# ORKL

building an archive for threat intelligence history

Robert Haist - FIRST CTI Symposium 2022

# About

**Robert Haist**

CISO @ TeamViewer

M.Sc. Advanced Security and Digital Forensics @ Edinburgh Napier University

Master Thesis: “*TIRAKL: an NLP assisted approach to curate OSINT Cyber Threat Intelligence News about Threat Actors*”

… foundation for this talk

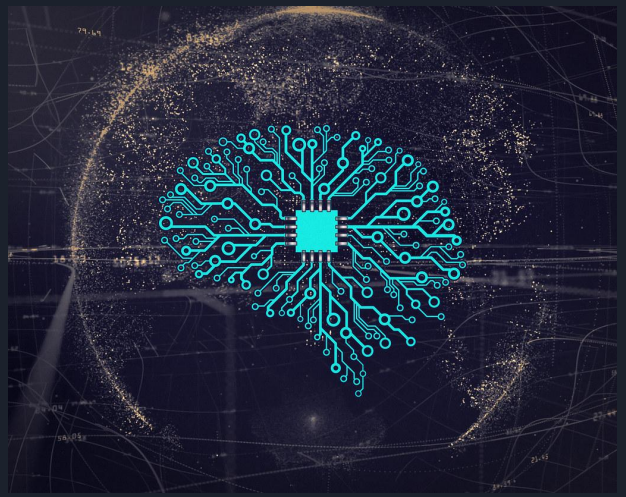
# Stupid Machines

AI / KI / NLP pipelines require clean data sets per knowledge domain for training / improvement

NLP frameworks come with pretrained models for Web / News - no Cyber

If we want to evolve from regex text matching to semantic research we need Cyber specific corpora

For academic verification they need to be public and reproducible



# TI Report Sources

There are many public (TLP:CLEAR) CTI Report sources with a largely varying degree of accessibility and context information.

Links to the original place of publication (i.e. cyber sec company blog) become inaccessible over time due to M&A etc.

**A lot of buried knowledge written by the sharpest minds of our community**

# Meet: ORKL


### Library Manager

Creates a reproducible file based corpus from different TI report sources

### ORKL API

Allows full-text searches on the corpus and related threat actor profiles

### ORKL Frontend

Basic web frontend to use the API interactively

*Disclaimer: I suck at JS*

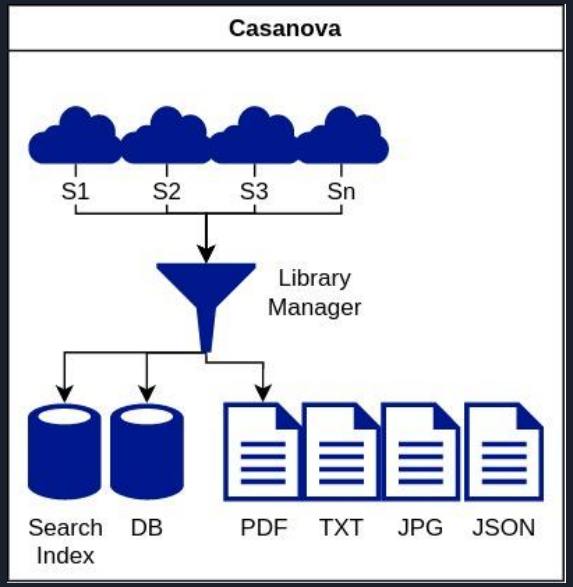
## ORKL Cyber Threat Intelligence Library

Search for relevant threat intelligence publication and see related threat actor profiles based on weighted synonyms

# The Casanova Quadrumvirate

For every report the library manager downloads from one of the sources it creates 4 files

- PDF → original file
- TXT → plain text representation
- JSON → metadata + plain text as JSON obj
- JPEG → image of the first page




If a report is in multiple sources it is only stored once in the library but with multiple source metadata records

Multiple report source metadata entries are merged into one library entry that is an intersection of all records

# Who are the baddies?

The library manager also acquires, stores and updates public Threat Actors profiles from a number of public sources.

Mainly interested in Threat Actor group {names, synonyms, aliases} - same for malicious tools.

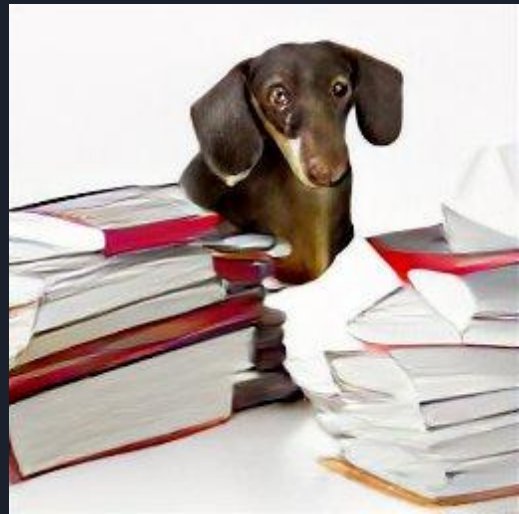
Reference to the original source is always kept.

Those {names, synonyms, aliases} are mapped to reports using the search index.

You can rank {names, synonyms, aliases} based on their frequency in the whole corpus (TF-IDF).

Hello
my name is

***Bureaucratic Dackel***



| | |
|---|---|
| **Malpedia** | https://malpedia.caad.fkie.fraunhofer.de |
| **Alienvault OTX** | https://otx.alienvault.com |
| **ETDA Threat Actor Library** | https://apt.etda.or.th |
| **CyberMonitor** | https://github.com/CyberMonitor/APT_CyberCriminal_Campagin_Collections |
| **APTNotes** | https://github.com/aptnotes |
| **SecureWorks** | https://www.secureworks.com/research/threat-profiles |
| **MITRE ATT&CK® Data** | https://github.com/mitre-attack/attack-stix-data |

# Sources for report leads and threat actor profiles

# ORKL API

Interact with the current library state

Full-Text search

Retrieve files (PLEASE BE REASONABLE)

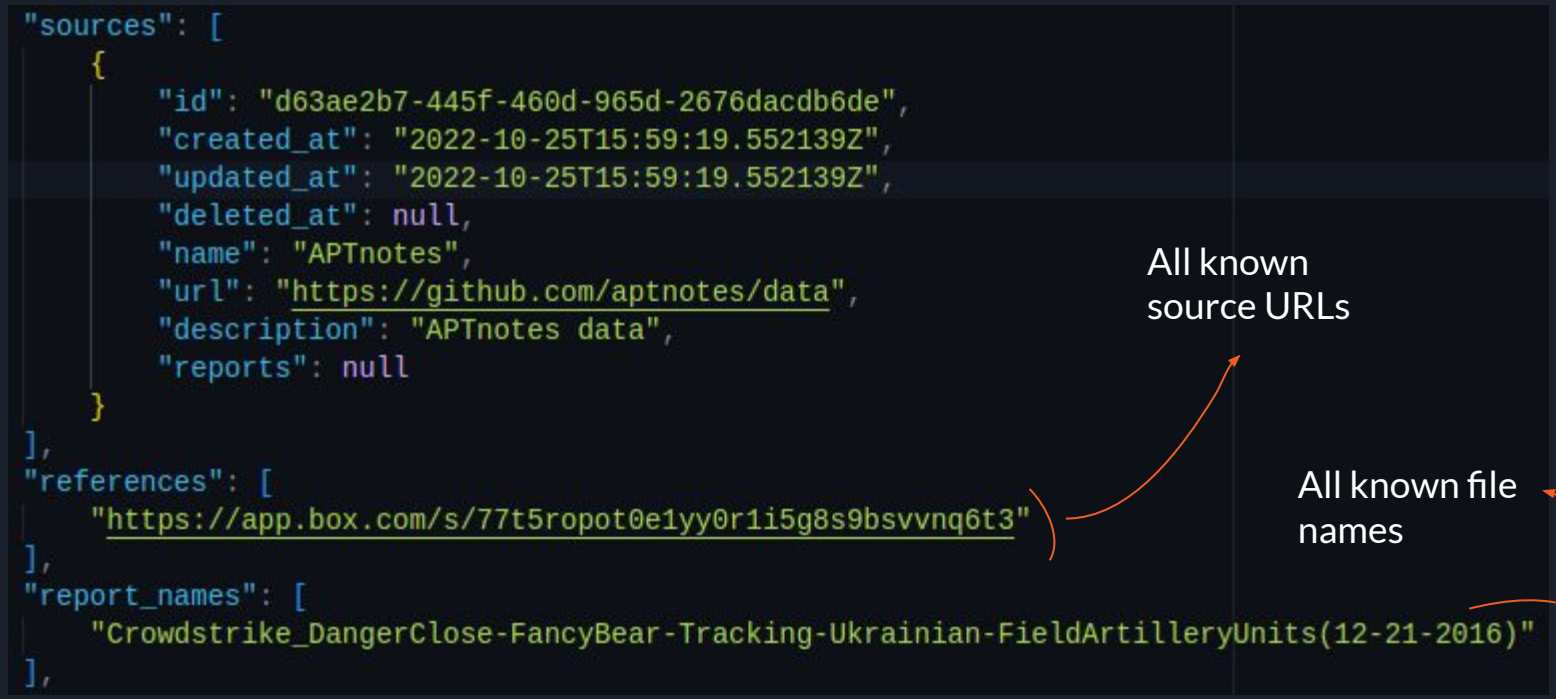
Get source information with each entry

Get threat actor matches with each entry

```
{
    "data": {
        "id": "67a2c542-0506-4eb8-8afd-20d0e757bf0c",
        "created_at": "2022-10-25T16:48:25.06851Z",
        "updated_at": "2022-10-28T13:16:04.976132Z",
        "deleted_at": null,
        "sha1_hash": "860387572ad036bfde33775ee89e7d92fa5d0aae",
        "title": "Danger Close: Fancy Bear Tracking of Ukrainian Field Artillery Units",
        "authors": "Crowdstrike",
        "file_creation_date": "2017-07-27T03:00:51Z",
        "file_modification_date": "0001-01-01T00:00:00Z",
        "file_size": 262427,
        "plain_text": "Danger Close: Fancy Bear Tracking of Ukrainian Field Artillery Units\n\n\n<SNIP>"
```


snip for readability

Example Library Entry: PDF/Source based Metadata



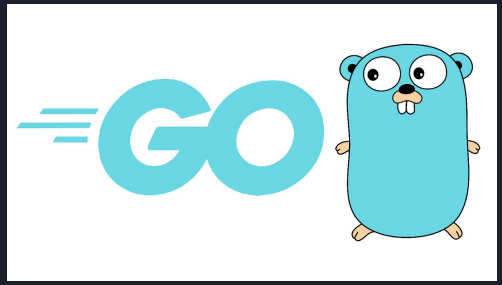

Example Entry: Continued


Combine Source and MainName to reference the Object throughout the App

Example Entry: Threat Actor association

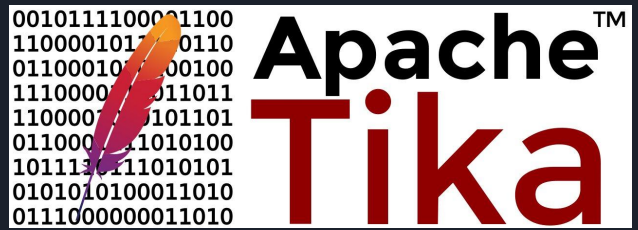

Example Entry: Threat Actor association

# Powered by Open Source ❤️







# Roadmap

## UI

Simple web UI to interact with the library content

- Search Reports
- Download Reports
- Threat Actors -> Reports

## Crowdsourced Librarian

Make ORKL it's own source for metadata and report uploads by the community

Allow the community to curate report metadata as a distributed effort

## NLP

Build cyber security centric NLP models for open source NLP frameworks based on the ORKL corpus (e.g. SpaCy, NLTK)

# Call to Action

**MORE SOURCES**

Which public CTI report sources am I missing?

**DISTRIBUTION**

… S3 storage / CDN that won’t bankrupt me :-)

**SHINY UI Help**

Frontend Wizards welcome

A logo would also be nice

# orkl.eu

Happy Testing :)

Follow for updates  @orkleu

# Contact

@RobertHaist

@rhaist

# Credits

Slide 3: Image via www.vpnsrus.com

Slide 4: Various front pages of sample CTI reports - copyright remains with the original authors

Slide 7: https://commons.wikimedia.org/wiki/File:Hello_my_name_is_sticker.svg

Slide 9: https://commons.wikimedia.org/wiki/File:Rijks_Museum_Library.jpg

Slide 14: Various Open Source Project logos - copyright remains with the original creators