UNINTENDED INFORMATION LEAKAGE THROUGH METADATA

(AT THE DUTCH GOVERNMENT)

BY AUKÉ ZWAAN
INTENTIONS

1. Doing science
2. Sincerely worried about the current state of security
3. Raising awareness
4. Having fun :-)
WHAT’S NOT THE GOAL?

HARMING ANY OF THE ORGANIZATIONS RESEARCHED
DISCLAIMER

For clarity, Information in this presentation can be:
- Redacted
- Added
- Modified
WHAT IS METADATA?
WHAT IS METADATA?

“Data about data”
ALBUM

The Stranger (30th Anniversary Legacy Edition)

By Billy Joel
1977 • 21 songs, 1 hr 46 min

<table>
<thead>
<tr>
<th>#</th>
<th>TITLE</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Movin’ Out (Anthony’s Song)</td>
<td>3:31</td>
</tr>
<tr>
<td>2</td>
<td>The Stranger</td>
<td>5:11</td>
</tr>
<tr>
<td>3</td>
<td>Just The Way You Are</td>
<td>4:51</td>
</tr>
</tbody>
</table>
**EXAMPLE: MUSIC**

**ALBUM**

**The Stranger (30th Anniversary Legacy Edition)**

*By Billy Joel*

1977 | 21 songs | 1 hr 46 min

<table>
<thead>
<tr>
<th>#</th>
<th>TITLE</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Movin' Out (Anthony's Song)</td>
<td>3:31</td>
</tr>
<tr>
<td>2</td>
<td>The Stranger</td>
<td>5:11</td>
</tr>
<tr>
<td>3</td>
<td>Just The Way You Are</td>
<td>4:51</td>
</tr>
</tbody>
</table>
RESEARCH QUESTION

WHAT TYPES OF METADATA DOES THE DUTCH GOVERNMENT LEAK THROUGH PUBLIC DOCUMENTS,

...AND

HOW COULD THIS INFORMATION BE USED IN THE RECONNAISSANCE PHASE OF A CYBER ATTACK?
WHAT IS “THE DUTCH GOVERNMENT”?

Open State Foundation: https://pulse.openstate.eu/https/domains/

• 1650 domains
STEPS IN THE METADATA ANALYSIS

1. Harvest URLs of public documents from the domains
2. Download the documents
3. Parse the metadata
4. Analyze and visualize the metadata
HARVESTING DATA

For analysis, all (or at least a reasonable number of) public documents are required. Where are they?
HARVESTING DATA

For analysis, all (or at least a reasonable number of) public documents are required. Where are they?

...let's use Google!
Ongeveer 39.100 resultaten (0,41 seconden)

[PDF] Reactienota Notitie Reikwijdte en Detailniveau Gebiedsontwikkeling ...
www.brabant.nl/reactienotatagol ▼

[PDF] Onderzoek borging externe veiligheid in Brabantse ...
www.brabant.nl/EVinbestemmingsplannen ▼

[PDF] Weerijs.psd
www.brabant.nl/vondstenweerijs ▼
Page 1. | | | | - | | | . Sporen van de eerste Rijsbergenmaren aangetroffen. | | | | - - - | | | tijdens herinrichting Weerijs. | | | | | | | | De bodem kan worden ...

[PDF] De geluidbelasting rondom de vliegbasis Woensdrecht voor het jaar ...
www.brabant.nl/~media/A1F19D0E9CDC42A9ACEE44A3F108A157.pdf
Ik ben geen robot

Verzenden

Over deze pagina

Onze systemen hebben ongebruikelijk verkeer van uw computernetwerk vastgesteld. Op deze pagina wordt gecontroleerd of de verzoeken daadwerkelijk door u worden verzonden en niet door een robot. Waarom gebeurt dit?

IP-adres: 145.100.103.43
Tijd: 2017-06-29T08:45:46Z
URL: https://www.google.nl/search?q=hack+the+planet!+hack+the+planet!&oq=hack+the+planet!+hack+the+planet!&aqs=chrome..69i57j69i60j0l4.2056j0j7&sourceid=chrome&ie=UTF-8
ACTIVE VS. PASSIVE SCANNING

PASSIVE SCANNING: using Google/Bing "dorks" and getting the file locations "passively"
ACTIVE VS. PASSIVE SCANNING

**Passive Scanning:** using Google/Bing “dorks” and getting the file locations “passively”

**Active Scanning:** scraping the entire domain through a script (by following all internal links)
LIMITATIONS ON ACTIVE/PASSIVE SCANNING

PASSIVE

• **Most important:** ROBOT DETECTION

• **Next, max. 1000 search results (Bing)** ➔ even manually
LIMITATIONS ON ACTIVE/PASSIVE SCANNING

PASSIVE
• Most important: robot detection
• Next, max. 1000 search results (Bing) \(\Rightarrow\) even manually

ACTIVE
• More workload on the servers, not stealthy
• “Backlinks” cannot be found (i.e. incoming links from other websites)
PRIOR WORK/EXISTING TOOLS

- **FOCA**
  - *(Windows-only, no command-line mode but only GUI, closed source)*

- **MeTagooFil**
  - *(Out of support for years, no longer working)*

- **Goolag Scanner**
  - *(Out of support for years, no longer working)*

- **Maltego**
  - *(Strict limitations on number of entities, not usable. Also, closed source. Can cost hundreds to thousands of US dollars)*
THE WAY TO GO: ACTIVE SCANNING

- Scrapy (Python)
THE WAY TO GO: ACTIVE SCANNING

- **Scrapy (Python)**
- **Follow all internal links and find the locations of public documents**
THE WAY TO GO: ACTIVE SCANNING

- Scrapy (Python)
- Follow all internal links and find the locations of public documents
- Took days to complete
“EDUCATING” THE SCRAPER

Use of a blacklist to prevent it from looping in:

• Calendars
• Badly designed internal search engines
• API documentation pages
• Thousands of pages of legislation
• Etc.
RESULTS OF THE DATA HARVESTING

- Found **461,356** public documents on **671** domains
- On average, **685** documents per domain
INTERESTING METADATA: FILE LOCATIONS
INTERESTING METADATA: FILE LOCATIONS

Manifest Reference File Path
: \\

Ingredients File Path
: \\

Tag dlc Doc Id Url
: http://

Tag dlc Doc Id Url
: http://

Dest Url
: smb://

Dest Url
: smb://

Dest Url
: smb://
VISUALIZING FILEPATHS

What to do with these filepaths?

Z:\Karst\Japanese-puzzles
Z:\Jaap\Vacation
O:\Arno\Dropbox
Z:\CdL\Transatlantic-cables
O:\Arno\unikernels\secrethackingstuff
O:\Mick\Koogle
EXAMPLE: FILEPATHS OF A DOMAIN
VISUALIZING SOCIAL NETWORKS

Interesting metadata in Microsoft Office files:

- Creator
- Author
- Last modified by
VISUALIZING SOCIAL NETWORKS: EDIT NETWORK

Last Modified By: B

Creator: A
VISUALIZING SOCIAL NETWORKS: EDIT NETWORK

Last Modified By

B

Creator

A

Last Modified By

C

Last Modified By
VISUALIZING SOCIAL NETWORKS: EDIT NETWORK

Last Modified By B

Creator

Last Modified By C
VISUALIZING SOCIAL NETWORKS: TRIADIC CLOSURE

Last Modified By

B

C

Last Modified By

Creator
VISUALIZING SOCIAL NETWORKS: TRIADIC CLOSURE

Last Modified By B

Last Modified By C

Creator A
APPLYING THE CLUSTERING COEFFICIENT (1/2)
APPLYING THE CLUSTERING COEFFICIENT (2/2)
VULNERABILITIES: CVES AND CPES
VULNERABILITIES: CVES AND CPES

Microsoft® Word 2016
Microsoft® Office Word 2007
Adobe Acrobat 11.0.0

cpe:/a:microsoft:word:2016
cpe:/a:microsoft:word:2007
cpe:/a:adobe:acrobat:11.0.0
VULNERABILITIES: CVES AND CPES

Microsoft® Word 2016
Microsoft® Office Word 2007
Adobe Acrobat 11.0.0

cpe:/a:microsoft:word:2016
cpe:/a:microsoft:word:2007
cpe:/a:adobe:acrobat:11.0.0

CVE-2017-8509
CVE-2017-0281
CVE-2017-0254
CVE-2017-0254
CVE-2015-5094
CVE-2015-5090
[...]
CVE-2015-5089
CVE-2015-5088
CVE-2014-8450
CVE-2014-1761
CVE-2014-0260
**Vulnerabilities**

Some metadata tags contain information on software or hardware used. This information can be used to get insight into potential vulnerabilities at the domain that produced the reviewed documents. On this page, an overview of these software and hardware versions can be found. To get the information, the Creator Tool tags were searched through.

Below, you find an overview for the information on potential vulnerabilities for each domain is shown per year.

*Note that CVEs may appear multiple times for a single domain. This is because, in those cases, the domain has more than one matching CPE.*

### 2017

<table>
<thead>
<tr>
<th>CVE ID</th>
<th>CPE string</th>
<th>CVSS score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVE-2017-8509</td>
<td>cpe:/a:microsoft:word:2016</td>
<td>9.3</td>
<td>A remote code execution vulnerability exists in Microsoft Office when the software fails to properly</td>
</tr>
</tbody>
</table>
ENRICHING THE DATA: FINDING EMAIL ADDRESSES

HOW CAN WE DERIVE EMAIL ADDRESSES FROM THE USER LIST?
Email address syntaxes

One source of email addresses includes data breaches. In this analysis, the Anti Public data breach was used to check for email addresses. In the output below, the occurrences of users from brabant.nl are listed.

<table>
<thead>
<tr>
<th>Email syntax</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>{f}{last}</td>
<td>100</td>
</tr>
<tr>
<td>{f}{prefix}{last}</td>
<td>29</td>
</tr>
<tr>
<td>{first}{prefix}{last}</td>
<td>13</td>
</tr>
</tbody>
</table>
### Email address syntaxes

One source of email addresses includes data breaches from brabant.nl are listed.

<table>
<thead>
<tr>
<th>Email syntax</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>{f}{last}</td>
<td>100</td>
</tr>
<tr>
<td>{f}{prefix}{last}</td>
<td>29</td>
</tr>
<tr>
<td>{first}{prefix}{last}</td>
<td>13</td>
</tr>
</tbody>
</table>

### Phishing Targets

The targets below are potential future acquaintances due to the concept of Triadic Closure.

This means they are likely to get to know each other at some point in the future. That makes the "User 1: "User 2") good source/target pairs for phishing attacks.

<table>
<thead>
<tr>
<th>User 1</th>
<th>User 2</th>
<th>Average clustering coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Email address syntaxes

One source of email addresses includes data breaches. from brabant nl are listed.

<table>
<thead>
<tr>
<th>Email syntax</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>f{last}</td>
<td>100</td>
</tr>
<tr>
<td>f{prefix}{last}</td>
<td>29</td>
</tr>
<tr>
<td>{first}{prefix}{last}</td>
<td>13</td>
</tr>
</tbody>
</table>

Modifiers

Below, you find a list of all unique modifiers (i.e. values of the Last Modified By metadata tags) for https://[blackacted] nl:

- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- A
- C
- D
- D
- D
- E
- E
- E
- G
- G
- G
- G
- G
- G
- H
- H
- H
- H
- H
Timelines

Creation and modification statistics

This page shows an overview of when files were created. Please note that only files with a date format of "yyyy-mm-dddd" are taken into account here.
Timelines

Creation and modification statistics

This page shows an overview of when files were created. Please note that only files with a date format of "yyyy-mm-dddd" are taken into account here.
Timelines

Creation and modification statistics

This page shows an overview of when files were created. Please note that only files with a date format of "yyyyMMdd" are taken into account here.
WEEKLY PATTERNS
CONCLUSIONS

- The Dutch government leaks metadata through public documents on almost all domains.

- This information turns out to be a valuable source of information for a hacker.

- It is not an attack tool on its own, but it does provide insight into an organization’s inner workings (filesystems, users, software, timelines, etc.).

- CVEs can be found for domains, but converting software versions found in metadata to valid CPEs turns out to be hard (resulting in false positives).
CONCLUSIONS

- The most interesting information in the metadata is about people, users and social networks;
- This information *seems valid*,
- Is present in *almost all documents*,
- ...and gives *most information about potential attack paths (infection spreading)*
FUTURE WORK

• Get even further:
  • Link to people’s Social Media accounts and harvest more information (interests, friends, personalities)

• Look for .docm files:
  • Indicates the organization uses Office Macros

• Make the scraper smarter (to avoid disturbance)

• Add offensive features to the dashboard
PLEASE ASK QUESTIONS!

FEEL FREE TO DROP BY FOR A LIVE DEMO