Extracting Malware Information from Cybersecurity Text

Michael Maiden, Casey Hanks, Leyton Lineburg
University of Maryland, Baltimore County

OnRamp II Symposium
13-14 October 2021

Acknowledgements

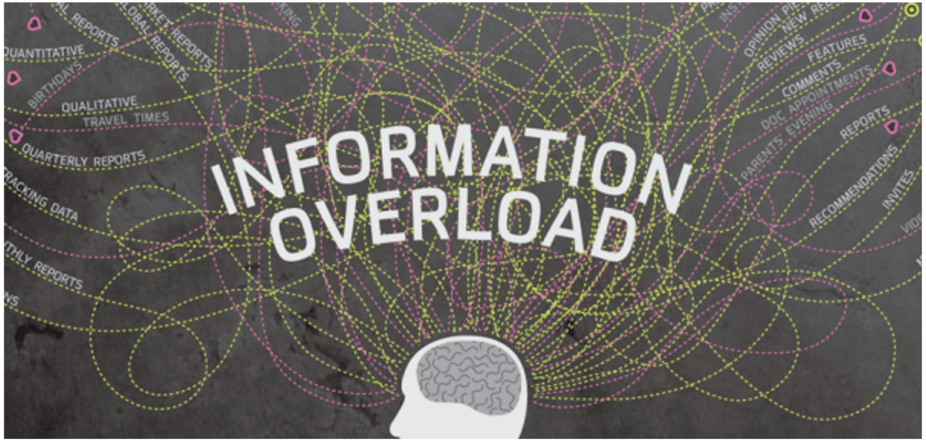
- Dr. Ahmad Ridley, NSA Moderator
- Dr. Anupam Joshi, Professor
- Dr. Tim Finin, Professor
- Priyanka Ranade, Graduate Student
- Aritran Piplai, Graduate Student

Thank you to the NSA for their support!

Research Goals and Background

Motivation

- Cyber Threat Intelligence (CTI) **aids cyber analysts** in their daily tasks of discovering and understanding evolving cybersecurity threats, exploits, and vulnerabilities
- Too much data to process!



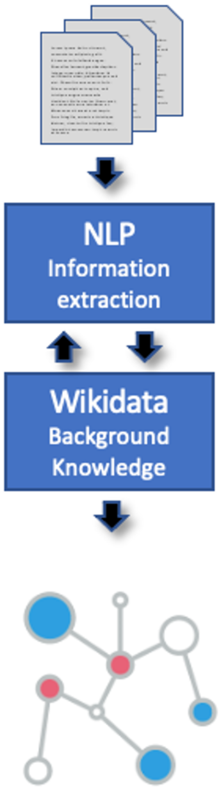
Research Question

Can we use state of the art NLP tools to process large corpora of CTI samples and provide *actionable insights* to analysts?

Overall Research Goal

Prototype a system to build and maintain a **knowledge graph** of current **cybersecurity data** extracted from public **text reports**

- Using state-of-the-art NLP systems, e.g., spaCy
- Using public knowledge graphs for background knowledge, e.g., Wikidata



Project Overviews

Use SpaCy for NLP Information Extraction:

- Data extraction and compilation
- SpaCy capabilities and utilization
- Progress, implementation, next steps

Use Wikidata for additional data:

- Catalog Wikidata's cyber knowledge and improve it
- Use its query service to access cyber information
- Improve our entity linker to link cyber entities to Wikidata

Background: Information Extraction for Cybersecurity

Types of Cybersecurity Text Data:

- Structured data - data that has a specific schema or set of rules for where data is stored, so it is easily searchable
- Unstructured data: Free-text, no standard content specification

Information Extraction is the process of extracting structured data from unstructured data

Background on Knowledge Graphs

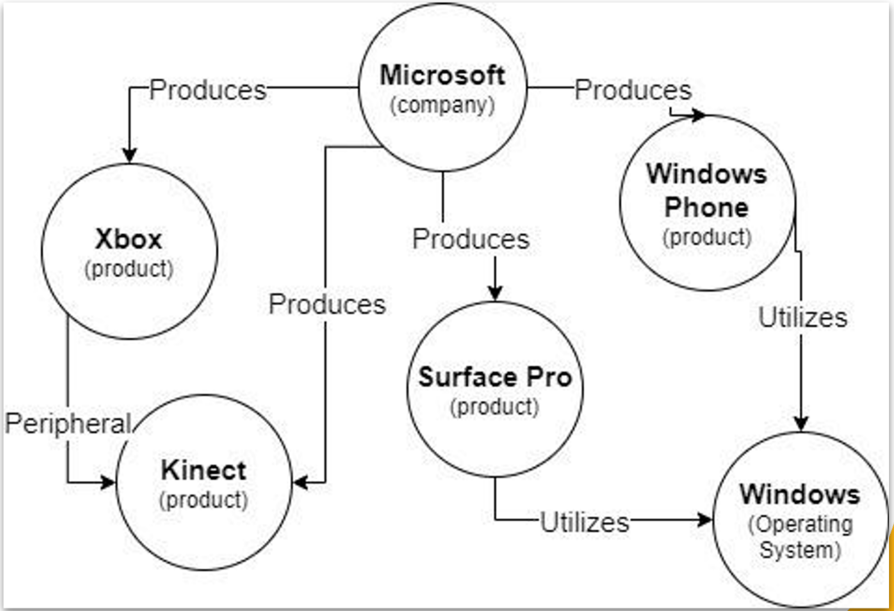
- A knowledge graph encodes data as a graph with entities as nodes and relations as edges
- Often represented as a set of ordered triples

  (head entity, relation, tail entity)

- Example:

  (Microsoft, Produces, XBOX)

  (XBOX, Peripheral, Kinect)



Background on Wikidata

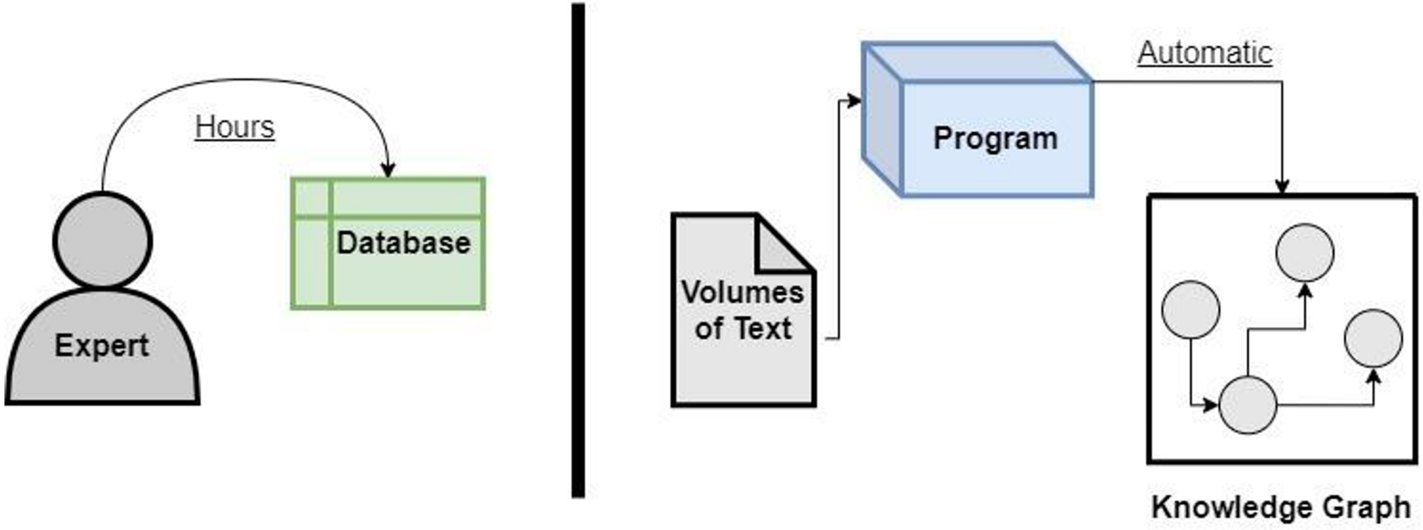
- What is Wikidata?
  - A knowledge graph derived from Wikipedias and more
  - A communal, open-source knowledge graph
- How is it accessed?
  - Has its own query service using SPARQL
- Cybersecurity knowledge
  - Has limited information available on cybersecurity
  - Each entity contains its own qualifiers

Michael Maiden, Casey Hanks

Using SpaCy to improve information extraction from cybersecurity text

Purpose

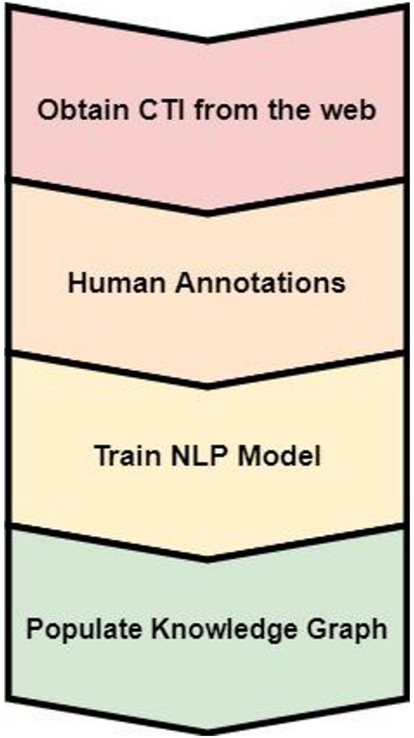
The purpose of this project is to create an automated method of transforming text based cybersecurity data to a computer friendly format



Knowledge Graphs help computers access the information faster and answer questions or Queries such as: What threat group utilized WannaCry malware?

Overview

- Cyber Threat Intelligence (CTI) is often disseminated as text
- Expressing CTI information in a knowledge graph allows integration, inference, and machine understanding
- We use synthesized data, language processing and machine learning to do this



Obtain CTI from the web

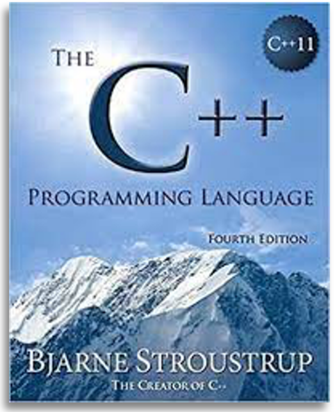
Human Annotations

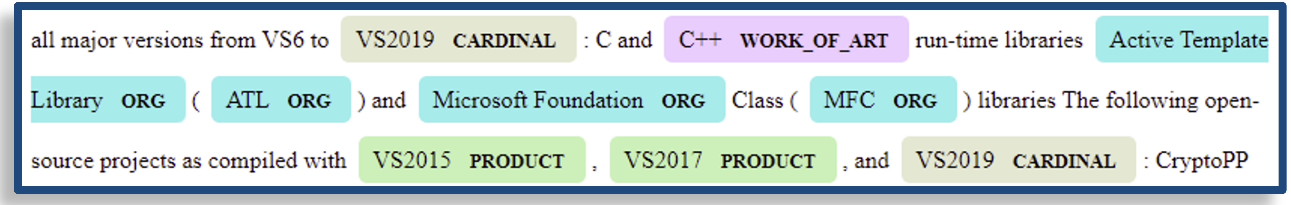
Train NLP Model

Populate Knowledge Graph

Extracting Information from Text

- Natural Language Processing (NLP) tools extract entities such as relations and events from text
- These tools are trained to recognize
  - People, Places, and Time
- Not trained for extracting cybersecurity entities
  - Malware
  - Software





all major versions from VS6 to VS2019 CARDINAL : C and C++ WORK_OF_ART run-time libraries Active Template Library ORG (ATL ORG) and Microsoft Foundation ORG Class (MFC ORG) libraries The following open-source projects as compiled with VS2015 PRODUCT , VS2017 PRODUCT , and VS2019 CARDINAL : CryptoPP

Extracting Cyber Information

Existing NLP tools have not seen cybersecurity entities before, so must be re-trained for the cybersecurity domain by

- Collecting relevant text examples
- Getting human annotations for entities & relations
- Training machine learning tools to recognize mentions of entities, relations, and events

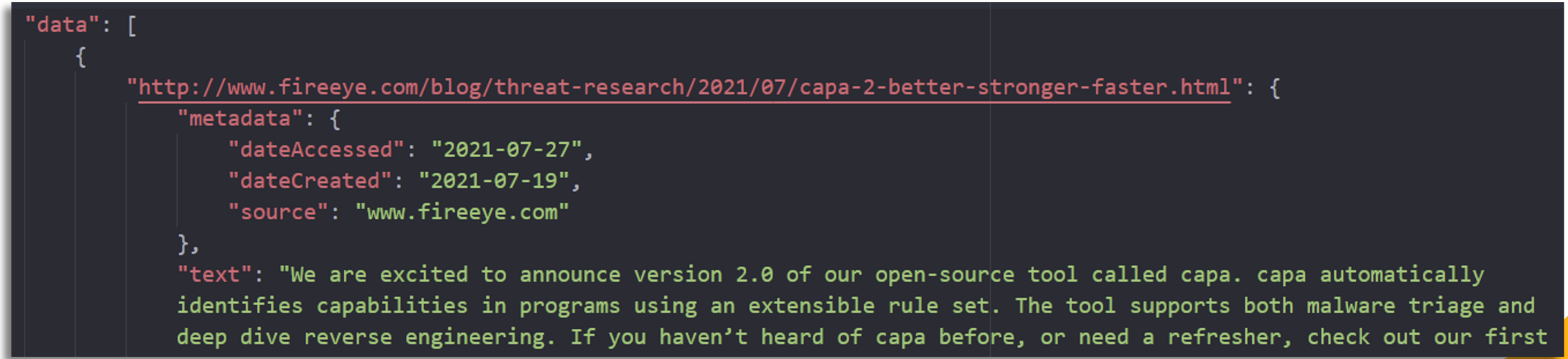
Methodology

Building a Cybersecurity Text Collection

- Collected open-source cyber-security text from a variety of sources
  - 3,269 paragraphs have been collected across 8 sources
- Developing tools to automatically update corpus as cyber exploits, vulnerabilities, and news sources continue to evolve

Blogs in Our Text Collection

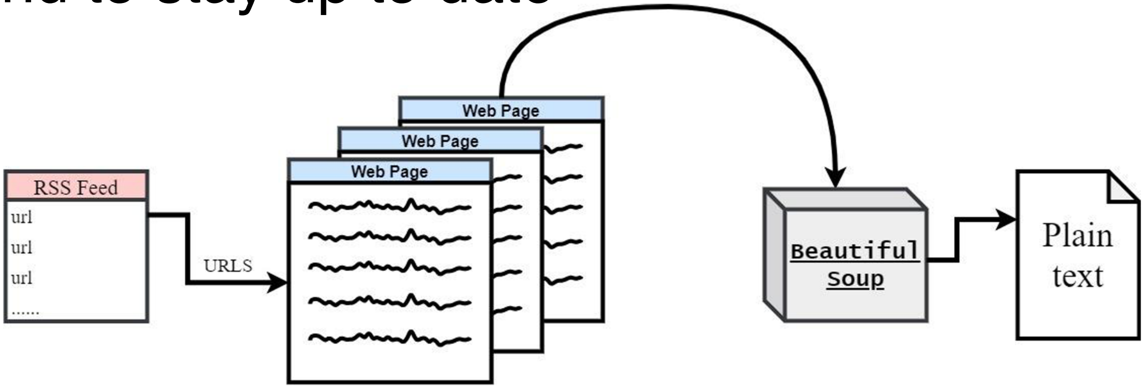
- IBM Security Intelligence
- McAfee
- Fortinet
- Kaspersky
- Securonix
- Palo Alto Networks
- Juniper Networks
- Fireeye



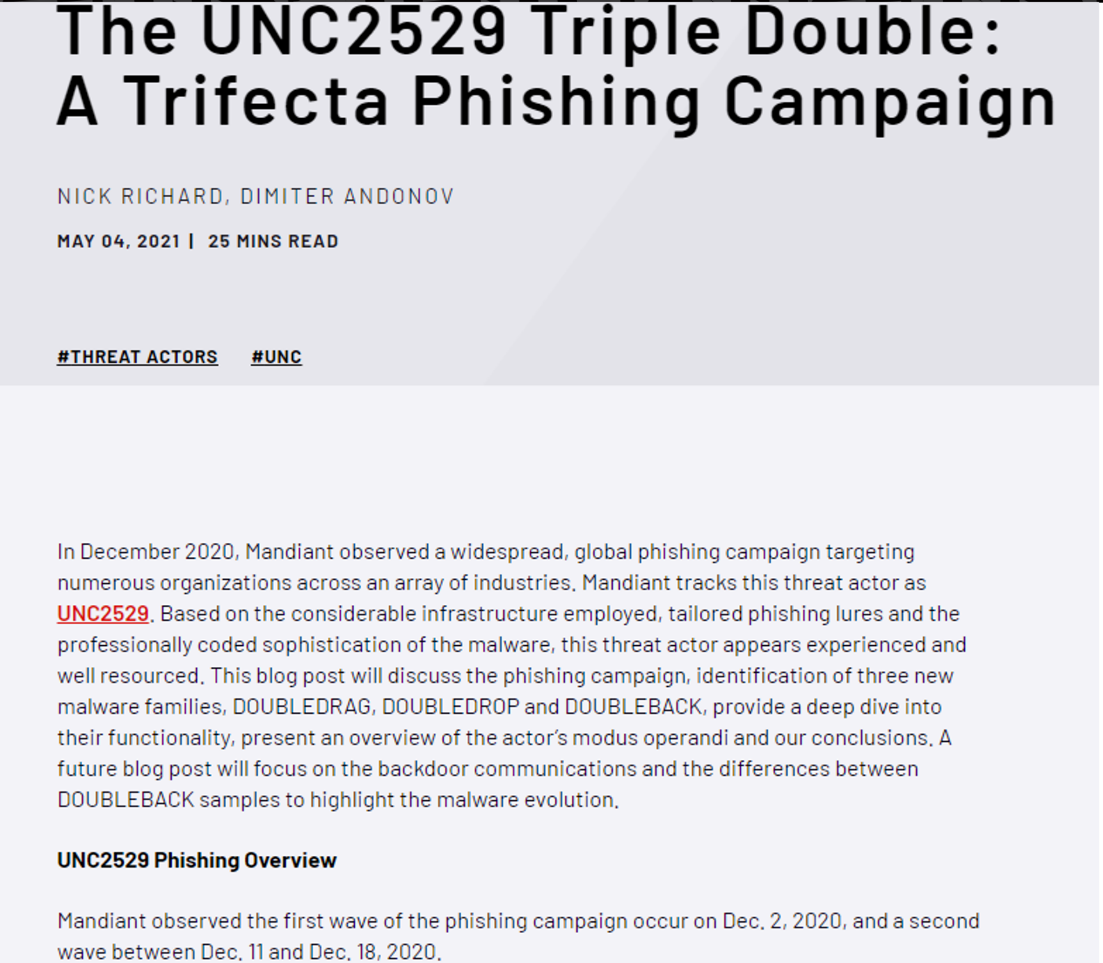
```
"data": [
    {
        "http://www.fireeye.com/blog/threat-research/2021/07/capa-2-better-stronger-faster.html": {
            "metadata": {
                "dateAccessed": "2021-07-27",
                "dateCreated": "2021-07-19",
                "source": "www.fireeye.com"
            },
            "text": "We are excited to announce version 2.0 of our open-source tool called capa. capa automatically
            identifies capabilities in programs using an extensible rule set. The tool supports both malware triage and
            deep dive reverse engineering. If you haven’t heard of capa before, or need a refresher, check out our first
```

Extracting text from web pages

- CTI information is extracted from multiple cybersecurity blogs using Beautiful Soup, a Python library for pulling data out of HTML sources
- RSS feeds of websites are used to extract URLs of new articles and to stay up to date



Example Article in Our Collection



The UNC2529 Triple Double: A Trifecta Phishing Campaign

NICK RICHARD, DIMITER ANDONOV

MAY 04, 2021 | 25 MINS READ

#THREAT ACTORS #UNC

In December 2020, Mandiant observed a widespread, global phishing campaign targeting numerous organizations across an array of industries. Mandiant tracks this threat actor as UNC2529. Based on the considerable infrastructure employed, tailored phishing lures and the professionally coded sophistication of the malware, this threat actor appears experienced and well resourced. This blog post will discuss the phishing campaign, identification of three new malware families, DOUBLEDRAG, DOUBLEDROP and DOUBLEBACK, provide a deep dive into their functionality, present an overview of the actor's modus operandi and our conclusions. A future blog post will focus on the backdoor communications and the differences between DOUBLEBACK samples to highlight the malware evolution.

UNC2529 Phishing Overview

Mandiant observed the first wave of the phishing campaign occur on Dec. 2, 2020, and a second wave between Dec. 11 and Dec. 18, 2020.

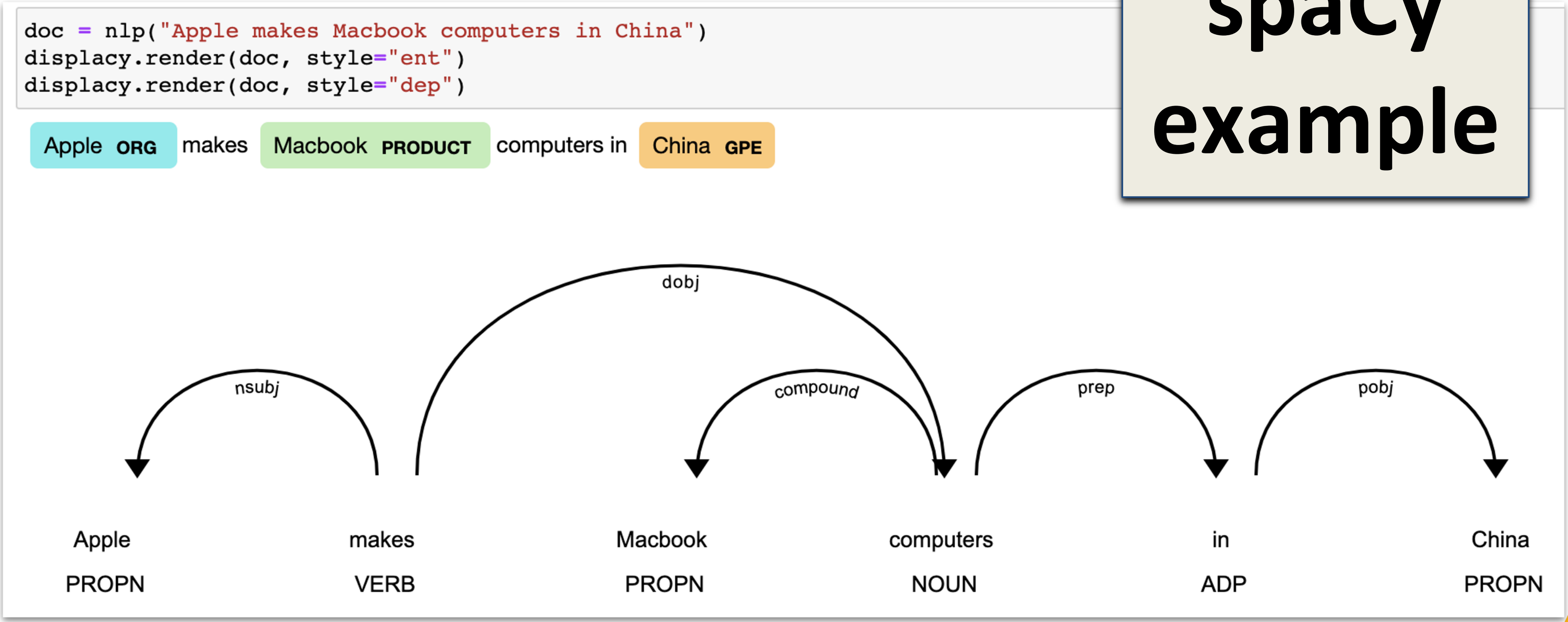
MANDIANT

spaCy NLP Tools

SpaCy is an state of the art, open source, NLP tool that supports

- Tokenization: segmenting text into sentences and words
- Syntactic analysis: parsing text to identify parts of speech, phrases, and sentence structure
- Text Similarity: computes similarity of text spans via embeddings
- Named Entity Recognition: Identify text spans for entities and their types, e.g., Apple is an entity of type Organization
- Training an NER model: with proper training data, SpaCy can be trained to identify new entity types and their entity labels
- Annotated text: the format of the training data

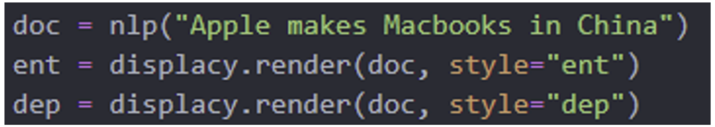
spaCy example



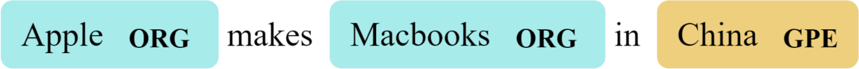
```
doc = nlp("Apple makes Macbook computers in China")
displacy.render(doc, style="ent")
displacy.render(doc, style="dep")
```

Apple **ORG** makes Macbook **PRODUCT** computers in China **GPE**

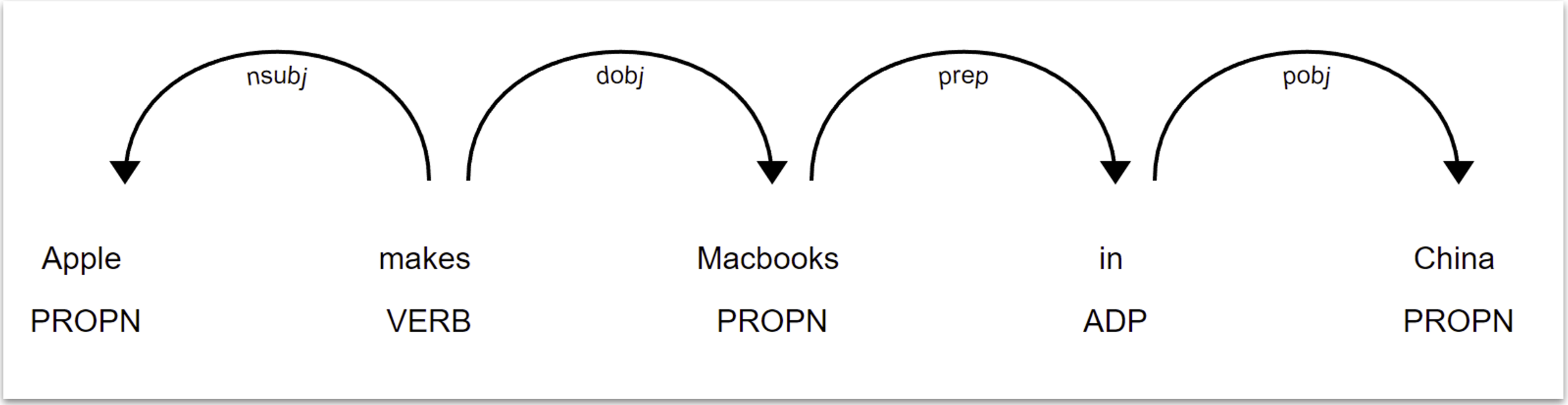
spaCy example



```
doc = nlp("Apple makes Macbooks in China")
ent = displacy.render(doc, style="ent")
dep = displacy.render(doc, style="dep")
```

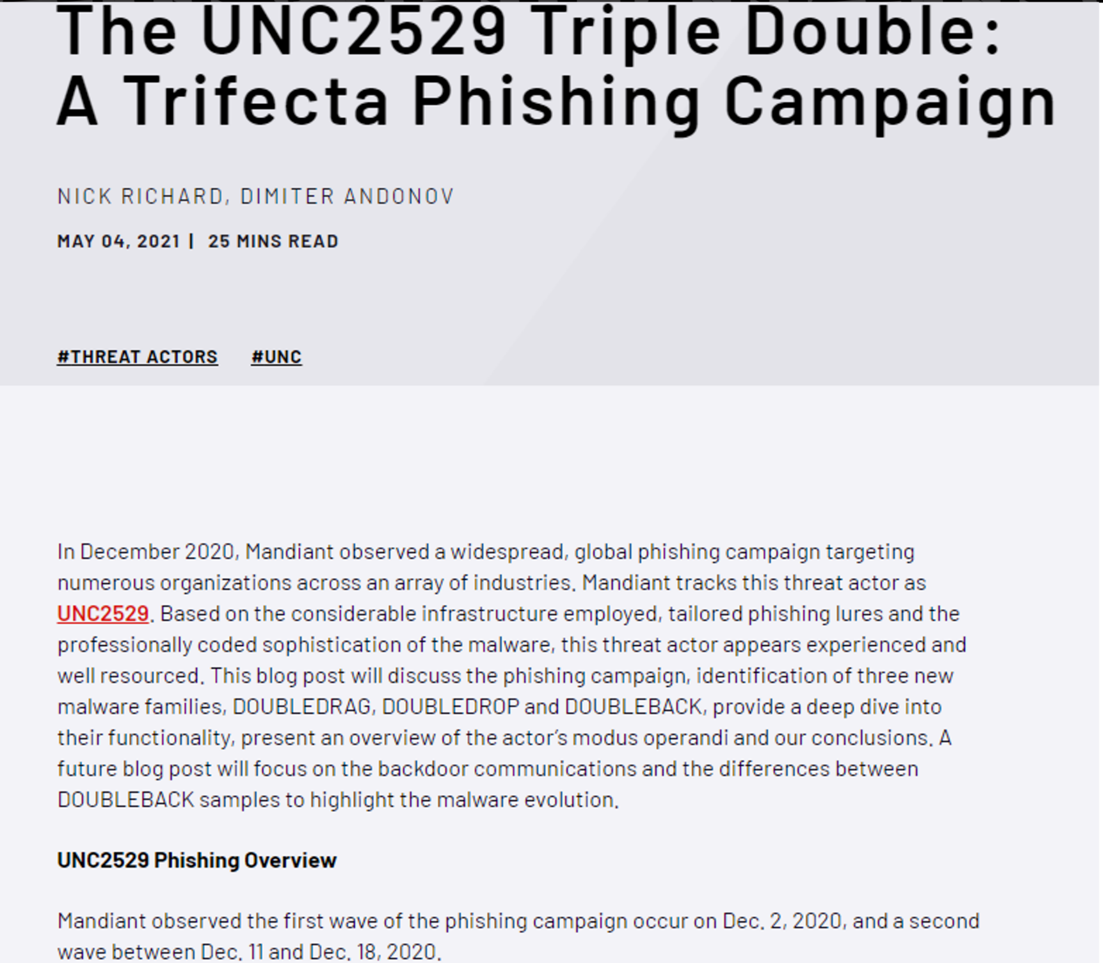


Apple **ORG** makes Macbooks **ORG** in China **GPE**



- spaCy’s generic model doesn’t recognize Macbook as a product or computer
- Need to train on cybersecurity text and/or link items to background knowledge like Wikidata (noun chunks)

Example Article in Our Collection



The UNC2529 Triple Double: A Trifecta Phishing Campaign

NICK RICHARD, DIMITER ANDONOV

MAY 04, 2021 | 25 MINS READ

#THREAT ACTORS #UNC

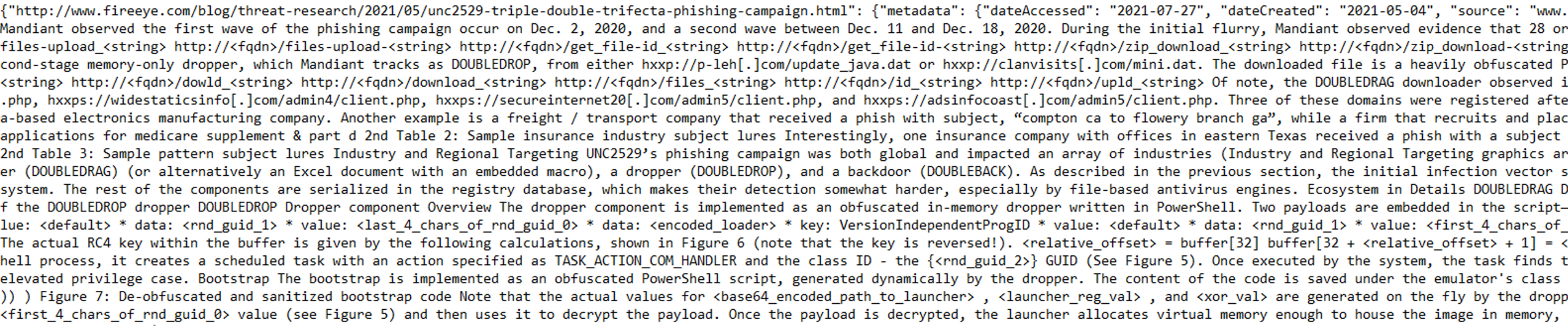
In December 2020, Mandiant observed a widespread, global phishing campaign targeting numerous organizations across an array of industries. Mandiant tracks this threat actor as UNC2529. Based on the considerable infrastructure employed, tailored phishing lures and the professionally coded sophistication of the malware, this threat actor appears experienced and well resourced. This blog post will discuss the phishing campaign, identification of three new malware families, DOUBLEDRAG, DOUBLEDROP and DOUBLEBACK, provide a deep dive into their functionality, present an overview of the actor's modus operandi and our conclusions. A future blog post will focus on the backdoor communications and the differences between DOUBLEBACK samples to highlight the malware evolution.

UNC2529 Phishing Overview

Mandiant observed the first wave of the phishing campaign occur on Dec. 2, 2020, and a second wave between Dec. 11 and Dec. 18, 2020.

MANDIANT

How the Text is Stored



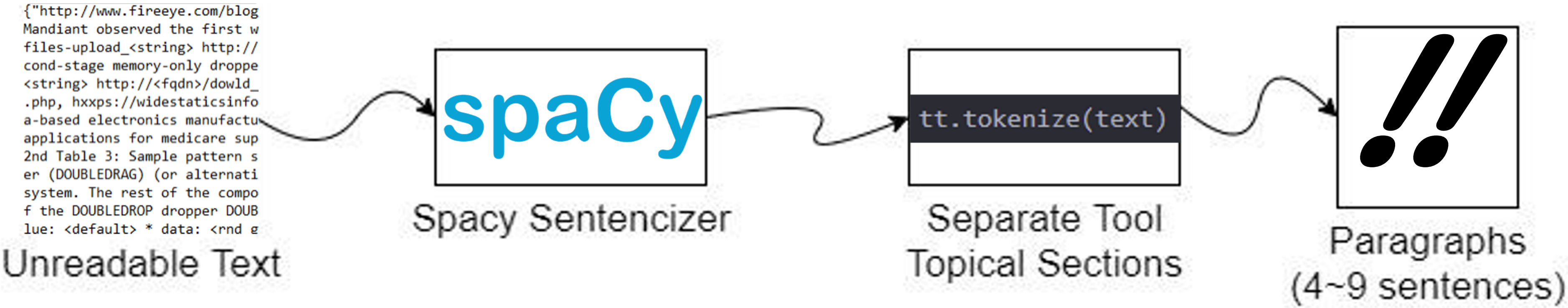
```
{"http://www.fireeye.com/blog/threat-research/2021/05/unc2529-triple-double-trifecta-phishing-campaign.html": {"metadata": {"dateAccessed": "2021-07-27", "dateCreated": "2021-05-04", "source": "www.
Mandiant observed the first wave of the phishing campaign occur on Dec. 2, 2020, and a second wave between Dec. 11 and Dec. 18, 2020. During the initial flurry, Mandiant observed evidence that 28 or
files-upload_<string> http://<fqdn>/files-upload-<string> http://<fqdn>/get_file-id_<string> http://<fqdn>/get_file-id-<string> http://<fqdn>/zip_download_<string> http://<fqdn>/zip_download-<string
cond-stage memory-only dropper, which Mandiant tracks as DOUBLEDROP, from either hxxp://p-leh[.]com/update_java.dat or hxxp://clanvisits[.]com/mini.dat. The downloaded file is a heavily obfuscated P
<string> http://<fqdn>/dowld_<string> http://<fqdn>/download_<string> http://<fqdn>/files_<string> http://<fqdn>/id_<string> http://<fqdn>/upld_<string> Of note, the DOUBLEDRAG downloader observed i
.php, hxxps://widestaticsinfo[.]com/admin4/client.php, hxxps://secureinternet20[.]com/admin5/client.php, and hxxps://adsinfocoast[.]com/admin5/client.php. Three of these domains were registered afte
a-based electronics manufacturing company. Another example is a freight / transport company that received a phish with subject, “compton ca to flowery branch ga”, while a firm that recruits and plac
applications for medicare supplement & part d 2nd Table 2: Sample insurance industry subject lures Interestingly, one insurance company with offices in eastern Texas received a phish with a subject
2nd Table 3: Sample pattern subject lures Industry and Regional Targeting UNC2529’s phishing campaign was both global and impacted an array of industries (Industry and Regional Targeting graphics ar
er (DOUBLEDRAG) (or alternatively an Excel document with an embedded macro), a dropper (DOUBLEDROP), and a backdoor (DOUBLEBACK). As described in the previous section, the initial infection vector s
system. The rest of the components are serialized in the registry database, which makes their detection somewhat harder, especially by file-based antivirus engines. Ecosystem in Details DOUBLEDRAG D
f the DOUBLEDROP dropper DOUBLEDROP Dropper component Overview The dropper component is implemented as an obfuscated in-memory dropper written in PowerShell. Two payloads are embedded in the script—
lue: <default> * data: <rnd_guid_1> * value: <last_4_chars_of_rnd_guid_0> * data: <encoded_loader> * key: VersionIndependentProgID * value: <default> * data: <rnd_guid_1> * value: <first_4_chars_of_
The actual RC4 key within the buffer is given by the following calculations, shown in Figure 6 (note that the key is reversed!). <relative_offset> = buffer[32] buffer[32 + <relative_offset> + 1] = <
hell process, it creates a scheduled task with an action specified as TASK_ACTION_COM_HANDLER and the class ID - the {<rnd_guid_2>} GUID (See Figure 5). Once executed by the system, the task finds t
elevated privilege case. Bootstrap The bootstrap is implemented as an obfuscated PowerShell script, generated dynamically by the dropper. The content of the code is saved under the emulator's class
)) ) Figure 7: De-obfuscated and sanitized bootstrap code Note that the actual values for <base64_encoded_path_to_launcher> , <launcher_reg_val> , and <xor_val> are generated on the fly by the dropp
<first_4_chars_of_rnd_guid_0> value (see Figure 5) and then uses it to decrypt the payload. Once the payload is decrypted, the launcher allocates virtual memory enough to house the image in memory,
```

- No newlines
- No paragraphs
- Hardly readable

Making it readable

- For human annotation we ideally have paragraphs
- Easily digestible, can easily annotate one or two and stop
- Humans identify paragraphs by white space, what if there is none?
- Spacy is good at identifying sentences, but not so much paragraphs.

The Solution



```
{"http://www.fireeye.com/blog
Mandiant observed the first w
files-upload_<string> http://
cond-stage memory-only droppe
<string> http://<fqdn>/dowld_
.php, hxxps://widestaticsinfo
a-based electronics manufactu
applications for medicare sup
2nd Table 3: Sample pattern s
er (DOUBLEDRAG) (or alternati
system. The rest of the compo
f the DOUBLEDROP dropper DOUB
lue: <default> * data: <rnd g
```

Unreadable Text

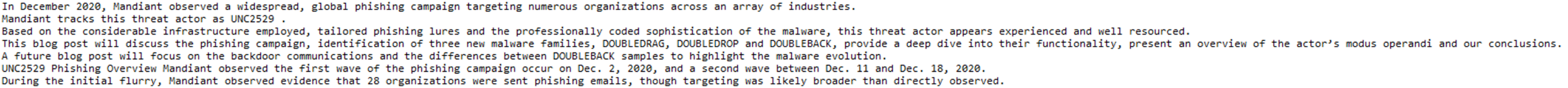
Spacy Sentencizer

```
tt.tokenize(text)
```

Separate Tool Topical Sections

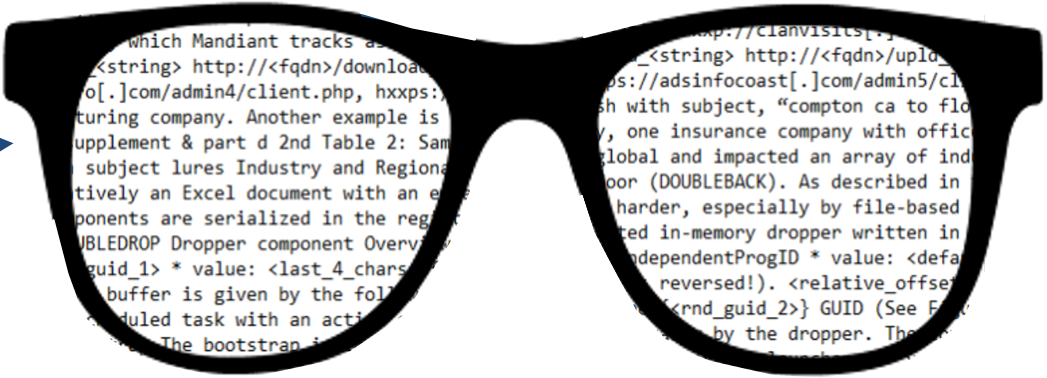
Paragraphs (4~9 sentences)

Much better!



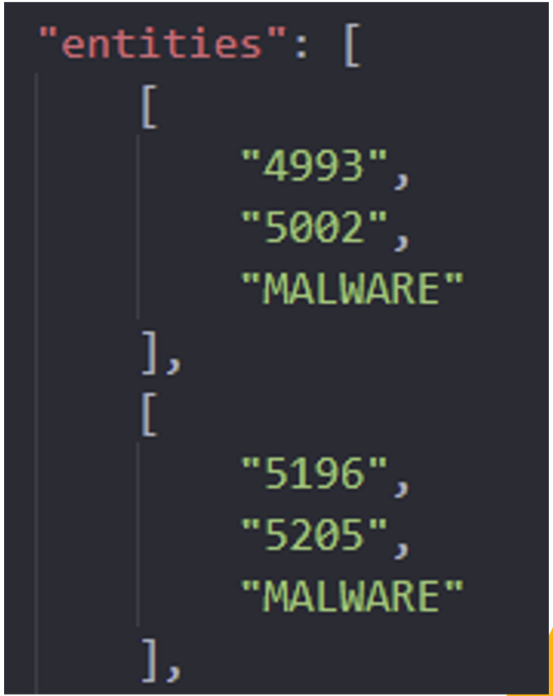
```
In December 2020, Mandiant observed a widespread, global phishing campaign targeting numerous organizations across an array of industries.
Mandiant tracks this threat actor as UNC2529 .
Based on the considerable infrastructure employed, tailored phishing lures and the professionally coded sophistication of the malware, this threat actor appears experienced and well resourced.
This blog post will discuss the phishing campaign, identification of three new malware families, DOUBLEDRAG, DOUBLEDROP and DOUBLEBACK, provide a deep dive into their functionality, present an overview of the actor’s modus operandi and our conclusions.
A future blog post will focus on the backdoor communications and the differences between DOUBLEBACK samples to highlight the malware evolution.
UNC2529 Phishing Overview Mandiant observed the first wave of the phishing campaign occur on Dec. 2, 2020, and a second wave between Dec. 11 and Dec. 18, 2020.
During the initial flurry, Mandiant observed evidence that 28 organizations were sent phishing emails, though targeting was likely broader than directly observed.
```

None of this


which Mandiant tracks a
<string> http://<fqdn>/downloa
o[.]com/admin4/client.php, hxxps:
turing company. Another example is
upplement & part d 2nd Table 2: Sam
subject lures Industry and Regiona
tively an Excel document with an e
ponents are serialized in the reg
BLEDROP Dropper component Overv
guid_1> * value: <last_4_chars
buffer is given by the fol
duled task with an acti
The bootstrap i
<string> http://<fqdn>/upld
ps://adsinfocoast[.]com/admin5/cl
sh with subject, “compton ca to flo
, one insurance company with offic
global and impacted an array of ind
oor (DOUBLEBACK). As described in
harder, especially by file-based
ted in-memory dropper written in
ndependentProgID * value: <defa
reversed!). <relative_offse
<rnd_guid_2>} GUID (See F
by the dropper. The

Human Annotations for Training Data

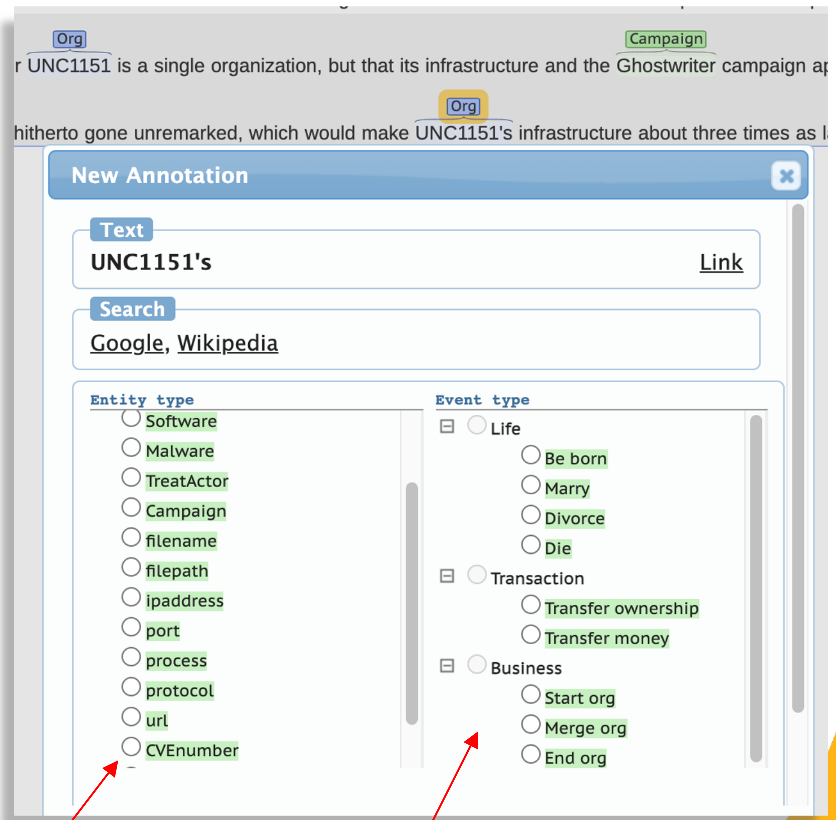
- Machines need many examples to learn
- Multiple humans act as entity recognizer and record occurrences of the desired label
  - (Ex. “MALWARE”)
- Annotation tools are often used to make the process more efficient



```
"entities": [
    [
        "4993",
        "5002",
        "MALWARE"
    ],
    [
        "5196",
        "5205",
        "MALWARE"
    ],
```

Brat Rapid Annotation Tool

- BRAT is a web-based tool that allows users to annotate text
- Annotations identify named entities, relations, and events
- BRAT outputs annotation data in a standard format
- Used to train SpaCy models to extract new entities and relations



r UNC1151 is a single organization, but that its infrastructure and the Ghostwriter campaign ap

hitherto gone unremarked, which would make UNC1151's infrastructure about three times as l

New Annotation

Text: UNC1151's — Link

Search: Google, Wikipedia

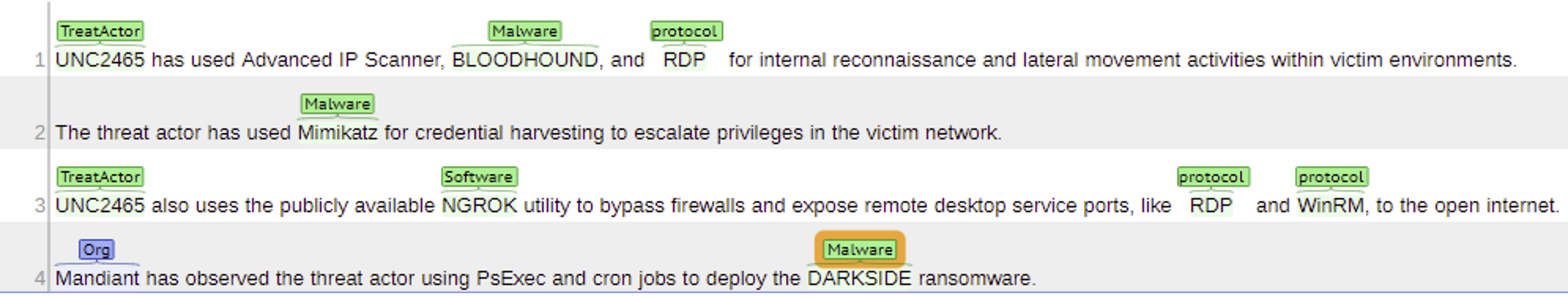
Entity type
○ Software
○ Malware
○ TreatActor
○ Campaign
○ filename
○ filepath
○ ipaddress
○ port
○ process
○ protocol
○ url
○ CVEnumber

Event type
⊟ ○ Life
○ Be born
○ Marry
○ Divorce
○ Die
⊟ ○ Transaction
○ Transfer ownership
○ Transfer money
⊟ ○ Business
○ Start org
○ Merge org
○ End org

possible entities

possible relations

BRAT Example



UNC2465 has used Advanced IP Scanner, BLOODHOUND, and RDP for internal reconnaissance and lateral movement activities within victim environments.

The threat actor has used Mimikatz for credential harvesting to escalate privileges in the victim network.

UNC2465 also uses the publicly available NGROK utility to bypass firewalls and expose remote desktop service ports, like RDP and WinRM, to the open internet.

Mandiant has observed the threat actor using PsExec and cron jobs to deploy the DARKSIDE ransomware.

A team of human annotators identify entities and relations and assign cybersecurity-relevant types using our central BRAT server

Training SpaCy tools

- SpaCy tools allow us to train Cybersecurity NER models using the BRAT annotations
- The trained model will be able to recognize cybersecurity entities like mentions of malware or threat actors
- Users are able to combine their own models with SpaCy's existing models
  - We can fine-tune an existing SpACY model to fit cybersecurity data

Creating CTI Knowledge Graphs

- Nodes in the knowledge graph will be filled with cybersecurity named entities
  - Ex. Malware, Software, Threat Actors, Campaigns
- Edges represent relations between two entities
  - Threat actors **utilize** malware
- Cybersecurity professionals can use the Knowledge Graphs as additional tools for understanding threat-related information

Preliminary Results & Next Steps

☑ Compiled collection of cybersecurity-related texts

☑ Compiling human annotation materials

□ Train SpaCY Cybersecurity NER model using labeled, annotated data

□ Evaluate the accuracy of the NER model

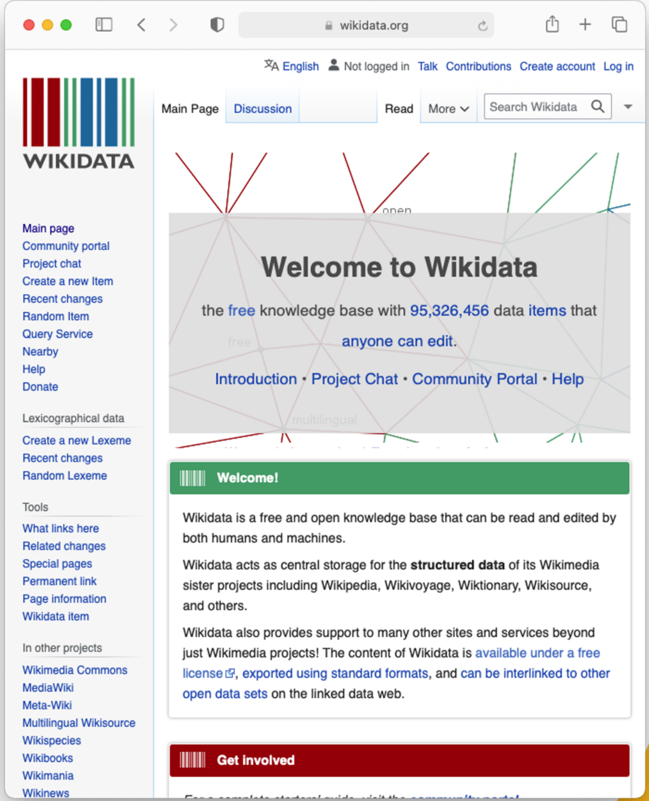
□ Use named entities detected with NER model to strengthen existing cybersecurity knowledge graph

Leyton Lineburg, Mike Anoruo

Surveying and Improving Cybersecurity Knowledge in Wikidata

Wikidata Overview

- Wikidata is an open-source knowledge graph of general background knowledge
- It has both general and useful cyber-security related knowledge
- We can use it to help train and improve our NLP information extraction systems
- And to develop plans to improve it and keep it current as a cybersecurity resource



Welcome to Wikidata

the free knowledge base with 95,326,456 data items that anyone can edit.

Introduction • Project Chat • Community Portal • Help

Welcome!

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others.

Wikidata also provides support to many other sites and services beyond just Wikimedia projects! The content of Wikidata is available under a free license, exported using standard formats, and can be interlinked to other open data sets on the linked data web.

Get involved

Methodology

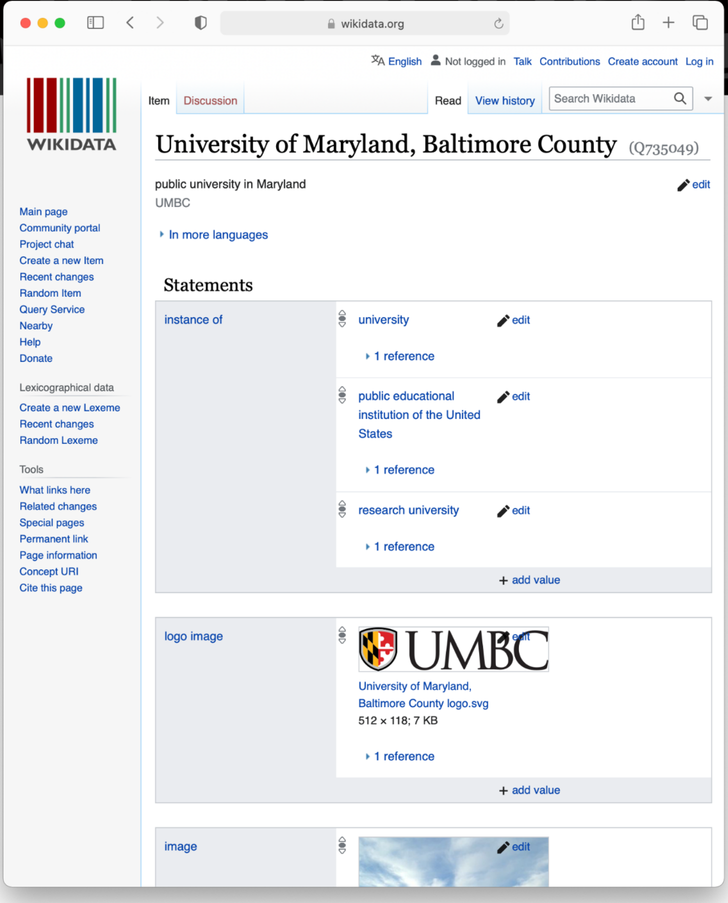
1. Query and retrieve cyber-related information
2. Manually search for errors and corrections
3. Update Wikidata periodically to keep information accurate and relevant
4. Link entities found in text to Wikidata entities to promote data integration

Knowledge Graphs (KGs)

- Google uses the largely private Google Knowledge Graph to better understand queries and web pages
- Wikidata is an open-source KG we can use for similar purposes
- It has a very rich schema and a powerful underlying semantic representation based on RDF and links to other semantic KGs, like DBpedia

Wikidata

- A community knowledge graph with ~1B statements about ~100M items
- Fine-grained ontology has ~2M types and ~9K properties
- Multilingual: all text values tagged with language id
- Has both a human and query interface
- Many community tools for editing, search, visualization, update



University of Maryland, Baltimore County (Q735049)

public university in Maryland

UMBC

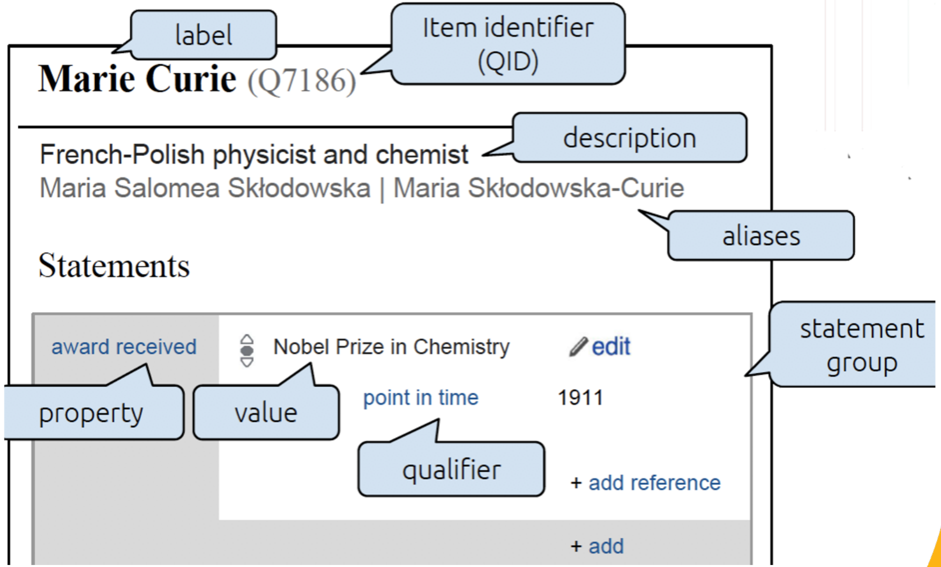
Statements

instance of	university 1 reference public educational institution of the United States 1 reference research university 1 reference
logo image	University of Maryland, Baltimore County logo.svg 512 × 118; 7 KB 1 reference
image	

Wikidata web interface for the UMBC entity, Q735049

Wikidata Items

- Unique ID, e.g. Q7186
- label (canonical name) in multiple languages
- Short description and some aliases in multiple languages
- One or more types, e.g., Q5 for Human
- Collection of statements with optional qualifiers, references
- Links to entries in WIkipedia sites & other knowledge graphs



Marie Curie (Q7186)

French-Polish physicist and chemist

Maria Salomea Skłodowska | Maria Skłodowska-Curie

Statements

award received | Nobel Prize in Chemistry | edit

point in time | 1911

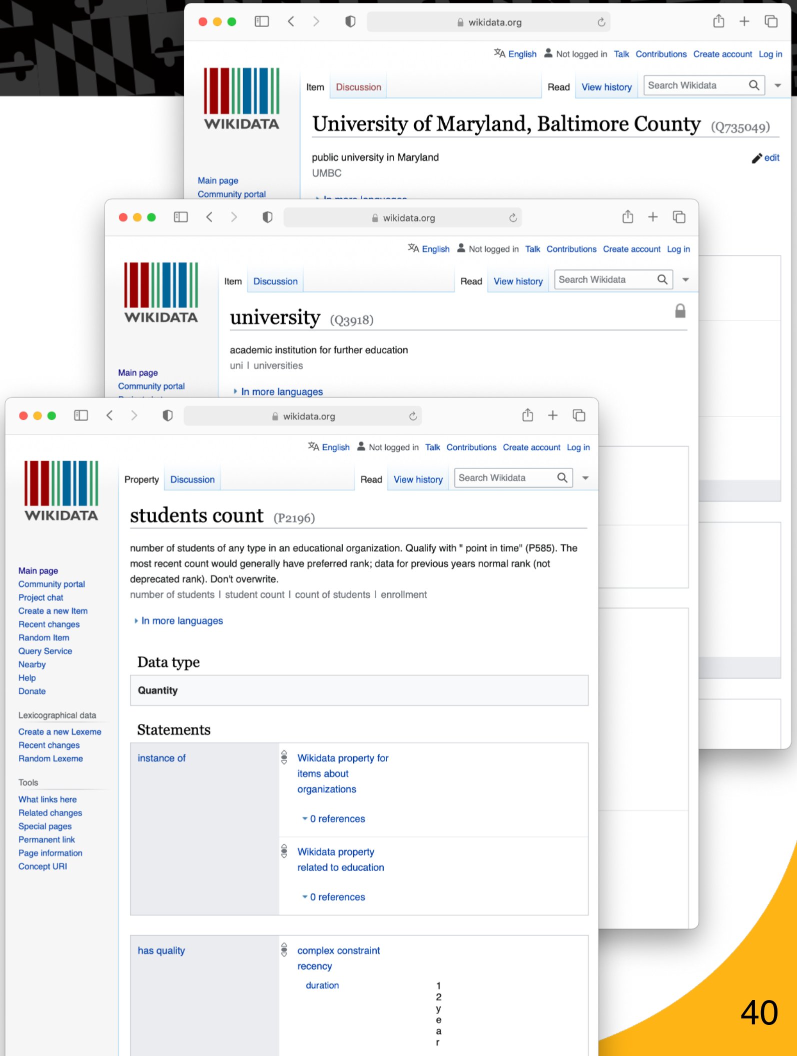
+ add reference

+ add

UMBC

Concepts & Properties

- Wikidata items like UMBC (Q735049) refer to entities, e.g., instances of a concept
- Some Wikidata refer to a concept or type, e.g., University (Q3918)
- Others refer to properties, e.g., students count (P2196)
- These concepts form Wikidata's rich ontology



University of Maryland, Baltimore County (Q735049)

public university in Maryland
UMBC

university (Q3918)

academic institution for further education
uni | universities

students count (P2196)

number of students of any type in an educational organization. Qualify with " point in time" (P585). The most recent count would generally have preferred rank; data for previous years normal rank (not deprecated rank). Don't overwrite.
number of students | student count | count of students | enrollment

Data type

Quantity

Statements

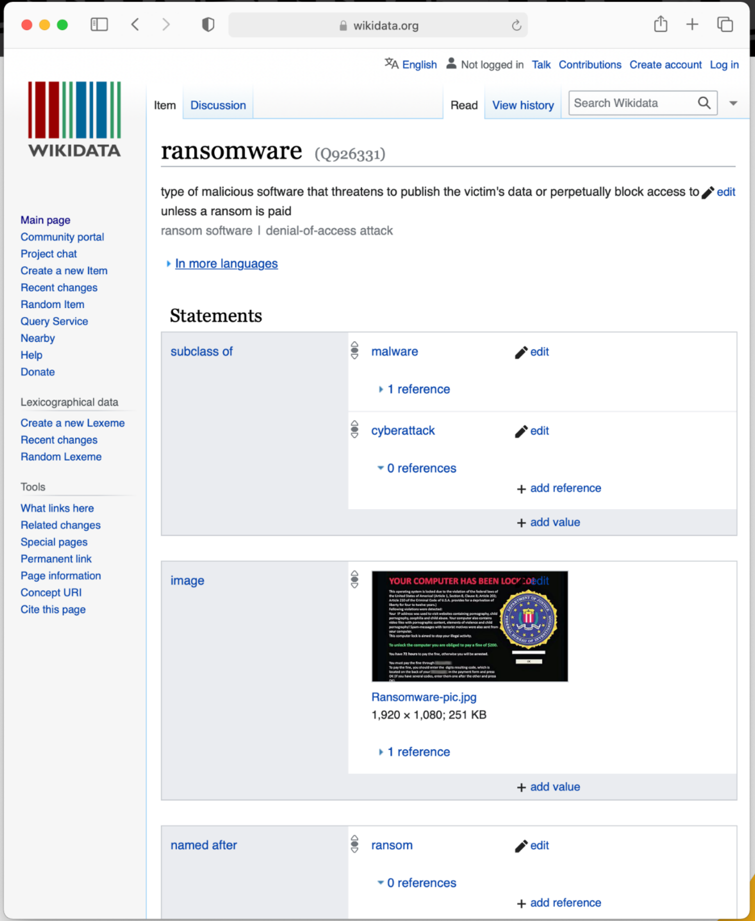
instance of: Wikidata property for items about organizations (0 references); Wikidata property related to education (0 references)

has quality: complex constraint recency
duration: 1 year

UMBC

Cybersecurity Concept

- Wikidata has data on many cyber-security related concepts
- Including, but not limited to:
  - Malware
  - Types of malware
  - Cyber attacks
  - Known threat actors
  - Operating system concepts
  - Key software products



ransomware (Q926331)

type of malicious software that threatens to publish the victim's data or perpetually block access to it unless a ransom is paid edit

ransom software | denial-of-access attack

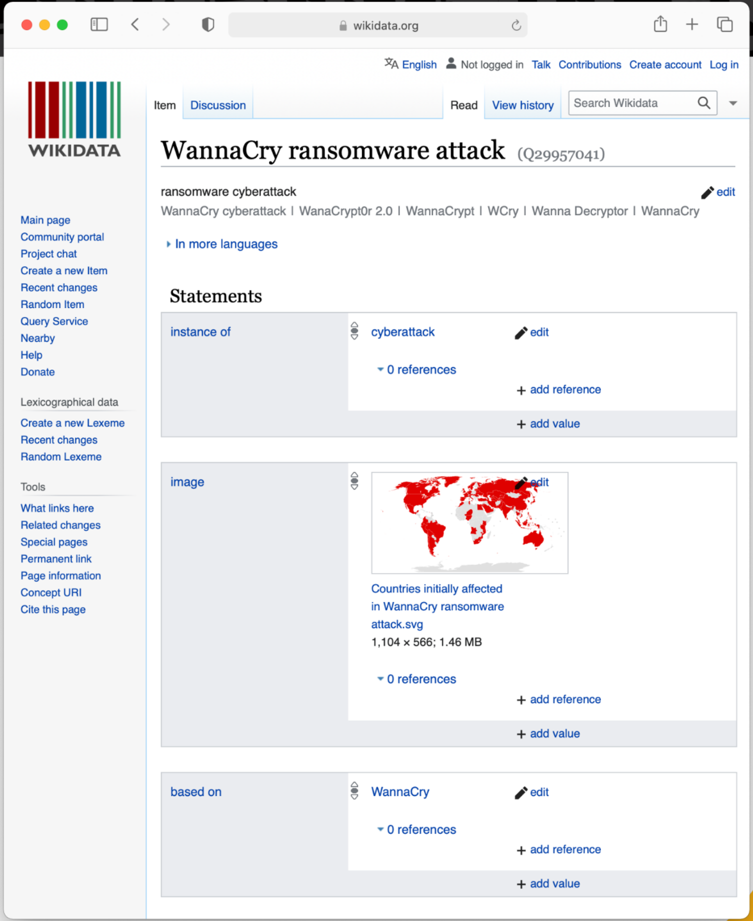
In more languages

Statements

subclass of	malware 1 reference	edit
	cyberattack 0 references + add reference	edit
	+ add value	
image	Ransomware-pic.jpg 1,920 × 1,080; 251 KB 1 reference	edit
	+ add value	
named after	ransom 0 references + add reference	edit

Cybersecurity Event

- Wikidata has data on many cybersecurity events, both generic and specific
- Has information of cyber attacks, such as a description of the attack, what areas were affected, when it occurred, outcomes, etc.



wikidata.org

English | Not logged in | Talk | Contributions | Create account | Log in

WIKIDATA

Item | Discussion | Read | View history | Search Wikidata

Main page
Community portal
Project chat
Create a new Item
Recent changes
Random Item
Query Service
Nearby
Help
Donate

Lexicographical data
Create a new Lexeme
Recent changes
Random Lexeme

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page

WannaCry ransomware attack (Q29957041)

ransomware cyberattack — edit

WannaCry cyberattack | WanaCrypt0r 2.0 | WannaCrypt | WCry | Wanna Decryptor | WannaCry

In more languages

Statements

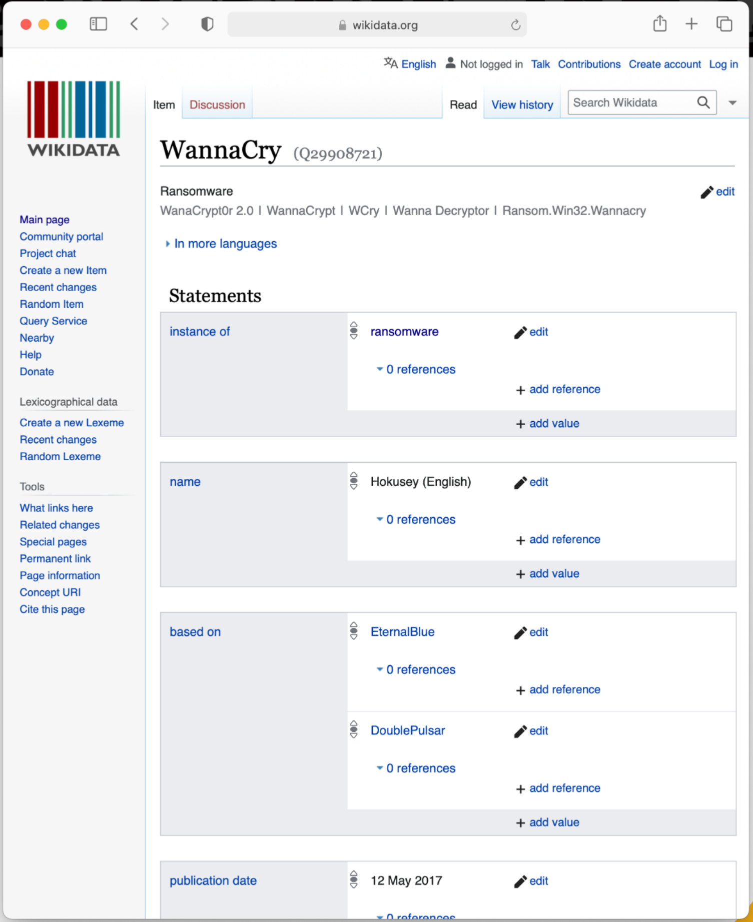
instance of — cyberattack — edit
0 references
+ add reference
+ add value

image — edit
Countries initially affected in WannaCry ransomware attack.svg
1,104 × 566; 1.46 MB
0 references
+ add reference
+ add value

based on — WannaCry — edit
0 references
+ add reference
+ add value

Cybersecurity Instance

- Wikidata has data on many specific cybersecurity instances
- A cybersecurity instance could be a specific instance of a virus
- It can have attributes including:
  - Name, aliases & short description
  - Discovery date
  - Operating systems affected
  - Links to other public knowledge graphs with more information



WannaCry (Q29908721)

Ransomware

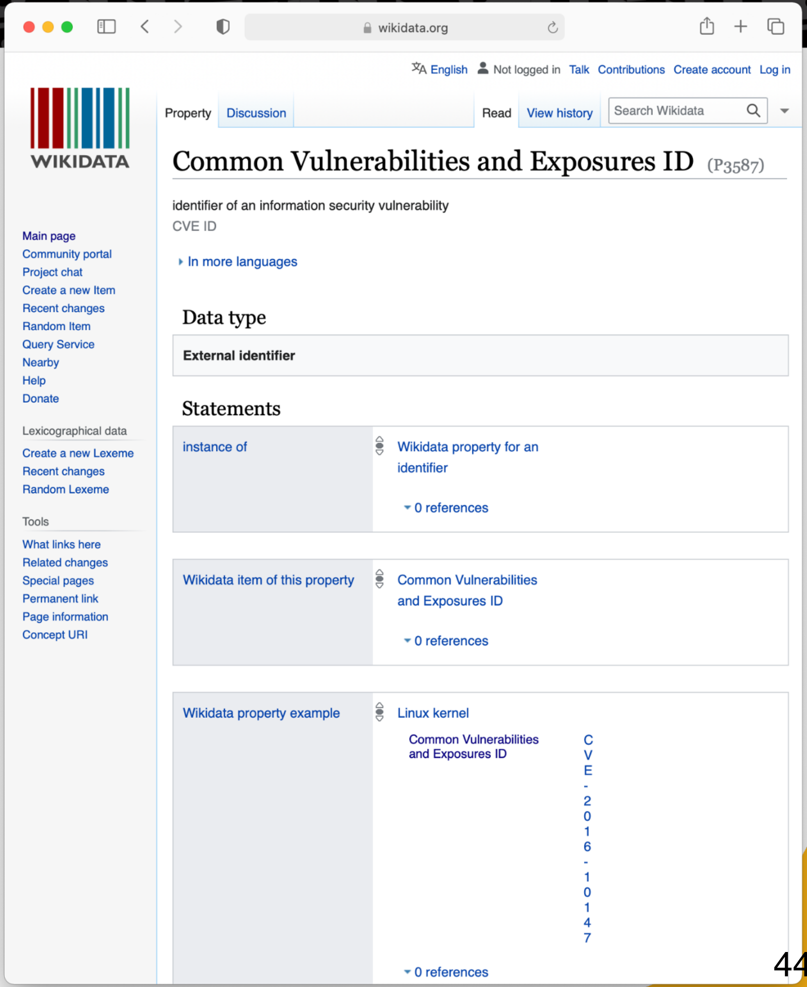
WanaCrypt0r 2.0 | WannaCrypt | WCry | Wanna Decryptor | Ransom.Win32.Wannacry

Statements

instance of	ransomware 0 references
name	Hokusey (English) 0 references
based on	EternalBlue 0 references DoublePulsar 0 references
publication date	12 May 2017 0 references

Cybersecurity Property

- Many properties of cybersecurity items are general and some very specific
- Operating System (P306) is a general property that’s useful in describing software
- CVE ID (P3587), on the other hand, is a property that only applies to CVE items



Common Vulnerabilities and Exposures ID (P3587)

identifier of an information security vulnerability

CVE ID

In more languages

Data type

External identifier

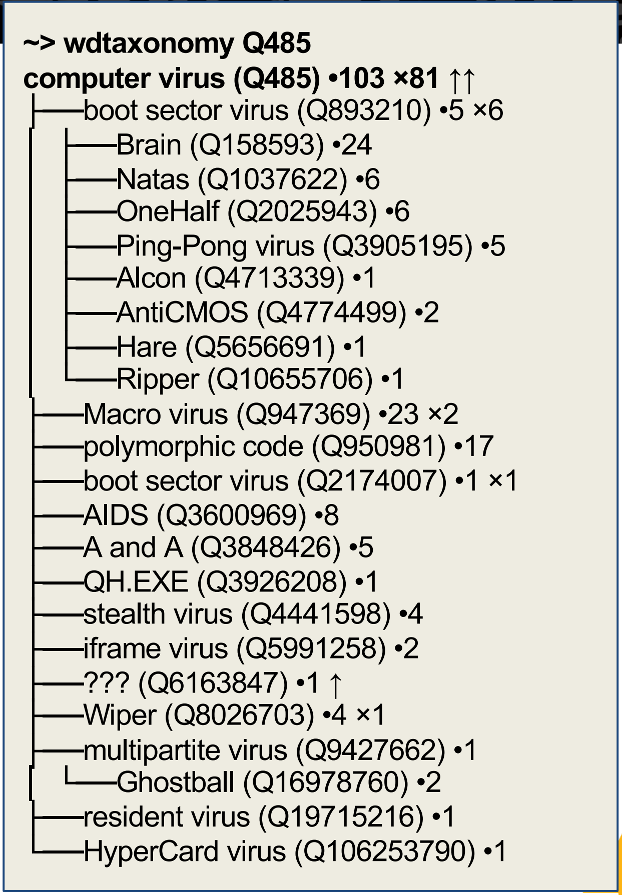
Statements

instance of	Wikidata property for an identifier 0 references
Wikidata item of this property	Common Vulnerabilities and Exposures ID 0 references
Wikidata property example	Linux kernel Common Vulnerabilities and Exposures ID CVE-2016-10147 0 references

Wikidata’s Cyber knowledge

We are exploring Wikidata’s cyber-security knowledge via:

- Its web interface for people
- SPARQL queries to retrieve and analyze its cybersecurity data
- Using other tools like wdtaxonomy and graph visualization systems



```
~> wdtaxonomy Q485
computer virus (Q485) •103 ×81 ↑↑
├──boot sector virus (Q893210) •5 ×6
│  ├──Brain (Q158593) •24
│  ├──Natas (Q1037622) •6
│  ├──OneHalf (Q2025943) •6
│  ├──Ping-Pong virus (Q3905195) •5
│  ├──Alcon (Q4713339) •1
│  ├──AntiCMOS (Q4774499) •2
│  ├──Hare (Q5656691) •1
│  └──Ripper (Q10655706) •1
├──Macro virus (Q947369) •23 ×2
├──polymorphic code (Q950981) •17
├──boot sector virus (Q2174007) •1 ×1
├──AIDS (Q3600969) •8
├──A and A (Q3848426) •5
├──QH.EXE (Q3926208) •1
├──stealth virus (Q4441598) •4
├──iframe virus (Q5991258) •2
├──??? (Q6163847) •1 ↑
├──Wiper (Q8026703) •4 ×1
├──multipartite virus (Q9427662) •1
│  └──Ghostball (Q16978760) •2
├──resident virus (Q19715216) •1
└──HyperCard virus (Q106253790) •1
```

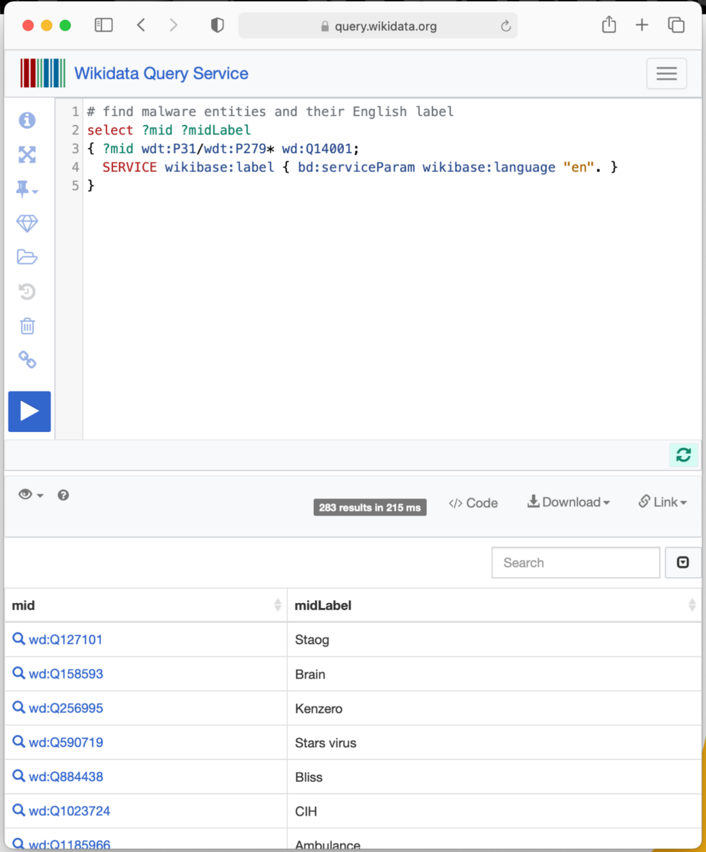
The wdtaxonomy command-line tool extracts data on WD concepts

Simple SPARQL query

- We query Wikidata with the SPARQL query language
- This query finds all 283 instances of malware (Q14001) and its sub-classes along with their names

```
select ?qid ?qidLabel {
  ?qid wdt:P31/wdt:P279* wd:Q14001;
  SERVICE wikibase:label
    {bd:serviceParam wikibase:language "en".}}
```

- The resulting triples can be easily ingested into a program or system



```
# find malware entities and their English label
select ?mid ?midLabel
{ ?mid wdt:P31/wdt:P279* wd:Q14001;
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

mid	midLabel
wd:Q127101	Staog
wd:Q158593	Brain
wd:Q256995	Kenzero
wd:Q590719	Stars virus
wd:Q884438	Bliss
wd:Q1023724	CIH
wd:Q1185966	Ambulance

Simple SPARQL query anatomy

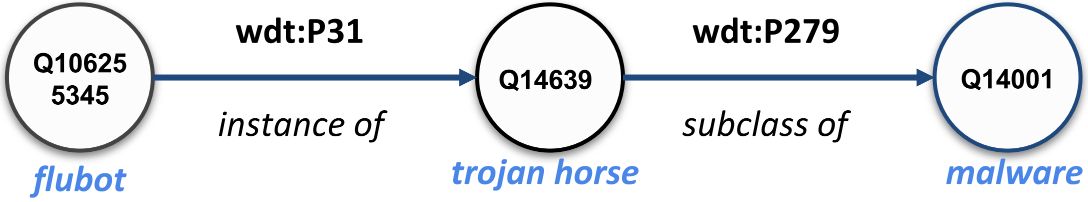
variables whose values to return

```
select ?qid ?qidLabel {

  ?qid wdt:P31/wdt:P279* wd:Q14001 ;

SERVICE wikibase:label
 {bd:serviceParam wikibase:language "en"}}
```

a path from ?qid consisting of a "instance of" (P31) followed by any number of "subClass of" (P279) edges ending at "malware" (Q14001)*

Languages we want any labels to be in

One match:



More complex SPARQL query

```
select ?qid ?name ?desc ?year
  (group_concat(DISTINCT ?alias; separator="|") as ?aliases)
  (group_concat(DISTINCT ?ava;separator="|") as ?avAliases)
  (group_concat(DISTINCT ?os;separator="|") as ?opSys)
{?qid wdt:P31/wdt:P279* wd:Q14001.
  optional {?qid rdfs:label ?name filter(lang(?name)='en')}
  optional {?qid schema:description ?desc filter(lang(?desc)='en')}
  optional {?qid skos:altLabel ?alias filter(lang(?alias)='en')}
  optional {?qid wdt:P1845 ?ava.}   # vendor antiVirusLabel
  optional {?qid wdt:P306/rdfs:label ?os filter(lang(?os)='en')}
  optional {?qid wdt:P575 ?date.  BIND (year(?date) as ?year) }
 }
GROUP BY ?qid ?name ?desc ?year
```

- **query finds seven key properties of all malware instances**
- **Asking only for English language text values**
- **use of optional means data not required (not all malware has an alias)**
- **Queries can be entered in website or sent via an API to WD**

More complex

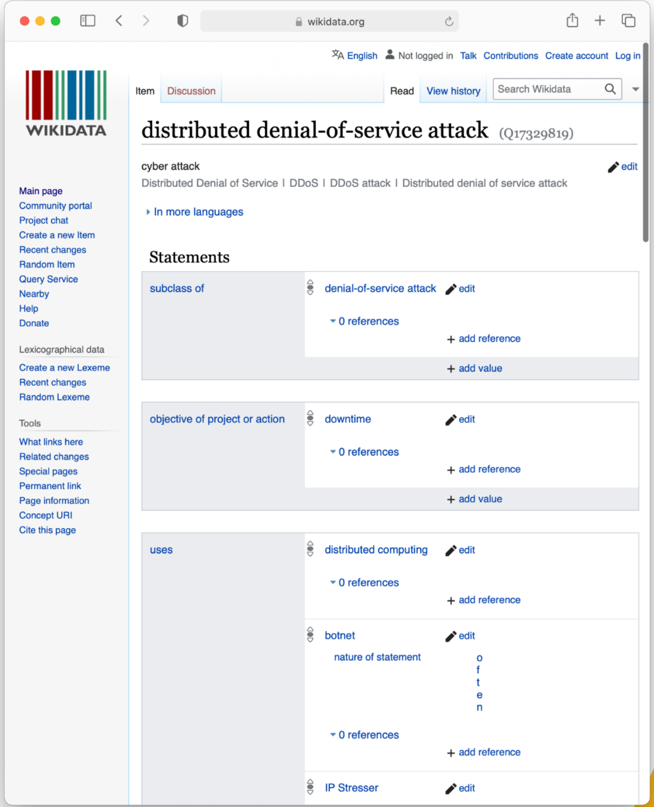
```
select ?qid ?name ?desc ?year
  (group_concat(DISTINCT ?alias; separator="
  (group_concat(DISTINCT ?ava;separator="|
  (group_concat(DISTINCT ?os;separator="|")
{?qid wdt:P31/wdt:P279* wd:Q14001.
  optional {?qid rdfs:label ?name filter(lang(?
  optional {?qid schema:description ?desc filt
  optional {?qid skos:altLabel ?alias filter(lang
  optional {?qid wdt:P1845 ?ava.}   # vendor a
  optional {?qid wdt:P306/rdfs:label ?os filter
  optional {?qid wdt:P575 ?date.  BIND (year(
 }
GROUP BY ?qid ?name ?desc ?year
```

```
[
 {
   "qid": "http://www.wikidata.org/entity/Q106255345",
   "name": "FluBot",
   "desc": "Android malware that is propagated using
                  fake SMS messages",
   "year": "2020",
   "aliases": "Cabassous",
   "avAliases":
         [
              "Trojan-Banker.AndroidOS.Flubot",
              "Trojan:Script/Wacatac.B!ml"
         ],
   "opSys": "Android"
 },
 ...
]
```

Query results available as triples, TSV, JSON, ...

Linking entities to Wikidata items

- Prototype system links text strings to Wikidata items
  - “ddos” => Q17329819: distributed denial-of-service attack
- Linking uses spaCy’s word embeddings and WD’s fine-grained type system
- Improves graph by merging nodes that refer to same concept, entity, or event:
  - DDoS, distributed denial of service, DDoS attack, …



distributed denial-of-service attack (Q17329819)

cyber attack

Distributed Denial of Service | DDoS | DDoS attack | Distributed denial of service attack

edit

In more languages

Statements

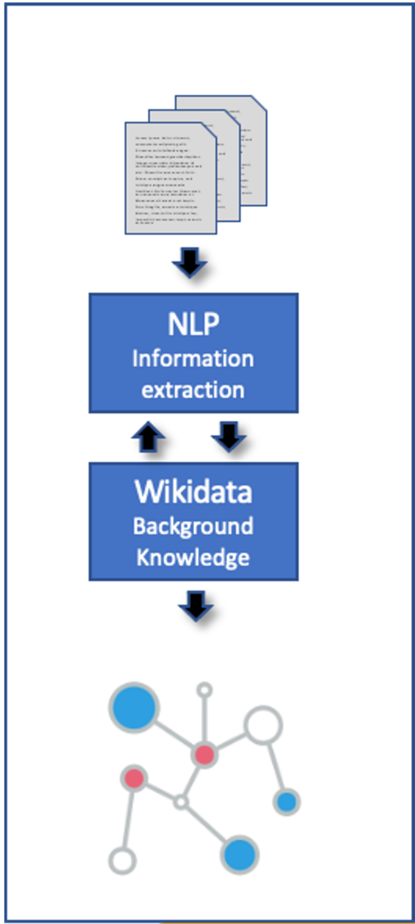
subclass of	denial-of-service attack	edit
	0 references	
	+ add reference	
	+ add value	
objective of project or action	downtime	edit
	0 references	
	+ add reference	
	+ add value	
uses	distributed computing	edit
	0 references	
	+ add reference	
	botnet	edit
	nature of statement	often
	0 references	
	+ add reference	
	IP Stresser	edit

Wikidata Next Steps

- Catalog and describe what Wikidata knows about cybersecurity
- Create Wikidata bots to identify and tag likely omissions and errors
- Improve our Wikidata entity linker for cybersecurity concepts, properties, and entities
- Explore ways to (semi-) automatically update Wikidata with new cybersecurity information

Conclusion and Next Steps

- We are working toward a system maintains a **knowledge graph** of **cybersecurity data** extracted from public **text reports**
- We are using the spaCy NLP tools and the Wikidata knowledge graph
- In the future we hope to extract and integrate data from semi-structured data in STIX, MISP and other formats



questions and discussion