Resolving toponyms in PubMed Central (PMC) Articles

This schema outlines the instructions followed to create a gold standard corpus for evaluating automatic toponym resolution in PMC articles. In this guide, we call a toponym mention any occurrence of a name of a location within a given PMC article. First, the annotators should find all toponym mentions in an article. This step is known as the toponym detection. Then, the annotators must associate unique coordinates to the toponyms that are identified, i.e., the latitude and longitude of their centroid. This step is referred as the toponym disambiguation. Both toponym detection and toponym disambiguation constitute the toponym resolution task.

A. Toponyms Detection

For the purpose of our task, the annotators should tag toponym mentions that are within the main text of the article. These toponym mentions can be located within titles, sub-headers, paragraphs, tables, figure captions, as well as other non-image text components of the papers. Locations that are in header and footer of pages, author affiliation, acknowledgement, and reference sections should not be tagged.

Toponym mentions consist of named geographic locations: continents, countries, states, provinces, regions, territories, counties, named lakes, named mountain ranges, named deserts, named bodies of water, etc. should be tagged.

Example 1: “…strain CO92-1356 was isolated in <Location>Colorado</Location> from Culex tarsalis mosquitoes in 1992” [Pubmed ID: 19446595].

This definition includes any two letter state abbreviations and common country abbreviations (such as USA or UK). Note that this definition excludes all indirect mentions of places such as my hometown, 30 km north of Boston, the Hong Kong strain, etc. Locations should be tagged as separate entities.

Example 2: “Genotype III strains of SLEV were isolated from Culex quinquefasciatus mosquitoes in <Location>Cordoba </Location>, <Location>Argentina </Location> in 2005, during the largest SLEV outbreak ever reported in <Location>South America </Location>.” [Pubmed ID: 21629729].

Example 3: “Viral RNA was reverse transcribed into cDNA with SuperScript III First - Strand Synthesis System (Invitrogen, Life Technologies, <Location>Carlsbad </Location>, <Location>CA </Location>, <Location>USA </Location>) with random hexamers.” [PubMed ID: 23920350]

B. Toponym Disambiguation

Once a toponym mention is tagged, a note indicating its unique latitude and longitude should be added in the Brat interface (see Figure 4). Latitudes and longitudes should contain 5 decimal places when available. Additional information, such as population size, can be included in the notes when available. The latitude and longitude should be obtained from GeoNames. If a toponym cannot be found in GeoNames, then additional searches can be carried out using Google Maps through their web interface. If the coordinates of the toponym are still unknown, a last search can be carried out through Wikipedia. If the coordinates remain unavailable for the toponym mention, “N/A” should be put in the notes section.
The annotators should use the context of the paragraph to disambiguate a toponym that can have more than one latitude and longitude. However, if another mention of this toponym previously occurred in the text, the annotators can disambiguate the new mention with the same coordinates, provided that no opposite evidence for this location is found in the paragraph. If a toponym mention can refer to a capital or a semi-independent political entity, such as the case with Hong Kong, the coordinates of the less specific location should be chosen (semi-independent political entity). This rule applies to other instances in which the location mention can be referring to more than one entity, with each entity having different specificities.

Example 4: “As shown in Table S1 and Figure S2, among these 833 viruses, 767 from dozens of provinces in China belonged to lineage h9.4.2, 48 from <Location> Hong Kong <LatLong> 22.25, 114.16667 </LatLong> </Location> and a city in <Location> Guangdong </Location> Province…” [PubMed ID: 23285143]

C. Resolution of Problematic Cases

During the preliminary pass of annotation, annotators have identified several problematic cases. In this section, we discuss the solutions they adopted and provide circumstantial examples along with their PubMed IDs for reference.

- Phrases that refer to multiple locations should not be annotated.
  - Example 5: “The place and time of origin of the 1918 influenza pandemic virus is unknown, with no evidence of excess respiratory disease or of excess mortality detected in United States military camps from May through September 1918.” [PubMed ID: 21930918]

- The official name should always be entirely annotated.
  - Example 6: “Support was provided in part by the <Location> Commonwealth of Kentucky </Location> as a Clinical and Translational Science Pilot Project at the University of Louisville to CBJ.” [PubMed ID: 23441208]
• The word *County* after a county name or *Province* after a province name should not be tagged.
  ◦ Example 7: “In this study, an H5N2 virus, A/duck/Jiangsu/m234/2012 (m234), was isolated from apparently healthy domestic mallard ducks in the *<Location> Jiangsu Province* of eastern *<Location> China* in January 2012.” [PubMed ID: 23087121]

• Descriptive words that are attached to named locations, such as *northern* or *mid-*-, should not be tagged. The only directional terms that should be tagged are those that are part of the official name of a location, e.g., *North America, South Carolina*.
  ◦ Example 8: “15 representative H9N2 viruses isolated from diseased chickens in *northern <Location> China* between 1998 and 2010 were characterized…” [PubMed ID: 22050764]

• Latitude and longitude values should be tagged.
  ◦ Example 9: “These birds had been sampled at the same lake in the *<Location> Hadejia-Nguru Wetlands* in northern *<Location> Nigeria* (<Location> Jigawa State <Location>); *<Location> latitude: 12°48’N; longitude: 10°44’E </Location>; Figure 1), respectively on the 14th (WFWD) and 17th (SWG) February 2007.” [PubMedID: 18704172]

• Street names should be tagged, with the word “Street” included, but street numbers and postal codes should not.

• General terms such as *the river, swamplands, in mountains*, that cannot be used to identify specific locations should not be tagged.

• Adjective locations should not be tagged. They frequently do not refer to an actual location.
  ◦ Example 10: “The first and most notable, known as the *Spanish* influenza occurred in 1918.1919 N1), remains unprecedented for its high mortality and attack rate [1]. [PubMedID: 24086762]

  In this example, *Spanish* is not tagged because it is used as an adjective that refers to a pandemic that occurred in 1918, not a virus in Spain.
  ◦ Example 12: “In particular, the clade 2.3.4 *Fujian*-like H5N1 viruses have gradually became endemic in *<Location> China* since 2005 (3, 5) and continue to evolve actively […]” [PubMed ID: 23087121]

  In this example, *Fujian* is not tagged because it is in an adjective position and it refers to a type of virus strain, not to the actual province in China.

• Location mentions that are embedded in strain names should not be tagged.
  ◦ Example 13: “Antigens for the assays were produced from the A/Nanjing/1/2013(H7N9) strain isolated from the confirmed case in *<Location> Jiangsu*.” [PubMed ID: 23920350]

• Names of organizations, such as companies or universities, as well as locations within organization names, should not be tagged.
  ◦ Example 14: “All testing was performed at the BSL-2 or BSL-3 laboratories of *Jiangsu Province Center for Disease Control and Prevention, <Location>Nanjing</Location>, <Location>China</Location>.” [PubMed ID: 23920350]