AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

ON THE APPLICATION OF NLP FOR ADVANCED OSINT ANALYSIS

16th International Conference on Cyber Conflict (CyCon)

Florian Skopik
Center for Digital Safety and Security
AIT Austrian Institute of Technology

TARANIS_AI
https://taranis.ai/

Tallinn, May 30th, 2024

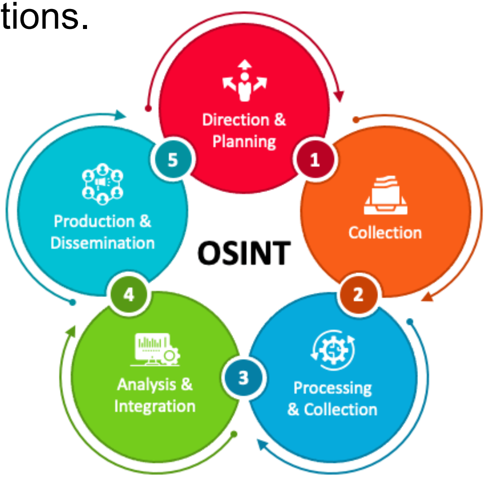
WHAT IS CYBER SECURITY OSINT?

- Open-source intelligence (OSINT) is the collection and analysis of data gathered from open sources to produce actionable intelligence.
- Technical Cyber Threat Intelligence (CTI) to configure detection systems:
  - Indicators to put into SIEMs
  - Domains to block in name-servers or proxies
  - Execution patterns to block in EDRs
- But also “soft” CTI:
  - News about threat actors
  - New (features of) security products
  - News about breaches, incidents, campaigns
  - News about vulnerabilities, patches, mitigations, counter-measures, exploitation, post-exploitation, …
  - Policy news: political/diplomatic initiatives, new EU policy documents, GDPR-related lawsuits
  - Updates on security standards (ISO, BSI, ANSI, CIS, OWASP, …)
  - Mergers, acquisitions, failures, … or other company news

Baseline today of one of our national stakeholders: approx 250 sources, up to 500 articles per day, more after longer weekends or large-scale events

WHAT IS OSINT GOOD FOR?

- Gather public information on potential security threats, vulnerabilities, trends, attacker TTPs, new risks etc. to maintain situational awareness and take early counter actions.
- Input for products
  - Advisories
  - Summaries (daily, weekly)
  - Situational reports, white papers, fact-sheets
- Awareness / Preparedness
  - Consulting / Answering calls for help
  - Media inquiries
  - “Boss/CEO/Politician asking questions”
  - Trigger for proactive activities

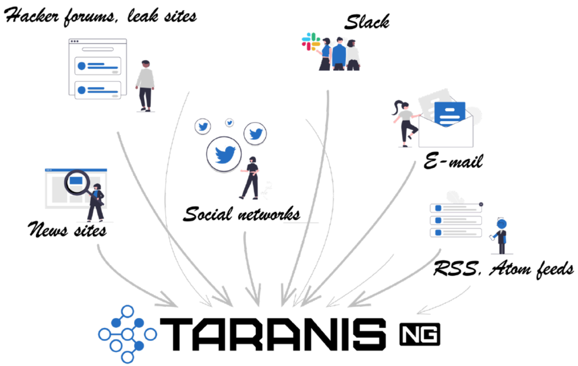


Sources: ICAC

- Challenge: Number of OSINT sources is high and the number of news items massive
  - Grasp quickly what’s relevant and omit the rest
  - Filter repetitive content
  - The workflow is actually pretty similar to a journalists work

TARANIS AI

- Based on *taranis3** and *taranis-ng***
  - Great tools to ingest raw unstructured data from various sources
  - Use human knowledge to identify relevant information
- Preserves the “taranis workflow” many CERTs are used to
  - Gather -> Assess -> Analyze -> Publish
- Introduces natural language processing (NLP) capabilities
  - Extraction of relevant **named entities**
  - **Clustering** of related **news items**
  - **Summaries** of “story clusters”
  - **Recommendations** of news items
  - Support for **creating OSINT products** (“reports”)



TARANIS_AI

https://taranis.ai/

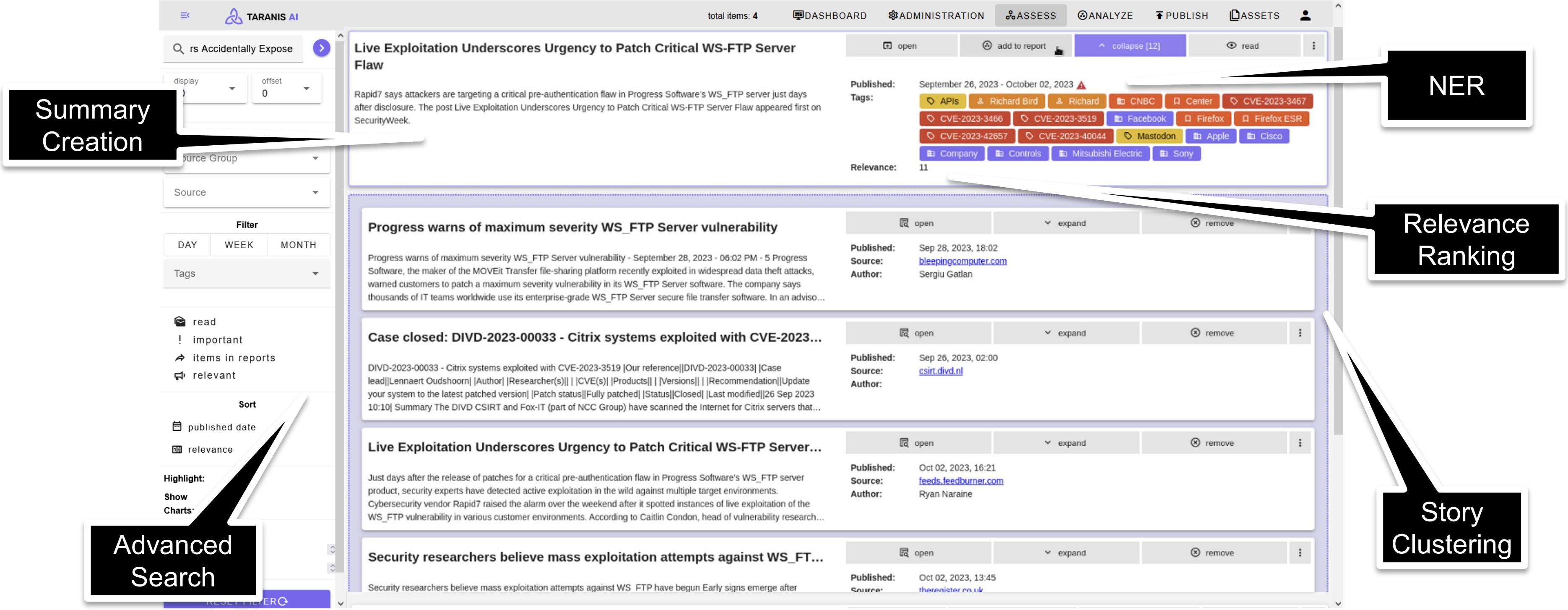
* https://github.com/NCSC-NL/taranis3

** https://github.com/SK-CERT/Taranis-NG

APPLYING THE POWER OF NLP

- **User Story 1:** *What are the 'hot topics' of the last 24 hours?*
- **User Story 2:** *What do we know about a specific entity? (e.g., a vulnerability, malware, company, product, person, location etc.)*
- **User Story 3:** *How can I find more related news items after reading this interesting article?*
- **User Story 4:** *Which news items are pertinent to my mission?*
- **User Story 5:** *How can I efficiently sum up my findings for my commander or operators of military IT services?*

NOVEL FEATURES AND DEVELOPMENTS



Live Exploitation Underscores Urgency to Patch Critical WS-FTP Server Flaw

Rapid7 says attackers are targeting a critical pre-authentication flaw in Progress Software's WS_FTP server just days after disclosure. The post Live Exploitation Underscores Urgency to Patch Critical WS-FTP Server Flaw appeared first on SecurityWeek.

Published: September 26, 2023 - October 02, 2023 ⚠

Tags: APIs, Richard Bird, Richard, CNBC, Center, CVE-2023-3467, CVE-2023-3466, CVE-2023-3519, Facebook, Firefox, Firefox ESR, CVE-2023-42657, CVE-2023-40044, Mastodon, Apple, Cisco, Company, Controls, Mitsubishi Electric, Sony

Relevance: 11

Progress warns of maximum severity WS_FTP Server vulnerability

Progress warns of maximum severity WS_FTP Server vulnerability - September 28, 2023 - 06:02 PM - 5 Progress Software, the maker of the MOVEit Transfer file-sharing platform recently exploited in widespread data theft attacks, warned customers to patch a maximum severity vulnerability in its WS_FTP Server software. The company says thousands of IT teams worldwide use its enterprise-grade WS_FTP Server secure file transfer software. In an adviso...

Published: Sep 28, 2023, 18:02
Source: bleepingcomputer.com
Author: Sergiu Gatlan

Case closed: DIVD-2023-00033 - Citrix systems exploited with CVE-2023...

DIVD-2023-00033 - Citrix systems exploited with CVE-2023-3519 |Our reference||DIVD-2023-00033| |Case lead||Lennaert Oudshoorn| |Author| |Researcher(s)|| | |CVE(s)| |Products|| | |Versions|| | |Recommendation||Update your system to the latest patched version| |Patch status||Fully patched| |Status||Closed| |Last modified||26 Sep 2023 10:10| Summary The DIVD CSIRT and Fox-IT (part of NCC Group) have scanned the Internet for Citrix servers that...

Published: Sep 26, 2023, 02:00
Source: csirt.divd.nl
Author:

Live Exploitation Underscores Urgency to Patch Critical WS-FTP Server...

Just days after the release of patches for a critical pre-authentication flaw in Progress Software's WS_FTP server product, security experts have detected active exploitation in the wild against multiple target environments. Cybersecurity vendor Rapid7 raised the alarm over the weekend after it spotted instances of live exploitation of the WS_FTP vulnerability in various customer environments. According to Caitlin Condon, head of vulnerability research...

Published: Oct 02, 2023, 16:21
Source: feeds.feedburner.com
Author: Ryan Naraine

Security researchers believe mass exploitation attempts against WS_FT...

Security researchers believe mass exploitation attempts against WS_FTP have begun Early signs emerge after

Published: Oct 02, 2023, 13:45
Source: theregister.co.uk

Summary Creation

NER

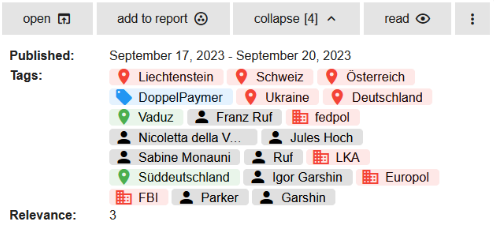
Relevance Ranking

Advanced Search

Story Clustering

FEATURES (1/5): NAMED ENTITY RECOGNITION

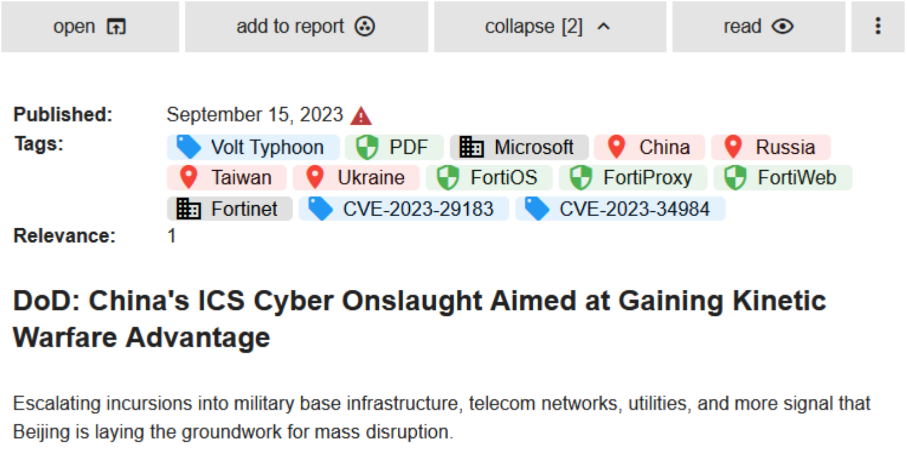
- Named-entity recognition (NER) seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, etc.
- Mix of pre-defined lists (e.g., countries), custom regex (e.g., CVEs), and trained language model on German/English standard text
- Custom extensions to recognize IT products, vendors, APT groups etc.
- *Future Extension:*
  - Additional use of domain-specific word lists



Published: September 17, 2023 - September 20, 2023

Tags: Liechtenstein, Schweiz, Österreich, DoppelPaymer, Ukraine, Deutschland, Vaduz, Franz Ruf, fedpol, Nicoletta della V..., Jules Hoch, Sabine Monauni, Ruf, LKA, Süddeutschland, Igor Garshin, Europol, FBI, Parker, Garshin

Relevance: 3



Published: September 15, 2023 ⚠

Tags: Volt Typhoon, PDF, Microsoft, China, Russia, Taiwan, Ukraine, FortiOS, FortiProxy, FortiWeb, Fortinet, CVE-2023-29183, CVE-2023-34984

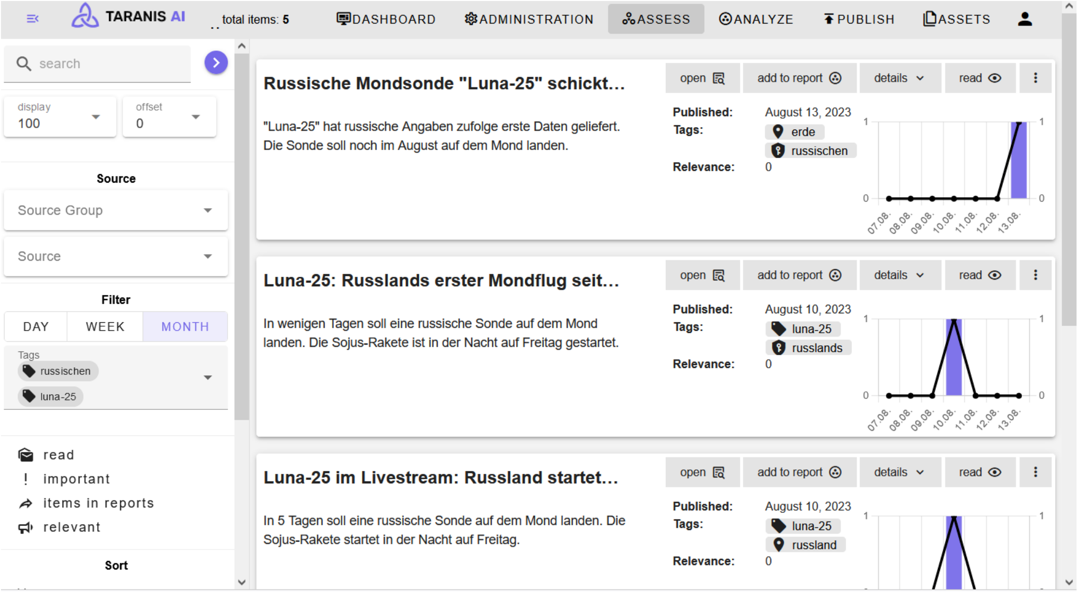
Relevance: 1

DoD: China's ICS Cyber Onslaught Aimed at Gaining Kinetic Warfare Advantage

Escalating incursions into military base infrastructure, telecom networks, utilities, and more signal that Beijing is laying the groundwork for mass disruption.

FEATURES (2/5): ADVANCED SEARCH & FILTERING

- Tags from NER can be used to filter and cluster items belonging to the same topic
- Additional free-text search
- Further filter and sort parameters
  - read/unread, relevance score etc.
  - Sources and source groups
  - Time spans
- *Future extension:*
  - Collaborative search and filtering
  - Sharing of filters



total items: 5

display: 100
offset: 0

Source Group
Source

Tags: russischen, luna-25

Russische Mondsonde "Luna-25" schickt...

"Luna-25" hat russische Angaben zufolge erste Daten geliefert. Die Sonde soll noch im August auf dem Mond landen.

Published: August 13, 2023
Tags: erde, russischen
Relevance: 0

Luna-25: Russlands erster Mondflug seit...

In wenigen Tagen soll eine russische Sonde auf dem Mond landen. Die Sojus-Rakete ist in der Nacht auf Freitag gestartet.

Published: August 10, 2023
Tags: luna-25, russlands
Relevance: 0

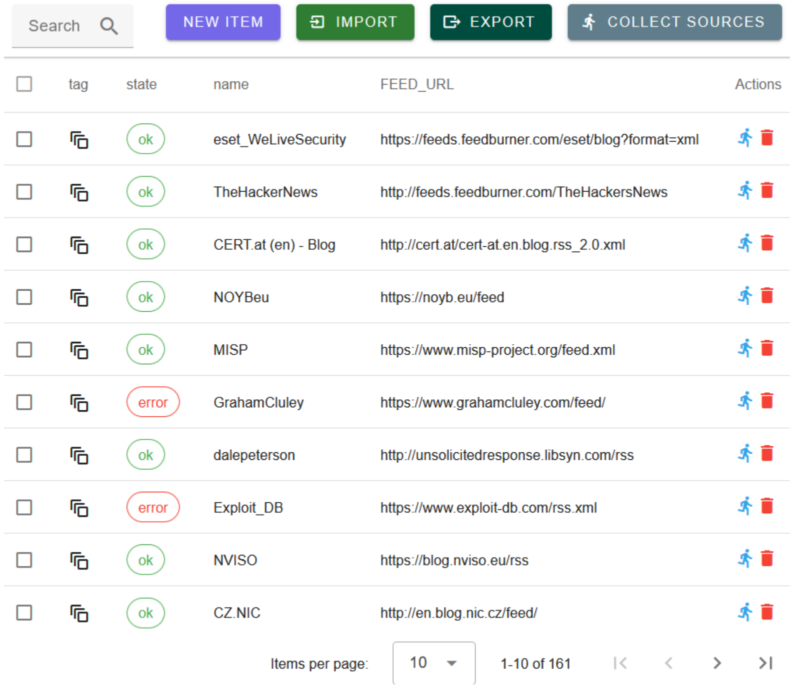
Luna-25 im Livestream: Russland startet...

In 5 Tagen soll eine russische Sonde auf dem Mond landen. Die Sojus-Rakete startet in der Nacht auf Freitag.

Published: August 10, 2023
Tags: luna-25, russland
Relevance: 0

FEATURES (3/5): RELEVANCE RANKING

- Relevance Ranking helps to identify “interesting items” (aka stories) based on general importance and personal preferences
  - Upvotes and downvotes from collaborators
  - Shared articles
  - Related news items
- *Future Extensions:*
  - Feedback Loop: Learn properties of often up-/downvotes items.
    - What properties do “good” articles have in common?
    - Which sources deliver such items?

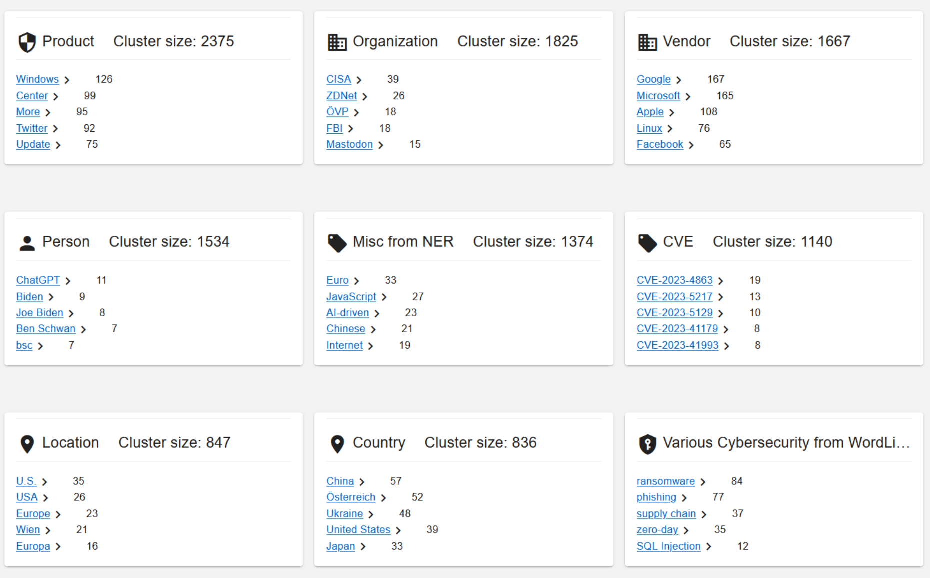


□	tag	state	name	FEED_URL	Actions
□	⧉	ok	eset_WeLiveSecurity	https://feeds.feedburner.com/eset/blog?format=xml	🏃 🗑
□	⧉	ok	TheHackerNews	http://feeds.feedburner.com/TheHackersNews	🏃 🗑
□	⧉	ok	CERT.at (en) - Blog	http://cert.at/cert-at.en.blog.rss_2.0.xml	🏃 🗑
□	⧉	ok	NOYBeu	https://noyb.eu/feed	🏃 🗑
□	⧉	ok	MISP	https://www.misp-project.org/feed.xml	🏃 🗑
□	⧉	error	GrahamCluley	https://www.grahamcluley.com/feed/	🏃 🗑
□	⧉	ok	dalepeterson	http://unsolicitedresponse.libsyn.com/rss	🏃 🗑
□	⧉	error	Exploit_DB	https://www.exploit-db.com/rss.xml	🏃 🗑
□	⧉	ok	NVISO	https://blog.nviso.eu/rss	🏃 🗑
□	⧉	ok	CZ.NIC	http://en.blog.nic.cz/feed/	🏃 🗑

Items per page: 10 1-10 of 161

FEATURES (4/5): SUMMARY & REPORT CREATION

- Summaries help to condense lengthy texts to their essential parts for quick decisions on their relevance
  - Summary of lengthy news items to quickly grasp its content
  - Summary of stories based on its collection of articles
- *Future Extensions*
  - Summary of Sharing Sets for reports
    - AI-assisted Pre-filling of report fields
  - Tuning of summaries regarding appropriate length, wording and content



Product Cluster size: 2375

Windows ›	126
Center ›	99
More ›	95
Twitter ›	92
Update ›	75

Organization Cluster size: 1825

CISA ›	39
ZDNet ›	26
ÖVP ›	18
FBI ›	18
Mastodon ›	15

Vendor Cluster size: 1667

Google ›	167
Microsoft ›	165
Apple ›	108
Linux ›	76
Facebook ›	65

Person Cluster size: 1534

ChatGPT ›	11
Biden ›	9
Joe Biden ›	8
Ben Schwan ›	7
bsc ›	7

Misc from NER Cluster size: 1374

Euro ›	33
JavaScript ›	27
AI-driven ›	23
Chinese ›	21
Internet ›	19

CVE Cluster size: 1140

CVE-2023-4863 ›	19
CVE-2023-5217 ›	13
CVE-2023-5129 ›	10
CVE-2023-41179 ›	8
CVE-2023-41993 ›	8

Location Cluster size: 847

U.S. ›	35
USA ›	26
Europe ›	23
Wien ›	21
Europa ›	16

Country Cluster size: 836

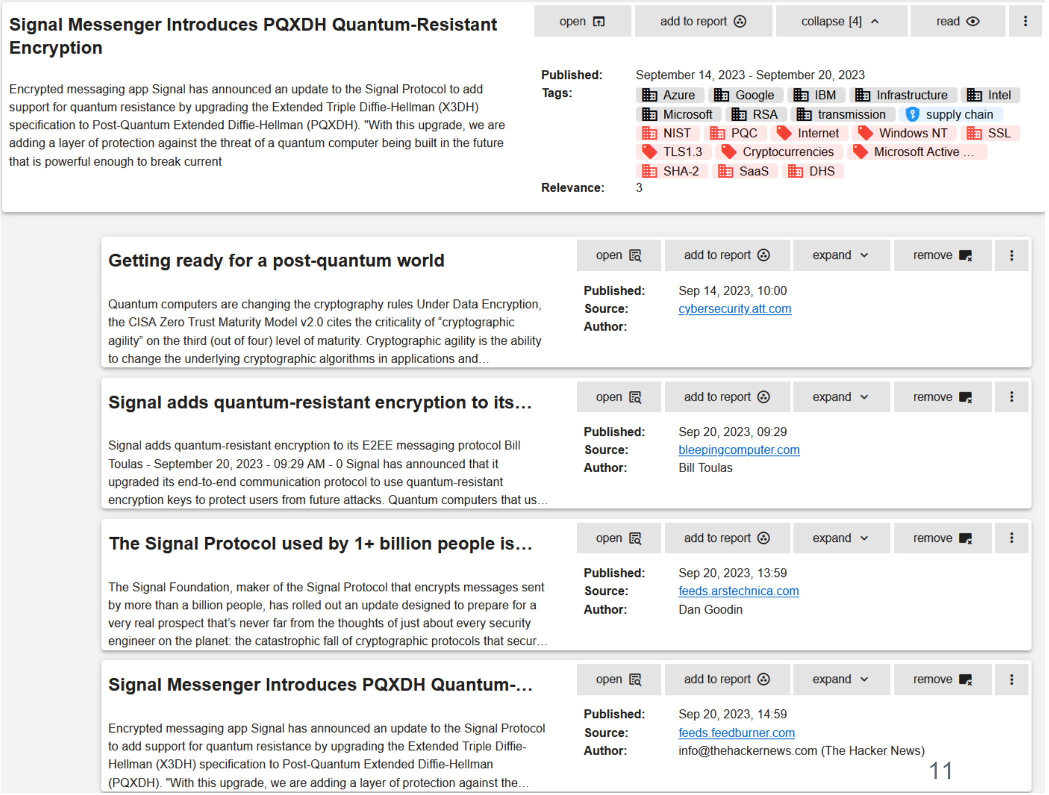
China ›	57
Österreich ›	52
Ukraine ›	48
United States ›	39
Japan ›	33

Various Cybersecurity from WordLi…

ransomware ›	84
phishing ›	77
supply chain ›	37
zero-day ›	35
SQL Injection ›	12

FEATURES (5/5): TOPIC & STORY CLUSTERING

- News Items are usually delivered not just by one, but multiple sources at approximately the same time with mostly similar content
- Cluster items and create “meta item” that summarizes important content (“story”)
  - Decrease human effort needed to ingest all news items!
  - Visualize development of a story over time
- Show development of “hot topics” over time



Signal Messenger Introduces PQXDH Quantum-Resistant Encryption

Encrypted messaging app Signal has announced an update to the Signal Protocol to add support for quantum resistance by upgrading the Extended Triple Diffie-Hellman (X3DH) specification to Post-Quantum Extended Diffie-Hellman (PQXDH). "With this upgrade, we are adding a layer of protection against the threat of a quantum computer being built in the future that is powerful enough to break current

Published: September 14, 2023 - September 20, 2023
Tags: Azure, Google, IBM, Infrastructure, Intel, Microsoft, RSA, transmission, supply chain, NIST, PQC, Internet, Windows NT, SSL, TLS1.3, Cryptocurrencies, Microsoft Active ..., SHA-2, SaaS, DHS
Relevance: 3

Getting ready for a post-quantum world

Quantum computers are changing the cryptography rules Under Data Encryption, the CISA Zero Trust Maturity Model v2.0 cites the criticality of “cryptographic agility” on the third (out of four) level of maturity. Cryptographic agility is the ability to change the underlying cryptographic algorithms in applications and...

Published: Sep 14, 2023, 10:00
Source: cybersecurity.att.com
Author:

Signal adds quantum-resistant encryption to its...

Signal adds quantum-resistant encryption to its E2EE messaging protocol Bill Toulas - September 20, 2023 - 09:29 AM - 0 Signal has announced that it upgraded its end-to-end communication protocol to use quantum-resistant encryption keys to protect users from future attacks. Quantum computers that us...

Published: Sep 20, 2023, 09:29
Source: bleepingcomputer.com
Author: Bill Toulas

The Signal Protocol used by 1+ billion people is...

The Signal Foundation, maker of the Signal Protocol that encrypts messages sent by more than a billion people, has rolled out an update designed to prepare for a very real prospect that's never far from the thoughts of just about every security engineer on the planet: the catastrophic fall of cryptographic protocols that secur...

Published: Sep 20, 2023, 13:59
Source: feeds.arstechnica.com
Author: Dan Goodin

Signal Messenger Introduces PQXDH Quantum-...

Encrypted messaging app Signal has announced an update to the Signal Protocol to add support for quantum resistance by upgrading the Extended Triple Diffie-Hellman (X3DH) specification to Post-Quantum Extended Diffie-Hellman (PQXDH). "With this upgrade, we are adding a layer of protection against the...

Published: Sep 20, 2023, 14:59
Source: feeds.feedburner.com
Author: info@thehackernews.com (The Hacker News)

TECHNOLOGY

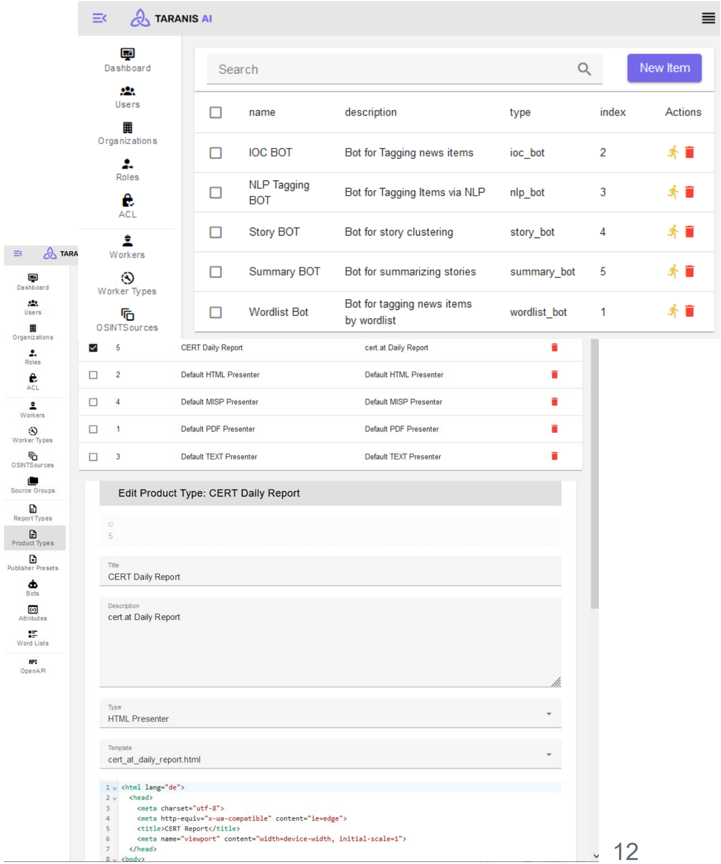
- Gathering from a **multitude of OSINT sources** via RSS, from the Web, e-Mail, and some APIs for common issue tracker
- **Asynchronous Pub-Sub architecture** with workers for flexible extensibility in terms of AI and NLP
  - Workers process items “best effort” in predefined order, e.g., extract IoCs, NER, storyclustering, summary creation etc.
- Resulting **products are text file, pdf, html** and pushed out via Mail, or to a **MISP** Server via API
- **100% open source** and free to use (EUPL license)
  - Please contribute! (issues, pull request, …)

OpenAPI

CELERY

python™

Vue.js



□	name	description	type	index	Actions
□	IOC BOT	Bot for Tagging news items	ioc_bot	2	
□	NLP Tagging BOT	Bot for Tagging Items via NLP	nlp_bot	3	
□	Story BOT	Bot for story clustering	story_bot	4	
□	Summary BOT	Bot for summarizing stories	summary_bot	5	
□	Wordlist Bot	Bot for tagging news items by wordlist	wordlist_bot	1	

☑	5	CERT Daily Report	cert.at Daily Report
□	2	Default HTML Presenter	Default HTML Presenter
□	4	Default MISP Presenter	Default MISP Presenter
□	1	Default PDF Presenter	Default PDF Presenter
□	3	Default TEXT Presenter	Default TEXT Presenter

Edit Product Type: CERT Daily Report

ID: 5

Title: CERT Daily Report

Description: cert.at Daily Report

Type: HTML Presenter

Template: cert_at_daily_report.html

```
<html lang="de">
  <head>
    <meta charset="utf-8">
    <meta http-equiv="x-ua-compatible" content="ie=edge">
    <title>CERT Report</title>
    <meta name="viewport" content="width=device-width, initial-scale=1">
  </head>
<body>
```

LESSONS LEARNED FROM FIRST PILOTS

The need for **high-quality gathering**

- „Garbage in – garbage out“
- Daily digests
- **Polluting texts** (menus, footers…)
- **Various formats**, changing layouts
- RSS v.s. HTML (in terms of tables, figures, etc.)
- **Near duplicates** & updated articles

The need for **adaptable and flexible workflows**

- Potentially **complex workflows** across different roles/departments
- **flexible use cases**
- „generic“ features, most important „**flags**“ (has been read, analyzed, reported, escalated, …)
- Preserve the „taranis workflow“, but be adaptable to different organizational structures

The need for **manually corrected AI-produced results**

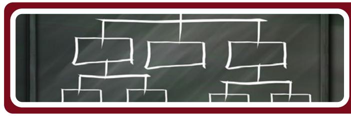
- AI can tremendously relieve human analysts from tiring tasks
- But **human element remains indispensable**
- Correction of AI-produced results is valuable **feedback to tune algorithms** (revising tags, removing items from story, splitting cluster…)
- Enhances the effectiveness of AI

The need for carefully **tuned recommender systems**

- Even with regular feedback through humans, **learned models remain intransparent**
- Importance of **explainability** – why was an item categorized in a certain way?
- Detect and **mitigate biases and concept drifts**
- **Re-evaluation of past decisions** is tricky (e.g., dismiss this story for good)

SCIENTIFIC CHALLENGES & NEW FEATURES AHEAD

improve **self-asset management** for alignment and matching of findings to own assets (prioritize/rank news items)



Taking further **context** into account, e.g., mission-specific information, to **provide targeted and easy to comprehend CTI** for efficient & informed decision making in tense and stressful situations

use of **LLMs to formulate situational awareness reports** for specific stakeholders in their own language, e.g. present information about vulnerabilities, major incidents, new risks for specific roles

automatic **notification of emerging threats**, e.g., because of dynamic growth of a story on a specific topic, **sentiment analysis** of social media posts that mention entities in a certain context

improve the **handling of classified information**, e.g., stories from OSINT that are being enriched with closed source information can only be partially disseminated and via pre-defined channels

improve **sharing capabilities** across instances of taranis-AI (currently analysts work all on one instance), standardized interfaces to common solutions in the domain, e.g., MISP

FINANCIAL SUPPORT AND COLLABORATIONS

European Commission

- The work in this presentation was mainly funded by the Connecting Europe Facility (CEF) program in course of the project **AWAKE** (2020- AT-IA-0254).
- The work in this paper has additionally received funding from the European Union - European Defence Fund under GA no. 101121418 (**EUCINF**) and 101121403 (**NEWSROOM**). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

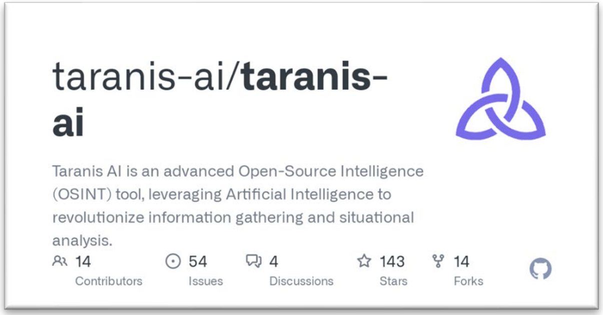
- Active collaboration with CERT.at, Austrian Ministry of Interior and Austrian MoD
  - Early adoption of taranis-ai by analysts
- Exchange with SK CERT, the developers of taranis-ng
- Open Source project on Github
  - Opportunity to contribute: https://taranis.ai/

Federal Ministry
Republic of Austria
Defence

Federal Ministry
Republic of Austria
Interior

CERT.at

SK CERT



taranis-ai/**taranis-ai**

Taranis AI is an advanced Open-Source Intelligence (OSINT) tool, leveraging Artificial Intelligence to revolutionize information gathering and situational analysis.

14 Contributors · 54 Issues · 4 Discussions · 143 Stars · 14 Forks

TARANIS_AI
https://taranis.ai/

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU!

Please contact:

Florian Skopik

florian.skopik@ait.ac.at

May, 30th 2024