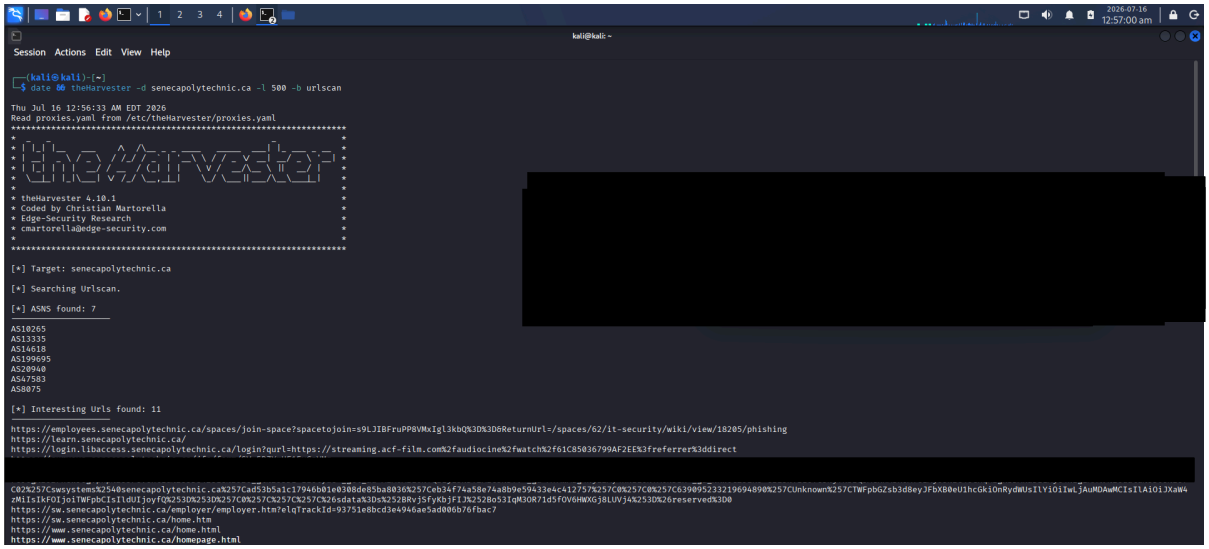


Figure 14 theHarvester urlscan search returning hosts, IP ranges, and indexed URLs

Figure 14 shows theHarvester using the urlscan source return 5 hosts, 27 IPs, 7 ASNs, and 11 indexed URLs. Two are notable: a link to an internal IT-security wiki page whose path ends in /phishing, and an email-campaign tracking link on sw.senecapolytechnic.ca that exposes how the institution's mass emails are structured. Both give an attacker material to imitate a legitimate internal message.

Yahoo Source

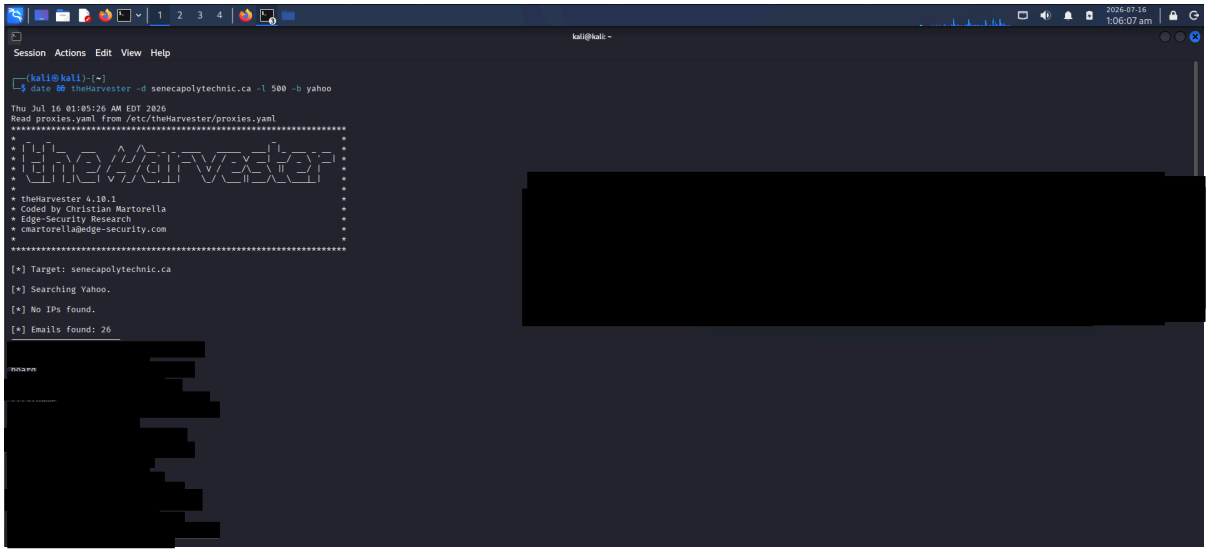


Figure 15 theHarvester yahoo search returning 26 email addresses and 30 hosts

Figure 15 shows theHarvester using the yahoo source return 26 email addresses and 30 hosts. The addresses confirm a firstname.lastname@senecapolytechnic.ca format through examples such as [redacted] [redacted] and [redacted] alongside role accounts like president, security, servicedesk, itso, and privacyoffice. Once the naming pattern is known, an attacker can construct almost any employee's address from their name.

This was the most productive tool of the lab, and there was plenty of interest to a social engineer, as Table 4 sets out. We recovered a large map of Seneca's external systems, including HR and finance ERPs, VPN gateways, mail servers, and staging environments, all named clearly enough to reference in a convincing pretext. We also recovered 26 email addresses that confirm the firstname.lastname@senecapolytechnic.ca format and mix named individuals with role accounts, which means an attacker could work out almost any employee's address. urlscan even surfaced an internal wiki page about phishing and an email-campaign tracking link, which together show how the institution's legitimate emails are built.

Table 4 theHarvester sources and what each search returned

Source	Data returned	Notable findings for a social engineer
Hackertarget	80 hosts with IPs	hrerp, financeerp, employeeservicehub, formeremployeesportal, VPN gateways (senecavpnnhegw1, studentvpnnhegw1), mail servers (mail1-nh, mail1-sy)
Urlscan	5 hosts, 27 IPs, 7 ASNs, 11 URLs	Internal IT-security wiki page ending in /phishing; email-campaign tracking link on sw.senecapolytechnic.ca
Yahoo	26 emails, 30 hosts	firstname.lastname format [redacted] [redacted] [redacted] role accounts (president, security, servicedesk, itso, privacyoffice)

Conclusion

In conclusion, across the four tools, the clearest lesson is how much useful information leaks out of ordinary files and public records. Office documents gave up author names, revision history, and exact software versions, while flat data files and platform-processed images gave up very little, which tells an attacker where the easy wins are. Metagoofil failed at its search stage, but the manual workaround still produced real names, a username, and a software profile from Seneca's own PDFs. theHarvester was the strongest of the set, handing over the email naming convention, individual staff addresses, and a detailed map of internal systems from nothing more than public sources. All of this sits in the reconnaissance stage of the attack cycle, and it shows why reducing metadata exposure and limiting what is publicly indexed is a real part of defending against social engineering.

References

[1] P. Harvey, "ExifTool by Phil Harvey," exiftool.org. [Online]. Available: https://exiftool.org/. [Accessed: Jul. 16, 2026].

[2] Metadata2Go, "Free online metadata viewer," metadata2go.com. [Online]. Available: https://www.metadata2go.com/. [Accessed: Jul. 16, 2026].

[3] Kali Linux, "metagoofil," Kali Linux Tools. [Online]. Available: https://www.kali.org/tools/metagoofil/. [Accessed: Jul. 16, 2026].

[4] opsdisk, "metagoofil: Search Google and download specific file types," GitHub. [Online]. Available: https://github.com/opsdisk/metagoofil. [Accessed: Jul. 16, 2026].

[5] Kali Linux, "theHarvester," Kali Linux Tools. [Online]. Available: https://www.kali.org/tools/theharvester/. [Accessed: Jul. 16, 2026].

Group Member Contributions

The work on this lab was divided as follows. completed the introduction, the online metadata task (Meta Data Tools: Online), the Metagoofil task (Meta Data Tools: Meta Data Collector), the conclusion, and the overall compilation and formatting of the report. Abdul-Ramon Adeyemi Folarin completed the desktop task (Meta Data Tools: Desktop). completed the hackertarget and urlscan portions of the theHarvestor task. completed the yahoo portion of the theHarvestor task and the references. (Since it was a short lab, everyone did a run through and did the lab on their own as well and we picked the best results/screenshots for the final report.