

Pedoman Pelabelan Entitas Bernama dengan Tautan Korespondensi di Wikidata

1. Nama kegiatan

Kegiatan ini merupakan kegiatan melabeli entitas bernama dengan tautan korespondensi (*corresponding link*) kemunculan entitas tersebut di basis pengetahuan (*knowledge base*), 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. Waktu pelaksanaan

Kegiatan ini dilaksanakan selama satu minggu, dimulai dari tanggal 20 Maret 2023 hingga 26 Maret 2023.

3. Pelabel (annotator)

Kegiatan ini dilaksanakan oleh empat orang pelabel yang merupakan mahasiswa/i semester 6 dan 8 di Program Studi Teknik Informatika, UIN Syarif Hidayatullah Jakarta.

4. Mekanisme pelabelan

- a. Pelabelan dilakukan pada dokumen yang berasal dari dataset pengenalan entitas bernama (*named entity recognition*) untuk Bahasa Indonesia, NER-UI. Dataset ini memiliki 2126 kalimat yang dilengkapi dengan label berupa tipe entitas bernama pada setiap entitas bernama di kalimat
- b. Pelabelan dilakukan secara paralel oleh dua kelompok pelabel, di mana masing-masing kelompok terdiri dari dua orang
- c. Dua pelabel di setiap kelompok akan melabeli 1063 kalimat yang sama secara terpisah (individu)
- d. Setelah pelabelan selesai dilakukan, *Inter-Annotation Agreement* (IAA) akan dihitung untuk mengetahui tingkat kesesuaian hasil pelabelan antara dua pelabel di setiap kelompok.
- e. 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 tekki 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 <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 <entity>Football Espana</entity>
<entity>Mega</entity> 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> 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> 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> <entity>Wawrinka</entity> andalan <entity>Swiss</entity> mencari trofi Piala Davis pertamanya dengan melawan <entity>Prancis</entity> pekan depan .
Banyak kritik yang dilontarkan dari berbagai pihak , termasuk dari <entity>PDIP</entity>
Penyerahan bantuan secara simbolis dilakukan <entity>Jokowi</entity> <entity>Desa Sanggeng</entity> <entity>Kecamatan Manokwari Barat</entity> <entity>Manokwari</entity> Selasa (5/4/2016) .
<entity>Squawka</entity> <entity>Morgan</entity> pemain yang berbahaya dalam tekel dan dominan dalam duel di udara
.

5. Tahapan Pelabelan

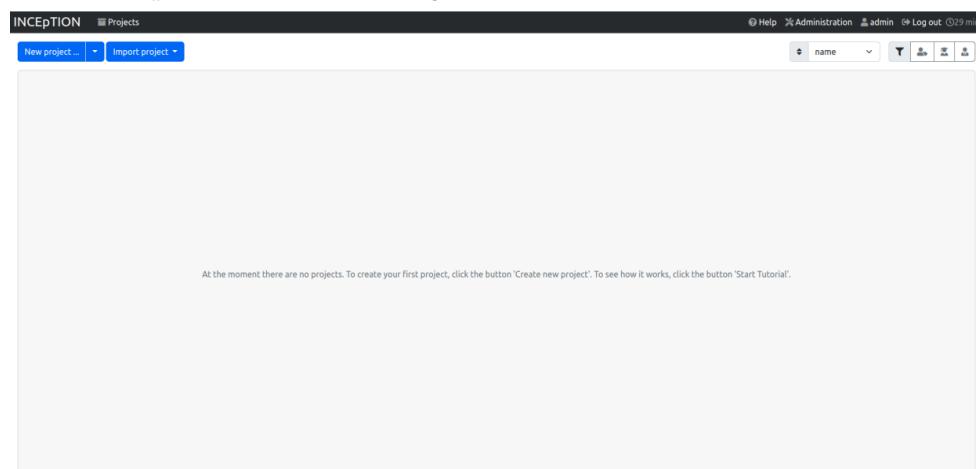
Pelabelan dilakukan melalui tahapan berikut ini:

- a. Unduh alat bantu pelabelan, INCEpTION, di <https://inception-project.github.io/downloads/>

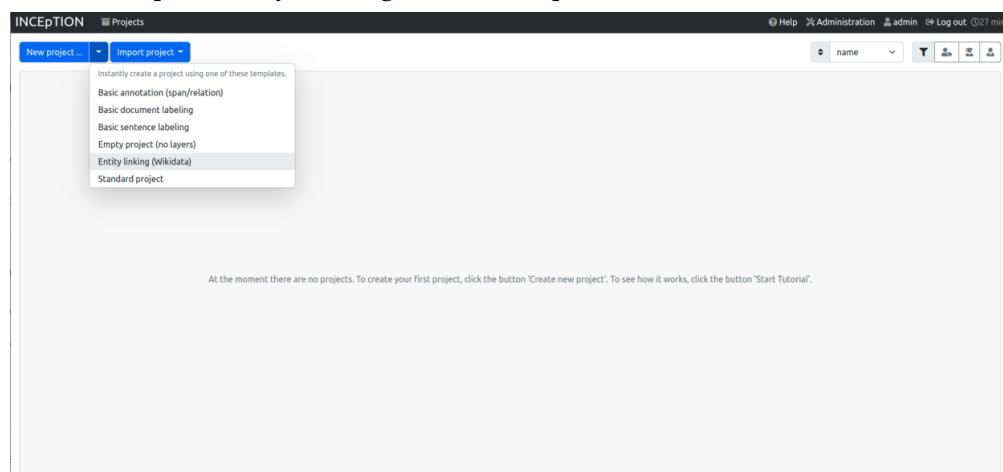
- b. Install dan jalankan aplikasi sesuai panduan yang ada di <https://inception-project.github.io/releases/27.1/docs/user-guide.html#section-installation>
- c. Jalankan aplikasi INCEpTION dengan memasukkan *admin* masing-masing untuk user name dan password.



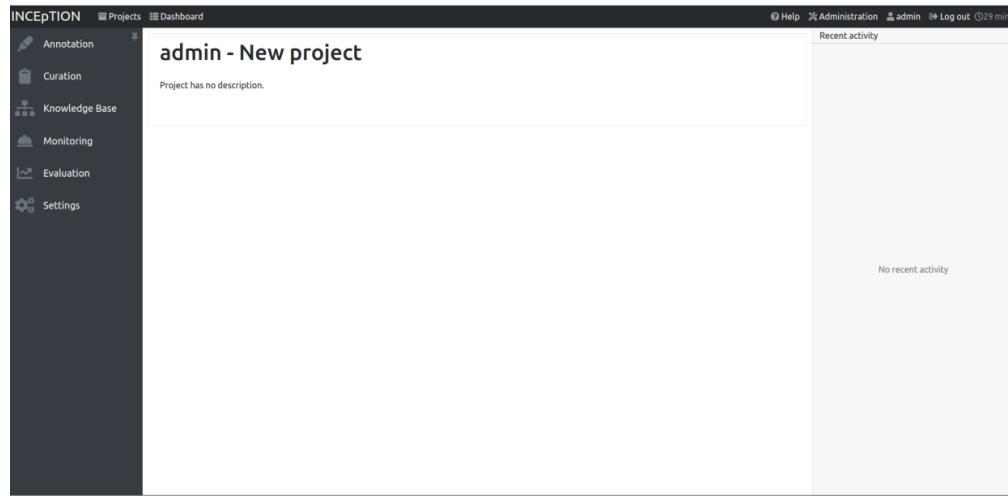
Berikut tampilan setelah berhasil *log in*.



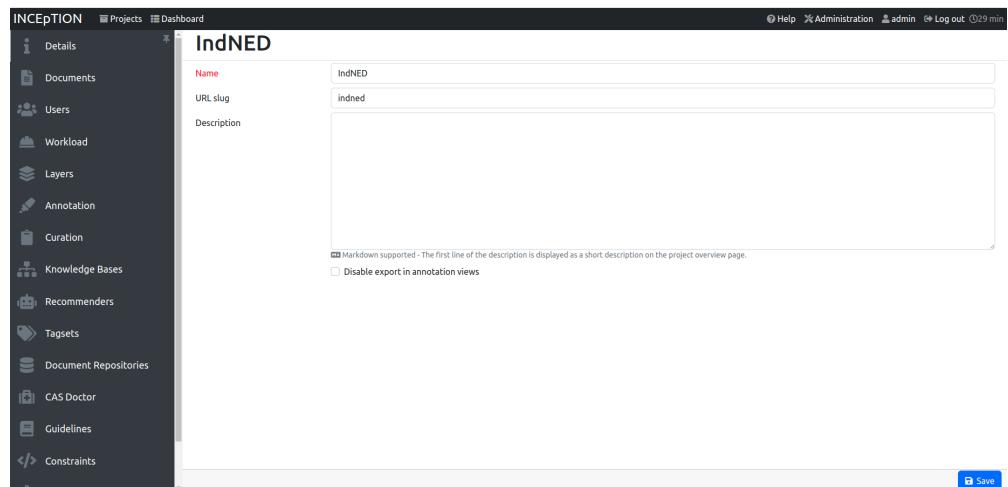
- d. Buat proyek baru dengan mengklik tombol panah di menu *New Project*, lalu pilih *Entity Linking (Wikidata)* pada sub menu.



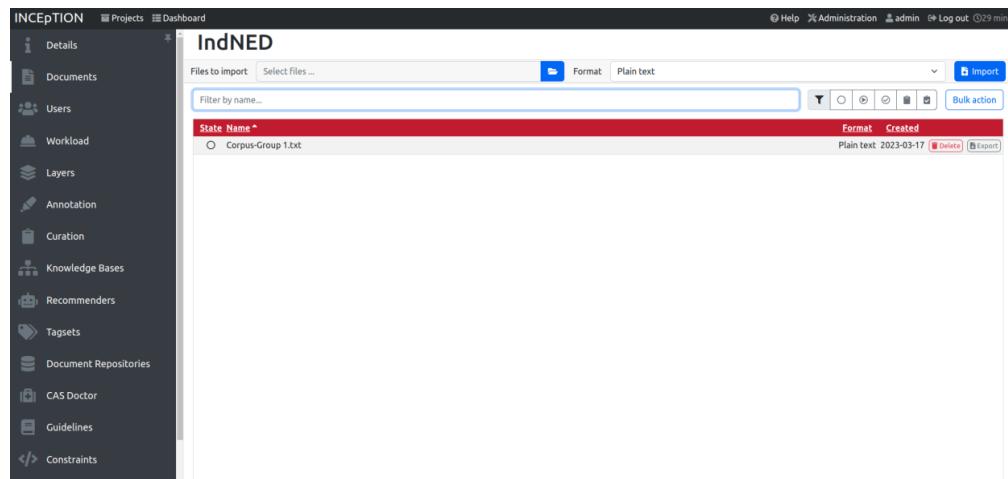
Berikut tampilan setelah pembuatan proyek baru berhasil dilakukan.



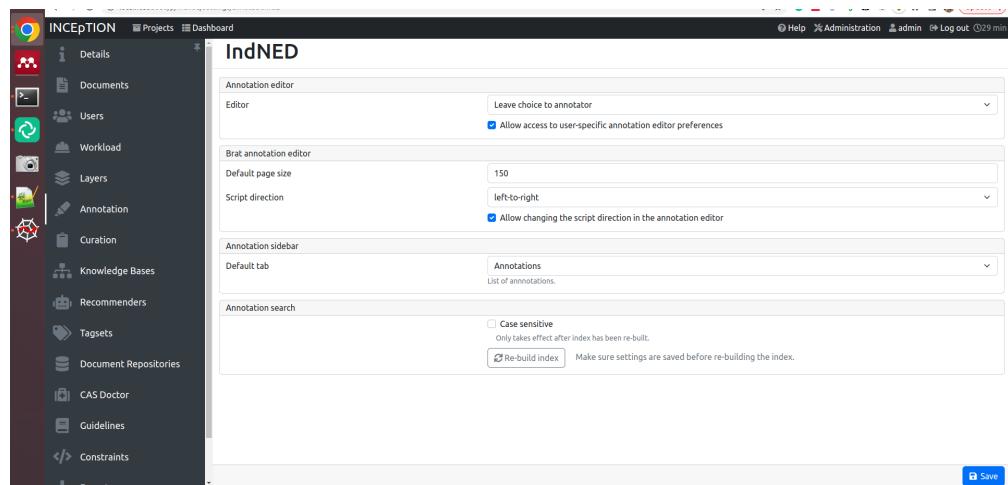
- e. Klik pada menu *Settings* untuk mengatur properti proyek.
 - i. Set nama proyek dengan *IndNED*
 - ii. Kosongkan bagian *URL slug* dan *Description*. Bagian *URL slug* akan otomatis diisi dengan *IndNED*.
 - iii. Klik tombol *Save* di bagian kanan bawah.



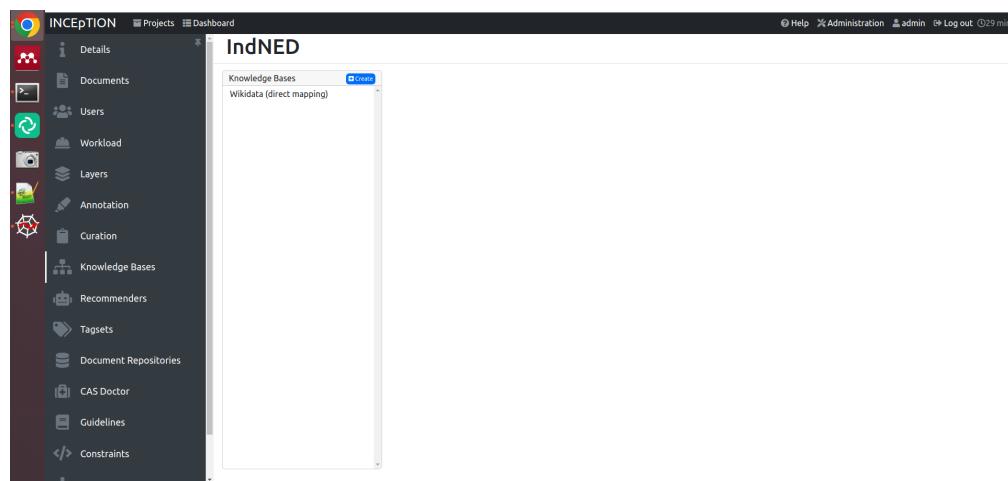
- f. Klik pada menu *Documents*. Lalu, upload file dokumen 1 (lihat seksi 5 bagian e). Berikut tampilan setelah mengupload dokumen 1 untuk kelompok 1 (Corpus-Group 1.txt).



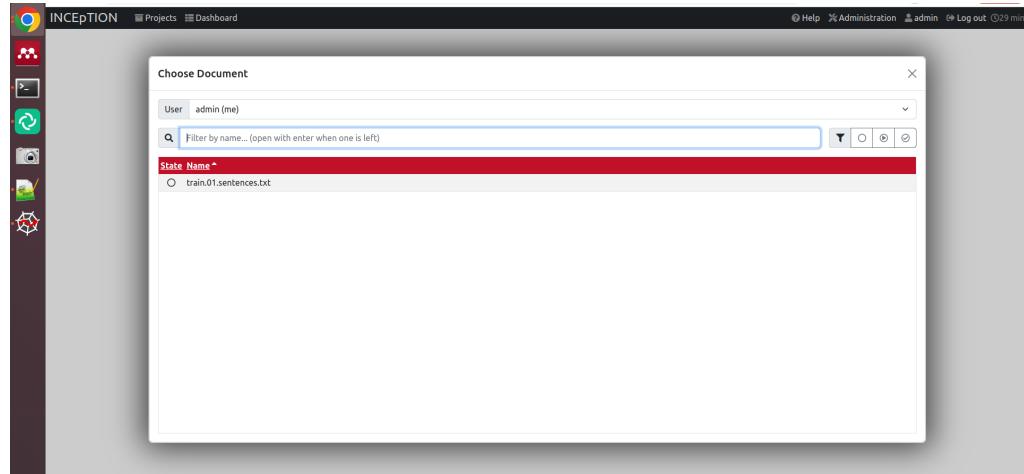
- g. Klik pada menu *Annotation*. Lalu, ubah nilai pada item *Default page size* menjadi 150. Centang checkbox *Allow changing the script direction in the annotation editor*. Lalu, pilih *Annotations* untuk item *Default tab*, hapus centang di checkbox *Case sensitive*, dan klik tombol *Save*.



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

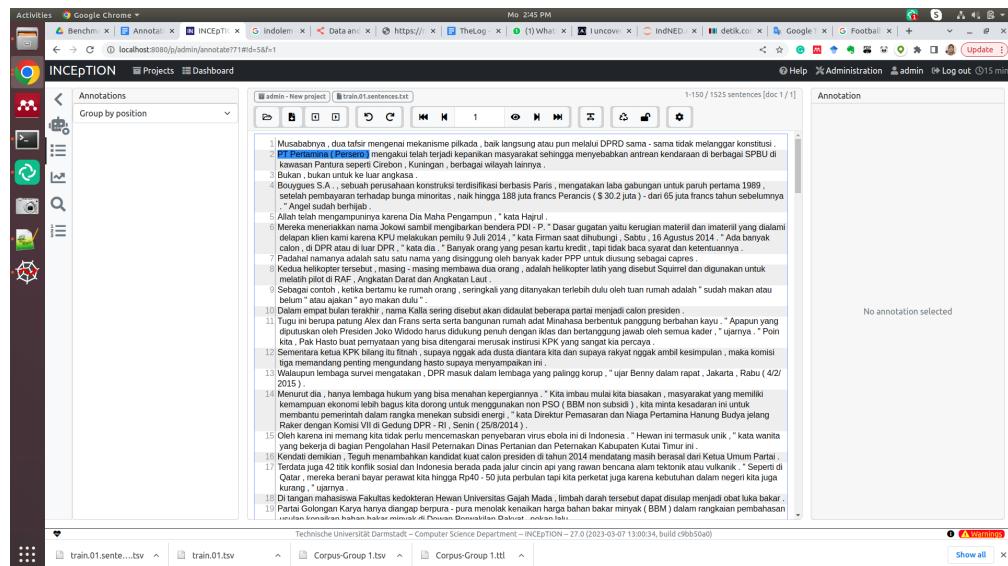


- i. Klik item *Dashboard*, lalu menu *Annotation*.

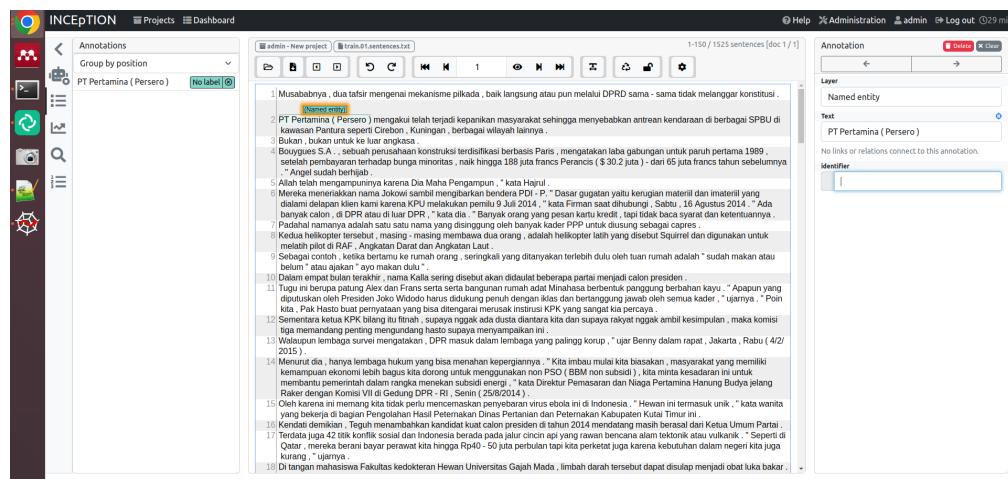


- j. Klik pada nama file dokumen yang telah diupload. Misalnya, train.01.sentences.txt. Tampilan berikutnya adalah editor pelabelan. Terdapat tiga bagian pada editor, yakni (kiri ke kanan), ringkasan hasil pelabelan (kiri), teks yang akan/telah dilabeli (tengah), dan tempat pendefinisian tautan entitas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

- k. Pilih entitas bernama di setiap kalimat berdasarkan informasi di dokumen 2 (seksi 4 poin e) 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. Berikut ini beberapa teknik untuk menemukan kemunculan entitas terkait di Wikidata:

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

The screenshot shows the INCEPTION annotation interface. On the left, there's a sidebar with icons for file operations like 'New project', 'Annotations', 'Group by position', and a search function. The main area displays a list of search results. One result is highlighted with an orange box around the word 'Pertamina'. The results include various news items and documents from different sources, all containing the name 'Pertamina'.

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 detailed view of the first search result for 'Pertamina'. The right panel displays the full text of a news article from 'train.01.sentences.txt'. The left panel shows the entity 'Pertamina' labeled with its URL, which is highlighted in orange. This indicates that the entity has been correctly identified and labeled in the text.

- 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:

The screenshot shows the INCEPTION annotation tool's search interface. On the left, there's a sidebar with project selection and annotation tools. The main area displays search results for 'Jusuf Kalla'. A specific result is highlighted in blue, showing its full text and a detailed annotation panel on the right. The annotation panel includes fields for 'Layer' (Named entity), 'Text' (Kalla), and 'Identifier' (Jusuf Kalla). Below the main search results, there's a list of numbered annotations corresponding to the search results.

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*)

iii. 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:

1. 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*).
2. Temukan nama lengkap atau profil dari entitas bernama di dokumen terpilih (relevan)
3. Lakukan langkah di poin i

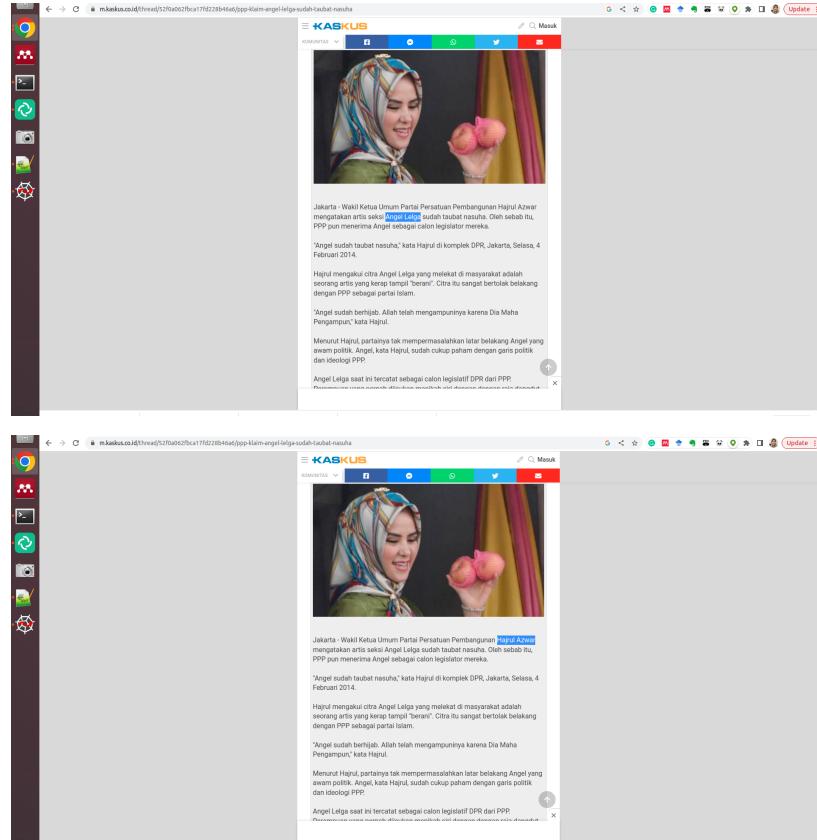
Contoh, *Angel* dan *Hajrul* merupakan entitas bernama yang tidak lengkap dan tidak ada informasi cukup tentang profil keduanya di kalimat berikut: “*Angel* sudah berhijab. Allah telah mengampuninya karena Dia Maha Pengampun , ” kata *Hajrul* .” Langkah yang dilakukan:

1. Pencarian dokumen relevan di Google:

The screenshot shows a Google search results page for the query "Angel sudah berhijab - Allah telah mengampuninya karena Dia Maha Pengampun". The results are as follows:

- 1. tempo.co.id/2014/09/09/angelsudahberhijab.html - Translate this page | ppp Klein Angel Leija Sudah Taubat Nasuha
- 2. www.temponews.co.id/read.aspx?... - Translate this page | Angel Leija Ogah Disebut Melakukan Taubat Nasuha
- 3. www.klikdokter.com/... - Translate this page | Angel Leija Sudah Tobat... | widsenbae
- 4. <https://id.pinterest.com/pin/> - Translate this page | Projek

- Dari *snippet* yang dikembalikan Google, kita dapat mengetahui bahwa dokumen keempat mengandung kalimat yang dimaksud.
- Buka dokumen terpilih. Isi dokumen membantu kita mengetahui bahwa nama Angel mengacu ke Angel Lelga (selebriti), dan nama Hajrul mengacu ke Hajrul Azwar.



- Lakukan langkah *i* sebagai berikut:

The image shows the interface of the INCEPTION annotation tool. On the left, there's a sidebar with project management options like 'Annotations', 'Projects', and 'Dashboard'. A main search bar at the top has the query 'Angel'. Below the search bar, a list of search results is displayed, each with a preview thumbnail, title, and URL. The first result is for 'Angel Lelga' with the URL <http://www.alexa.org/q/14740991>. The second result is for 'Angel Lelga Anggriani' with the URL <http://www.alexa.org/q/10308944>. The third result is for 'Angel' with the URL <http://www.alexa.org/q/13140495>. The fourth result is for 'Angel' with the URL <http://www.alexa.org/q/14740445>. The fifth result is for 'Angel' with the URL <http://www.alexa.org/q/13141445>. The sixth result is for 'Angel' with the URL <http://www.alexa.org/q/13140496>. The seventh result is for 'Angel' with the URL <http://www.alexa.org/q/13140497>. The eighth result is for 'Angel' with the URL <http://www.alexa.org/q/13140498>. The ninth result is for 'Angel' with the URL <http://www.alexa.org/q/13140499>. The tenth result is for 'Angel' with the URL <http://www.alexa.org/q/13140500>.

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. On the left, there's a sidebar with a tree view of annotations under 'PT Pertamina (Persero)'. In the main area, several annotations for the name 'Hajrul' are listed. One annotation for 'Hajrul' has a 'Delete' button next to it. To the right, a detailed view of one annotation is shown, which includes a link to a Wikidata page for Hajrul.

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

This screenshot is similar to the previous one, showing the INCEPTION annotation interface. It displays annotations for the name 'Hajrul'. One annotation for 'Hajrul' has a 'Delete' button next to it. A detailed view of an annotation for 'Hajrul' is shown on the right, linking to its Wikidata page.

- n. 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 seksi 5 poin e.
- o. 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)*. Lalu, kirimkan hasilnya ke ria.hari.gusmita@uni-paderborn.de.

