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. Additionally, annotators must tag certain sections within the article to facilitate their detection and removal by the system

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.

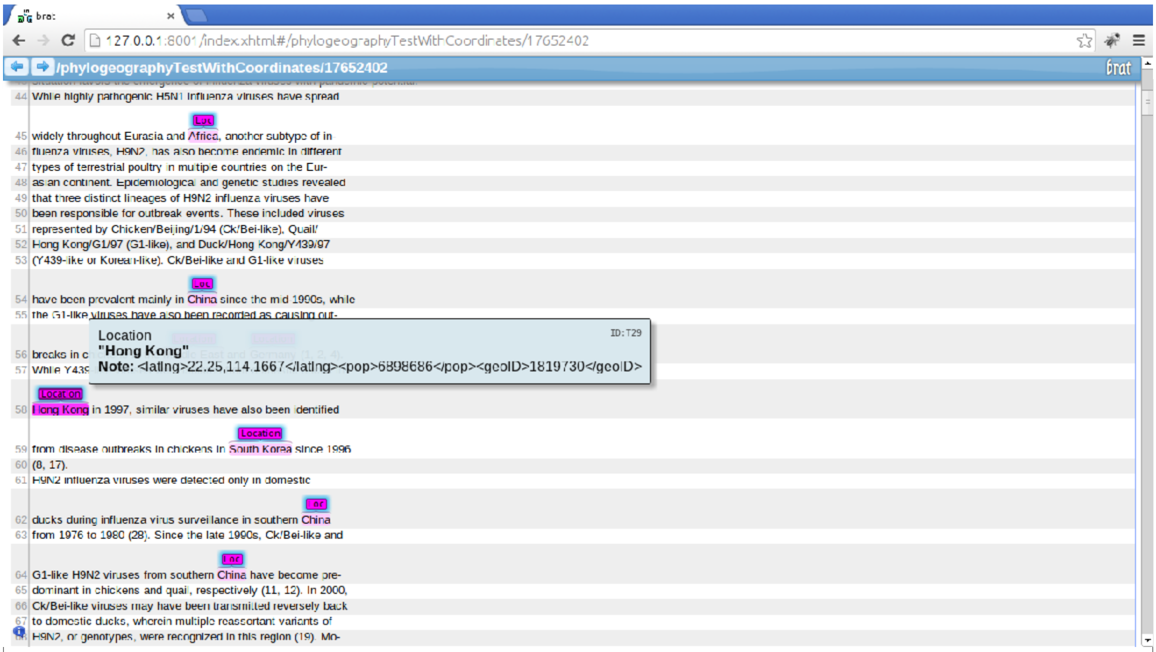
> Example 1: "…strain CO92-1356 was isolated in **<Location>Colorado</Location>** from Culex tarsalis mosquitoes in 1992" [Pubmed ID: 19446595].

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.



bra:

127.0.0.1:8001/index.xhtml#/phylogeographyTestWithCoordinates/17652402

/phylogeographyTestWithCoordinates/17652402

While highly pathogenic H5N1 influenza viruses have spread
widely throughout Eurasia and Africa, another subtype of in-
fluenza viruses, H9N2, has also become endemic in different
types of terrestrial poultry in multiple countries on the Eur-
asian continent. Epidemiological and genetic studies revealed
that three distinct lineages of H9N2 influenza viruses have
been responsible for outbreak events. These included viruses
represented by Chicken/Beijing/1/94 (Ck/Bei-like), Quail/
Hong Kong/G1/97 (G1-like), and Duck/Hong Kong/Y439/97
(Y439-like or Korean-like). Ck/Bei-like and G1-like viruses
have been prevalent mainly in China since the mid-1990s, while
the G1-like viruses have also been recorded as causing out-
breaks in c
While Y439
Hong Kong in 1997, similar viruses have also been identified
from disease outbreaks in chickens in South Korea since 1996
(8, 17).
H9N2 influenza viruses were detected only in domestic
ducks during influenza virus surveillance in southern China
from 1976 to 1980 (28). Since the late 1990s, Ck/Bei-like and
G1-like H9N2 viruses from southern China have become pre-
dominant in chickens and quail, respectively (11, 12). In 2000,
Ck/Bei-like viruses may have been transmitted reversely back
to domestic ducks, wherein multiple reassortant variants of
H9N2, or genotypes, were recognized in this region (19). Mo-

Location
"Hong Kong"
Note: <latlng>22.25,114.1667</latlng><pop>6898686</pop><geoID>1819730</geoID>

Figure 1: Annotation tool Brat used for tagging toponyms.

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 </Location>** **Province** of eastern <Location> China </Location> 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 </Location>** 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 </Location> in northern <Location> Nigeria </Location> (<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]

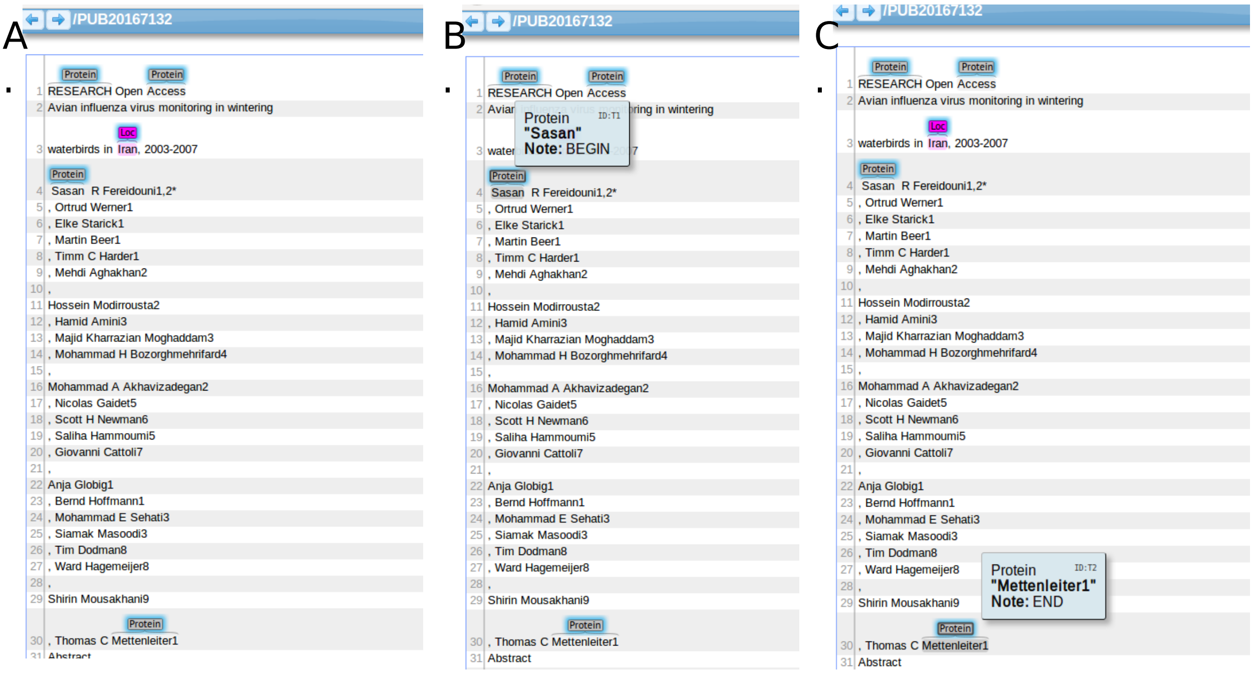
    In this example, *Spanish* is not tagged because it is used as an adjective that refers to a pandemic that occurred in 1918, not a virus in Spain.
  - Example 11: "The 2009 pandemic H1N1 influenza viruses arose through reassortment of two preexisting swine influenza viruses, a **Eurasian** avian-like virus and a **North American** triple reassortant virus." [PubMed ID: 23441208]
  - Example 12: "In particular, the clade 2.3.4 **Fujian**-like H5N1 viruses have gradually became endemic in <Location> China </Location> 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 </Location>." [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]
- Due to limitations with the BRAT annotation tool, location mentions that are split by sections removed (see Section D) should not be annotated as they cannot be resolved.
  - Example 15: "….**North** <Protein>Journal</Protein><Protein>.</Protein> **America**…"

D. Special Tagging for Section Removal within text

There are certain sections within articles that need to be removed by the system before performing automated toponym detection and disambiguation. To facilitate this removal, annotators will tag the beginning and end of these sections. The sections include any heading before the article title, author names and their affiliations, headers and footers, and references. Unfortunately, the PDF to text conversion process is not perfect and these sections can be interspersed throughout the document.

For this task, annotator will select the “protein’ tag in BRAT to annotate the beginning and end of the section. Annotators should tag the first word of the section and add the word “BEGIN” in the notes section. The last word of the section should then be tagged with the word “END’ being placed in the notes (Figure 2).



A.

RESEARCH Open Access
Avian influenza virus monitoring in wintering
waterbirds in Iran, 2003-2007
Sasan R Fereidouni1,2*
, Ortrud Werner1
, Elke Starick1
, Martin Beer1
, Timm C Harder1
, Mehdi Aghakhan2
,
Hossein Modirrousta2
, Hamid Amini3
, Majid Kharrazian Moghaddam3
, Mohammad H Bozorghmehrifard4
,
Mohammad A Akhavizadegan2
, Nicolas Gaidet5
, Scott H Newman6
, Saliha Hammoumi5
, Giovanni Cattoli7
,
Anja Globig1
, Bernd Hoffmann1
, Mohammad E Sehati3
, Siamak Masoodi3
, Tim Dodman8
, Ward Hagemeijer8
,
Shirin Mousakhani9
, Thomas C Mettenleiter1
Abstract

B.

Protein ID:T1
"Sasan"
Note: BEGIN

RESEARCH Open Access
Avian influenza virus monitoring in wintering
waterbirds in Iran, 2003-2007
Sasan R Fereidouni1,2*
, Ortrud Werner1
, Elke Starick1
, Martin Beer1
, Timm C Harder1
, Mehdi Aghakhan2
,
Hossein Modirrousta2
, Hamid Amini3
, Majid Kharrazian Moghaddam3
, Mohammad H Bozorghmehrifard4
,
Mohammad A Akhavizadegan2
, Nicolas Gaidet5
, Scott H Newman6
, Saliha Hammoumi5
, Giovanni Cattoli7
,
Anja Globig1
, Bernd Hoffmann1
, Mohammad E Sehati3
, Siamak Masoodi3
, Tim Dodman8
, Ward Hagemeijer8
,
Shirin Mousakhani9
, Thomas C Mettenleiter1
Abstract

C.

Protein ID:T2
"Mettenleiter1"
Note: END

RESEARCH Open Access
Avian influenza virus monitoring in wintering
waterbirds in Iran, 2003-2007
Sasan R Fereidouni1,2*
, Ortrud Werner1
, Elke Starick1
, Martin Beer1
, Timm C Harder1
, Mehdi Aghakhan2
,
Hossein Modirrousta2
, Hamid Amini3
, Majid Kharrazian Moghaddam3
, Mohammad H Bozorghmehrifard4
,
Mohammad A Akhavizadegan2
, Nicolas Gaidet5
, Scott H Newman6
, Saliha Hammoumi5
, Giovanni Cattoli7
,
Anja Globig1
, Bernd Hoffmann1
, Mohammad E Sehati3
, Siamak Masoodi3
, Tim Dodman8
, Ward Hagemeijer8
,
Shirin Mousakhani9
, Thomas C Mettenleiter1
Abstract

Figure 2: A) The first and last word of the unwanted section is annotated with the protein tag. B) “BEGIN” is placed in the Note of the first word of the section. C) “END” is placed in the Note on the last word.