Jurnal Sains Informatika Terapan (JSIT)
E-ISSN: 2828-1659, Volume: 5, Issue: 2, Month: Juni, Year: 2026
Page: 417 - 426, Available online at: https://rcf-indonesia.org/home/

PENERAPAN TEKNIK *GOOGLE DORKING* UNTUK IDENTIFIKASI KERENTANAN *SENSITIVE DATA EXPOSURE* MENGGUNAKAN *GOOGLE HACKING DATABASE (GHDB)* DENGAN METODE *FOCUSED CRAWLING*: STUDI KASUS SITUS NASA

Francesco Jeremy Topol[1)], Quido C. Kainde[2)], Kristofel Santa[3)],
[1,2,3]Program Studi Teknik Informatika,Fakultas Teknik,Universitas Negeri Manado
Corresponding Author: [1] 21210064@unima.ac.id

Article Info

Article history:

Received: Feb 17, 2026
Revised: Feb 19, 2026
Accepted: Mei 31, 2026
Published: Jun 01, 2026

Keywords:

Sensitive Data Exposure
Google Dorking
Google Hacking Database (GHDB)
Focused Crawling
Penetration Testing
Keamanan Siber

ABSTRACT

Di era digital saat ini, kebocoran data sensitif (*Sensitive Data Exposure*) menjadi ancaman serius bagi integritas organisasi, sering kali disebabkan oleh kesalahan konfigurasi izin akses misalnya studi kasus ini yaitu pada layanan penyimpanan awan seperti *Google Drive* dan Spreadsheet. Penelitian ini bertujuan untuk merancang dan mengimplementasikan sistem audit keamanan otomatis bernama DorkWatch untuk mengidentifikasi potensi kebocoran data pada domain nasa.gov. Penelitian ini memanfaatkan sumber data primer berupa aset digital yang terindeks pada domain nasa.gov dan subdomain terkait, yang dikumpulkan secara *real-time* melalui pemanfaatan 5 kueri dork spesifik dari *Google Hacking Database* (GHDB) yang menargetkan ekosistem Google Workspace. Dalam proses pengujian yang dilakukan, dengan limit 25 hasil pencarian sistem berhasil mengekstraksi dataset yang mencakup tautan layanan penyimpanan awan publik (Google Drive dan Docs) dari konten HTML serta metadata file PDF menggunakan *Regular Expression (Regex)* dengan metode *Focused Crawling*. Dataset yang dihasilkan kemudian diklasifikasikan secara otomatis ke dalam 2 kategori utama, yaitu *Docs Exposure* dan *Drive Exposure.* Sistem ini dibangun menggunakan bahasa pemrograman Python dengan kerangka kerja Flask, dan menerapkan teknik *Google Dorking* berbasis *Google Hacking Database (GHDB)* untuk pencarian dokumen publik dan metode *Focused Crawling* untuk penelusuran aset digital secara lebih terarah, serta melakukan analisis risiko otomatis berdasarkan validasi status HTTP dan deteksi kata kunci sensitif. Hasil pengujian menunjukkan bahwa sistem DorkWatch efektif dalam menemukan dan mengklasifikasikan aset NASA yang terekspos ke publik, termasuk tautan file *Google Docs* serta folder dan file *Google Drive* yang tidak muncul pada hasil pencarian standar, sehingga membuktikan bahwa integrasi metode ini efisien dalam memitigasi risiko kebocoran informasi.

1. PENDAHULUAN

Saat ini, siapa pun dapat dengan mudah mengirim dan menerima data seperti video atau email, ke seluruh dunia hanya dengan satu klik [1]. Namun, keamanan data yang dikirimkan tersebut sering kali menjadi pertanyaan penting. Internet telah menjadi infrastruktur yang sangat cepat berkembang dalam kehidupan kita, namun ancaman terhadap keamanan siber juga meningkat pesat [2].

Di era digital saat ini, keamanan siber memiliki peran krusial untuk melindungi data, privasi, dan integritas berbagai sistem serta proses dari ancaman siber yang terus berkembang [3]. Ancaman seperti malware, ransomware, dan serangan phishing dapat menyebabkan akses ilegal, perubahan atau penghapusan data sensitif, serta mengganggu operasional bisnis sehari-hari [4].

Ada 90% populasi dunia memiliki akses ke internet pada tahun 2024, 90% penduduk menggunakan smartphone pada tahun 2023 [5]. Meskipun lebih dari setengah populasi dunia telah menggunakan mesin pencari Google sejak tahun 2019, hampir tidak ada di antara mereka yang pernah mendengar istilah "*Google Dorking*". *Google Dorking* adalah metode khusus dalam memanfaatkan mesin pencari Google untuk menemukan data sensitif yang tanpa sengaja terbuka di internet. Teknik ini dapat

memberikan manfaat dan tetap aman jika digunakan secara bertanggung jawab oleh peneliti, jurnalis, developer atau pengguna yang 3 ingin tahu. Namun, teknik ini dapat menjadi sangat berbahaya jika disalahgunakan oleh pihak yang berniat jahat [6]. *Google Dorking* juga dapat mengungkapkan potensi akses tidak sah ke sistem, seperti portal login yang lemah atau aplikasi web yang tidak dilindungi dengan baik, sehingga memberikan gambaran menyeluruh tentang kerentanan yang ada dan membuka peluang untuk eksploitasi lebih lanjut [7]. Siapa pun yang memiliki akses internet dapat melakukan *Google Dorking*.

Teknik ini sangat relevan dalam mengidentifikasi berbagai kerentanan pada situs web dan server, seperti file sensitif yang seharusnya dilindungi namun dapat diakses publik, data yang terlupakan seperti folder backup yang tidak terkunci, serta informasi terkait infrastruktur dan perangkat lunak yang digunakan. Contohnya, sebuah studi kasus pada situs NASA yang saya temukan. Kondisi sebelumnya pada situs NASA ini menunjukkan bahwa sistem atau domain-nya sudah aman namun beberapa file publik, termasuk file spreadsheet atau folder *Drive* yang ada dan terindeks oleh mesin pencari tidak dilindungi dan dapat diakses melalui pencarian sederhana, yang isi filenya mengungkapkan informasi atau datadata penting dan dapat di EDIT dan DELETE data-data tersebut.

Situasi ini menegaskan adanya gap antara pengelolaan akses file atau folder publik yang seharusnya dapat mencegah *Sensitive Data Exposure* dengan realitas di lapangan, yaitu masih adanya kesalahan konfigurasi dari pemilik penyimpanan awan milik NASA yang secara tidak sengaja membagikan atau mengatur Akses *Google Spreadsheet* dan *Google Drive* dalam mode publik sehingga dapat diakses secara luas sehingga menyebabkan adanya *Vulnerability Sensitive Data Exposure*.

Penelitian ini hadir dengan pendekatan baru yang mengombinasikan *Google Dorking* berbasis GHDB dengan metode *Focused Crawling* yang otomatis dan terarah, yang ditujukan secara khusus untuk mengidentifikasi potensi kebocoran *Sensitive Data Exposure* pada situs resmi NASA. Melalui integrasi *Google Dorking* berbasis *GHDB* dan *Focused Crawling*, penelitian ini menghadirkan solusi deteksi kerentanan yang otomatis. Pengembangan *DorkWatch* didasarkan pada efisiensi fungsi sesuai teori *Form Follows Function* [8]. Dengan memprioritaskan akurasi deteksi di atas kompleksitas desain, sistem ini dirancang untuk secara efektif mengidentifikasi paparan data sensitif *(Sensitive Data Exposure)* pada domain NASA.

1.1. Landasan Teori

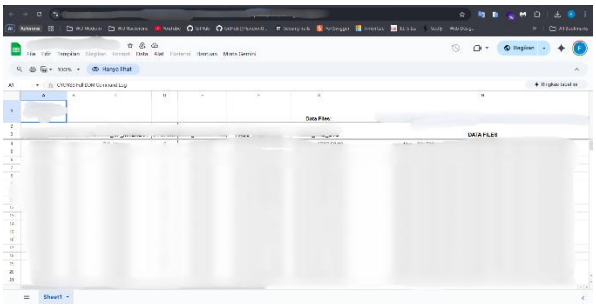
1.1.1. *Penetration Testing*

Penetration Testing adalah upaya untuk mengeksploitasi sistem secara sah guna mengidentifikasi potensi celah keamanan. Dalam pengujian ini, penguji memiliki izin untuk melakukan pengujian penetrasi dan secara sengaja mencoba mengeksploitasi sistem guna menemukan kerentanan yang mungkin ada [9].

1.1.2. *Google Spreadsheet* dan *Google Drive*

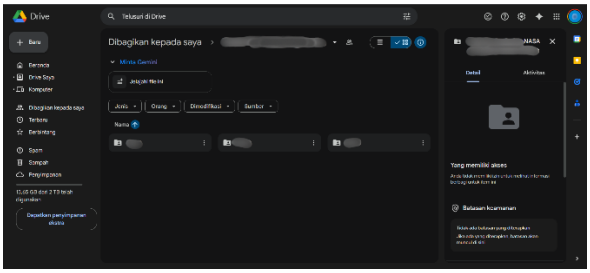
Google Spreadsheet adalah salah satu alat yang dapat diakses secara gratis kapan saja dan dari mana saja melalui layanan Google. Pengguna juga 9 dapat mengakses *Google Spreadsheet* secara offline menggunakan komputer, tablet, atau perangkat seluler [10].

Pada *Google Spreadsheet* (sheets) ini memiliki fitur yang unik dan keren yaitu, seperti berbagi dokumen google sheets ini dan kolaborasi secara realtime, memungkinkan para kolabolator atau orang-orang yang diberikan akses pada dokumen spreadsheet ini dari berbagai lokasi untuk dapat bekerja pada dokumen yang sama secara bersamaan [11].



Gambar 1. *Interface Google Spreadsheets*

Google Drive adalah layanan penyimpanan berbasis cloud yang memungkinkan pengguna menyimpan, berbagi, dan mengakses file dari mana saja. Banyak orang memakai *Google Drive* karena praktis untuk membagikan dokumen, foto, atau video. Namun setiap file yang dibagikan tetap memakan ruang penyimpanan pemiliknya, sehingga pengguna perlu menghapus file yang sudah dibagikan agar kapasitas *Google Drive* tetap lega [12].



Gambar 2. *Interface Google Drive*

1.1.3. *Google Dorking* atau Google Hacking

Google Dork atau *Google Dorking* belum memiliki definisi atau klasifikasi yang jelas sebagai aplikasi atau perangkat lunak. Tetapi, istilah ini lebih merujuk pada teknik Advanced Search yang dijalankan melalui search engine yaitu Google. *Google Dorking* merupakan metode teknis yang

digunakan untuk mencari informasi sensitif yang tidak mudah diakses oleh pengguna umum (publik). Informasi sensitif ini hanya dapat ditemukan dengan menggunakan perintah khusus untuk mendapatkan data yang bersifat rahasia atau sensitif [13].

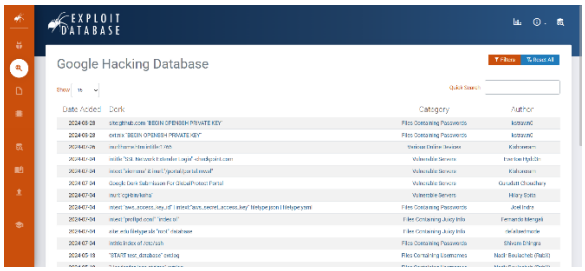
1.1.4. *Sensitive Data Exposure* (*Cryptographic Failures*)

Sensitive Data Exposure merujuk keadaan atau situasi dimana data-data sensitif, seperti informasi pribadi, keuangan, atau kesehatan tidak dilindungi dengan baik atau terkekspose secara publik sehingga siapa saja atau penyerang dapat meengakses data-data tersebut. Masalah ini sering terjadi karena kesalahan konfigurasi, keamanan yang lemah, atau eksploitasi celah dalam sistem [14].

Sensitive Data adalah informasi yang dilindungi agar tidak diungkapkan secara sembarangan, mencakup Informasi Identitas Pribadi (*Personally Identifiable Information/PII*), Informasi Kesehatan yang Dilindungi (*Protected Health Information/PHI*), atau data lain yang bersifat pribadi dan rahasia [15].

1.1.5. *Google Hacking Database (GHDB)*

Untuk mendukung penggunaan teknik *Google Dorking*/Hacking yang digunakan dengan tujuan Pentest untuk menemukan kerentanan dalam struktur sebuah situs web, file database yang terekspos, serta dokumen-dokumen rahasia, tersedia di internet yang namanya *Google Hacking Database (GHDB)* yang merupakan sebuah basis data yang berisikan banyak *query dork* yang telah dievaluasi dan divalidasi oleh *Offensive Security* [16].



Gambar 3. *Interface Google Hacking Database (GHDB)*

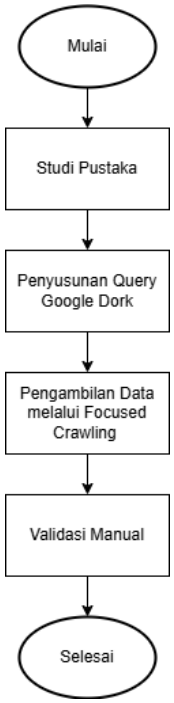
1.1.6. *Focused Crawling*

Focused Crawling merupakan suatu metode yang di mana melakukan crawler secara otomatis dengan menjelajahi halaman web yang relevan atau yang terkait langsung denga topik yang dicari. *Focused Crawling* merupakan metode pencarian halaman web secara terarah dengan mengandalkan kata kunci yang relevan. Dalam pendekatan ini, setiap tautan web yang berkaitan dengan suatu halaman atau kata kunci yang telah ditentukan akan dijelajahi lebih lanjut dan terarah [17].

2. METODE PENELITIAN

2.1. Tahapan Penelitian

Dalam penelitian ini, penelusuran dilakukan dengan menggunakan teknik *Google Dorking* yang query-nya bersumber dari *GHDB (Google Hacking Database)*, lalu dikombinasikan dengan metode *Focused Crawling* untuk menemukan kemungkinan kebocoran data sensitif (*Sensitive Data Exposure*/SDE) pada *Google Spreadsheet* dan *Google Drive*. Tahapan penelitian disusun untuk memastikan proses identifikasi kerentanan berjalan secara terukur dan etis. Berikut Adalah tahapan penelitian yang dilakukan:



Gambar 4. Tahapan Penelitian

2.1.1. Studi Pustaka

Melakukan kajian mendalam terhadap jurnal, artikel ilmiah, whitepaper, dan dokumentasi terkait teknik Google Dorking, GHDB (Google Hacking Database), serta Focused Crawling. Dengan tujuan adalah untuk membangun pemahaman konseptual dan teknis tentang metode yang akan diimplementasikan dalam sistem, serta mengidentifikasi teknik terbaik untuk mendeteksi sensitive data exposure pada Google Spreadsheet dan Google Drive.

2.1.2. Penyusunan Query Google Dork

Tahap ini membuat kata kunci pencarian khusus yang disebut Google Dork, dengan bantuan database GHDB, untuk mencari celah keamanan di situs NASA. Kata kunci ini dibuat dengan pola mencari link atau file spreadsheet maupun file-file pada folder google drive yang berisikan data-data sensitif.

2.1.3. Pengambilan Data melalui Focused Crawling

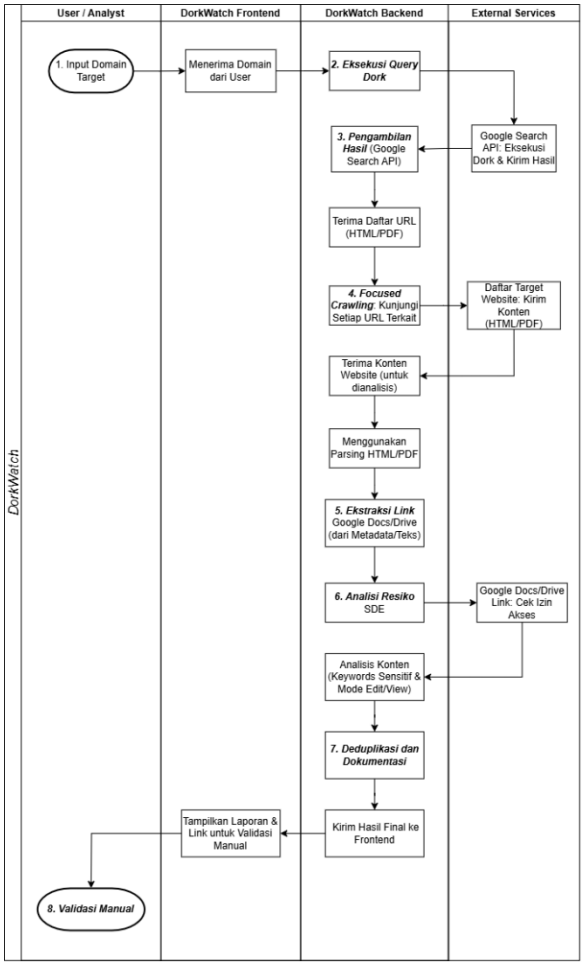
Pengambilan data melalui Focused Crawling merupakan tahap bagi penulis dalam proses pengambilan data yang dilakukan secara otomatis menggunakan metode Focused Crawling yang berjalan melalui beberapa Langkah untuk menemukan kerentanan Senstive Data Exposure.

2.1.4. Validasi Manual

Sebagian kecil hasil crawling akan diperiksa secara manual untuk memastikan apakah halaman tersebut benar-benar mengandung indikasi kebocoran informasi. Ini dilakukan pada data yang diklasifikasikan dengan tingkat risiko tinggi atau memiliki struktur unik.

2.2. Alur Kerja DorkWatch

Penelitian ini menggunakan pendekatan eksploratif yang bertujuan untuk mengidentifikasi potensi kebocoran data sensitif pada file *Google Spreadsheet* maupun file-file di folder *Google Drive* yang dapat diakses secara publik di situs resmi NASA melalui eksplorasi teknik Google Dork berbasis GHDB, serta menggunakan metode *Focused Crawling* untuk mempersempit cakupan pencarian data relevan [17].



Gambar 5. *The Architecture of Focused Crawler*

2.2.1. Inisialisasi Domain Target

Proses audit dimulai dengan tahap inputan domain target oleh pengguna, yang di mana pengguna mendefinisikan domain spesifik (dalam hal ini yaitu: nasa.gov dan subdomain-subdomain terkait) yang akan menjadi objek pengujian.

Langkah ini berfungsi untuk pembatasan ruang lingkup crawling, supaya memastikan bahwa mekanisme pencarian sistem bekerja secara terfokus dan terarah pada aset digital yang relevan dengan entitas pemilik domain tersebut

2.2.2. Eksekusi *Query dork*

Pada tahap ini, setelah pengguna menginputkan domain target serta limit output yang ingin dihasilkan, sistem secara otomatis akan mengeksekusi *query dork* yang telah diintegrasikan ke dalam logika aplikasi, berdasarkan referensi dari *Google Hacking Database (GHDB)* yang bertujuan untuk mengidentifikasi indeks dokumen atau direktori tersembunyi yang terekspos ke publik.

Berikut ini *Query dork* yang sudah diintegrasikan ke dalam kode:

1) site:{domain} *Docs*.google.com/spreadsheets
2) site:{domain} *Docs*.google.com/document/d/
3) site:{domain} drive.google.com/file
4) site:{domain} drive.google.com/drive/folders
5) site:{domain} drive.google.com/open?id=

2.2.3. Pengambilan Hasil

Sistem menggunakan Search API untuk menjalankan pencarian dan mengambil daftar URL yang sesuai dengan kueri yang dijalankan. Pada tahap ini, sistem menerapkan filter awal untuk menyeleksi URL yang memiliki indikasi kuat memuat tautan ke layanan penyimpanan cloud (*Google Drive*/*Docs*). URL yang lolos penyaringan tersebut kemudian disimpan sebagai kumpulan URL awal (seed URLs) sebelum crawling dimulai.

2.2.4. *Focused Crawling*

Setelah seed URL terkumpul, DorkWatch melakukan *Focused Crawling* terhadap setiap halaman yang ditemukan. Crawling dilakukan secara mendalam namun terbatas, hanya pada halaman yang berasal dari domain target dan mengandung potensi tautan ke dokumen Google Spreasheet dan *Google Drive*. Sistem juga mampu menembus konten statis seperti file PDF dan HTML untuk mengekstrak informasi yang tersembunyi.

Berbeda dengan general crawling, *Focused Crawling* pada DorkWatch tidak menjelajah seluruh web, melainkan hanya mengikuti jalur yang telah disaring sebelumnya. Hal ini membuat proses lebih cepat, hemat sumber daya, dan tetap relevan terhadap tujuan utama yaitu mendeteksi eksposur data sensitif dari domain yang diaudit.

2.2.5. Ekstraksi Link

Pada tahap ini, sistem mengekstrak tautan *Google Docs* dan *Google Drive* dari setiap halaman yang telah dikunjungi. Proses ekstraksi dilakukan dengan parsing HTML untuk halaman web biasa, dan parsing teks dari file PDF menggunakan pustaka seperti PyPDF2. Hanya tautan yang sesuai pola *Google Docs*/*Drive* yang akan disimpan. Ini dilakukan dengan teknik crawling terbatas (focused) hanya pada halaman yang relevan atau terkait.

2.2.6. Analisis Resiko SDE

Pada tahap ini, setiap tautan yang berhasil diekstrak kemudian dianalisis untuk mengevaluasi status akses dan potensi risiko kebocoran data. Tool memeriksa apakah tautan dapat diakses publik, serta mode aksesnya (view, comment, edit). Selain itu, sistem juga mendeteksi keberadaan kata kunci sensitif seperti *"password", "confidential", "internal"*, dan sebagainya sebagai langkah informasi apakah ada potensi risiko kebocoran data.

Analisis ini dilakukan secara otomatis untuk memberikan penilaian awal terhadap tingkat risiko dari setiap dokumen. Dokumen dengan akses edit dan mengandung kata kunci sensitif akan diberi peringatan risiko tinggi.

2.2.7. Deduplikasi dan Dokumentasi

Setelah semua tautan dianalisis, sistem ini akan melakukan deduplikasi berdasarkan ID unik dari setiap dokumen Google. Hal ini penting untuk menghindari pengulangan hasil yang sama akibat munculnya tautan identik dari berbagai sumber. Proses ini juga mencatat log deduplikasi yang bisa didownload untuk keperluan audit dan transparansi.

Hasil akhir kemudian disimpan dalam format TXT dan CSV yang dapat diunduh oleh pengguna. Format ini dipilih karena mudah dibaca, dapat diolah lebih lanjut, dan cocok untuk dokumentasi resmi.

2.2.8. Validasi Manual

Tahap terakhir adalah pengecekan manual oleh pengguna. Walaupun DorkWatch sudah melakukan analisis otomatis, pemeriksaan langsung tetap dibutuhkan untuk memastikan apakah dokumen yang ditemukan benar-benar berisi data sensitif. Langkah ini penting agar tidak terjadi kesalahan deteksi (false positive) dan supaya hasil audit tetap akurat.

3. HASIL DAN PEMBAHASAN

3.1. Implementasi Sistem

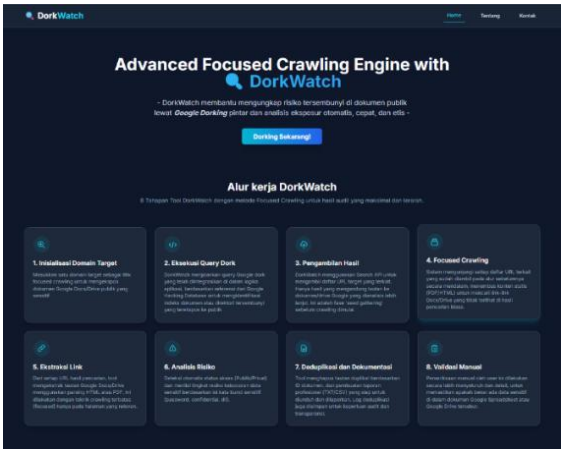
Tahapan Implementasi ini merupakan proses realisasi dari rancangan sistem yang sudah dijelaskan pada bab sebelumnya. Pada tahap ini, DorkWatch dibangun sebagai aplikasi berbasis web dengan menggunakan Python sebagai bahasa pemrograman. Bagian backend dibuat dengan framework Flask, sementara tampilan antarmuka dibuat menggunakan Tailwind CSS.

3.1.1. Implementasi Antarmuka Pengguna (User Interface)

Tampilan antarmuka dibuat agar peneliti dapat dengan mudah memasukkan domain target NASA dan melihat hasil pengecekan potensi kerentanan yang ditemukan oleh sistem.

3.1.1.1. Halaman awal (Dashboard)

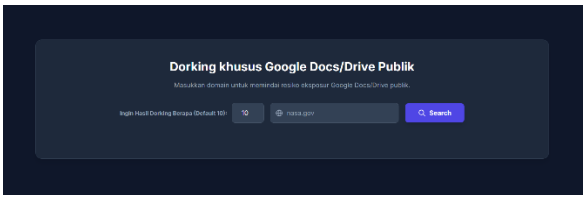
Halaman utama menjadi informasi awal aplikasi. Sesuai alur kerja sistem, halaman ini menyediakan informasi ringkas serta menampilkan 8 alur kerja yang dijalankan sistem, mulai dari Inisialisasi Domain Target hingga Validasi Manual.



Gambar 6. Tampilan Halaman Awal DorkWatch

3.1.1.2. Halaman Utama DorkWatch

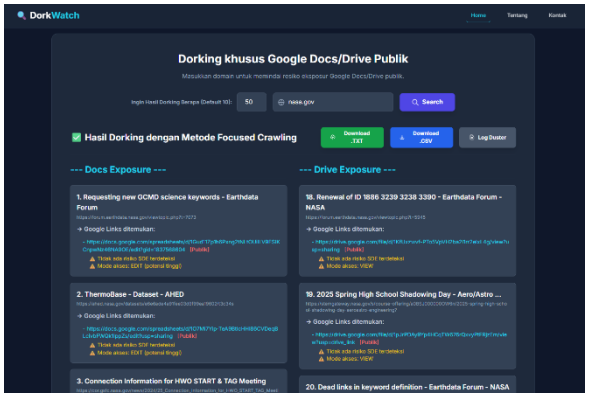
Halaman utama berfungsi sebagai pusat kendali. Pada halaman ini, pengguna memasukkan domain target dan batas hasil pencarian (limit) sebelum proses crawling dimulai.



Gambar 7. Tampilan Halaman Utama DorkWatch

3.1.1.3. Tampilan Hasil Audit (Result Page)

Hasil temuan ditampilkan secara dinamis menjadi dua kolom yaitu *Docs Exposure* dan *Drive Exposure*.



Gambar 8. Tampilan Hasil Audit DorkWatch

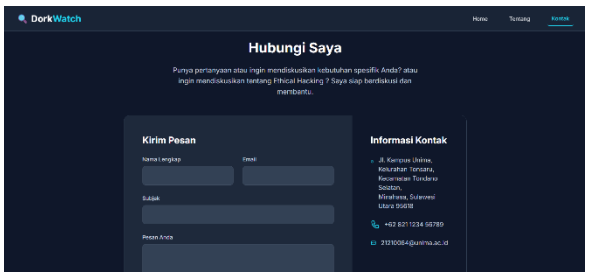
3.1.1.4. Tampilan Halaman About DorkWatch

Halaman ini dirancang untuk memberikan penjelasan transparansi mengenai metodologi dan tujuan pengembangan sistem DorkWatch ini.

Gambar 9. Tampilan Halaman Tentang Aplikasi DorkWatch

3.1.1.5. Tampilan Halaman Kontak DorkWatch

Halaman kontak dibuat agar pengguna bisa berkomunikasi langsung dengan pengembang, misalnya untuk melaporkan bug atau berdiskusi soal masalah teknis. Halaman ini bekerja dengan menghubungkan tampilan di frontend dengan proses pengiriman pesan di bagian backend, sehingga setiap pesan yang dikirim dapat diproses dengan baik.



Gambar 10. Tampilan Formulir Kontak

3.2. Hasil Pengujian Sistem

Pengujian dilakukan dengan studi kasus langsung pada domain nasa.gov untuk melihat seberapa efektif sistem dalam mengidentifikasi kerentanan *Sensitive Data Exposure* di file *Google Spreadsheet* maupun folder dan file *Google Drive* yang terindeks di dalam halaman domain serta subdomain NASA.

3.2.1. Skenario Pengujian

Pengujian dilakukan menggunakan parameter berikut:

Tabel 1. Skenario Pengujian Domain NASA

Target Domain	Limit Pencarian	Waktu Pengujian
nasa.gov	25	28/11/2025

3.2.2. Temuan Audit Kerentanan

Sebelum membahas semua hasil temuan, sistem DorkWatch perlu menetapkan terlebih dahulu instrumen atau acuan penilaian yang dipakai untuk menentukan tingkat risiko pada setiap URL yang ditemukan. Penyusunan instrumen ini tidak hanya didasarkan pada logika algoritma sistem, tetapi juga diselaraskan dengan pedoman resmi NASA *Vulnerability Disclosure Program (VDP)*.

Berdasarkan kebijakan VDP NASA, meskipun layanan pihak ketiga (third-party services) berada di luar ruang lingkup pengujian penetrasi sistem, keberadaan data non-publik NASA pada layanan pihak ketiga publik (seperti *Google Drive*/*Docs*) secara eksplisit dinyatakan sebagai kerentanan yang wajib dilaporkan. Berikut adalah tabel instrumen klasifikasi risiko audit yang digunakan sebagai standar penilaian:

Tabel 2. Instrumen Klasifikasi Audit Kerentanan [18] https://bugcrowd.com/engagements/nasa-vdp

Parameter	Indikator	Klasifikasi Risiko	Keterangan
Aksesibilitas	Status HTTP 200 (OK)	Valid (Publik)	Tautan dapat diakses oleh publik tanpa autentikasi login.
	Status HTTP 403/404	Invalid	Tautan terkunci (Private) atau file sudah dihapus.
Izin Akses (URL)	Mengandung segmen /edit	Tinggi (Critical)	Publik memiliki izin untuk mengubah, mengedit, atau menghapus data.
	Mengandung segmen /view atau /comment	Sedang/Rendah	Publik hanya memiliki akses lihat (View Only) atau komentar.

Sensitivitas Konten	Ditemukan kata kunci sensitif (misal: password, email, phone, confidential, internal, credential, dsb)	Tinggi (High)	Judul atau isi dokumen mengandung indikasi kuat data rahasia.
	Tidak ditemukan kata kunci sensitif	Info/Low	Dokumen publik biasa.
Kesesuaian Scope (NASA VDP)	Data Publik di Layanan Pihak Ketiga (*Google Drive*/*Docs*)	Valid (In-Scope)	Sesuai klausul NASA VDP: "Nonpublic NASA data residing on public third-party services."

Untuk memvalidasi akurasi sistem, berikut disajikan analisis mendalam pada salah satu sampel temuan yang terdeteksi memiliki risiko tertinggi namun terbukti aman setelah dilakukan validasi manual:

3.2.2.1. Satu Contoh Temuan

File Spreadsheet berjudul "*Requesting new GCMD science keywords - Earthdata Forum*"

3.2.2.2. Analisis Berdasarkan Instrumen

a. Cek Aksesibilitas

Sistem menerima respon HTTP 200 OK, yang berarti file valid dan terbuka untuk publik.

b. Cek Izin Akses

Pada URL ditemukan segmen /edit (.../edit?gid=1837568604), sehingga sistem secara otomatis melabeli sebagai "Mode akses: EDIT (potensi tinggi)". Setelah dilakukan pemeriksaan manual (Tahap Validasi Manual), diketahui bahwa meskipun URL mengandung parameter /edit, konfigurasi akses dokumen sebenarnya telah dibatasi menjadi "hanya lihat" (View Only) oleh owner dari dokumen spreadsheet tersebut.

c. Cek Sensitivitas Konten

Berdasarkan pemindaian judul dan cuplikan konten, sistem memberikan hasil "Tidak ada risiko SDE terdeteksi". Hal ini karena tidak ditemukan kata kunci sensitif seperti password atau confidential di dalam dokumen tersebut.

d. Cek Scope VDP

File berada di layanan pihak ketiga (*Docs*.google.com) dan terkait domain nasa.gov. Ini memenuhi kriteria pelaporan NASA VDP.

e. Kesimpulan

Temuan ini valid sebagai dokumen publik, namun risiko edit yang dideteksi sistem terbukti aman setelah dilakukan validasi manual. Hal ini menunjukkan pentingnya tahap validasi manual untuk memverifikasi temuan otomatis dari sistem agar tidak adanya temuan yang false positive.

Dengan menggunakan instrumen penilaian tersebut, berikut ini adalah rangkuman 3 dari 25 hasil audit kerentanan yang ditemukan pada domain NASA. Data yang ditampilkan pada tabel merupakan data yang tidak berisi data penting ataupun data sensitif untuk menjaga pedoman *ethical hacking* dalam program NASA VDP serta data tersebut merupakan ringkasan dari file CSV yang secara otomatis dibuat oleh sistem.

Tabel 3. Hasil Identifikasi Kerentanan pada Domain dan Subdomain NASA

Kategori	**Judul Halaman Sumber**	**URL Sumber**	**Link *Google Docs*/Drive**	**Status**	**Risiko SDE**
Docs Exposure	Requesting new GCMD science keywords - Earthdata Forum	https://forum.earthdata.nasa.gov/viewtopic.php?t=7073	https://*Docs*.google.com/spreadsheets/d/1Gud11Zp1h6Pxng2tNLtOUilLV9ESIKCnpwNz46NA9OE/edit?gid=1837568604	Publik	Tidak ada risiko SDE terdeteksi | Mode akses: EDIT (potensi tinggi)
Docs Exposure	Thermo Bas e - Dataset - AHED	https://ahed.nasa.gov/datasets/e6e6ade4a91fee03d9f99ea19602/t3c34s	https://*Docs*.google.com/spreadsheets/d/1O7Mi7Yip-TeA9BtIcHHI86CVDeqBLcivbFWQk1IppZs/edit?usp=sharing	Publik	Tidak ada risiko SDE terdeteksi | Mode akses: EDIT (potensi tinggi)

Drive Exposure	Renewal of ID 1886 3239 3238 3390 - Earthdata Forum - NASA	https://forum.earthdata.nasa.gov/viewtopic.php?t=5945	https://drive.google.com/file/d/1KfUxzwvf-PTo5VpVH2ba2l1rr2elxL4g/view?usp=sharing	Publik	Tidak ada risiko SDE terdeteksi | Mode akses: VIEW

Berikut akan membahas secara detail arti dari setiap kolom dan penjelasan untuk masing-masing temuan:

a. Kategori

Kolom ini digunakan untuk membedakan jenis temuan berdasarkan *query dork* yang dipakai saat proses pencarian. Ada dua kategori utama, yaitu *Docs Exposure* dan *Drive Exposure*. *Docs Exposure* menunjukkan kebocoran pada dokumen kerja seperti Spreadsheet atau *Google Docs* yang biasanya berisi data tabel atau teks yang terus diperbarui. Sementara itu, *Drive Exposure* menandakan kebocoran pada jenis file lain seperti PDF, gambar, atau bahkan satu folder penuh yang bisa diunduh sekaligus. Pembagian ini memudahkan auditor memahami jenis dan sumber kebocoran data tanpa harus mengecek tipe file satu per satu.

b. Judul Halaman Sumber

Kolom ini menampilkan judul meta (meta title) dari halaman web tempat tautan *Google Drive* tersebut muncul atau terdeteksi. Dalam proses investigasi digital, judul halaman membantu memberikan gambaran awal tentang isi dokumen tanpa harus membukanya. Misalnya, judul seperti *"Internal Team Contact List"* atau *"Budget Planning 2024"* sudah cukup menunjukkan bahwa dokumen tersebut memuat informasi sensitif. Judul ini diambil langsung dari hasil crawling melalui snippet pencarian Google, yang kadang masih tersimpan di cache mesin pencari meskipun halaman aslinya mungkin sudah berubah.

c. URL Sumber

Kolom ini menampilkan alamat lengkap (URL) dari halaman nasa.gov atau subdomainnya tempat tautan *Google Drive* atau *Google Docs* ditemukan. Informasi ini sangat penting untuk mengetahui dari URL NASA mana kebocoran itu berasal, Apakah dari halaman resmi, portal internal yang tanpa sengaja terbuka, atau dari folder arsip lama yang sudah tidak dipantau. Dengan mengetahui sumber URL tersebut, admin NASA bisa memperbaiki masalah tidak hanya dengan menutup akses file *Google Drive* yang bocor, tetapi juga dengan menghapus tautannya dari halaman web agar tidak lagi bisa diakses atau muncul di hasil pencarian.

d. Link *Google Docs*/Drive

Kolom ini berisi tautan asli menuju *Google Docs* spreadsheet atau *Google Drive* yang berhasil ditemukan oleh DorkWatch lewat proses *Focused Crawling* dan pembacaan dokumen. Jika URL sumber adalah lokasi ditemukannya tautan, maka bagian ini adalah link langsung menuju file yang berpotensi bocor. Tautan tersebut sudah dibersihkan dan disaring oleh sistem agar formatnya benar, dan inilah link yang nantinya akan dicek untuk melihat apakah file tersebut bisa diakses publik atau sebenarnya bersifat privat.

e. Status

Kolom status menunjukkan hasil pengecekan teknis yang dilakukan dengan mengirim permintaan HTTP ke tautan *Google Docs* atau Google 48 *Drive* yang ditemukan. Tujuannya adalah untuk memastikan apakah file tersebut bisa dibuka oleh siapa saja tanpa perlu login. Jika muncul status "Publik (200 OK)", itu berarti Google membiarkan dokumen tersebut diakses bebas, dan kondisi ini menjadi tanda kuat bahwa telah terjadi kebocoran data (*Sensitive Data Exposure*). Sebaliknya, status seperti "403 Forbidden" atau halaman login menandakan file tersebut sudah terlindungi. Proses pengecekan otomatis ini membantu memisahkan mana temuan yang benar-benar berisiko dan mana yang aman, sehingga proses audit menjadi jauh lebih efisien.

f. Risiko SDE

Kolom terakhir berisi penilaian tentang seberapa berbahaya dokumen yang terbuka tersebut, berdasarkan hasil analisis otomatis. Sistem akan memeriksa judul dan potongan isi dokumen untuk mencari kata-kata sensitif seperti password, confidential, internal, dsb, lalu mengecek juga jenis akses pada URL-nya, misalnya /edit atau /view. Jika muncul informasi seperti "Mode akses: EDIT" atau "Kata sensitif: Password", berarti dokumen tersebut masuk kategori risiko tinggi dan harus segera ditangani tapi harus di cek dahulu secara manual apakah benar-benar ada data sensitif dan apakah bisa di edit. Kondisi ini dapat menyebabkan kebocoran informasi penting atau bahkan memungkinkan orang lain mengubah isi dokumen tanpa izin.

3.3. Pembahasan

3.3.1. Komparasi dengan Penelitian Terdahulu dan Validasi Kebaruan (Novelty)

Untuk memastikan bahwa hasil penelitian ini benar-benar memiliki nilai penting dan berbeda dari penelitian lain, dilakukan perbandingan dengan studi terdahulu yang memiliki topik serupa. Langkah ini digunakan untuk menunjukkan bagian mana yang menjadi keunggulan dan kebaruan (novelty) dari sistem DorkWatch, terutama dalam penerapannya

untuk audit keamanan pada situs milik instansi pemerintah seperti NASA.

Peneliti & Referensi	Metode & Fokus Utama	Keterbatasan Penelitian Terdahulu (Gap)	Keunggulan/ Novelty Sistem DorkWatch
[19]	Analisis privasi pada situs web pemerintah (Government Websites) menggunakan crawling dan analisis manual.	Proses deteksi kebocoran data spesifik pada layanan cloud (*Google Docs*/Drive) belum terotomatisasi penuh dan analisis konten dokumen masih terbatas.	Mengimplementasikan *Focused Crawling* yang secara spesifik menargetkan tautan *Google Drive*/*Docs* yang tersembunyi, termasuk kemampuan parsing isi file PDF untuk menemukan tautan yang tidak terindeks mesin pencari.
[7]	Menggunakan *Google Dorking* sebagai salah satu tahap footprinting dalam kerangka kerja Vulnerability Assessment.	*Google Dorking* hanya digunakan sebagai alat pencarian awal tanpa mekanisme validasi risiko otomatis (apakah file benar-benar bisa diakses atau tidak).	Dilengkapi fitur Validasi HTTP Status Otomatis (200 OK, 403 Forbidden, 404 Not Found) dan deteksi risiko berbasis kata kunci (keyword matching), sehingga meminimalkan false positive sebelum validasi manual.
[17]	Penerapan algoritma *Focused Crawling* untuk mengklasifikasikan konten pada Dark Web.	Fokus algoritma diarahkan pada konten Dark Web yang strukturnya berbeda dengan Surface Web dan layanan Cloud.	Mengadaptasi metode *Focused Crawling* khusus untuk ekosistem Google Workspace (*Docs*, Sheets, Drive) pada Surface Web, dengan logika ekstraksi ID file unik untuk deduplikasi data yang lebih akurat.

3.3.2. Analisis Efektivitas *Google Dorking* Berbasis GHDB

Penggunaan query site:nasa.gov *Docs*.google.com/spreadsheets dan sebagainya seperti pada pembahasan 3.2. terbukti bekerja dengan sangat baik untuk menyaring hasil pencarian. Dengan kombinasi tersebut, sistem bisa langsung melewati halaman-halaman umum dan menemukan dokumen yang memang relevan. Temuan ini menunjukkan bahwa memakai parameter site: bersama URL layanan Google yang spesifik untuk *Docs* dan *Drive* (seperti yang ada di GHDB) bisa membuat pencarian jauh lebih focus, relevan dan efisien.

3.3.3. Peran Metode *Focused Crawling* dalam Deteksi Mendalam

Keunggulan utama dari penelitian ini dibandingkan dengan dorking manual adalah penggunaan metode *Focused Crawling*. Dari hasil uji coba pada 4.2.2. Tabel 3, terbukti bahwa sistem bisa menemukan link *Google Docs* dan *Google Drive* yang tidak muncul di halaman hasil pencarian Google (SERP). Link-link tersebut ternyata tersembunyi di dalam isi dokumen lain, seperti file PDF laporan teknis NASA atau pada parameter-parameter url yang tersembunyi, dan hanya bisa terdeteksi karena crawler menelusuri isi file tersebut secara lebih mendalam. Metode *Focused Crawling* yang diterapkan menunjukkan keunggulannya dalam menangani file PDF juga.

4. KESIMPULAN

Berdasarkan penelitian yang dibahas terkait sistem otomatisasi DorkWatch, dapat ditarik beberapa kesimpulan utama sebagai berikut:

- Penelitian ini berhasil mengembangkan aplikasi web crawler yang bernama DorkWatch yang menerapkan teknik *Google Dorking* berbasis *Google Hacking Database (GHDB)* dengan metode *Focused Crawling*. Sistem ini mampu mengotomatisasi proses pengecekan keamanan yang biasanya dilakukan secara manual, terutama untuk menemukan file *Google Spreadsheet* serta folder dan file *Google Drive* yang tanpa sengaja terbuka untuk publik pada domain nasa.gov.
- Penggunaan metode *Focused Crawling* terbukti mampu memberikan hasil yang lebih dalam dibandingkan pencarian biasa. Dengan kemampuan membaca isi dokumen, termasuk mengambil teks dari file PDF, sistem dapat menemukan tautan sensitif yang sebenarnya tersembunyi di dalam laporan dan tidak tampil di hasil pencarian Google (Search Engine Results Page). Temuan ini menunjukkan bahwa metode tersebut efektif untuk mengungkap kerentanan yang tidak terlihat secara langsung. Hal ini menjawab rumusan masalah mengenai efektivitas identifikasi kerentanan yang tersembunyi.
- Sistem ini mampu mendeteksi dan mengelompokkan tingkat risiko kebocoran data sensitif (SDE) secara otomatis. Dengan mengecek

apakah sebuah file benar-benar bisa diakses (melalui status HTTP 200 OK) dan pencocokan kata kunci sensitif (seperti password, internal, confidential, dll), sistem ini dapat membedakan mana dokumen publik yang aman dan mana yang berpotensi membocorkan informasi penting. Hasilnya, auditor jadi lebih mudah menentukan temuan mana yang harus diprioritaskan.

REFERENSI

[1] A. Hasibuan and E. Dalimunthe, "Implementasi Metode Client Server pada Penerapan Aplikasi Simulasi Ujian Akhir," vol. 5, no. 2, pp. 152–161, 2020.

[2] J. E. Lantu, K. Santa, F. I. Sangko, and O. Kembuan, "Development of a Web-Based File Encryption System Using the Advanced Encryption Standard Method," 2025.

[3] Q. C. Kainde *et al.*, "BONGKAR RAHASIA CYBERCRIME," 2024.

[4] S. Ghundare, A. Patil, and R. Lad, "Importance of Cyber Security," *Int. J. Eng. Res. Technol.*, vol. 8, 2020, [Online]. Available: https://consensus.app/papers/importance-of-cyber-security-ghundare-patil/9cbac46a2d4a506e88b32294ab64c6ae/

[5] A. Kenap, E. Kembuan, E. Usoh, and H. Tondo, *Optimizing the Digital Education Technology in Learning Management System Design During and Post-Covid-19 Pandemic in Society 5 . 0*, vol. 1. Atlantis Press SARL, 2023. doi: 10.2991/978-2-494069-35-0.

[6] S. Kashman, "*Google Dorking* or Legal Hacking: From the Cia *Google Dorking* or Legal Hacking: From the Cia Compromise To Your Cameras At Home, We Are Not As Compromise To Your Cameras At Home, We Are Not As Safe As We Think Safe As We Think," *Technol. Arts Washingt. J. Law*, vol. 18, no. February, pp. 1–2, 2023, [Online]. Available: https://digitalcommons.law.uw.edu/wjltahttps://digitalcommons.law.uw.edu/wjlta/vol18/iss2/1Electroniccopyavailableat:https://ssrn.com/abstract=4369984

[7] P. S. S. Kiran Gandikota, D. Valluri, S. B. Mundru, G. K. Yanala, and S. Sushaini, "Web Application Security through Comprehensive Vulnerability Assessment," *Procedia Comput. Sci.*, vol. 230, no. 2023, pp. 168–182, 2023, doi: 10.1016/j.procs.2023.12.072.

[8] F. E. Kawatu, R. Sumenge, and M. F. Suharto, "Perancangan Oceanarium dengan Pendekatan Arsitektur Kontemporer di Manado," vol. 11, no. 02, pp. 39–48, 2023.

[9] Vegesna, "Utilising VAPT Technologies (Vulnerability Assessment & *Penetration testing*) as a Method for Actively Preventing Cyberattacks," *Int. J. Manag. Technol. Eng.*, vol. XII, no. Vii, pp. 81–94, 2022, [Online]. Available: https://ssrn.com/abstract=4612524

[10] M. S. Ikbal, B. D. Utami, and A. Halimah, "Development of Practicum Assessment Rubric Assisted by *Google Spreadsheet* in Basics Electronics Material," *J. Ilm. Pendidik. Fis.*, vol. 7, no. 3, p. 471, 2023, doi: 10.20527/jipf.v7i3.9698.

[11] A. Santos, "Doing peer feedback in a high school EFL writing class via *Google Docs* and Sheets: A workshop.," *JALTCALL Publ.*, vol. PCP2021, no. 1, 2022, doi: 10.37546/jaltsig.call.pcp2021-08.

[12] A. R. Paul, "Analyzing the User Experience of *Google Drive* Storage Management with Alert Notification," *2025 Intermt. Eng. Technol. Comput.*, no. May, pp. 1–5, 2025, doi: 10.1109/IETC64455.2025.11039452.

[13] S. A. S. A. Abdulaziz Mohammed Ali Al Bin Yahya, "Discovering Security Gaps Using the Google Dorks," pp. 87–92, 2023.

[14] P. van der Linden, "*Sensitive Data Exposure* &Web Scraping with Python," 2022, [Online]. Available: https://scholarworks.calstate.edu/downloads/4t64gt62b

[15] J. Fox, E. H. Il, and G. Z. Il, "MINIMIZING *SENSITIVE DATA EXPOSURE* DURING PREPARATION OF REDACTED DOCUMENTS," vol. 2, 2019.

[16] J. R. G. Evangelista, R. J. Sassi, and M. Romero, "Google Hacking Database Attributes Enrichment and Conversion to Enable the Application of Machine Learning Techniques," *Indian J. Sci. Technol.*, vol. 16, no. 42, pp. 3771–3777, 2023, doi: 10.17485/ijst/v16i42.1799.

[17] P. R. Yunelfi, Y. Purwanto, M. F. Ruriawan, A. S. Popalia, and F. Fahrani, "DarkWeb Crawling using Focused and Classified Algorithm," *[CEPAT] J. Comput. Eng. Progress, Appl. Technol.*, vol. 1, no. 02, p. 1, 2022, doi: 10.25124/cepat.v1i02.4879.

[18] National Aeronautics and Space Administration (NASA), "National Aeronautics and Space Administration (NASA) - Vulnerability Disclosure Program," Bugcrowd. [Online]. Available: https://bugcrowd.com/engagements/nasa-vdp

[19] N. Samarasinghe, A. Adhikari, M. Mannan, and A. Youssef, *Et tu, Brute? Privacy Analysis of Government Websites and Mobile Apps*, vol. 1, no. 1. Association for Computing Machinery, 2022. doi: 10.1145/3485447.3512223.