AnnoCTR: Cyber Threat Reports
General-Layer Annotation Guidelines

Lukas Lange, Marc Mueller, Patrick Grau, Dragan Milchevski, Ghazaleh Torbati, Annemarie Friedrich
Bosch Center for Artificial Intelligence, Renningen, Germany

Task 1: Text Cleansing

Each file name follows the pattern `zscaler_date_title-in-url.txt`. For example, use `zscaler_2021-11-16_return-emotet-malware.txt` for the post at https://www.zscaler.com/blogs/security-research/return-emotet-malware.

We provide you with a set of raw text files that are supposed to be cleaned before annotation. They have been extracted from the ZScaler blog and split into sentences (one-sentence-per-line) automatically. The data is hence likely to be somewhat noisy, and the first task consists in cleaning the texts in the raw text format. In these texts, use a one-sentence-per-line format and **read through the texts manually to ensure that the text is clean and correct** (comparing to the original web page as necessary). CodeSnippets are considered as separate sentences (unless they are very small and embedded into what really feels like a natural sentence).

For this task, please use a text editor where you can make sure that no "carriage return" symbols are added, e.g., NotePad++.

Task 2: Annotation of Named Entities and Entity Linking

Mark the texts with the following information:

- CLICommand/CodeSnippet: CLI commands and code snippets
- Organisations (ORG): mentions of companies, etc. → classify into industry sectors
- Locations (LOC): mentions of cities, places, countries, continents, …
- Dates

The CLICommand/CodeSnippet, ORG, and LOC tags are annotated on the **NamedEntityAndLinking** layer.

1. Select the NamedEntityAndLinking layer in the drop-down list on the right-hand panel (if it is not already there).
2. Select a text span on the left-hand side (via double-clicking or marking a span via click+drag).
3. In the "value" dropdown on the right-hand panel, select the tag that applies to the span.

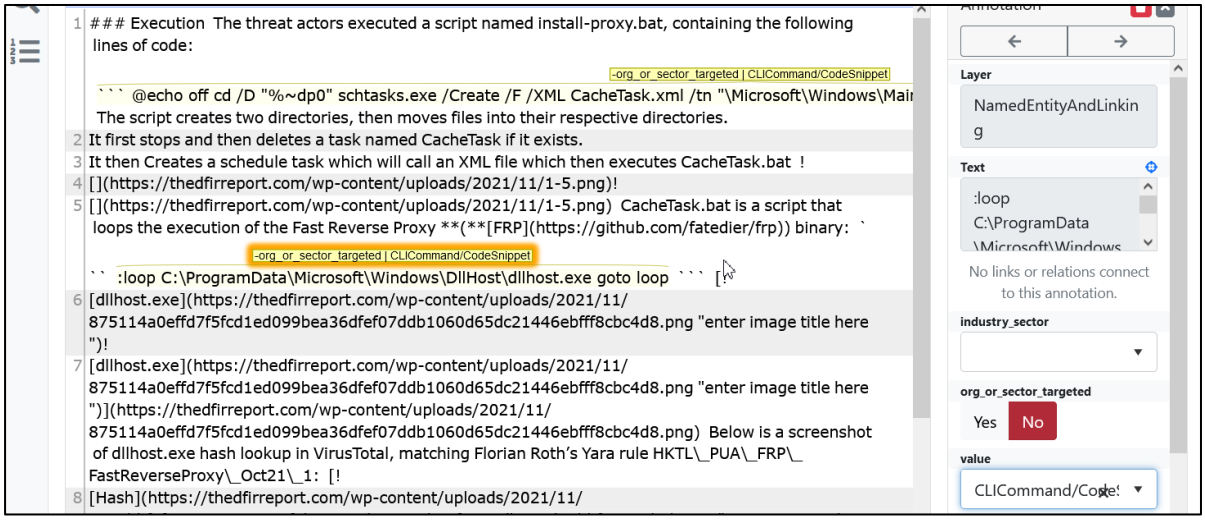
Dates are annotated on the **Timex** layer, which will be explained below.

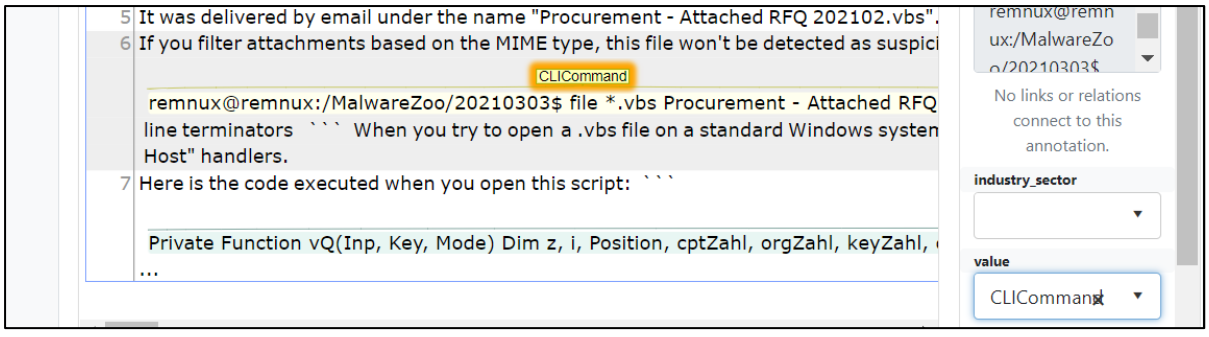
For locations and organisations, please also indicate the WikiData identifier of the specific entity in the **wiki_identifier** field. The reason is that the same name may apply to several entities, e.g., "Paris" may refer either to the city or to "Paris Hilton". We call this step "entity linking" because we connect the *mentions* of the entities in the text with the *knowledge graph* that contains unique entries for the entities (here, the knowledge graph is WikiData).

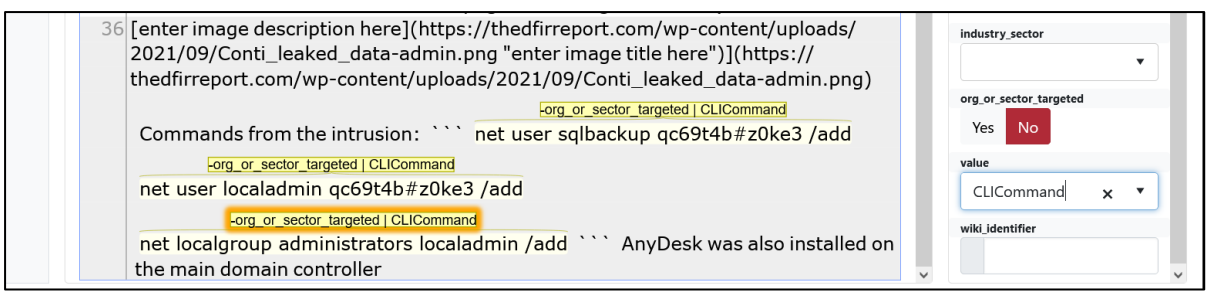
If the search within Inception is too slow or doesn't give the results you expect, please go to https://www.wikidata.org/wiki and search there. You can then put the "code" for the entity (the last part of the page's URL, if you found the correct entry) into the **wiki_identifier** field. In the example below, this identifier is "Q408", also displayed next to the title (=name of the entity).

CLICommand/CodeSnippet

Use the **CLICommand/CodeSnippet** tag to mark any command line (CLI) arguments and code snippets in the text. However, as you can see below, if just hash numbers or file names are mentioned, you do not need to mark them. There is no need to fill out wiki_identifier or industry_sector for code snippets and CLICommands.







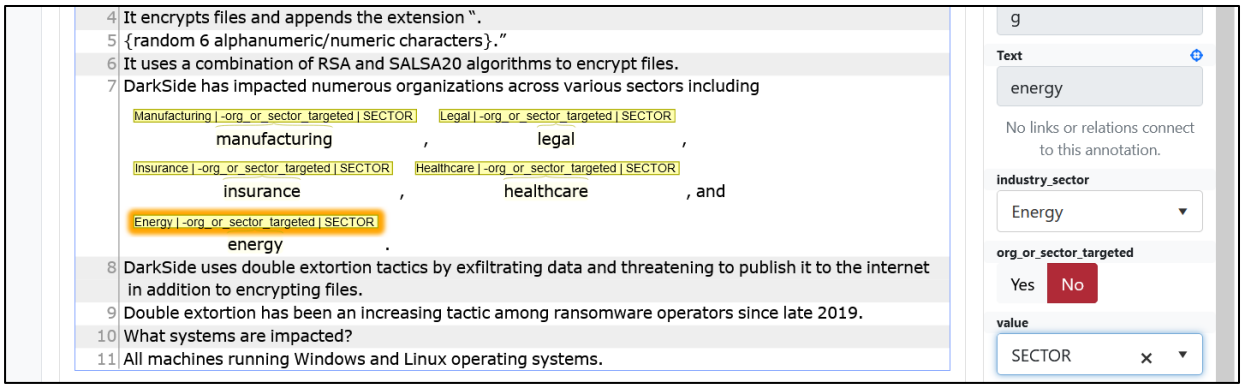
Organizations (ORG) and SECTOR

Organizations are companies (“Microsoft”, “Bosch”) or institutes (“Humboldt University”).

If the organization is a company and can be categorized as belonging to one of the industry sectors in the following list, please choose the appropriate label and enter it into the field “industry_sector”, where you can choose from a dropdown list.

In addition, this also goes for industry sectors which are explicitly mentioned.

Example for SECTOR

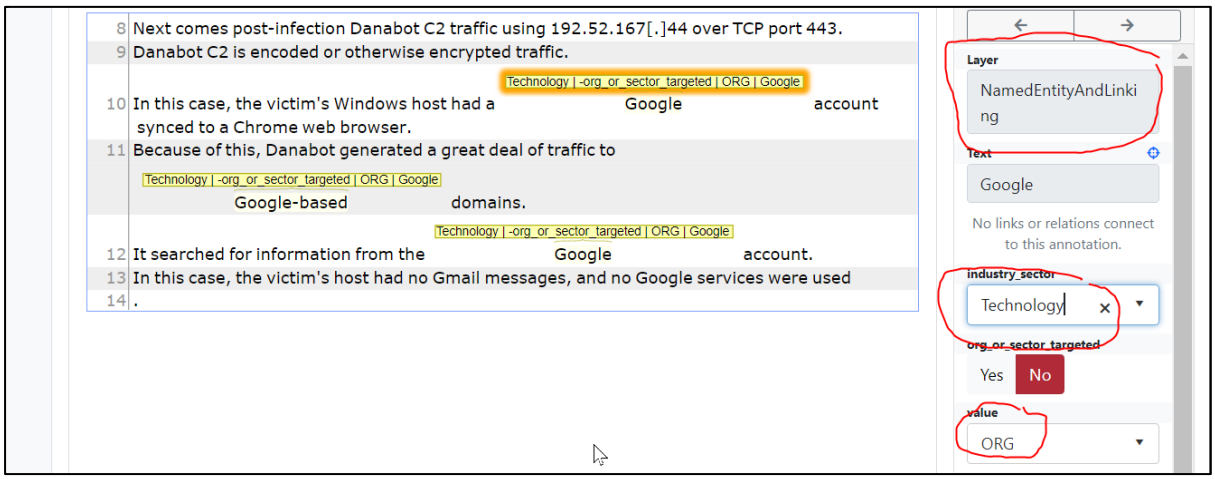


Category	Items (Detail)
Agriculture	
Automotive	Car, Motor
Civil Society	NGO, Citizens, Political Party, Political Organizations, Opposition, Dissidents, Social Networks, Separatists, Journalist, Activists, Private Charitable Organization, Non-Profit, Human Rights
Construction	Architecture
Defense	Defence, Intelligence, Military, Ministry of Defense, Defense Contractors, Defense R&D
Education	Research, Universities, Think Tanks, Laboratories, Schools, Academia, Higher Education
Energy	Gas, Oil, Oilfield, Power, Electricity, Electric, Scada, Petrochemical, Nuclear, Renewable Energies, Water
Entertainment	Culture, Sports, Movies, Arts, Game, Gaming, Video, Streaming, Casino, Music, Photography
Finance	Bank, Banking, Market, Credit Card, Accounting, Financial Services, Payment, Investment
Government	Administration, Immigration, Intergovernmental organizations (IGOs), Central Administration and Ministries, Parliamentary chambers, Local Administrations, Diplomacy, Police, Law Enforcement, Justice, International Relations
High-Tech	Semiconductors, Electronics, Innovation, Optoelectronics, Robotics, Satellite, Navigation
Infrastructure	Pipeline, Power Plant, Critical Infrastructure
Insurance	
Legal	Law, Law Firm, Lawyer
Manufacturing	Industry, Machinery Building, Industrial, Chemical, Metal, Steel, Iron, Plastic, Hydraulics
Media	Audiovisual, News, Press, Publishing, Television, Broadcast, Marketing, Advertising, Media Production
Mining	
Healthcare	Health, Healthcare Services, Pharmacy, Drug Manufacturing, Healthcare Research, Hospital, Pharmaceuticals, Biomedical, Medical, Neuroscience, Biotechnology
Retail	Trade, Food, Beverage, E-Commerce, Consumer
Religion	Religious Organization
Technology	Engineering, Development, Programming, IT, Software, Computer, Life Science, Materials Science, Hardware, Security Systems, Information Technology
Telecommunication	Internet Service, Telecom, Network, Internet Infrastructure Service, Communication Service Provider, Hosting, IXP, ISP, Phone

Transport	Transportation, Aerospace, Aeronautics, Space, Aircraft, Aviation, Rail, Railway, Maritime, Shipping, Road, Logistic, Airlines, Trucking, Railroad
Tourism	Hotels, Hospitality, Leisure, Restaurants, Travel
Supply Chain	MSP, Managed Service Provider
Unknown	
Other	Utility
Multi-Sector	Conglomerat

Example for ORG

In the following example, mark Google as an ORG, select “Technology” as the relevant industry_sector, and choose the appropriate wiki_identifier. Note that org_or_sector_targeted is “No” because in this context, Google is not the target of the threat.



ORG_OR_SECTOR_TARGETED

Set the org_or_sector_targeted flag to Yes if the sector or organization you have marked is the target of this attack. For example:

The ORG is targeted: The attackers gained initial access to the Google infrastructure by using the new log4J exploit.

The ORG is <u>NOT</u> targeted: The attackers sent messages with an malicious link via Facebook messenger, to get the bank credentials of their victims.

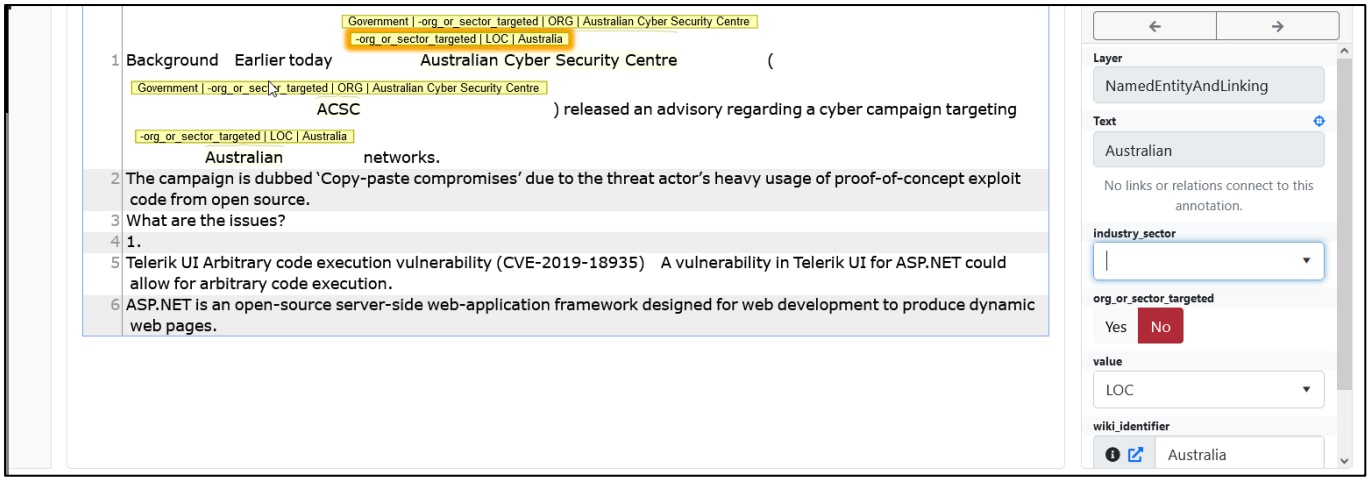
The attackers abused an vulnerability in the google chrome browser, to get initial access on the victim host.

Locations (LOC)

Locations can be cities, countries, continents, … basically all places.

In the following example, you can note the following:

- We annotate only specific organization or locations, e.g., “Australian Cyber Security Centre”, but if the mention is too broad, e.g., “multiple Australian government organizations”, we do not add an annotation.
- Sometimes, mentions may be nested, e.g., as in “Australian Cyber Security Centre”, where the entire span refers to an ORG, but the span “Australian” within the larger span refers to a LOC.



Dates (Timex)

For dates, when referring to a specific year, day, or month, please select the **Timex** layer, create an annotation on the part of the text referring to the date, and enter the normalized date into the "value" box on the right-hand-side panel. For deciding on the date, you might have to take the publication date of the document into account.

Please enter the normalized dates in one of the following formats:

- For specific days YYYY-MM-DD
- for a month YYYY-MM
- for a year YYYY

If a mention clearly refers to a day or a month, but parts are unknown and cannot be resolved via context, e.g., in "April 1" and without any context we do not know the year, would be "XXXX-04-01".

Please annotate any expressions that refer to a year, month, or day. This might be explicit mentions or mentions such as "in the past year", which you need to resolve with regard to the publication date of the document.

Please do not include "in/on" into the span. For example, if a mention says "in last month" → only "last month" is marked as the span. "On April 5, 2020" → Mark only "April 5, 2020".

We are following these guidelines:

https://sharedtasksinthedh.github.io/assets/howto-annotation/timeml-1.2.1.pdf

(italic parts: not included in annotation span, bold parts: included in annotation span)

Mr. Smith left **Friday, October 1, 1999**
the second of December
yesterday
in **October of 1963**
in **the summer of 1964**
on **Tuesday 18th**
in **November 1943**
this year's summer

Special cases:

If a country is mentioned within a word, e.g., anti-Russian or non-Iranian, mark only the part that refers to the country.