

# **Pedoman Pelabelan Entitas Bernama pada Pembangunan Dataset Pedoman Disambiguasi Entitas Bernama Berbahasa Indonesia untuk Domain Umum dan Khusus**

## **1. Nama kegiatan**

Kegiatan ini merupakan kegiatan melabeli entitas bernama (EB) dengan tautan korespondensi (*corresponding link*) kemunculan entitas tersebut di basis pengetahuan (*knowledge base*), Wikidata Bahasa Indonesia (selanjutnya disebut Wikidata). Entitas bernama merupakan nama dari sebuah obyek spesifik yang ada di dunia. Contoh entitas bernama antara lain:

- a. UIN Syarif Hidayatullah Jakarta (nama obyek universitas spesifik)
- b. Jl. Ir. H. Juanda (nama obyek lokasi spesifik)
- c. BEM UIN Jakarta (nama obyek organisasi spesifik)

Hasil pelabelan akan menjadi dataset pedoman untuk proses disambiguasi entitas bernama menggunakan Wikidata untuk teks berbahasa Indonesia. Disambiguasi entitas bernama adalah proses mengidentifikasi profil entitas bernama tersebut berdasarkan konteks di kalimat tempat entitas bernama tersebut muncul. Contoh proses disambiguasi ditunjukkan pada kalimat berikut:

“Kemacetan parah terjadi di Juanda sejak pagi hari, kata satpam yang bertugas di kampus UIN Jakarta dan turut membantu mengurai kemacetan.”

Kalimat di atas memiliki dua entitas bernama, yakni *Juanda* dan *kampus UIN Jakarta*. Berdasarkan konteks di kalimat, *Juanda* dan *kampus UIN Jakarta* adalah nama tempat spesifik (jalan untuk *Juanda* dan bangunan tempat UIN Jakarta beroperasi untuk *kampus UIN Jakarta*). Proses disambiguasi akan menentukan bahwa *Juanda* mengacu ke nama jalan yang melewati kampus UIN Jakarta, bukan nama bandara di Surabaya atau nama stasiun kereta api di Jakarta. Hal ini dikarenakan, berdasarkan konteks kalimat, yang menyampaikan informasi tentang kemacetan di *Juanda* adalah satpam yang bertugas di kampus UIN Jakarta yang turut serta mengurai kemacetan. Sementara, disambiguasi pada *kampus UIN Jakarta* menghasilkan kampus tempat UIN Jakarta beroperasi.

## **2. Mekanisme pelabelan**

- a. Pelabelan dilakukan pada dokumen yang berasal dari dataset pengenalan entitas bernama (*named entity recognition*) untuk Bahasa Indonesia, NER-UI, untuk domain umum dan IndQNER untuk domain spesifik. Dataset NER-UI mengandung 2114 kalimat dan 5055 EB dari kelas Orang, Lokasi, dan Organisasi. IndQNER mengandung 3117 kalimat dan 2475 EB, mewakili 18 kelas EB.

- b. Pelabelan dilakukan secara paralel oleh dua dan satu kelompok pelabel masing-masing pada domain umum dan khusus, di mana masing-masing kelompok terdiri dari dua orang.
- c. Pada domain umum, masing-masing dari dua pelabel dalam satu kelompok akan melabeli 1063 kalimat, sementara semua pelabel pada domain spesifik akan melabeli 3117 kalimat.
- d. Setiap pelabel akan memperoleh dua jenis dokumen yang dibutuhkan untuk melakukan pelabelan. Dokumen 1 (Corpus-Group 1.txt/Corpus-Group 2.txt) berisi daftar kalimat yang tidak memiliki label apapun. Dokumen ini merupakan dokumen yang setiap entitas bernama di dalamnya akan dilabeli dengan tautan korespondensi terkait di Wikidata. Berikut adalah contoh isi dokumen pertama.

" 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 .

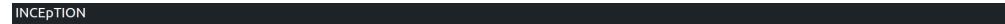
Dokumen 2 (Labeled corpus-Group 1.txt/Labeled corpus-Group 2.txt) berisi daftar kalimat yang ada di dokumen 1, namun setiap entitas bernama dilengkapi dengan label berupa tipe entitas bernama tersebut. Label yang dimaksud akan membantu pelabel mengetahui entitas bernama apa saja yang muncul di kalimat serta tipe/jenis entitas bernama tersebut. Berdasarkan informasi tersebut, pelabel dapat memilih bagian di kalimat yang akan dilabeli dengan tautan korespondensi di Wikidata yang tepat. Berikut ini contoh isi dokumen kedua.

" 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 mengkritik wartawan yang selalu menyebutkan Laksamana sebagai mantan orang PDIP  
" Jika kami tidak tampil bagus maka itu akan memudahkan Dortmund  
Kalau Davenport 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 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 Wawrinka andalan Swiss mencari trofi Piala Davis pertamanya dengan melawan Prancis pekan depan .  
Banyak kritik yang dilontarkan dari berbagai pihak , termasuk dari PDIP  
Penyerahan bantuan secara simbolis dilakukan Jokowi Desa Sanggeng Kecamatan Manokwari Barat Manokwari Selasa ( 5/4/2016 ) .  
Squawka Morgan pemain yang berbahaya dalam tekel dan dominan dalam duel di udara .

### 3. Tahapan Pelabelan

Pelabelan dilakukan melalui tahapan berikut ini:

- a. Unduh alat bantu pelabelan, INCEpTION, di <https://inception-project.github.io/downloads/>
- b. Install dan jalankan aplikasi sesuai panduan yang ada di [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. Jalankan aplikasi INCEpTION dengan memasukkan *admin* masing-masing untuk user name dan password.



Welcome!  
Log in using credentials

User ID: admin  
Password:

SIGN IN

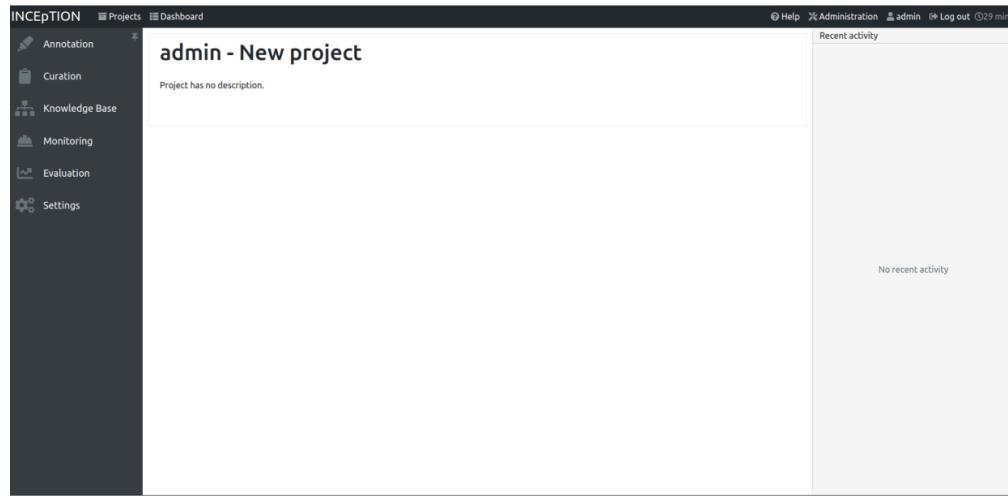
Berikut tampilan setelah berhasil *log in*.

This screenshot shows the main project management screen. At the top, there's a navigation bar with 'INCEPTION' and 'Projects'. Below it, buttons for 'New project ...' and 'Import project ...'. On the right, there are filters for 'name' and other search options. The central area displays a message: 'At the moment there are no projects. To create your first project, click the button 'Create new project'. To see how it works, click the button 'Start Tutorial'.' There are also icons for creating, deleting, and editing projects.

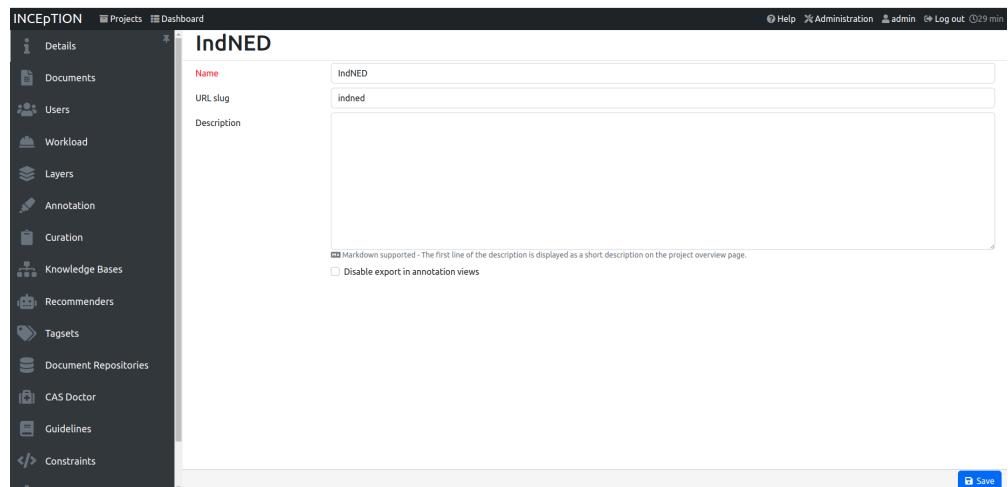
- d. Buat proyek baru dengan mengklik tombol panah di menu *New Project*, lalu pilih *Entity Linking (Wikidata)* pada sub menu.

A modal dialog box titled 'New project' is open. It contains a list of project templates: 'Basic annotation (span/relation)', 'Basic document labeling', 'Basic sentence labeling', 'Empty project (no layers)', 'Entity linking (Wikidata)' (which is highlighted), and 'Standard project'. Below the dialog, the same message as in the previous screenshot is visible: 'At the moment there are no projects. To create your first project, click the button 'Create new project'. To see how it works, click the button 'Start Tutorial'.'

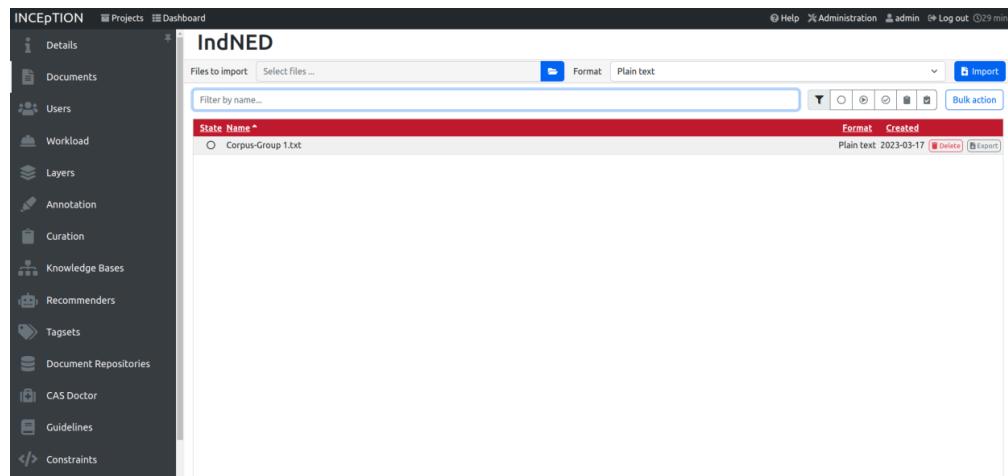
Berikut tampilan setelah pembuatan proyek baru berhasil dilakukan.



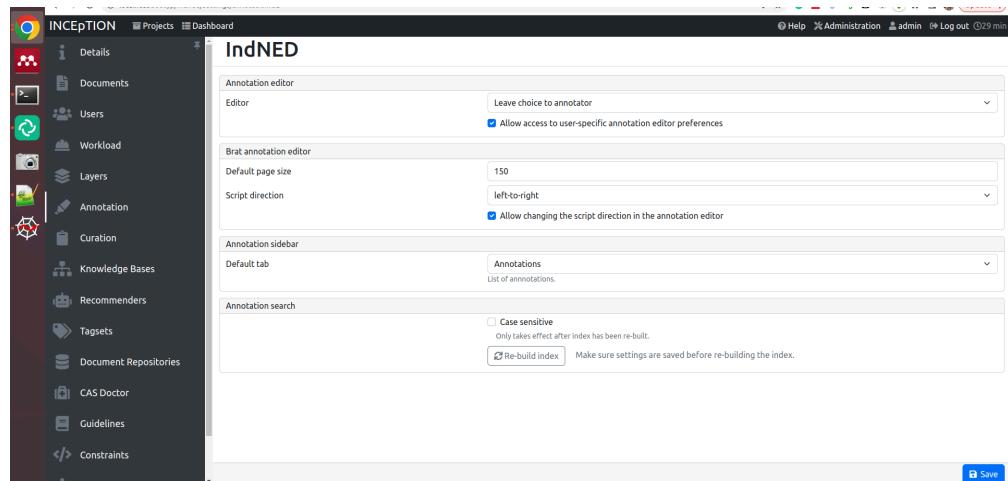
- e. Klik pada menu *Settings* untuk mengatur properti proyek.
  - i. Set nama proyek dengan, misalnya: *IndNED*
  - ii. Kosongkan bagian *URL slug* dan *Description*. Bagian *URL slug* akan otomatis diisi dengan *IndNED*.
  - iii. Klik tombol *Save* di bagian kanan bawah.



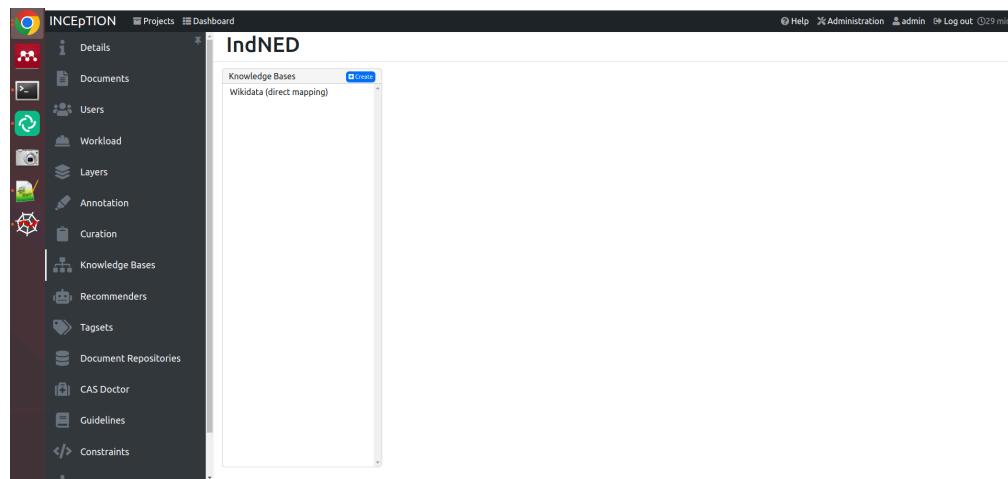
- f. Klik pada menu *Documents*. Lalu, upload file dokumen 1 (lihat seksi 2 bagian d). Berikut tampilan setelah mengupload dokumen 1 untuk kelompok 1 (Corpus-Group 1.txt).



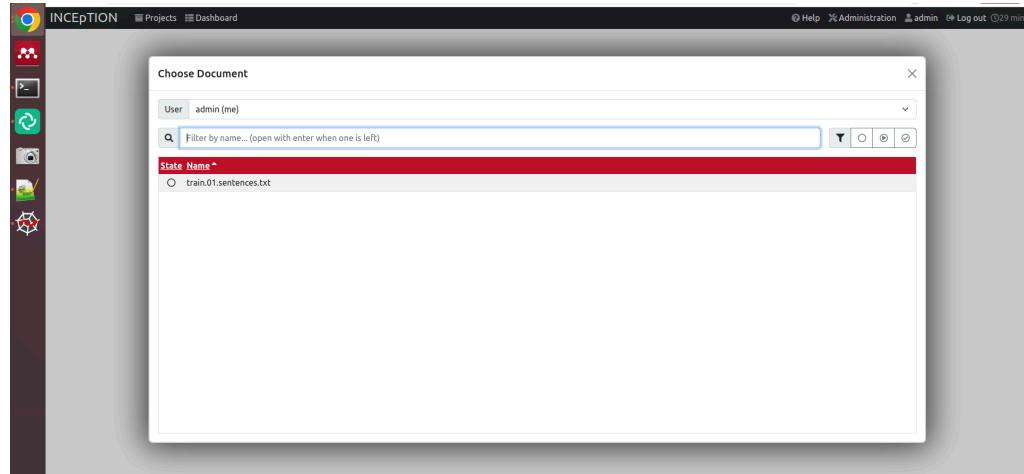
- g. Klik pada menu *Annotation*. Lalu, ubah nilai pada item *Default page size* menjadi 150. Centang checkbox *Allow changing the script direction in the annotation editor*. Lalu, pilih *Annotations* untuk item *Default tab*, hapus centang di checkbox *Case sensitive*, dan klik tombol *Save*.



- h. Klik pada menu *Knowledge Bases*. Pastikan bahwa *Wikidata (Direct Mapping)* telah terpilih.



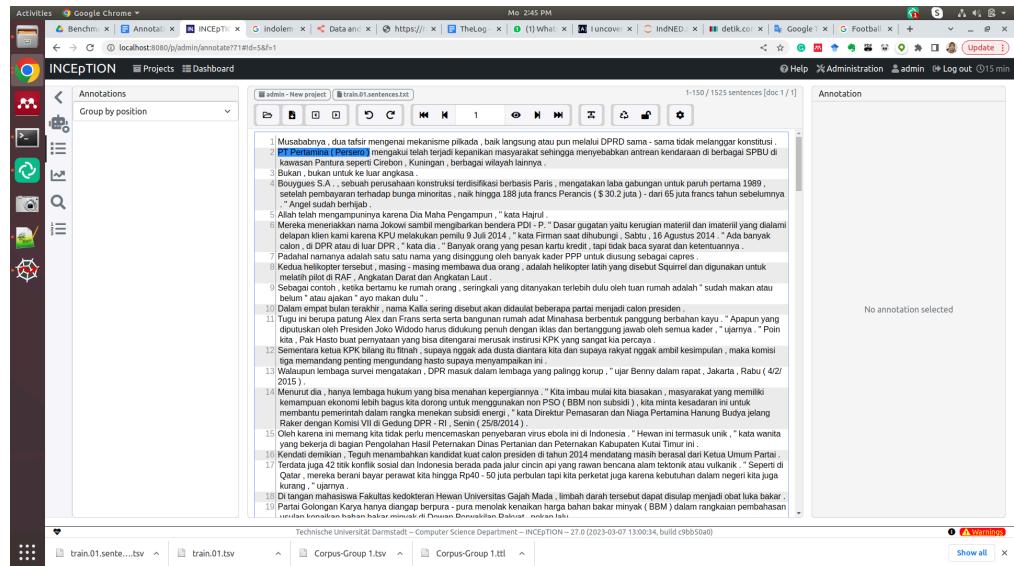
- i. Klik item *Dashboard*, lalu menu *Annotation*.



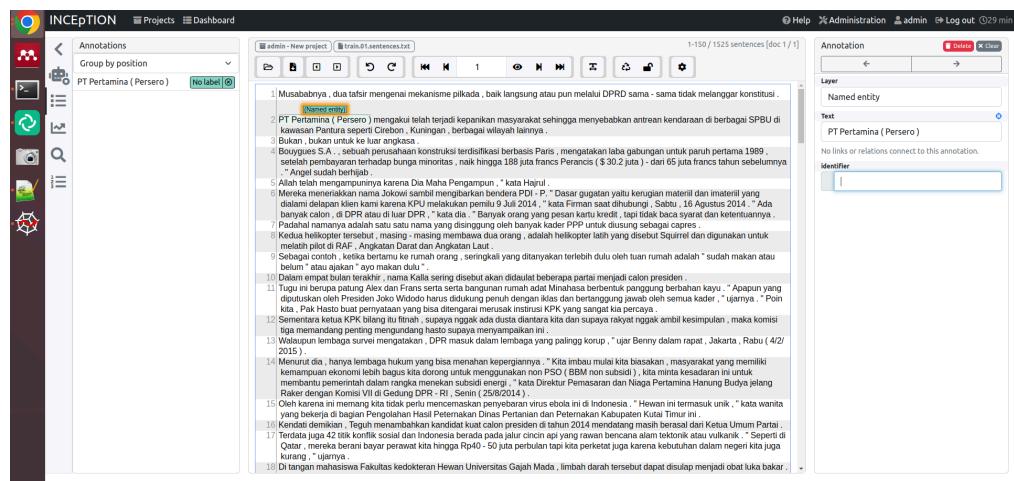
- j. Klik pada nama file dokumen yang telah diupload. Misalnya, train.01.sentences.txt. Tampilan berikutnya adalah editor pelabelan. Terdapat tiga bagian pada editor, yakni (kiri ke kanan), ringkasan hasil pelabelan (kiri), teks yang akan/telah dilabeli (tengah), dan tempat pendefinisian tautan entitas bernama terkait di Wikidata (kanan). Pilih *Group by position* di bagian *summary* untuk mendapatkan tampilan informasi yang lebih baik.

The screenshot shows the INCEPTION annotation interface. On the left, there's a sidebar with icons for file operations and a dropdown menu. The main area is titled 'Annotations' and has a sub-menu 'Group by position'. Below this, there's a list of sentences from the file 'train.01.sentences.txt'. Each sentence is displayed in a box with various colored highlights (blue, red, green) over specific words or phrases, indicating entity mentions. To the right of the list, there's a detailed view of one sentence with annotations highlighted. The bottom right corner of the interface says 'Annotation'.

- k. Pilih entitas bernama di setiap kalimat berdasarkan informasi di dokumen 2 (seksi 2 bagian d) dengan cara memblok (*click and drag*) seluruh bagian entitas bernama tersebut. Misalnya kita menemukan sebuah entitas bernama di kalimat kedua, yakni, *PT. Pertamina (Persero)*. Seluruh bagian dari entitas tersebut, dimulai dari karakter pertama (*P*) hingga karakter terakhir (*tanda kurung tutup*), diblok seperti tampilan berikut ini.



- l. Bagian kanan editor secara otomatis menampilkan *interface* untuk memberi tautan entitas terkait di Wikidata.



- m. Pastikan bahwa entitas bernama yang telah diblok telah muncul di item *Text*. Lalu, setelah pelabel mengetahui obyek tertentu mana yang diwakili oleh entitas terpilih, masukkan kata kunci terkait di kotak pencarian pada item *Identifier* untuk menemukan kemunculan entitas terkait di Wikidata. Ketika menemukan hasil pencarian, pastikan profil dari yang terpilih (ditulis di bawah nama entitas) sesuai dengan profil obyek yang dimaksud. Berikut ini beberapa teknik untuk menemukan kemunculan entitas terkait di Wikidata:

1. Menggunakan bagian yang sering digunakan dari entitas bernama sebagai kata kunci. contoh: *Pertamina* pada entitas *PT. Pertamina (Persero)*

The screenshot shows the INCEpTION annotation tool interface. On the left, there's a sidebar with icons for file operations, search, and annotations. The main area displays a list of sentences from a file named 'train.01.sentences.txt'. A specific sentence is highlighted in yellow: "1 Musababnya , dua tafsir mengenai mekanisme pilk [Pertamina]". To the right of the list, there's an annotation panel titled 'Annotation' with fields for 'Text' (containing 'PT Pertamina (Persero)') and 'Identifier' (containing 'Pertamina'). Below these fields, it says 'No links or relations connect to this annotation.'

Hasil pencarian pertama adalah Pertamina dengan profil: *Indonesian state-owned oil and gas company*, di mana profil ini sama dengan profil obyek yang dimaksud, yakni perusahaan minyak milik negara Indonesia. Dengan demikian, hasil pencarian pertama dinyatakan sebagai entitas Wikidata yang diacu oleh entitas *PT. Pertamina (Persero)* pada kalimat kedua. Klik hasil pencarian pertama tersebut, dan pelabelan secara otomatis akan dilakukan. Untuk memastikan pelabelan telah dilakukan, cek bagian kiri editor. Gambar di bawah ini menunjukkan bahwa entitas *PT. Pertamina (Persero)* telah dilabeli dengan tautan korespondensi (URL) di Wikidata yang tepat (dengan nama *Pertamina*).

This screenshot shows the same INCEpTION interface as the previous one, but with a notable difference: the entity 'Pertamina' is now explicitly labeled in the list of results. The sentence "1 Musababnya , dua tafsir mengenai mekanisme pilk [Pertamina]" has 'Pertamina' highlighted in a yellow box. This indicates that the system has successfully identified and labeled the entity within the context of the sentence.

2. Jika entitas bernama berbentuk nama yang tidak lengkap, namun pelabel dapat mengetahui obyek spesifik mana yang diacunya berdasarkan konteks kalimat, maka gunakan nama lengkap entitas tersebut sebagai kata kunci. Contoh, entitas bernama *Kalla* di kalimat: "*Dalam empat bulan terakhir, nama Kalla sering disebut akan didaulat beberapa partai menjadi calon presiden.*" merupakan nama yang tidak lengkap. Namun, konteks kalimat membantu pelabel untuk mengetahui bahwa *Kalla* yang dimaksud adalah *Jusuf Kalla*. Dengan demikian, *Jusuf Kalla* digunakan sebagai kata kunci seperti berikut ini:

Catatan: Hasil pencarian pertama adalah yang benar, karena profilnya sama dengan profil obyek terkait, yakni wakil presiden ke-10 dan ke-12 RI (*10th and 12th Vice President of Indonesia, businessman*)

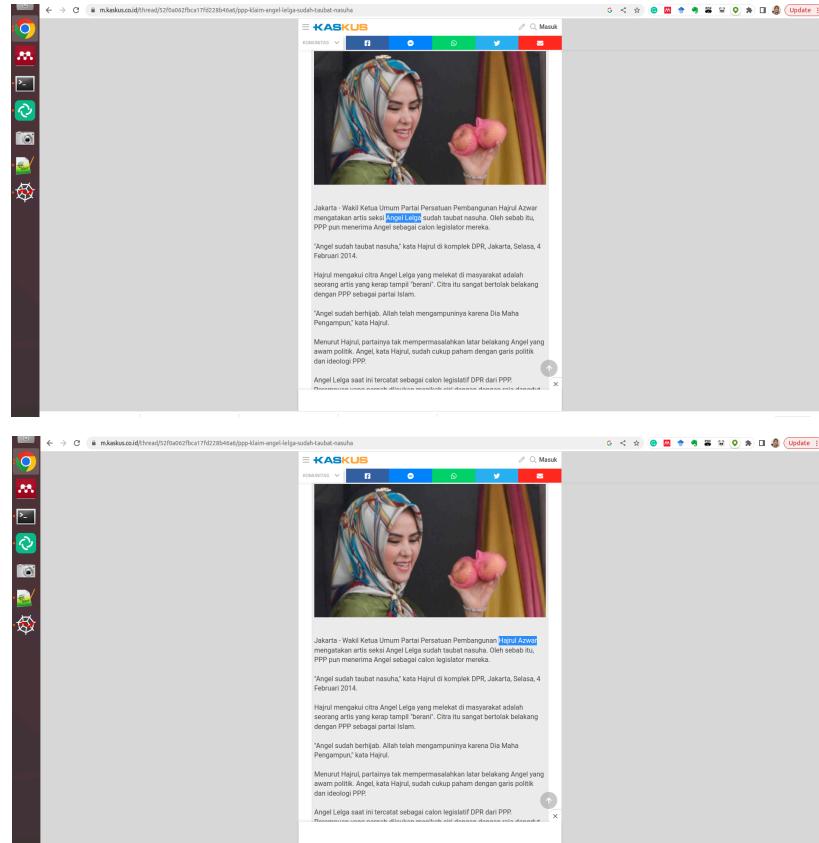
- Jika entitas bernama berbentuk nama yang tidak lengkap dan tidak ada informasi yang cukup di kalimat tentang obyek mana yang diwakilinya, maka lakukan langkah berikut ini:
  - Temukan dokumen web yang mengandung kalimat tersebut di Google. Berikan tanda petik ganda di bagian awal dan akhir kalimat untuk melakukan pencarian tepat (*exact matching*).
  - Temukan nama lengkap atau profil dari entitas bernama di dokumen terpilih (relevan)
  - Lakukan langkah di poin *m.1*

Contoh, *Angel* dan *Hajrul* merupakan entitas bernama yang tidak lengkap dan tidak ada informasi cukup tentang profil keduanya di kalimat berikut: “*Angel sudah berhijab. Allah telah mengampuninya karena Dia Maha Pengampun,*” kata *Hajrul*. Langkah yang dilakukan:

- Pencarian dokumen relevan di Google:

Dari *snippet* yang dikembalikan Google, kita dapat mengetahui bahwa dokumen keempat mengandung kalimat yang dimaksud.

- b. Buka dokumen terpilih. Isi dokumen membantu kita mengetahui bahwa nama *Angel* mengacu ke *Angel Lelga* (selebriti), dan nama *Hajrul* mengacu ke *Hajrul Azwar*.



- c. Lakukan langkah *m.1* sebagai berikut:

Catatan: Hasil pencarian kedua adalah yang dipilih, karena profilnya menyatakan Indonesian politician (born 1977 in Pontianak). Hal ini sesuai dengan konteks kalimat yang sedang membicarakan *Angel Lelga* sebagai politisi PPP, bukan sebagai selebritis Indonesia (profil hasil pencarian pertama).

The screenshot shows the INCEPTION annotation interface. A list of sentences from a document is displayed, with various entities annotated. One entity, 'Hajrul', is highlighted with a yellow box. The annotation details show its Wikidata ID (Q1158072) and a link to its Wikidata page.

Khusus untuk nama *Hajrul*, tidak ditemukan entitas terkait di Wikidata. Dalam kasus ini, pelabelan pada *Hajrul* tidak dilakukan. Klik tombol *Delete* pada bagian kanan atas *editor* untuk menggagalkan pelabelan.

The screenshot shows the INCEPTION annotation interface. A list of sentences from a document is displayed, with various entities annotated. One entity, 'Hajrul', is highlighted with a yellow box. The annotation details show its Wikidata ID (Q1158072) and a link to its Wikidata page.

- Jika sebuah entitas bernama tidak memiliki tautan terkait di Wikidata, maka entitas tersebut tidak perlu dilabeli. Selanjutnya, pelabel perlu mencatat entitas bernama tersebut di daftar *NIL entities*, yakni entitas yang tidak muncul di Wikidata.
- Pada dokumen kedua, terdapat kemungkinan entitas bernama yang tidak dilabeli (*missing entity*). Dalam hal ini, pelabel tetap perlu melabelinya dengan tautan korespondensi terkait di Wikidata. Gunakan informasi di poin *k*.
- Setelah pelabelan selesai dilakukan, unduh hasil pelabelan melalui menu *Export Document* (urutan kedua dari kiri) di bagian tengah editor. Pilih format *WebAnno TSV v3.3 (WebAnno v3.x)*.

The screenshot shows the INCEPTION annotation software interface. On the left, there's a sidebar with project management options like 'New project', 'Import', 'Export', 'Dashboard', and 'Logout'. The main area is titled 'Annotations' and shows a list of sentences from a file named 'Train.01.sentences.txt'. Each sentence has a unique ID, a text snippet, and an annotation status (e.g., 'No label'). A context menu icon is visible next to each sentence. An 'Annotation' dialog box is open in the top right, showing details for a specific sentence. The 'Text' field contains 'Kalla', and the 'Identifier' field contains 'Jusuf Kalla'. The 'Layer' dropdown is set to 'Named entity'. The 'Format' dropdown is set to 'CoNLL 2000'. The 'Annotation' dialog also includes sections for 'Text', 'Layer', 'Identifier', and 'Annotation' itself. At the bottom of the interface, there are navigation buttons for 'First', 'Previous', 'Next', 'Last', and 'Search'.