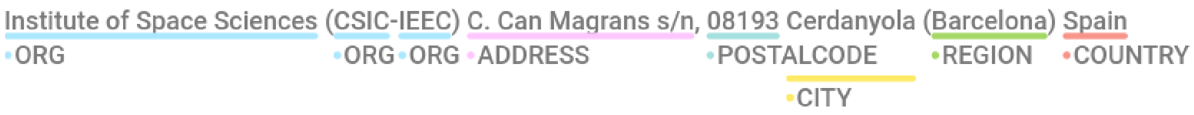
Annotation guidelines for affilgood-NER

Objective

To create an annotated corpus for Named Entities Recognition to identify Named Entities in a raw author affiliation string and to classify these named entities based on the context into a set of classes, 6 classes in the case of affilgood-NER.

An example of the tasks is the following, given a raw affiliation string, to classify sequences of information with the 6 categories:



Institute of Space Sciences (CSIC-IEEC) C. Can Magrans s/n, 08193 Cerdanyola (Barcelona) Spain

This allows a smarter parsing of all pieces of information in a raw affiliation string, for example when there are no punctuation separators, and to extract an organisation name specifically for linking with a database of organisations or the city or the country.

Classes

The list of NER classes with examples are given in the following table.

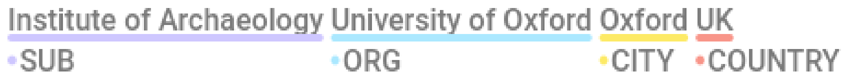
Class name	Class code	Description
Sub-organisation	SUB	This represents subordinate entities within an affiliation string, such as departments, divisions, university schools, sections, faculties, laboratories, research groups, or specific units associated with the main organisation mentioned. For instance, in the affiliation "Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768, USA.", "Biological Sciences Department" would be annotated as a SUB entity. However, this class can be ambiguous or challenging in some cases, because an institution called "Institute of Archaeology" could part of a university in "Institute of Archaeology, Oxford University, Oxford, UK". Or in the case of "Institute of Archaeology, Barcelona, Spain". For this reason, we recommend following as a criterion the internal grammar and role of the entity in an affiliation, trying to follow as much as possible the intention of the author. When the annotator has doubts, can also search in internet if that entity is an independent
Organisation	ORG	This is designated to identify the organisations mentioned in the affiliation string. In an example like "Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA", "California Institute of Technology" would be annotated as an ORG entity.

		Organisation type refers to “high-level” organisations that researchers cite as an affiliation. To meet the definition of a “high-level” entity, an organisation should have a reasonable degree of independence from any parent or related organisations. Common types of entities that are in scope: Universities and colleges, companies, private foundations, government agencies and public administration, hospitals and healthcare centres, laboratories, non-profits, research institutes, and research facilities [REF].
City	CITY	This class pertains to named entities that represent cities within the affiliation string. For example, in the affiliation "New York University, New York", "New York" would be annotated as a CITY entity.
Country	COUNTRY	The class is used for annotating named entities denoting countries in the affiliation string. In the affiliation “University of Oxford, United Kingdom” “United Kingdom” would be labelled as a COUNTRY entity. To add list with exceptions. All countries in Geonames list of countries, as well as other countries with limited recognition, will be considered as countries.
Address	ADDR	This class is assigned to entities that represent specific addresses within the affiliation string. For instance, in "University of Latvia, Jelgavas Street 3, Riga, LV-1004, Latvia" "Jelgavas Street 3" would be annotated as an ADDR entity. Common types of entities that are in scope: street and number, postal box, campus or building name.
Postal code	POSTAL	The POSTAL class is used for annotating postal code entities within the affiliation string. In an example like “University of Latvia, Jelgavas Street 3, Riga, LV-1004, Latvia,” “LV-1004” would be labelled as a POSTAL entity.
Region	REGION	The REGION class is applied to named entities representing regions within the affiliation string. This class includes toponyms between city and country, which could include province, district, county, region, state, or island within a country. In the affiliation “University of Southern California, Los Angeles, California 90089, USA.” “California” would be annotated as a REGION entity.

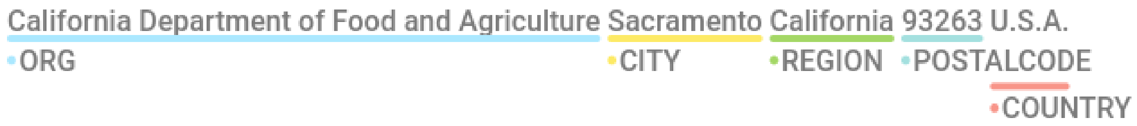
Annotation examples and cases

We have split cases between classical examples and exceptions. Hereafter, we introduce some practical examples with the explanation of the rationale and of how we have interpreted those cases, which can be used as a practical extension of the guidelines.

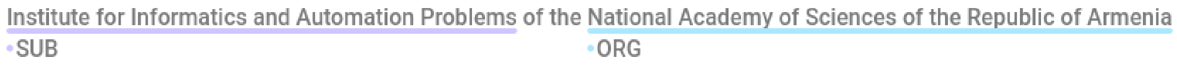
In the following case the institute is considered part of the university structure, which follows a classical affiliation structure: *subsection, university, city, country*.


Institute of Archaeology University of Oxford Oxford UK

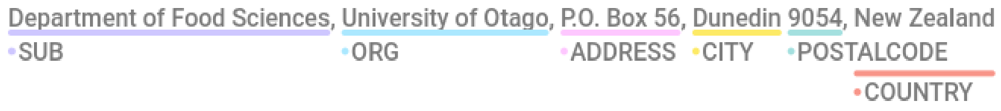
In the following case, region and postal code also appear:


California Department of Food and Agriculture Sacramento California 93263 U.S.A.

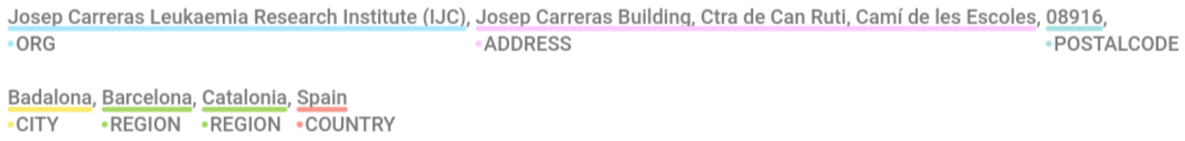
In the following case, the author uses prepositional connectors instead of punctuation to separate pieces of information. However, there could also be some ambiguity between considering “Republic of Armenia” as a country or as part of the organisation name. In these cases, we propose to take the version used on Wikipedia, ROR, or the web of the organisation. As in this case the complete name is “National Academy of Sciences of Armenia”, we consider the country name as part of the organisation name.


Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia

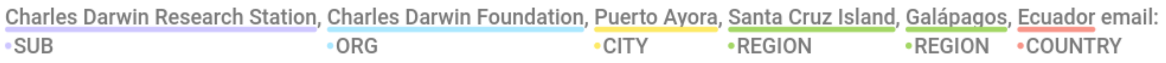
In the following case, the affiliation string follows the classic structure, adding postal code and postal box, and as we describe in the annotation guidelines, we will treat postal boxes as addresses.


Department of Food Sciences, University of Otago, P.O. Box 56, Dunedin 9054, New Zealand

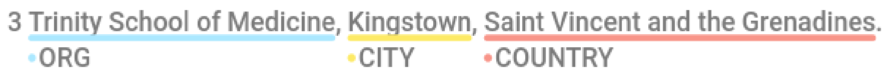
In the following case, the organisation name also includes an acronym of the same name, and this is to be considered part of the same entity. Street, road, and building, are considered as addresses of the organisation. In this case, two regional entities are used, one is “Barcelona” which plays a role of “province”, and the other is “Catalonia” which is the region. As expressed in the guidelines, REGION means all toponyms between CITY and COUNTRY.


Josep Carreras Leukaemia Research Institute (IJC), Josep Carreras Building, Ctra de Can Ruti, Camí de les Escoles, 08916, Badalona, Barcelona, Catalonia, Spain

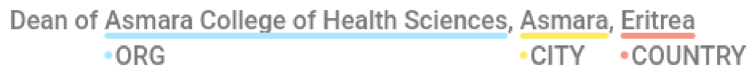
In the following case, another meaning of REGION is expressed to refer to an island and to the archipelago. Furthermore, in this example there is the text “email:” which does not correspond to any named entity.

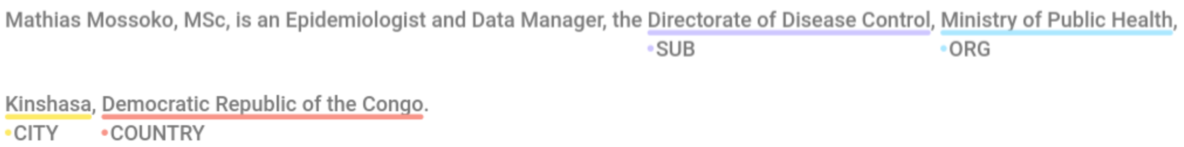

Charles Darwin Research Station, Charles Darwin Foundation, Puerto Ayora, Santa Cruz Island, Galápagos, Ecuador email:

In the following case, a super index of affiliation number of the author has been processed as part of the affiliation; this is not part of any entity.

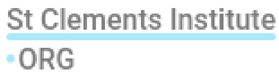

3 Trinity School of Medicine, Kingstown, Saint Vincent and the Grenadines.

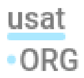
In the following two cases, there is some text that does not correspond to any of the entities about author name, role, or position of the author in the organisation.

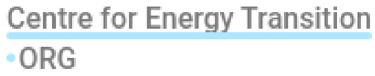

Dean of Asmara College of Health Sciences, Asmara, Eritrea


Mathias Mossoko, MSc, is an Epidemiologist and Data Manager, the Directorate of Disease Control, Ministry of Public Health, Kinshasa, Democratic Republic of the Congo.

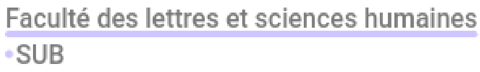
In the following three cases, there only appears limited information about the name of the institution, which in this context plays the semantic role of main institution. The three could be unique and organisations with a certain degree of independence. The second one seems to be an acronym, although it is difficult to know without more context, but this is indicative of an organisation. The third one refers to an organisation; in this case, however, the same phrase with different context like "Center for Energy Transition, University of Siena, Siena, Italy" could be labelled as sub-organisation.

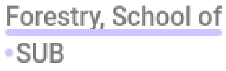

St Clements Institute


usat


Centre for Energy Transition

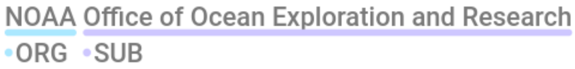
In the following two examples, the authors have only expressed information about sub-organisation, or part of the affiliation string disappeared during data extraction by the data provider. These two cases have to be labelled as SUB-ORGANISATION because they are not "high-level" organisations.


Faculté des lettres et sciences humaines



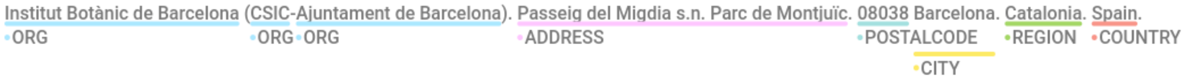
Forestry, School of

In the following case, order of organisation and sub-organisation are expressed in different order.

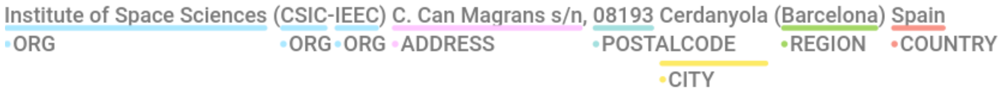


NOAA Office of Ocean Exploration and Research

In the following case, “Institut Botànic de Barcelona” is expressed as a primary organisation, because it seems to have a certain degree of independence, although it belongs to two organisations that are also indicated between brackets. In this case, for instance, pieces of entities are divided by points and not with commas, and these are good examples to show that authors use different types of separators, and sometimes only white spaces.

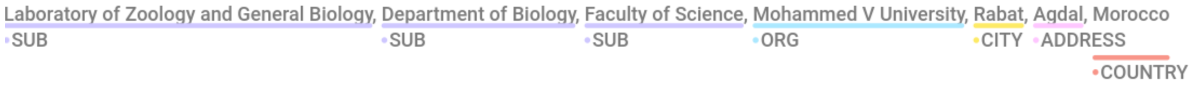


Institut Botànic de Barcelona (CSIC-Ajuntament de Barcelona). Passeig del Migdia s.n. Parc de Montjuïc. 08038 Barcelona. Catalonia. Spain.



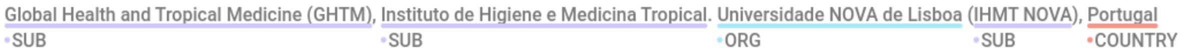
Institute of Space Sciences (CSIC-IEEC) C. Can Magrans s/n, 08193 Cerdanyola (Barcelona) Spain

In the following case, three sub-organisation entities are mentioned, something frequent when authors refer to the specific research group, indicating it within the hierarchy of the university. In this case, “Rabat” is the CITY, and “Agdal” is a neighbourhood in Rabat, which should be labelled as ADDRESS.



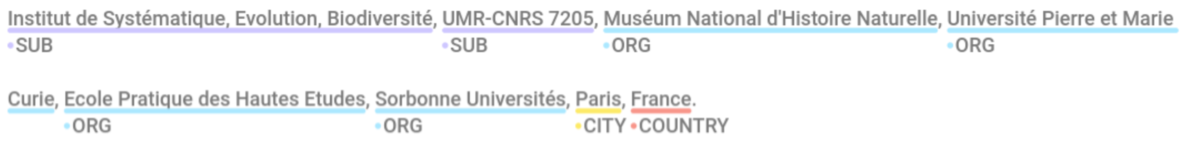
Laboratory of Zoology and General Biology, Department of Biology, Faculty of Science, Mohammed V University, Rabat, Agdal, Morocco

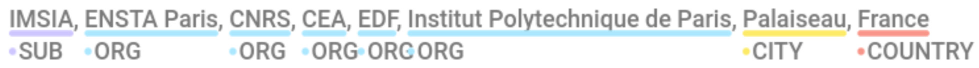
In the following case, two subunits of the university as mentioned, and between brackets appear the abbreviations of the second sub-organisation, and for this reason they are labelled as SUB-ORGANISATION.



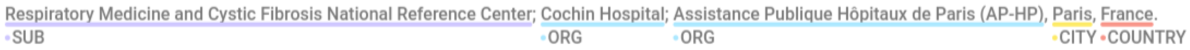
Global Health and Tropical Medicine (GHTM), Instituto de Higiene e Medicina Tropical. Universidade NOVA de Lisboa (IHMT NOVA), Portugal

In the following two cases, there are two typical French affiliations of mixed research units, which indicate the research unit with all the parent organisations of the unit. In these cases, the name of the unit and the code are considered as SUB-ORGANISATION. However, in the second example, most of the names appear as acronyms and abbreviations, and it can be difficult for the annotator to differentiate between sub-organisation and organisation; therefore, in very ambiguous cases, we recommend annotating them with organisation.

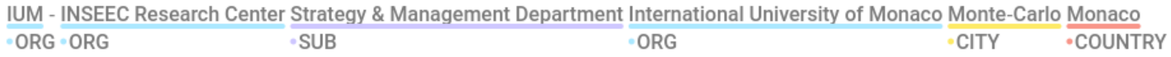

Institut de Systématique, Evolution, Biodiversité, UMR-CNRS 7205, Muséum National d'Histoire Naturelle, Université Pierre et Marie Curie, Ecole Pratique des Hautes Etudes, Sorbonne Universités, Paris, France.


IMSIA, ENSTA Paris, CNRS, CEA, EDF, Institut Polytechnique de Paris, Palaiseau, France

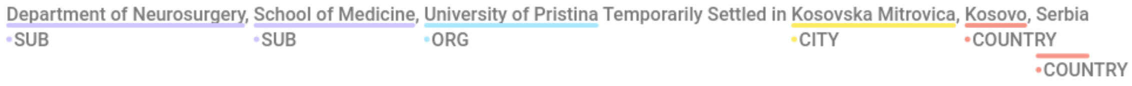
In the following case, a hospital is mentioned together with also its parent administrative organisation, however, hospitals are "high-level" institutions.

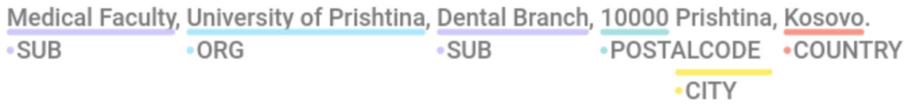

Respiratory Medicine and Cystic Fibrosis National Reference Center; Cochin Hospital; Assistance Publique Hôpitaux de Paris (AP-HP), Paris, France.

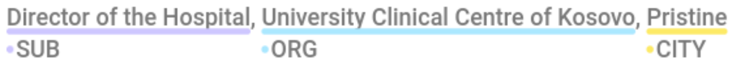
In the following case, because there are no separators, the split can be ambiguous when the annotator doesn't know the name of organisations mentioned, because starting and ending of the entities can be interpreted in different ways. In these cases, we recommend exploring on the internet for the complete names of the organisations.


IUM - INSEEC Research Center Strategy & Management Department International University of Monaco Monte-Carlo Monaco

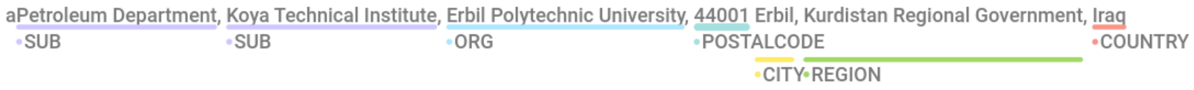
In the following cases, we see one of the exceptions that we indicate in the definition of the category country. We take into consideration as country those countries that appear in Geonames list of countries, as well as other countries with limited recognition. In the first example, "Kosovo" appears as well as "Serbia", both are names of countries, and they have to be annotated as such.

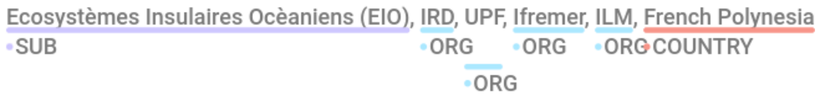

Department of Neurosurgery, School of Medicine, University of Pristina Temporarily Settled in Kosovska Mitrovica, Kosovo, Serbia

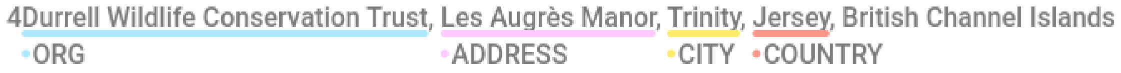

Medical Faculty, University of Prishtina, Dental Branch, 10000 Prishtina, Kosovo.

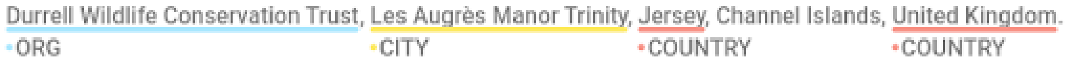

Director of the Hospital, University Clinical Centre of Kosovo, Pristine

In the following cases, we can find some slightly more complex examples, or ones which could be interpreted in different ways. The first one, "Kurdistan Regional Government" is annotated as Region, because it doesn't appear in the list of countries, however, it is an autonomous region in Iraq. In the second, "French Polynesia" is considered as a COUNTRY because it is in the list of Geonames, although this is still a French territory. In the last two examples, the country is Jersey, and also the archipelago name is written. However, we don't tag it, because it refers to a geographical entity and not a political entity. In the last example, the "United Kingdom" is also mentioned, and it is annotated as COUNTRY, without interpreting the relation between Jersey as British Crown Dependency island, because both are names of countries.

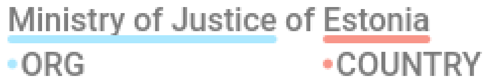

aPetroleum Department, Koya Technical Institute, Erbil Polytechnic University, 44001 Erbil, Kurdistan Regional Government, Iraq

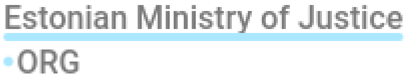

Ecosystèmes Insulaires Ocèaniens (EIO), IRD, UPF, Ifremer, ILM, French Polynesia


4Durrell Wildlife Conservation Trust, Les Augrès Manor, Trinity, Jersey, British Channel Islands


Durrell Wildlife Conservation Trust, Les Augrès Manor Trinity, Jersey, Channel Islands, United Kingdom.

In the following case, there is a prepositional connector between the ORGANISATION and the COUNTRY. In this case, we have interpreted the ORG only as “Ministry of Justice” as appears on the web. However, the same affiliation written as “Estonian Ministry of Justice” should be annotated as ORGANISATION, because of the location adjective.


Ministry of Justice of Estonia


Estonian Ministry of Justice

Format and annotation procedure

We will use Doccano[1], which is an open-source text annotation tool for human annotations.

The objective of annotation is to annotate the entities in the affiliations, and not to interpret if they make sense or not. As, for instance, in an affiliation like “MIT, Madrid, France”. The relation of the following entities doesn’t make sense. But what we want is to identify entities appearing in an affiliation, without overinterpreting too much.

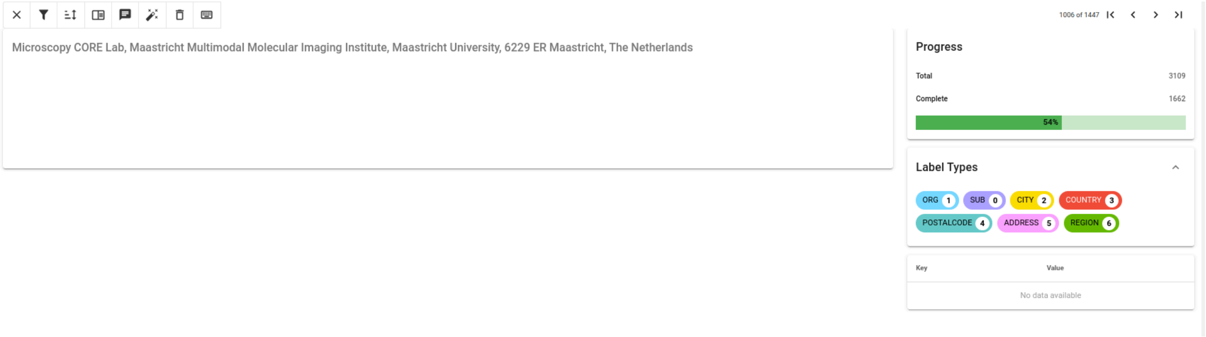
During the annotation we recommend following the annotation guidelines, the definition of categories and the examples reported. Please, try to follow the author's intention, considering the internal grammar and hierarchy of the affiliation.

[1] https://github.com/doccano/doccano

It can be useful for the annotator to search in the internet, Wikipedia, or ROR, to confirm if a string is a city, country, among others, for those cases the annotator does not have knowledge on the country.

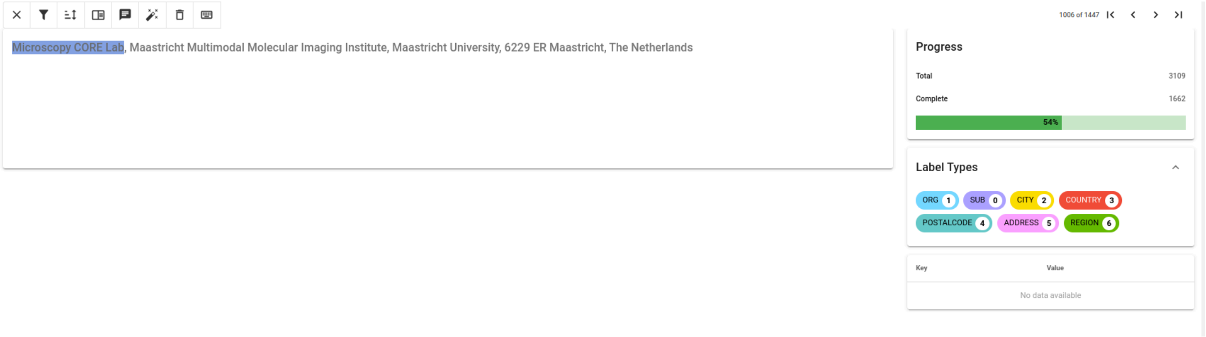
The annotation process must be done by following different steps. The next lines describe the step-by-step process that the annotator must accomplish for each raw affiliation string:

1. Read the affiliation string and analyse hierarchy and Named Entities.



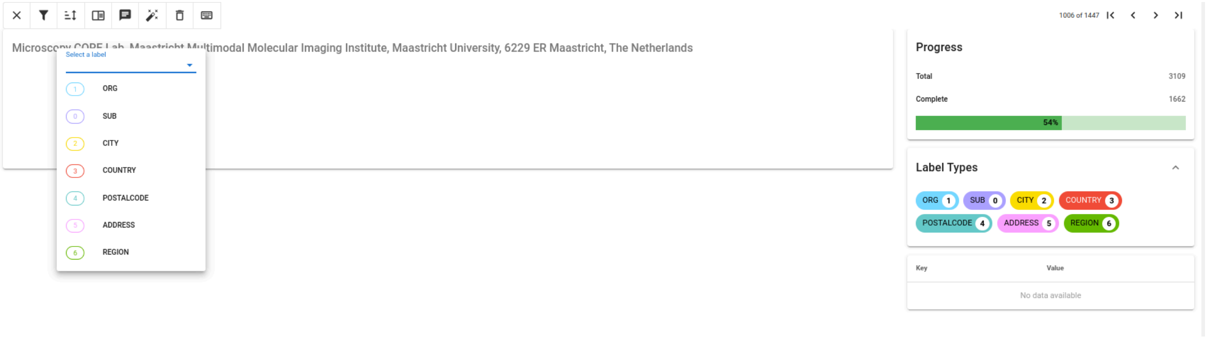
Microscopy CORE Lab, Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

2. For annotating a sequence, you should select the set of strings.



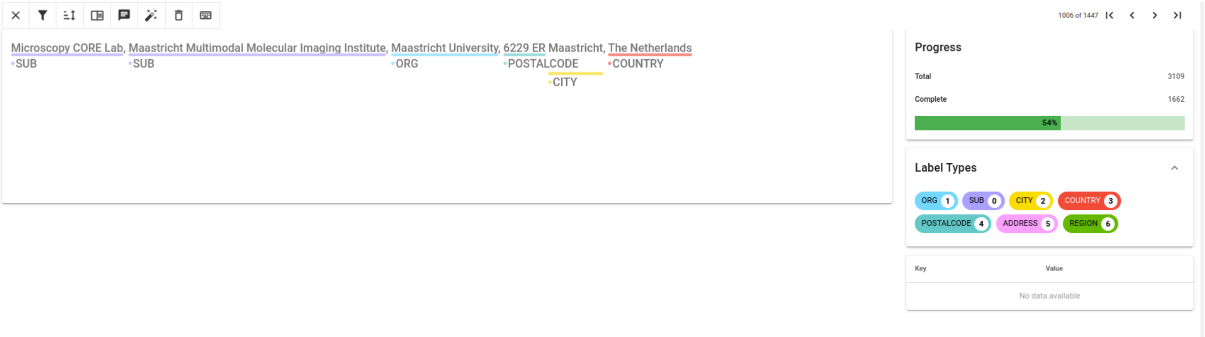
Microscopy CORE Lab, Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

3. The “Select label” drop-down menu will appear.



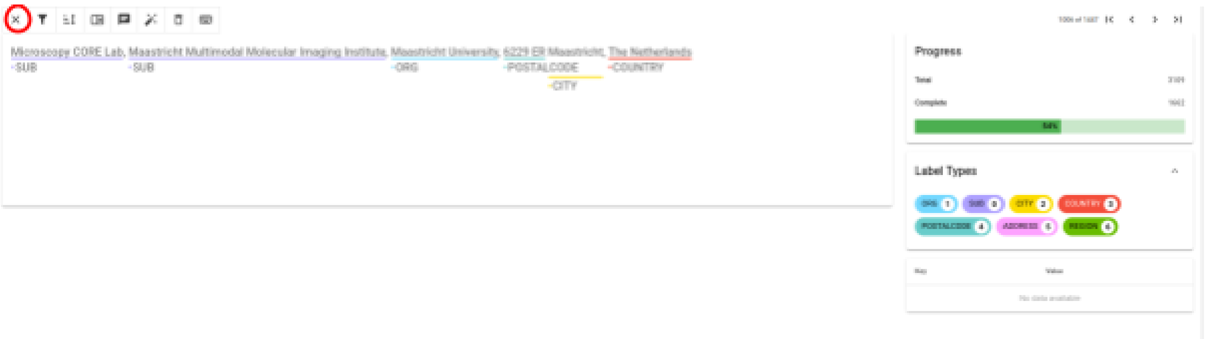
Microscopy CORE Lab, Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

4. Once you have annotated all the Named Entities.



Microscopy CORE Lab, Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

5. Select the “To Check” button to validate the annotation.



Microscopy CORE Lab, Maastricht Multimodal Molecular Imaging Institute, Maastricht University, 6229 ER Maastricht, The Netherlands

6. Automatically, Doccano will pass to the next affiliation to annotate. But if you find an affiliation that you don’t know how to annotate, because it is written in another alphabet, or you have doubts about the entities, it’s normal, but don’t select “To Check”. The affiliation will remain “in progress” and will not be considered in the training data. Some affiliations can be strange and weird, or written in an alphabet that is not possible to annotate.