Annotation Guidelines of the Creation of Indonesian Entity Linking Benchmark Dataset for General and Specific Domains

1. Introduction

These guidelines are designed to providing direction in manual annotation of Indonesian Entity Linking (EL) benchmark dataset creation. The dataset is purposed for general and specific domains. The manual annotation will label Named Entities (NEs) with their corresponding entries link in a Knowledge Base (KB), where we use Indonesian Wikidata. Named entities are a name on text that refers to a specific object in the world, where the name distinguished an object from the others. NEs exist in the form of proper nouns, and they belong to particular classes, such as, *Person, Location*, *Organization*, etc. The followings are examples of NEs:

a. UIN Syarif Hidayatullah Jakarta (a name of particular university in Indonesia)
b. Jl. Ir. H. Juanda (a name of particular street in South Tangerang, Indonesia)
c. DPR RI (a name of particular organization in Indonesia)

EL or NE disambiguation is a task in Natural Language Processing (NLP) that aims to connect NEs to their corresponding entries in a KB. In the below sentence,

> "*Kemacetan parah terjadi di Juanda sejak pagi hari, kata satpam yang bertugas di kampus UIN Jakarta dan turut membantu mengurai kemacetan.*" ("Severe congestion has occurred on Juanda since the morning, according to the security guard on duty at UIN Jakarta campus who also helped to alleviate the traffic jam.")

there are two NEs, including *Juanda* and *kampus UIN Jakarta* (UIN Jakarta campus). According to the sentence's context, a disambiguation process or EL will set *Juanda* as a street name located in South Tangerang, not as an airport name in Surabaya city or a train station name in South Tangerang. This is because a security guard who was on duty at UIN Jakarta was the one who reported the severe congestion where he also helped to alleviate the traffic jam. UIN Jakarta is a university that is located in South Tangerang. Furthermore, the disambiguation process also determines *kampus UIN Jakarta* as the area at which UIN Jakarta located.

2. Time

The annotation process will be performed within one week, i.e. from March 20, 2023 to March 26, 2023.

3. Annotators

The annotation will be performed by four persons in the general domain and 2 persons in the specific domain. The annotators are Informatics Engineering and Quran and Tafseer students at the State Islamic University Syarif Hidayatullah Jakarta.

4. Annotation Mechanism

a. The annotation will be performed on entities that are obtained from one of Named Entity Recognition (NER) benchmark datasets for Indonesian, NER UI, for the general domain and IndQNER for the specific domain. NER UI dataset contains 2114 sentences and 5055 NEs from *Person*, *Location*, and *Organization* NE classes. IndQNER contains 3117 sentences and 2475 NEs, representing 18 NE classes.

b. The annotation will be performed by two groups of annotators in the general domain and one group of annotators in the specific domain. Each group contains two annotators.

c. In the general domain, each of two annotators in one group will annotate 1063 sentences, while all annotators in the specific domain will annotate 3117 sentences.

d. Every annotator will be provided with two types of documents: one containing raw text with sentences and another with the same text pre-tagged with entities. The raw text serves as the workspace for annotators to identify and tag NEs, while the pre-tagged document is intended to guide the annotators by highlighting the specific sections of text that are NEs. This is a snippet of the first document.

```
"Ini lebih mudah daripada yang saya duga," kata Federer seperti dilansir Reuters.
"Menyedihkan kalah di laga seperti ini tapi kami tahu kami harus berpikir bagaimana caranya comeback," ujar bek Madrid Pepe seperti dikutip Football Espana.
Mega malah mengkritik wartawan yang selalu menyebutkan Laksamana sebagai mantan orang PDIP.
"Jika kami tidak tampil bagus maka itu akan memudahkan Dortmund."
Kalau Davenport harus menjalani jalanan berbatu kerikil, maka tidak demikian dengan unggulan utama putra Roger Federer.
Pada Senin malam saat kejadian, petugas tower yang berkomunikasi dengan pesawat Batik Air yang akan tinggal landas, menggunakan saluran radio VHF yang tidak dapat didengar oleh petugas towing yang menggunakan Handy Talky (HT) dengan saluran frekuensi yang berbeda.
Pasalnya, Federer dan Wawrinka jadi andalan Swiss dalam mencari trofi Piala Davis pertamanya dengan melawan Prancis pada pekan depan.
Banyak kritik yang dilontarkan dari berbagai pihak, termasuk dari PDIP.
Penyerahan bantuan secara simbolis dilakukan Jokowi di Desa Sanggeng Kecamatan Manokwari Barat Manokwari Selasa (5/4/2016).
Squawka mencatat Morgan sebagai pemain yang berbahaya dalam tekel dan dominan dalam duel di udara.
```

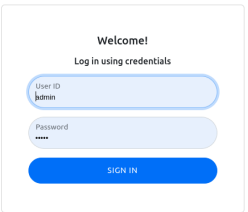
And the following is a snippet of the second document:

"Ini lebih mudah daripada yang saya duga," kata <entity>Federer</entity>, seperti dilansir <entity>Reuters</entity>.
"Menyedihkan kalah di laga seperti ini tapi kami tahu kami harus berpikir bagaimana caranya comeback," ujar bek <entity>Madrid</entity>, <entity>Pepe</entity>, seperti dikutip Football Espana.
<entity>Mega</entity> malah mengkritik wartawan yang selalu menyebutkan Laksamana sebagai mantan orang <entity>PDIP</entity>.
"Jika kami tidak tampil bagus maka itu akan memudahkan <entity>Dortmund</entity>.
Kalau <entity>Davenport</entity> harus menjalani jalanan berbatu kerikil, maka tidak demikian dengan unggulan utama putra <entity>Roger Federer</entity>.
Pada Senin malam saat kejadian, petugas tower yang berkomunikasi dengan pesawat <entity>Batik Air</entity> yang akan tinggal landas, menggunakan saluran radio VHF yang tidak dapat didengar oleh petugas towing yang menggunakan Handy Talky (HT) dengan saluran frekuensi yang berbeda.
Pasalnya, <entity>Federer</entity> dan <entity>Wawrinka</entity> jadi andalan <entity>Swiss</entity> dalam mencari trofi Piala Davis pertamanya dengan melawan <entity>Prancis</entity> pada pekan depan.
Banyak kritik yang dilontarkan dari berbagai pihak, termasuk dari <entity>PDIP</entity>.
Penyerahan bantuan secara simbolis dilakukan <entity>Jokowi</entity> di <entity>Desa Sanggeng</entity>, <entity>Kecamatan Manokwari Barat</entity>, <entity>Manokwari</entity>, Selasa (5/4/2016).
<entity>Squawka</entity> mencatat <entity>Morgan</entity> sebagai pemain yang berbahaya dalam tekel dan dominan dalam duel di udara.

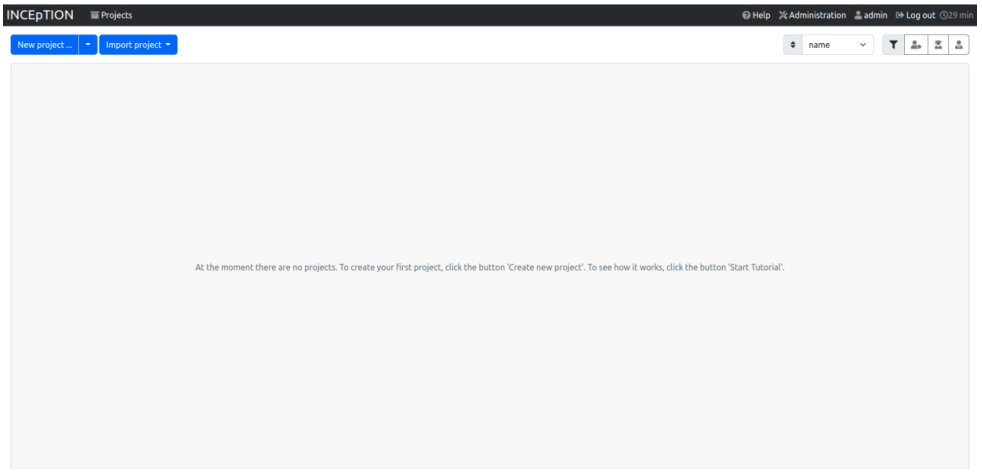
5. How to Annotate

The annotation will be performed by following these steps:

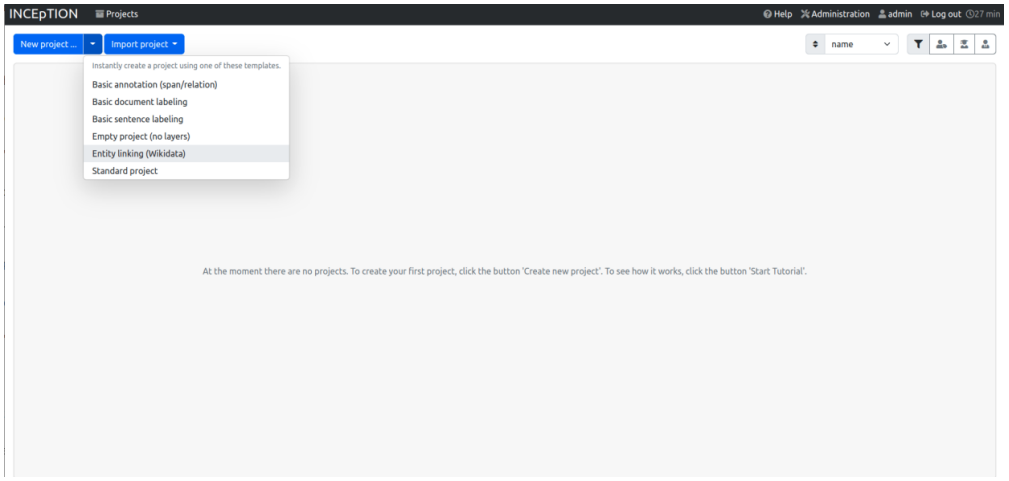
a. Download the annotation tool, INCEpTION, at https://inception-project.github.io/downloads/
b. Install and run the application following the steps explained at: https://inception-project.github.io/releases/27.1/docs/user-guide.html#sect_installation
c. Log in to the application with *admin* as the user name and password.



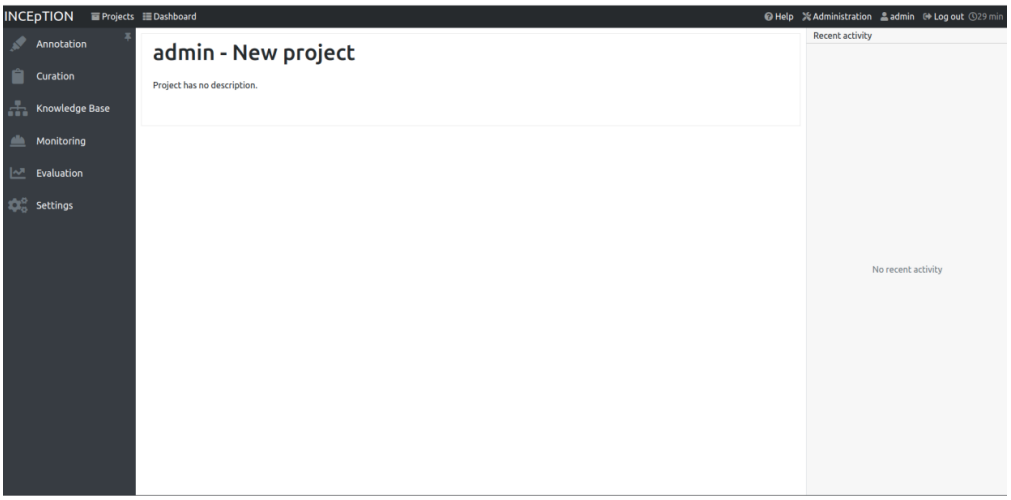
This is the display once annotators have logged in successfully.



d. Create a new project by clicking the arrow icon on *New Project* menu, then choose *Entity Linking (Wikidata)* from the sub-menus.



This is the display once annotators have created the new project successfully.

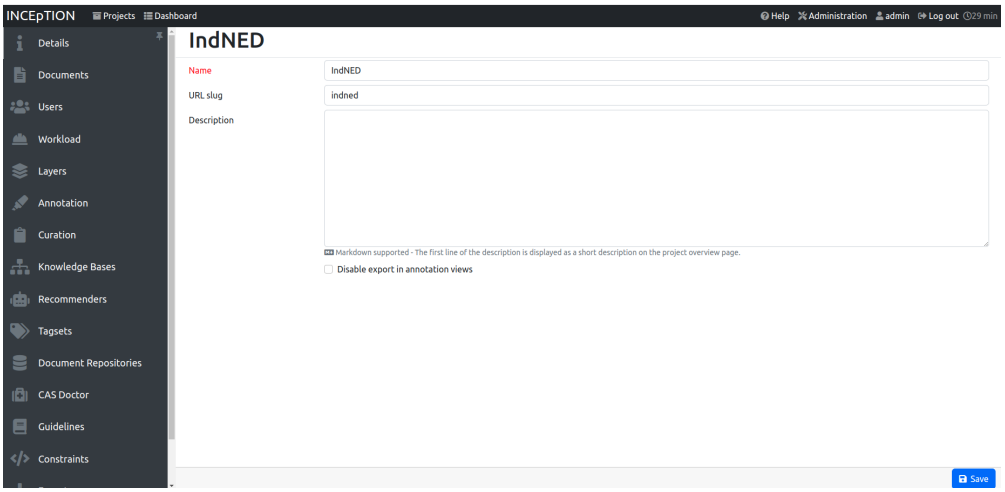


e. Click on *Settings* to set up the properties of the project. They include:

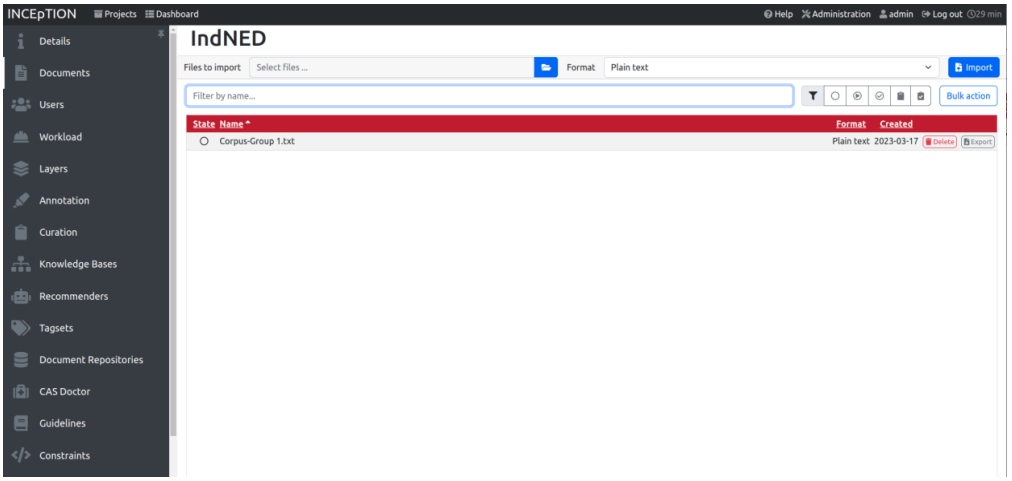
i. The project's name: *IndNED*

ii. Ignore the *URL slug* and *Description*. It will be automatically filled with *IndNED*.

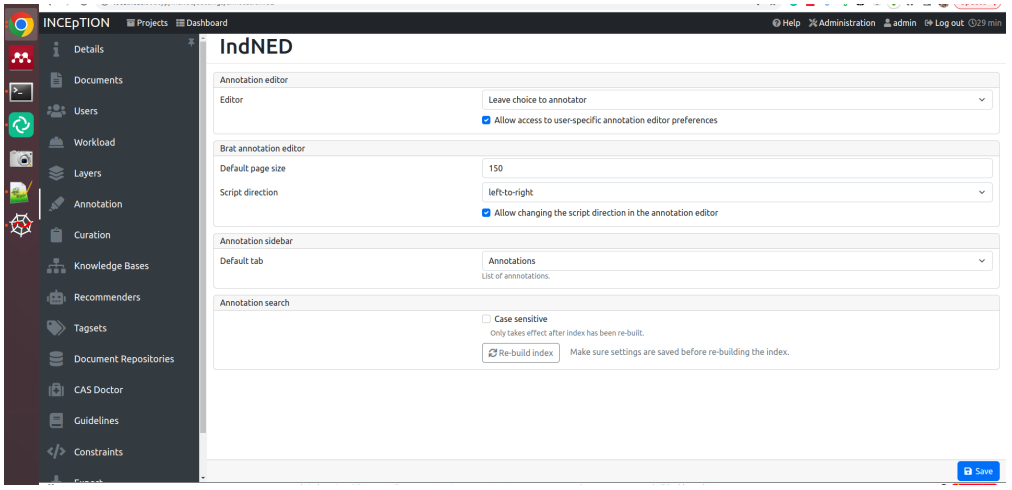
iii. Click *Save*.



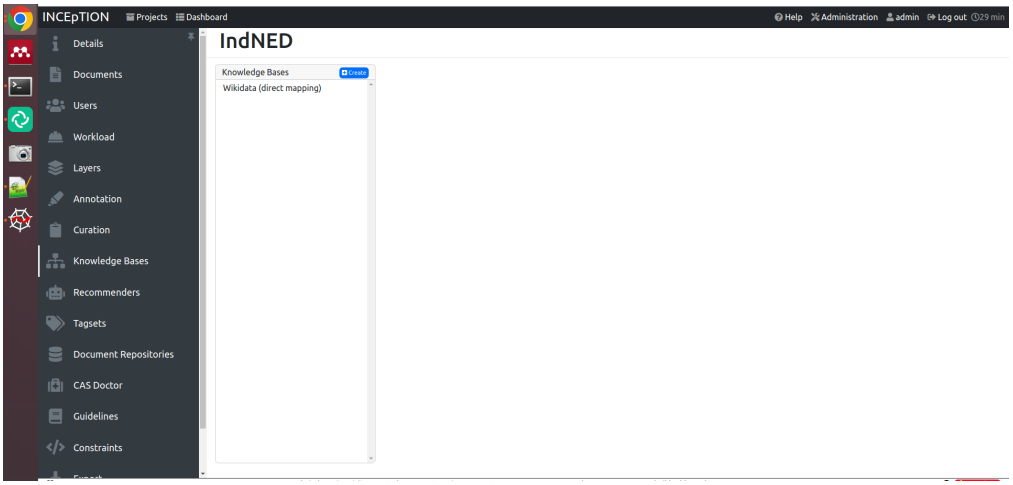
f. Click on *Documents*. Then, upload the first document (refer to point 4, and sub-point *d*). This is the display after annotators have selected the document.



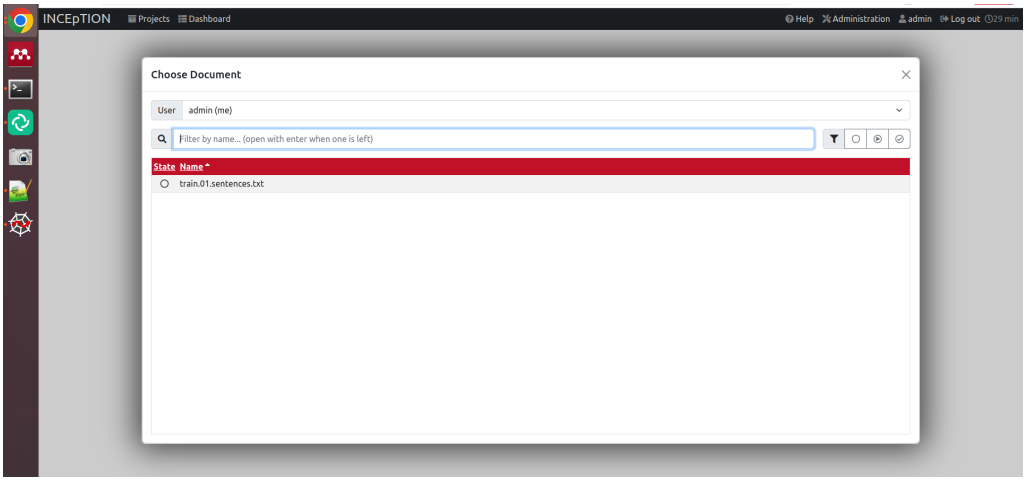
g. Click on *Annotation*. Then, set the value of *Default page size* to 150. Enable the checkbox *Allow changing the script direction in the annotation editor*. Then, select *Annotations* at *Default tab* item, disable checkbox *Case sensitive*, and click *Save*.



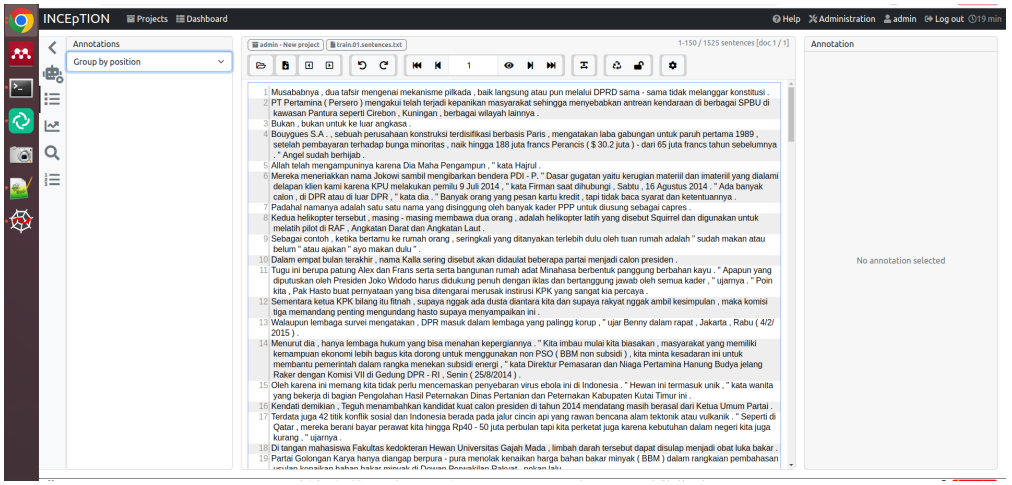
h. Click on *Knowledge Bases*. Make sure that *Wikidata (Direct Mapping)* has been automatically chosen.



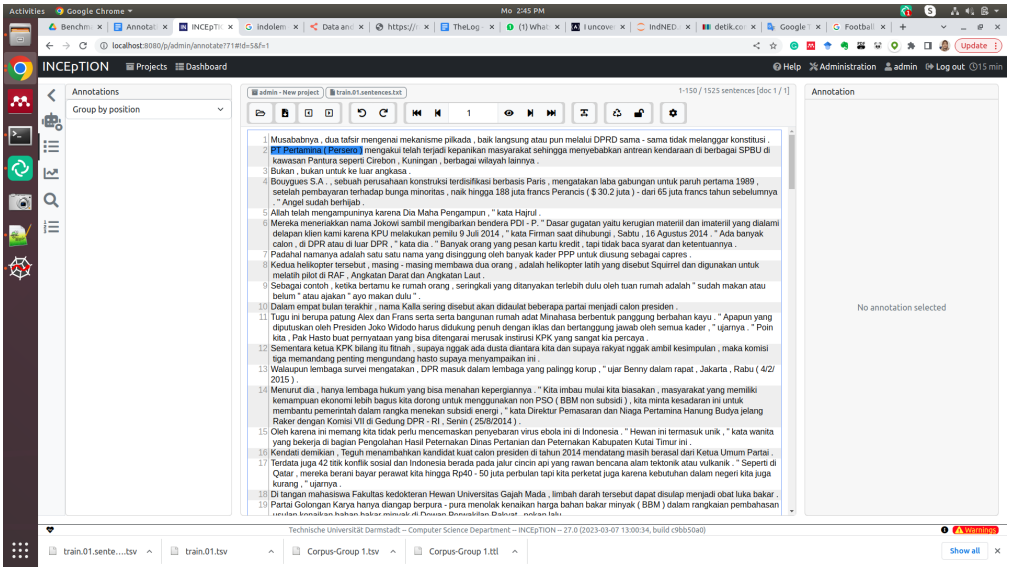
i. Click on *Dashboard* item, then choose sub-menu *Annotation*.



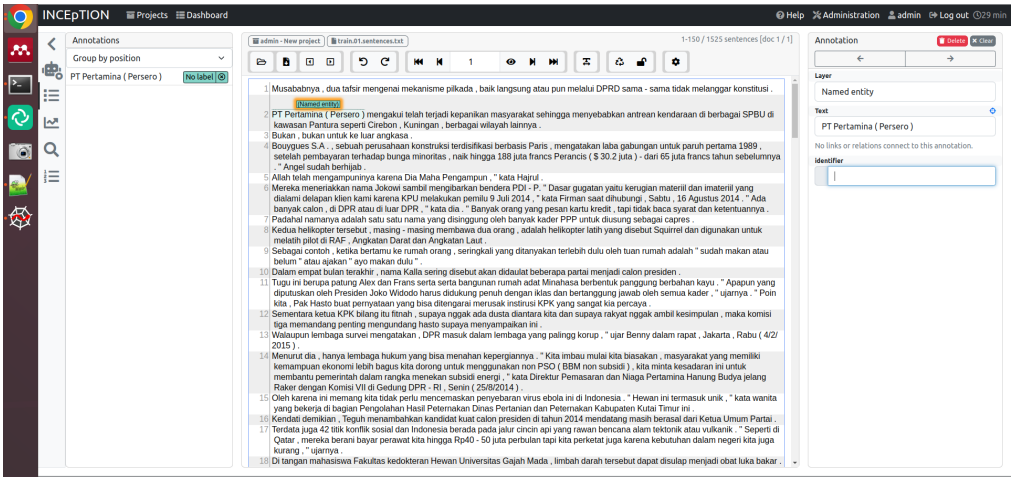
j. Click on the document's name that has been uploaded, for instance, train.01.sentences.txt. You will be directed to the annotation editor page. On this page, there are three parts (left-to-right): Summary of annotation results, text as the annotation workspace, and the places where the corresponding links are searched within Indonesian Wikidata. You can choose *Group by position* on *summary* item to get a better display of information.



k. Select NEs in a sentence according to information in the second document (refer to point *4* sub-point *d*) by *clicking and dragging* the whole span of the NEs. For example, there is an NE in the second sentence, i.e., *PT. Pertamina (Persero)*. The whole span of the NE is blocked, as shown below.

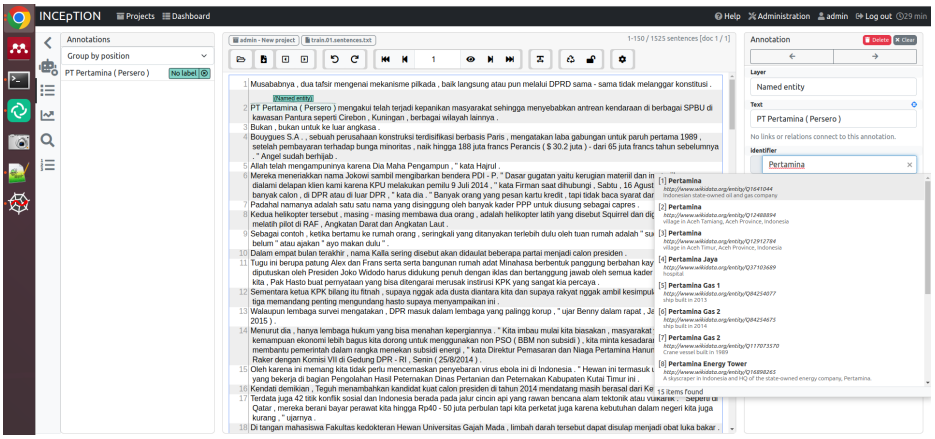


l. The right part of the editor page automatically display the *interface* to define the corresponding link in Indonesia Wikidata.

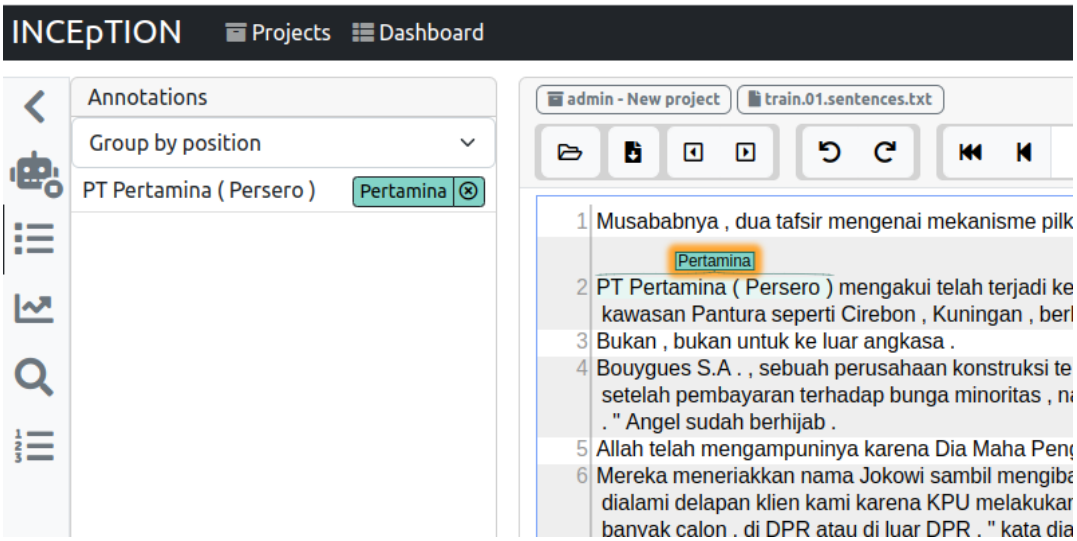


m. Make sure the selected NE appears on *Text* item. Then, if annotators are certain of which specific object is being represented by the selected NE, annotators can enter the keyword into the search menu on *Identifier* item to find the corresponding entity in Wikidata. When annotators find a list of names, make sure to choose the one with the same profile with the selected NE's profile. The followings are steps to find the selected NE in Wikidata:

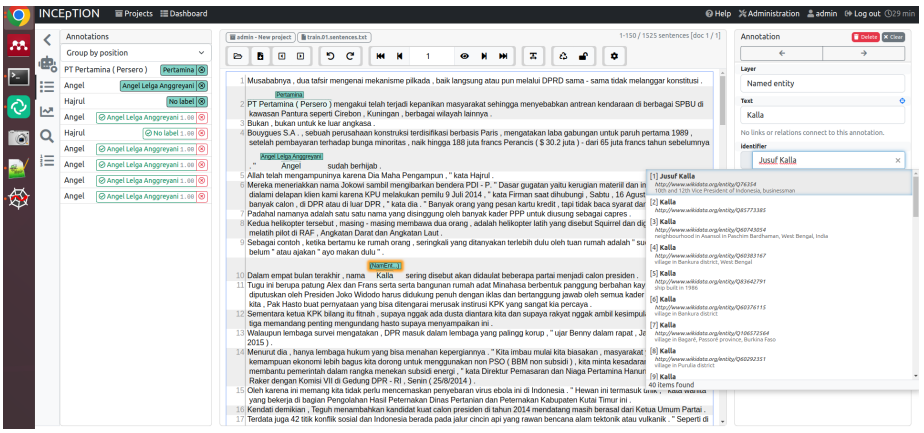
   i. Use the common term/name of the selected NE as the keyword. For instance, *Pertamina* is the common name of *PT. Pertamina (Persero)*.



As seen on the above display, the first entry of the results is *Pertamina* with profile: *Indonesian state-owned oil and gas company*, where this is the profile of the selected NE. Subsequently, annotators must make sure that the selected NE is the Indonesian label on the entry. If it is the case, annotators can choose that name as the corresponding entry in Indonesian Wikidata.



ii. If the selected NE is not complete, but annotators are sure of which specific entity it refers to according to the sentence's context, annotators can use the complete name as the keyword. For instance, *Kalla* in sentence "*Dalam empat bulan terakhir , nama Kalla sering disebut akan didaulat beberapa partai menjadi calon presiden*." (In the last four months, Kalla has been frequently mentioned as being nominated by several political parties as a presidential candidate.) is an incomplete name. According to the sentence's context, we can know that the mentioned Kalla refers to *Jusuf Kalla* (the vice president at that moment). Therefore, annotators can use *Jusuf Kalla* as the keyword, as shown below.



Note: The correct entry in Wikidata is the first one, as its profile is the *10th and 12th Vice President of Indonesia, businessman*.

iii. If the selected entity is incomplete and there is no sufficient information on it in the sentence, annotators can do these followings steps:

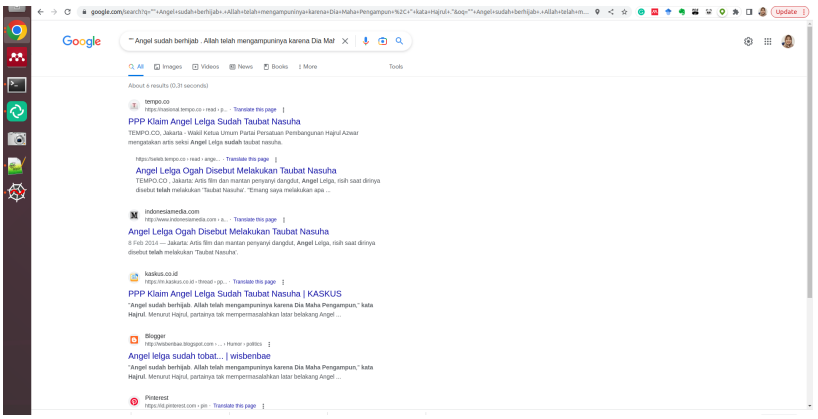
a. Find relevant web documents of the entity by performing Google's documents retrieval. To do an exact matching, enclose the keyword with double quotes.

b. Find the complete name of the selected entity in the relevant documents.

c. Repeat step described at point *i*.

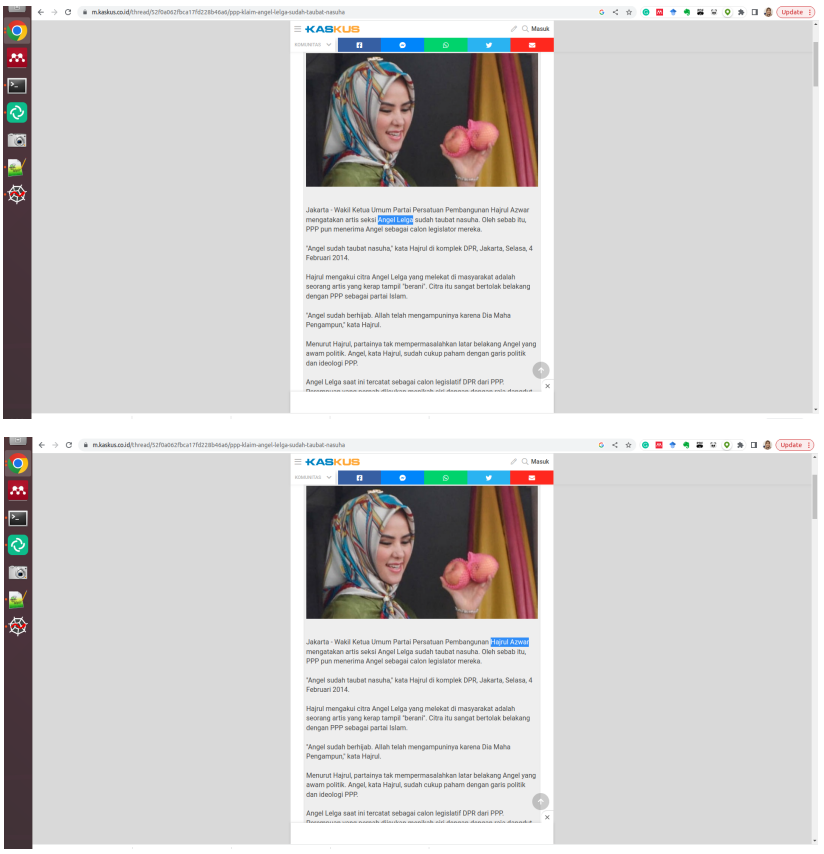
For example, *Angel* and *Hajrul* are incomplete names and no further information about them in this sentence "*"Angel sudah berhijab. Allah telah mengampuninya karena Dia Maha Pengampun , " kata Hajrul .*" (Angel has worn a hijab. Allah has forgiven her because He is the Most Forgiving," said Hajrul). Therefore, annotators can do these steps:

a. Find relevant documents that contain *Angel* and *Hajrul* from Google's retrieval process.

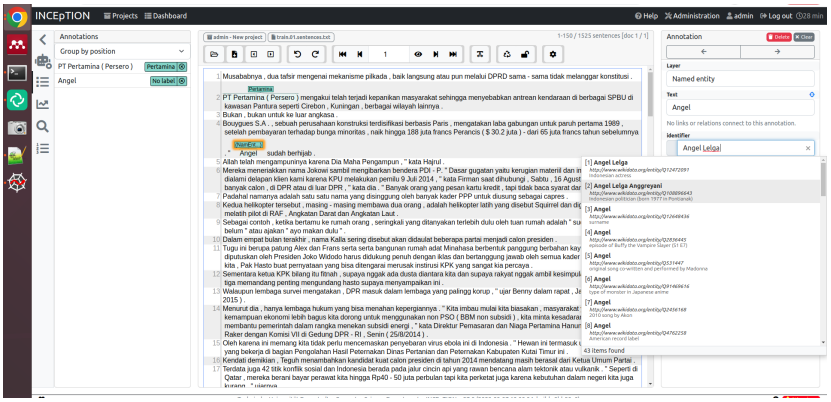


b. According to *snippets* returned by Google, we can know which ones contain the keyword and it is the fourth one.

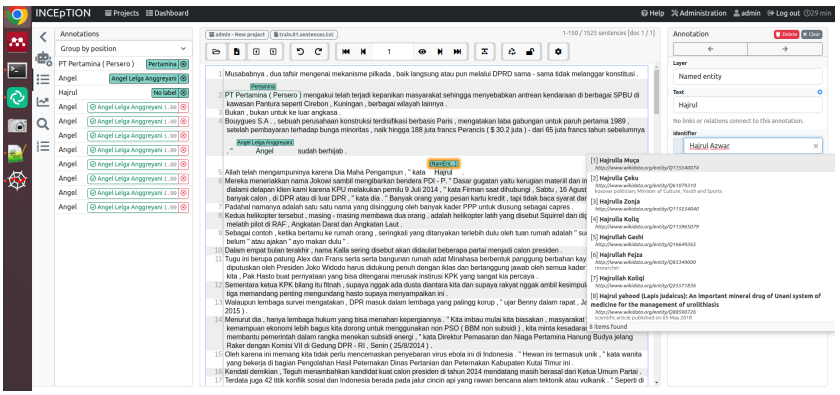
c. Select the snippet and find the complete name of *Angel* and *Hajrul*. In this case, *Angel* refers to *Angel Lelga*, and *Hajrul* refers to *Hajrul Azwar*.



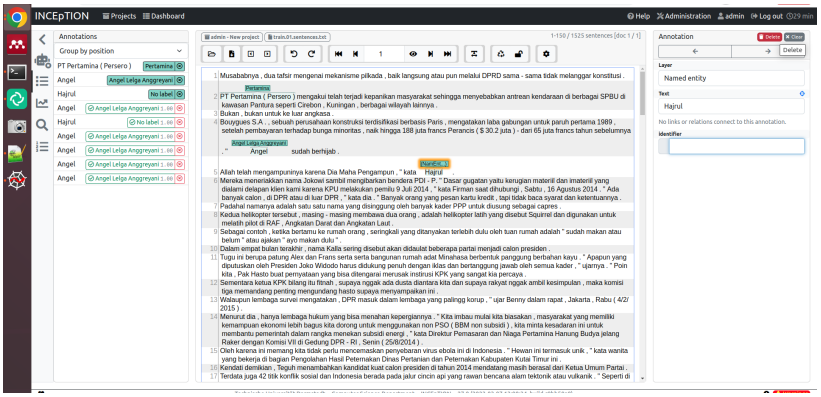
d. Repeat step at point i.



Note: The second entry is the right entry of *Angel*, because its profile is an *Indonesian politician (born 1977 in Pontianak)*. Even though *Angel Lelga* is recognized as a celebrity, in this context, she is mentioned as a politician. Therefore, rather than selecting the second option that labels her as a celebrity, the appropriate action is to choose the first option that identifies her as a politician.



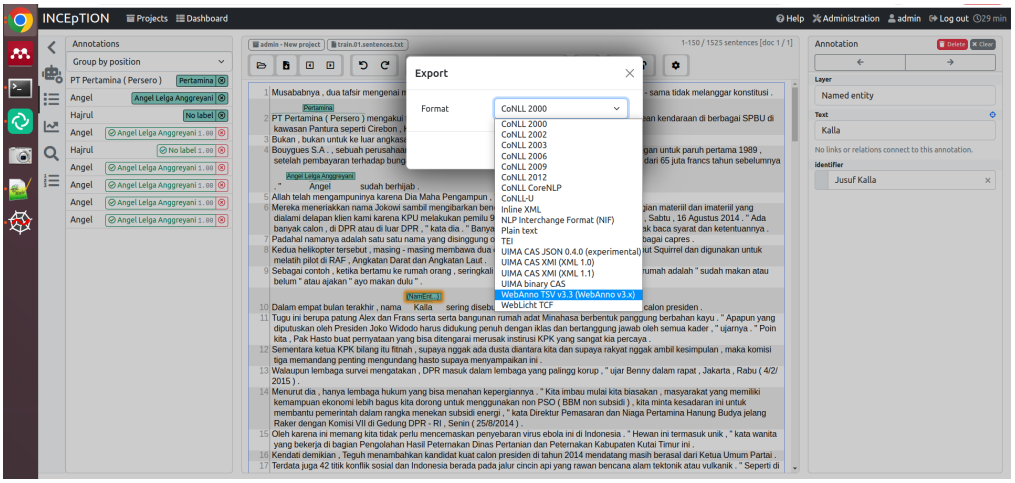
According to the above picture, *Hajrul* does not exist in Wikidata. In this case, *Hajrul* remains unlabeled. Annotators can click *Delete* button in the top right corner of the *editor* to cancel the labeling.



n. If the selected entity does not have a corresponding entry in Wikidata, then it remains unlabeled. Annotators are required to log this entity in the *NIL entities* list. NIL (Not in Lexicon) entities are those whose the corresponding links do not exist in Wikidata.

o. In the second provided document, it's possible that some NEs were not identified or missed. If annotators notice these omissions, they should proceed to label these entities according to the outlined steps.

p. After the annotation process is finished, annotators need to export the results in the format of *WebAnno TSV v3.3 (WebAnno v3.x)*. This can be done by selecting menu *Export Document* (second from the left) in the middle part of the editor. Send the results to ria.hari.gusmita@uni-paderborn.de.


INCEpTION
Projects
Dashboard
Help
Administration
admin
Log out
Annotations
Group by position
Export
Format
CoNLL 2000
CoNLL 2000
CoNLL 2002
CoNLL 2003
CoNLL 2006
CoNLL 2009
CoNLL 2012
CoNLL CoreNLP
CoNLL-U
Inline XML
NLP Interchange Format (NIF)
Plain text
TEI
UIMA CAS JSON 0.4.0 (experimental)
UIMA CAS XMI (XML 1.0)
UIMA CAS XMI (XML 1.1)
UIMA binary CAS
WebAnno TSV v3.3 (WebAnno v3.x)
WebLicht TCF
Annotation
Layer
Named entity
Text
Kalla
No links or relations connect to this annotation.
Identifier
Jusuf Kalla