

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

## **1. Introduction**

These guidelines are designed to providing direction in manual annotation of the 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. 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.

"Menyediakan 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>.

"Menyediakan 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, 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.

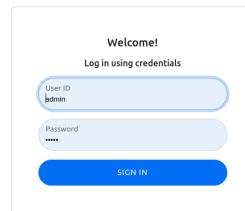
Penyerahan bantuan secara simbolis dilakukan Jokowi di Desa Sanggeng, Kecamatan Manokwari Barat, Manokwari, Selasa (5/4/2016).

Squawka mencatat Morgan sebagai pemain yang berbahaya dalam tekel dan dominan dalam duel di udara.

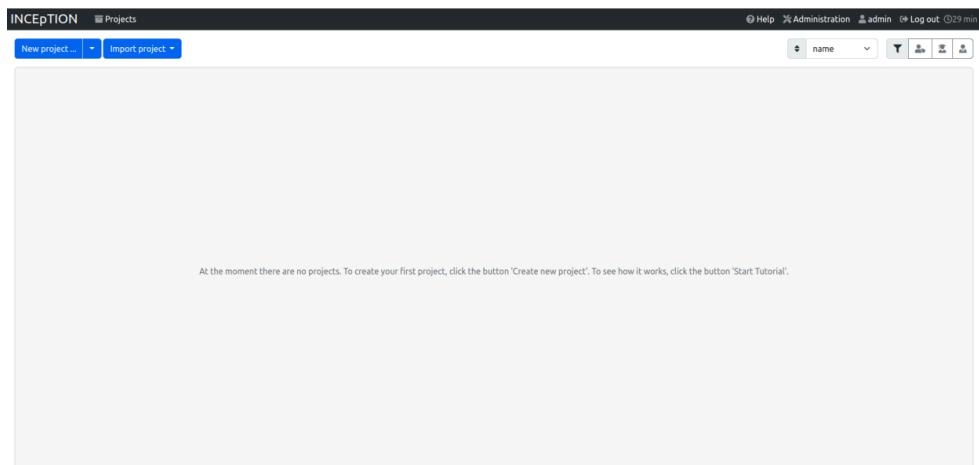
### 3. How to Annotate

The annotation will be performed by following these steps:

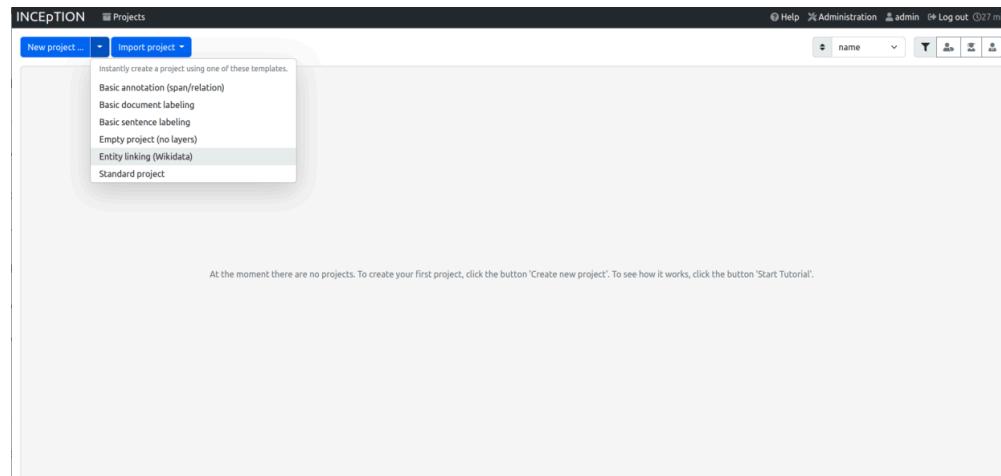
- a. Download the annotation tool, INCEpTION, at <https://inception-project.github.io/downloads/>
- b. Install and run the application following the steps explained at [https://inception-project.github.io/releases/27.1/docs/user-guide.html#sect\\_installation](https://inception-project.github.io/releases/27.1/docs/user-guide.html#sect_installation)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



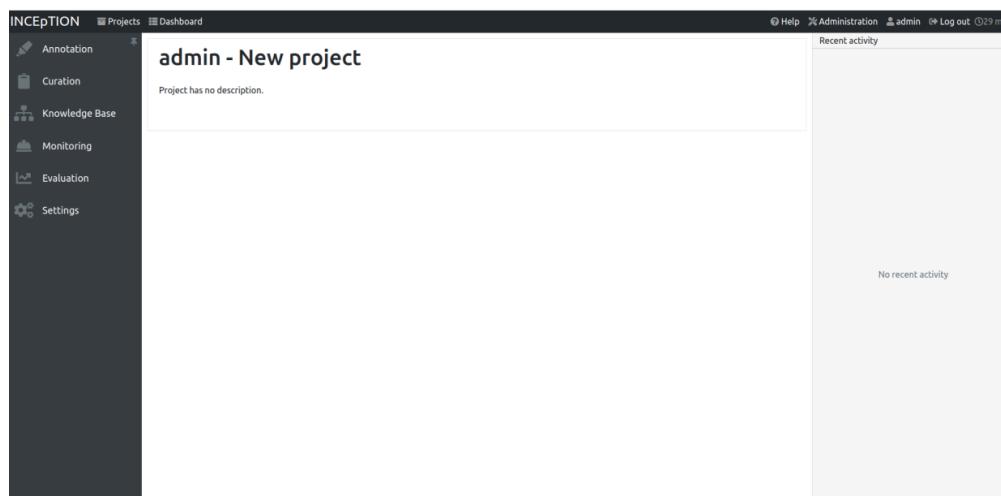
This is the display once annotators have logged in successfully.



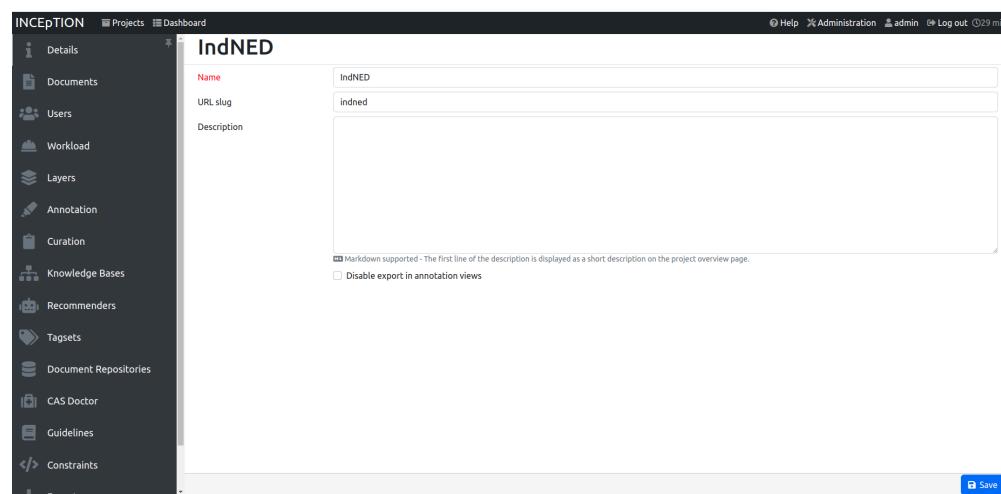
- d. Create a new project by clicking the arrow icon on *New Project* menu, then choose *Entity Linking (Wikidata)* from the sub-menus.



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:
1. The project's name, for example: *IndNED*
  2. Ignore the *URL slug* and *Description*. It will be automatically filled with *IndNED*.
  3. Click *Save*.



- f. Click on *Documents*. Then, upload the first document (refer to point 2, and sub-point d). This is the display after annotators have selected the document.

The screenshot shows the INCEPTION IndNED interface. On the left, there is a sidebar with various icons and labels: Details, Documents, Users, Workload, Layers, Annotation, Curation, Knowledge Bases, Recommenders, Tagsets, Document Repositories, CAS Doctor, Guidelines, and Constraints. The main area is titled 'IndNED' and shows a table with one row. The table has columns for 'State Name' and 'Format'. The row contains 'Corpus-Group 1.txt' under 'State Name', 'Plain text' under 'Format', and '2023-03-17' under 'Created'. There are also 'Delete' and 'Export' buttons at the bottom of the table.

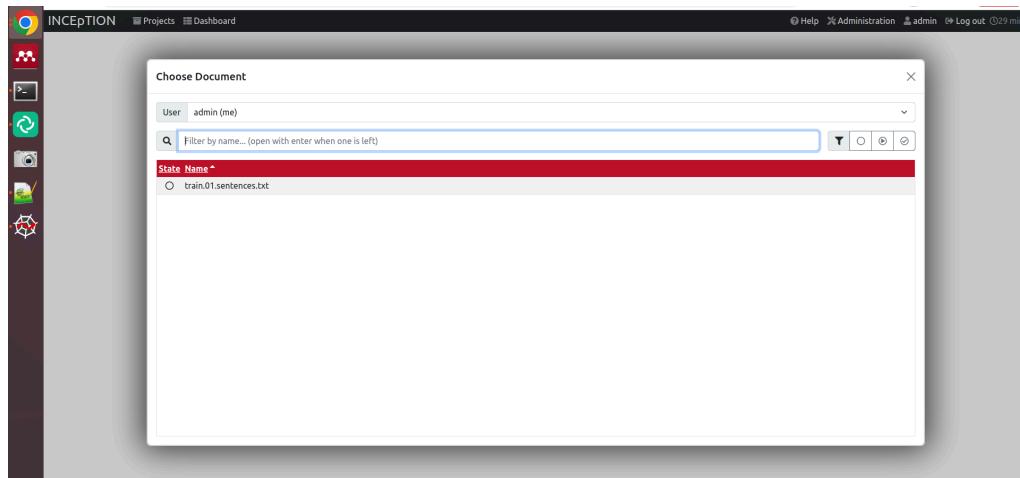
- g. Click on *Annotation*. Then, set the value of *Default page size* to 150. Enable the checkbox *Allow changing the script direction in the annotation editor*. Then, select *Annotations* at *Default tab* item, disable checkbox *Case sensitive*, and click *Save*.

The screenshot shows the 'Annotation' settings page. It includes sections for 'Annotation editor', 'Brat annotation editor', 'Annotation sidebar', and 'Annotation search'. In the 'Brat annotation editor' section, the 'Default page size' is set to '150' and the 'Script direction' is set to 'left-to-right'. The 'Annotation sidebar' section shows 'Default tab' set to 'Annotations'. In the 'Annotation search' section, the 'Case sensitive' checkbox is unchecked. At the bottom right, there is a blue 'Save' button.

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

The screenshot shows the 'Knowledge Bases' section. It displays a table with one row. The table has columns for 'Knowledge Bases' and 'Create'. The row contains 'Wikidata (direct mapping)' under 'Knowledge Bases' and a small 'Create' button under 'Create'.

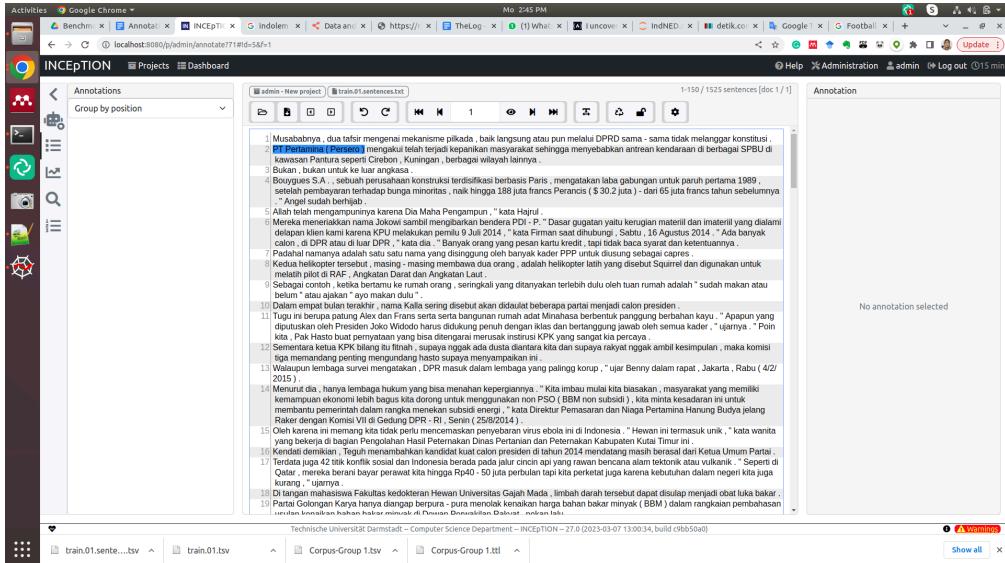
- i. Click on *Dashboard* item, then choose sub-menu *Annotation*.



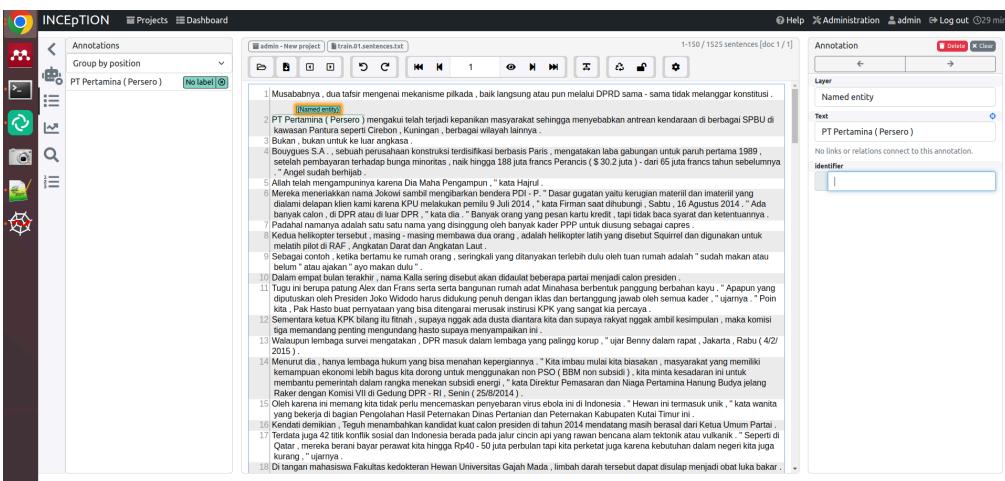
- j. Click on the document's name that has been uploaded, for instance, *train.01.sentences.txt*. You will be directed to the annotation editor page. On this page, there are three parts (left-to-right): Summary of annotation results, text as the annotation workspace, and the places where the corresponding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

Annotation ID	Text	Description
1	Masababnya, dua tafir mengenai mekanisme pilkada, baik langsung atau pun melalui DPRD sama-sama tidak melanggar konstitusi.	Menyatakan bahwa dua tafir mengenai mekanisme pilkada, baik langsung atau pun melalui DPRD sama-sama tidak melanggar konstitusi.
2	PT Pertamina (Persero) mengaku telah terjadi kepenaan masyarakat sehingga menyebabkan antrean kendaraan di berbagai SPBU di kawasan Pantura dan Cirebon, Kuningan, berbagai wilayah lainnya.	Menyatakan bahwa PT Pertamina (Persero) mengaku telah terjadi kepenaan masyarakat sehingga menyebabkan antrean kendaraan di berbagai SPBU di kawasan Pantura dan Cirebon, Kuningan, berbagai wilayah lainnya.
3	Banyak orang yang belum lama ini	Menyatakan bahwa banyak orang yang belum lama ini
4	Boxygues S.A., sebuah perusahaan konstruksi terdirifikasi berbasis Paris, mengatakan laba gabungan untuk pertama 1989, setelah pembayaran terhadap bunga minoritas, naik hingga 188 juta francs Perancis (\$ 30,2 juta) - dari 65 juta francs tahun sebelumnya.	Menyatakan bahwa Boxygues S.A., sebuah perusahaan konstruksi terdirifikasi berbasis Paris, mengatakan laba gabungan untuk pertama 1989, setelah pembayaran terhadap bunga minoritas, naik hingga 188 juta francs Perancis (\$ 30,2 juta) - dari 65 juta francs tahun sebelumnya.
5	Angel sudah berpindahnya karena Dia Mata Pengampun , kata Hajri .	Menyatakan bahwa Angel sudah berpindahnya karena Dia Mata Pengampun , kata Hajri .
6	Mereka menantakan nama Jokowi sembilan bendera PDF - P . Dasar gugatan yaitu kerugian materil dan imateril yang dialami dalam klien kami karena KPU melakukan pemilihan 9 Juli 2014 , " kata Firman saat diinterviu , Sabtu , 16 Agustus 2014 . " Ada banyak calon di DPR atau di luar DPR , " kata dia . " Banyak orang yang pesan kartu kredit , tapi tidak bisa bayar dan keterutamanya .	Menyatakan bahwa mereka menantakan nama Jokowi sembilan bendera PDF - P . Dasar gugatan yaitu kerugian materil dan imateril yang dialami dalam klien kami karena KPU melakukan pemilihan 9 Juli 2014 , " kata Firman saat diinterviu , Sabtu , 16 Agustus 2014 . " Ada banyak calon di DPR atau di luar DPR , " kata dia . " Banyak orang yang pesan kartu kredit , tapi tidak bisa bayar dan keterutamanya .
7	Dua helikopter itu adalah dua helikopter yang biasanya digunakan oleh banyak kader PPP untuk diliusung sebagai capres .	Menyatakan bahwa dua helikopter tersebut adalah dua helikopter yang biasanya digunakan oleh banyak kader PPP untuk diliusung sebagai capres .
8	Kedua helikopter tersebut, misang - misang membawa dua orang, adalah helikopter milik yang disebut Sugarni dan digunakan untuk melatih pilot di RAF, Angkatan Darat dan Angkatan Laut .	Menyatakan bahwa kedua helikopter tersebut, misang - misang membawa dua orang, adalah helikopter milik yang disebut Sugarni dan digunakan untuk melatih pilot di RAF, Angkatan Darat dan Angkatan Laut .
9	Sebagian cerita , ketika berbicara ke rumah orang , sering kali yang ditanyakan terlebih dulu oleh tuan rumah adalah " Sudah makan atau belum ? " atau " Sudah minum atau belum ? "	Menyatakan bahwa sebagian cerita , ketika berbicara ke rumah orang , sering kali yang ditanyakan terlebih dulu oleh tuan rumah adalah " Sudah makan atau belum ? " atau " Sudah minum atau belum ? "
10	Dalam empat bulan terakhir , nama Kolia sering disebut akan diduduki beberapa partai menjadi calon presiden .	Menyatakan bahwa dalam empat bulan terakhir , nama Kolia sering disebut akan diduduki beberapa partai menjadi calon presiden .
11	Tugu ini berupa patung Alex dan Frans serta serta bengungan rumah adat Minahasa berbentuk panggung berbahan kayu . " Apapun yang dipersoalkan oleh Presiden Joko Widodo harus diadukan perih dengan benar dan bertanggung jawab oleh semua kader , " ujarnya . " Poin kita , Presiden Joko Widodo yang bersama-sama dengan kita mengadakan persidangan di KPK yang sangat kita percaya .	Menyatakan bahwa tugu ini berupa patung Alex dan Frans serta serta bengungan rumah adat Minahasa berbentuk panggung berbahan kayu . " Apapun yang dipersoalkan oleh Presiden Joko Widodo harus diadukan perih dengan benar dan bertanggung jawab oleh semua kader , " ujarnya . " Poin kita , Presiden Joko Widodo yang bersama-sama dengan kita mengadakan persidangan di KPK yang sangat kita percaya .
12	Sementara ketua KPK bilang itu fitnah , supaya ngletek ada dusya diantara kita dan supaya nylek nylek ambil kesimpuln , maka komisi tiga memandang pending mengundang hasil supaya menyampaikan ini .	Menyatakan bahwa sementara ketua KPK bilang itu fitnah , supaya ngletek ada dusya diantara kita dan supaya nylek nylek ambil kesimpuln , maka komisi tiga memandang pending mengundang hasil supaya menyampaikan ini .
13	Walaupun lembaga survei mengatakan , DPR masuk dalam lembaga yang paling korup , " ujar Benny dalam rapat , Jakarta , Rabu (4/2/2014) .	Menyatakan bahwa walaupun lembaga survei mengatakan , DPR masuk dalam lembaga yang paling korup , " ujar Benny dalam rapat , Jakarta , Rabu (4/2/2014) .
14	Menurut dia , hanya lembaga hukum yang bisa menahan kepergiannya . " Kita imbau mulai kita biseksan , masyarakat yang memiliki kepercayaan akan lembaga hukum kita dan untuk masyarakat non PSO (BBM non subsidi ) , kita minta kesadaran ini untuk mendukung pemerintah dalam rangka mencegah adat istiadat negara Direktorat Pemasaran dan Niaga Pertamina Hanur Budya jelang Raker dengan Komisi VII di Gedung DPRP - RI , Senin (5/2/2014) .	Menyatakan bahwa hanya lembaga hukum yang bisa menahan kepergiannya . " Kita imbau mulai kita biseksan , masyarakat yang memiliki kepercayaan akan lembaga hukum kita dan untuk masyarakat non PSO (BBM non subsidi ) , kita minta kesadaran ini untuk mendukung pemerintah dalam rangka mencegah adat istiadat negara Direktorat Pemasaran dan Niaga Pertamina Hanur Budya jelang Raker dengan Komisi VII di Gedung DPRP - RI , Senin (5/2/2014) .
15	Oleh karena ini memang kita tidak perlu mencemaskan penyebaran virus ebola ini di Indonesia , " Hewan ini termasuk unk , " kata wanita yang bekerja di bagian Pengelahan Hasil Pertambahan Dinas Pertambahan dan Penerbangan Kabupaten Kutai Kartanegara di Samarinda .	Menyatakan bahwa oleh karena ini memang kita tidak perlu mencemaskan penyebaran virus ebola ini di Indonesia , " Hewan ini termasuk unk , " kata wanita yang bekerja di bagian Pengelahan Hasil Pertambahan Dinas Pertambahan dan Penerbangan Kabupaten Kutai Kartanegara di Samarinda .
16	Terdata sekitar 42 ribu ton kerikil sosial di Indonesia yang pada jahr circa 400 ton yang merupakan bencana alam tektonik atau vulkanik . " Seperti di Qatar , mereka berani bayar perawat kita hingga Rp40 - 50 juta perbulan tapi kita perkelet juga karena kebutuhan dalam negeri kita juga kurang , " ujarnya .	Menyatakan bahwa terdata sekitar 42 ribu ton kerikil sosial di Indonesia yang pada jahr circa 400 ton yang merupakan bencana alam tektonik atau vulkanik . " Seperti di Qatar , mereka berani bayar perawat kita hingga Rp40 - 50 juta perbulan tapi kita perkelet juga karena kebutuhan dalam negeri kita juga kurang , " ujarnya .
17	Ditangkap mahasiswa Fakultas kedokteran Hasan Universitas Cagak Mada , limbah dari tersebut dapat disulap menjadi obat luka bakar .	Menyatakan bahwa ditangkap mahasiswa Fakultas kedokteran Hasan Universitas Cagak Mada , limbah dari tersebut dapat disulap menjadi obat luka bakar .
18	Petra Golongan Karya hanya dianggap berputus - putus memakai kerikil karena bahan bakar minyak ( BBM ) dalam rangkaian pembatasan	Menyatakan bahwa Petra Golongan Karya hanya dianggap berputus - putus memakai kerikil karena bahan bakar minyak ( BBM ) dalam rangkaian pembatasan

- k. Select NEs in a sentence according to information in the second document (refer to point 2 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.



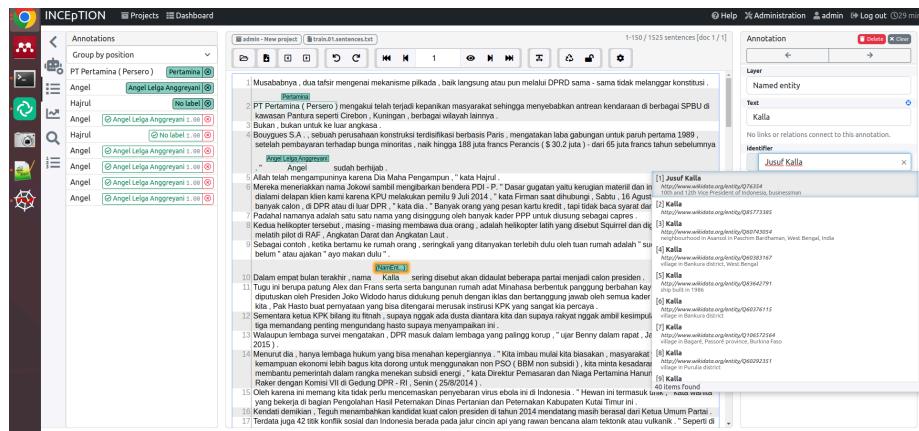
- m. Make sure the selected NE appears on *Text* item. Then, if annotators are certain of which specific object is being represented by the selected NE, annotators can enter the keyword into the search menu on *Identifier* item to find the corresponding entity in Wikidata. When annotators find a list 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:
  1. Use the common term/name of the selected NE as the keyword. For instance, *Pertamina* is the common name of *PT. Pertamina (Persero)*.

The screenshot shows the INCEPTION annotation interface. On the left, there's a sidebar with icons for file operations, search, and navigation. The main area is titled 'Annotations' and shows a list of sentences. The first sentence is highlighted with a blue background and contains the text: 'Musababnya, dua tafsir mengenai mekanisme pilk'. To the right of the text, there's an annotation panel with a 'Text' field containing 'PT Pertamina (Persero)' and a 'Identifier' field containing 'Pertamina'. Below the identifier, it says 'No links or relations connect to this annotation.'

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.

This screenshot shows the same INCEPTION interface as the previous one, but with a different set of sentences. The first sentence is again highlighted with a blue background and contains the text: 'Musababnya, dua tafsir mengenai mekanisme pilk'. The annotation panel on the right shows the same information: 'Text' field with 'PT Pertamina (Persero)' and 'Identifier' field with 'Pertamina'. The sidebar on the left is visible, showing icons for file operations, search, and navigation.

2. 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.

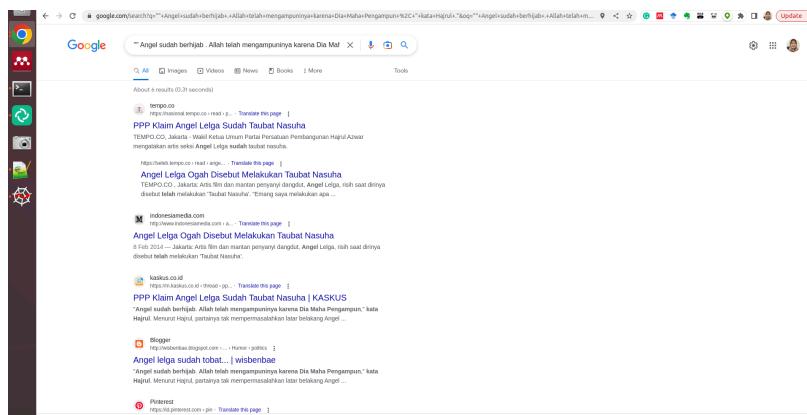


Note: The correct entry in Wikidata is the first one, as its profile is the *10<sup>th</sup> and 12<sup>th</sup> Vice President of Indonesia, businessman*.

3. 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 *m.1*.

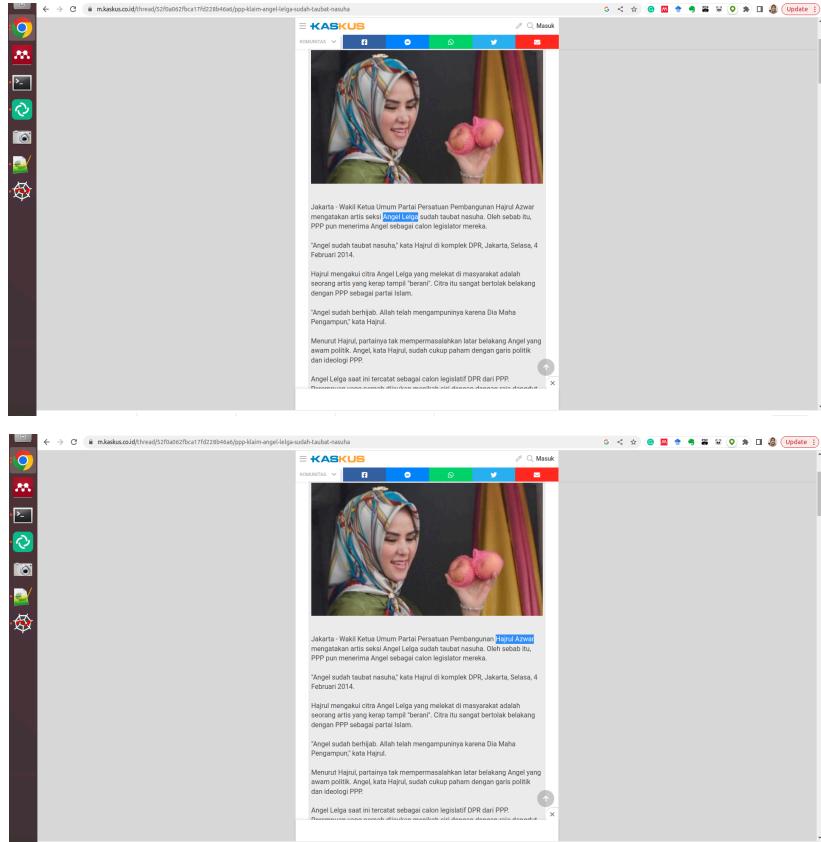
For example, *Angel* and *Hajrul* are incomplete names and no further information about them in this sentence “*“Angel sudah berhijab. Allah telah mengampuninya karena Dia Maha Pengampun , ” kata Hajrul .*” (*Angel* has worn a hijab. Allah has forgiven her because He is the Most Forgiving,” said Hajrul).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

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



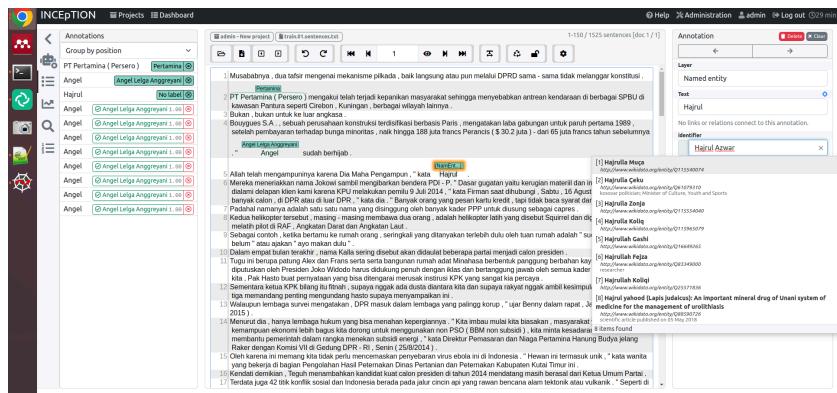
According to *snippets* returned by Google, we can know which ones contain the keyword and it is the fourth one.

- b. 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*.

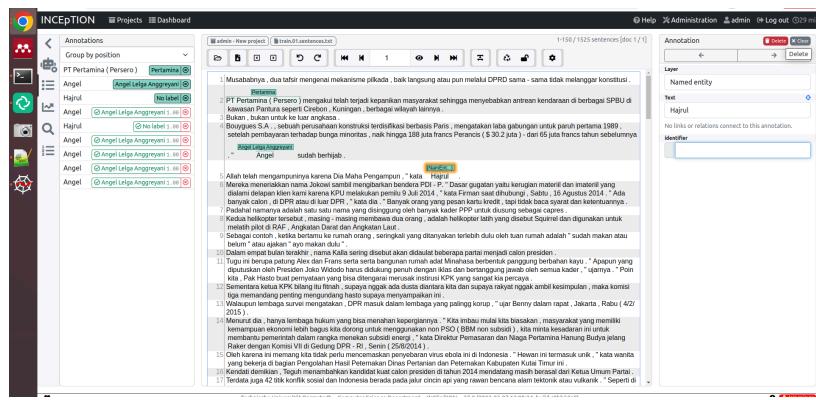


c. Repeat step at point m.1.

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.

The screenshot shows the INCEption annotation software interface. On the left, there's a sidebar with icons for file operations like Open, Save, and Print. Below that is a tree view of annotations under 'Annotations' with categories like 'Group by position' and 'Named entity'. A main pane displays a document with several annotations. An annotation for 'Pertamina' is selected, showing its details: 'Angel' is the label, 'Angel Leiga Anggryani' is the text, and 'Pertamina' is the entity. A modal dialog titled 'Export' is open, showing a dropdown menu for 'Format' with options like CoNLL 2000, CoNLL 2009, and CoNLL 2012. The right side of the interface includes a header with 'Help', 'Administration', 'admin', 'Log out', and a timer ('029 min'). It also has sections for 'Annotation' (with a 'Delete' button), 'Layer' (with 'Named entity'), 'Text' (containing 'Kalla'), and 'Identifier' (containing 'Jusuf Kalla'). The bottom status bar indicates '1/150 | 1525 sentences [doc 1 / 1]'.