# Natural Language Processing Methods for Enhancing Geographic Metadata for Phylogeography of Zoonotic Viruses

**Tasnia Tahsin**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
ttahsin@asu.edu

**Rachel Beard**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
rachel.beard@asu.edu

**Robert Rivera**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
rdriver1@asu.edu

**Rob Lauder**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
rlauder@asu.edu

**Davy Weissenbacher**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
dweissen@asu.edu

**Garrick Wallstrom**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
gwallstrom@asu.edu

**Matthew Scotch**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
mscotch@asu.edu

**Graciela Gonzalez**
Department of Biomedical Informatics
Arizona State University
13212 E Shea Blvd
Scottsdale, AZ 85259
Graciela.gonzalez@asu.edu

## Abstract

Zoonotic viruses, viruses that are transmittable between animals and humans, represent emerging or re-emerging pathogens that pose significant public health threats throughout the world. It is therefore crucial to advance current surveillance mechanisms for these viruses through outlets such as phylogeography. Phylogeographic techniques may be applied to trace the origins and geographical distribution of these viruses using sequence and location data, which are often obtained from publicly available databases such as GenBank. Despite the abundance of zoonotic viral sequence data in GenBank records, phylogeographic analysis of these viruses is greatly limited by the lack of adequate geographic metadata. Although more detailed information may often be found in the related articles referenced in these records, manual extraction of this information presents a severe bottleneck. In this work, we propose an automated system for extracting this information using Natural Language Processing (NLP) methods. In order to validate the need for such a system, we first determine the percentage of GenBank records with "insufficient" geographic metadata for seven well-studied zoonotic viruses. We then evaluate four different named entity recognition (NER) systems which may help in the automatic extraction of information from related articles that can be used to improve the GenBank geographic metadata. This includes a novel dictionary-based location tagging system that we introduce in this paper.

*Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014)*, pages 1–9,
Baltimore, Maryland USA, June 26-27 2014.

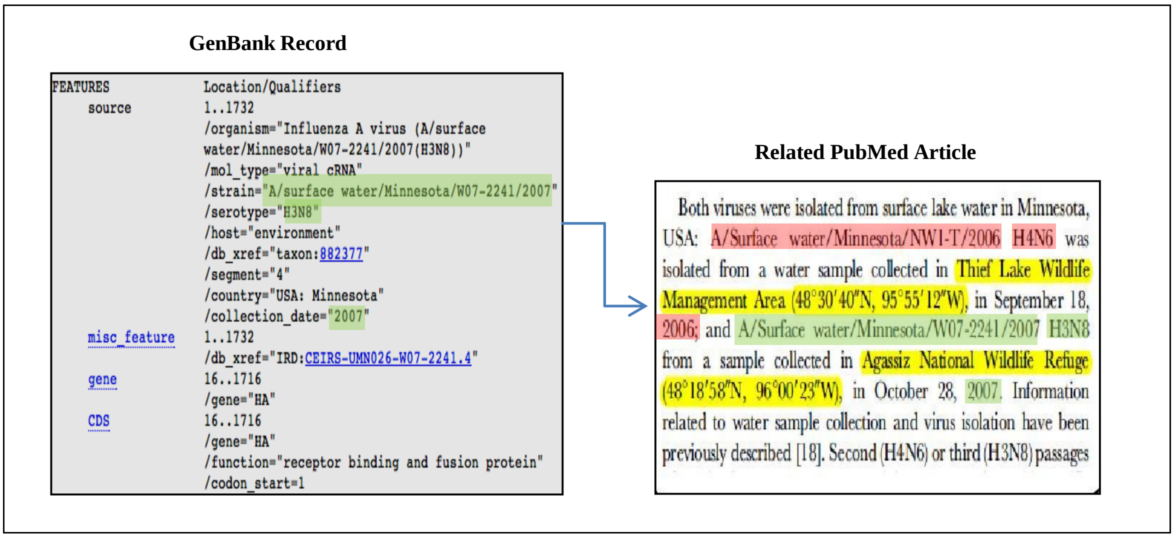
# 1 Introduction

Zoonotic viruses, viruses that are transmittable between animals and humans, have become increasingly prevalent in the last century leading to the rise and re-emergence of a variety of diseases (Krauss, 2003). In order to enhance currently available surveillance systems for these viruses, a better understanding of their origins and transmission patterns is required. This need has led to a greater amount of research in the field of phylogeography, the study of geographical lineages of species (Avise, 2000). Population health agencies frequently apply phylogeographic techniques to trace the evolutionary changes within viral lineages that affect their diffusion and transmission among animal and human hosts (Ciccozzi et al., 2013; Gray and Salemi, 2012; Weidmann et al., 2013). Prediction of virus migration routes enhances the chances of isolating the viral strain for vaccine production. In addition, if the source of the strain is identified, intervention methods may be applied to block the virus at the source and limit outbreaks in other areas.

Phylogeographic analysis depends on the utilization of both the sequence data and the location of collection of specific viral sequences. Researchers often use publicly available databases such as GenBank for retrieving this information. For instance, Wallace and Fitch (2008) used data from GenBank records to study the migration of the H5N1 virus in various animal hosts over Europe, Asia and Africa, and were able to identify the Guangdong province in China as the source of the outbreak. However, the extent of phylogeographic modeling is highly dependent on the specificity of available geospatial information and the lack of geographic data more specific than the state or province level may limit phylogeographic analysis and distort results. In the previous example, Wallace and Fitch (2008) had to use town-level information to identify the source of the H5N1 outbreak; without specific location data, they would not have been able to identify the Guangdong province as the source. Unfortunately, while there is an abundance of sequence data in GenBank records, many of them lack sufficient geographic metadata that would enable specific identification of the isolate's location of collection. A prior study conducted by Scotch et al. (2011) showed that the geographic information of 80% of the GenBank records associated with single or double stranded RNA viruses within tetrapod hosts is less specific than 1st level administrative boundaries (ADM1) such as state or province.

Though many of the records lack specific geographic metadata, more detailed information is often available within the journal articles referenced in them. However, manual extraction of this information is time-consuming and cumbersome and presents a severe bottleneck on phylogeographic analysis. In this work, we investigate the potential of NLP techniques to enhance the geographic data available for phylogeographic studies of zoonotic viruses using NER systems. In addition to geographic metadata and sequence information, GenBank records also contain several other forms of metadata such as host, collection date and gene for each isolate. Journal articles that are referenced in these records often mention the location of isolation for the viral sample in conjunction with related metadata (Figure 1 provides an example of such a case). Therefore, by allowing identification of location mentions along with mentions of related GenBank metadata in these articles, we believe that NER systems may help to accurately link each GenBank record to its corresponding location of isolation and distinguish it from other location mentions.

Previously Scotch et al. (2011) evaluated the performance of BANNER (Leaman and Gonzalez, 2008) and the Stanford NER tool (Finkel et al., 2005) for automated identification of gene and location mentions respectively, in 10 full-text PubMed articles, each related to a specific GenBank record. They were both found to achieve f-scores of less than 0.45, thereby establishing the need for NER systems with better performance and/or a larger test corpus (Scotch et al, 2011). In this study, we start by evaluating the state of geographic insufficiency for zoonotic viruses in GenBank records using a new automated approach. Next, we further expand upon the work done by Scotch et al. (2011) by building our own dictionary-based location-tagging system and evaluating its performance on a larger corpus corresponding to over 8,500 GenBank records for zoonotic viruses. In addition, we also evaluate the performance of three other state-of-the-art NER tools for tagging gene, date and species mentions in this corpus. We believe that identification of these entities will be useful for the future development of a system for extracting the location of collection of viral isolates from articles related to their respective GenBank records.



GenBank Record

```
FEATURES             Location/Qualifiers
     source          1..1732
                     /organism="Influenza A virus (A/surface
                     water/Minnesota/W07-2241/2007(H3N8))"
                     /mol_type="viral cRNA"
                     /strain="A/surface water/Minnesota/W07-2241/2007"
                     /serotype="H3N8"
                     /host="environment"
                     /db_xref="taxon:882377"
                     /segment="4"
                     /country="USA: Minnesota"
                     /collection_date="2007"
     misc_feature    1..1732
                     /db_xref="IRD:CEIRS-UMN026-W07-2241.4"
     gene            16..1716
                     /gene="HA"
     CDS             16..1716
                     /gene="HA"
                     /function="receptor binding and fusion protein"
                     /codon_start=1
```

Related PubMed Article

Both viruses were isolated from surface lake water in Minnesota, USA: A/Surface water/Minnesota/NW1-T/2006 H4N6 was isolated from a water sample collected in Thief Lake Wildlife Management Area (48°30′40″N, 95°55′12″W), in September 18, 2006; and A/Surface water/Minnesota/W07-2241/2007 H3N8 from a sample collected in Agassiz National Wildlife Refuge (48°18′58″N, 96°00′23″W), in October 28, 2007. Information related to water sample collection and virus isolation have been previously described [18]. Second (H4N6) or third (H3N8) passages

Figure 1. Example of how the date, gene, and strain metadata within a GenBank record may be used to differentiate between two potential locations in a related article

## 2 Methods

The process undertaken to complete this study can be divided into three distinct stages: selection of the zoonotic viruses and extraction of relevant GenBank data related to each virus, computation of "sufficiency" statistics on the extracted data, and development/evaluation of NER systems for tagging location, gene, date and species mentions in full-text PubMed Central articles. A detailed description of each phase is given below.

### 2.1 Virus Selection and GenBank Data Extraction

The domain of this study has been limited to zoonotic viruses that are most consistently documented and tracked by public health, agriculture and wildlife state departments within the United States. These viruses include influenza, rabies, hantavirus, western equine encephalitis (WEE), eastern equine encephalitis (EEE), St. Louis encephalitis (SLE), and West Nile virus (WNV). The Entrez Programming Utilities (E-Utilities) was used to download the following fields from 59,595 GenBank records associated with these viruses: GenBank Accession ID, PubMed Central ID, Strain name, Collection date and Country. These records were the result of a query performed to retrieve all accession numbers related to the selected viruses which had at least one reference to a PubMed Central article. The results from the query was retrieved on August 22nd, 2013.

### 2.2 Sufficiency Analysis

**Database Integration:** The data extracted from Genbank was used to compute the percentage of GenBank records that had insufficient geographic information for each of the selected viruses. In order to perform this computation, we used data from the ISO 3166-1 alpha-2 [1] table and the GeoNames database. The ISO 3166-1 alpha-2 is the International Standard for representing country names using two-letter codes. The GeoNames[2] database contains a variety of geospatial data for over 10 million locations on earth, including the ISO 3166-1 alpha-2 code for the country of each location and a feature code that can be used to determine the administrative level of each location. To allow for efficient querying, we downloaded the main GeoNames table and the ISO alpha-2 country codes table from their respective websites and stored them in a local SQL database. Prior to adding the ISO data to the database, some commonly used country names and their corresponding country codes were added to the table since it only included a single title for each country. For example, the ISO table included the country name "United States" but not alternate names such as "USA", "United States of America", or "US". Using the created database in conjunction with a parser written in Java, we were able to retrieve most

[1] Iso.org. [Internet]. Genève. c2013. Available from http://www.iso.org/iso/home/standards/country_codes.htm

[2] Geonames.org. [Internet]. Egypt. c2013. [updated 2013 Apr 30] Available from http://www.geonamesorg/EG/administrative-division-egypt.html

of the geographic information present within the records and classify each of them as sufficient or insufficient.

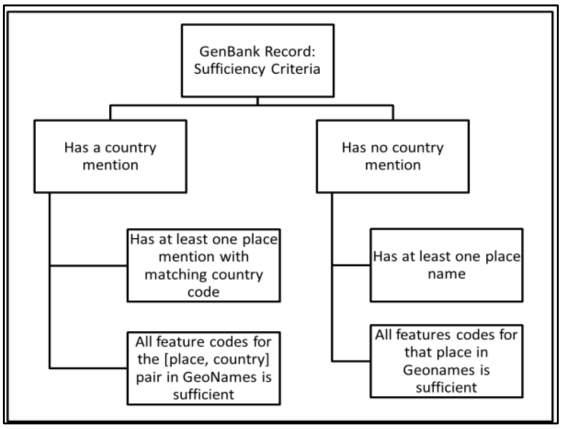

GenBank Record: Sufficiency Criteria
Has a country mention
Has no country mention
Has at least one place mention with matching country code
All feature codes for the [place, country] pair in GeoNames is sufficient
Has at least one place name
All features codes for that place in Geonames is sufficient

Figure 2. Sufficiency Criteria

**Sufficiency Criteria:** For the purpose of this project, we considered any geographical boundary more specific than ADM1 to be "sufficient". Based on this criterion, a feature code in GeoNames was categorized as sufficient only if it was absent from the following list of feature codes: ADM1, ADM1H, ADMD, ADMDH, PCL, PCLD, PCLF, PCLH, PCLI and PCLS. Evaluation of the geographical sufficiency of a GenBank record was dependent upon whether the record included a country name. A GenBank record with a country mention was called sufficient if the geographic information extracted from that record included another place mention whose feature code fell within the class of sufficient feature codes and whose ISO country code matched that of the retrieved country. For instance, a GenBank record with the geographic metadata "Orange County, United States" will be called sufficient since the place "Orange County" has a sufficient feature code of "ADM2" and a country code of "US" which matches the country code of the retrieved country, "United States". Place mentions with matching country codes often had several different feature codes in GeoNames. Such places were only called sufficient if all feature codes corresponding to the given pair of place name and country code were classified as sufficient. In cases where the GenBank record had no country mention, the record was called sufficient only if all matching GeoNames entries for any of the places mentioned in it had sufficient feature codes. The sufficiency criteria were designed to ensure that a geographic location is only called sufficient if its administrative level was found to be more specific than ADM1 without any form of ambiguity. Figure 3 illustrates the pathways of geographical sufficiency for GenBank records in a diagram.

**Sufficiency Computation:** In order to obtain the geographic information for each Genbank record, we used a Java parser which automatically extracted data from the "country" field of each record. Since the "country" field typically contained multiple place mentions divided by a set of delimiters consisting of comma, colon and hyphen, we first split this field using these delimiters. We then checked each string obtained through this process against the ISO country code table to determine whether it was a potential country name for the record's location. If the query returned no results, then the locally stored GeoNames table was searched and for each match found, the corresponding ISO country code and feature code were extracted. Figure 4 shows a diagram of this process.

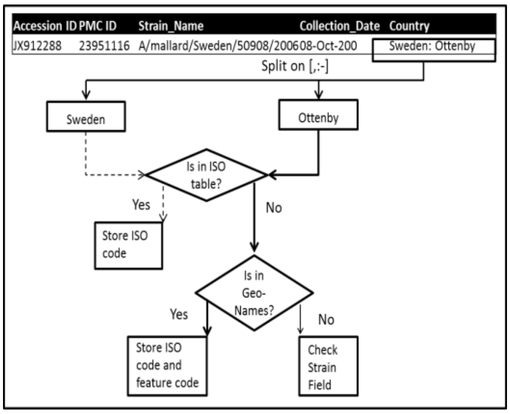
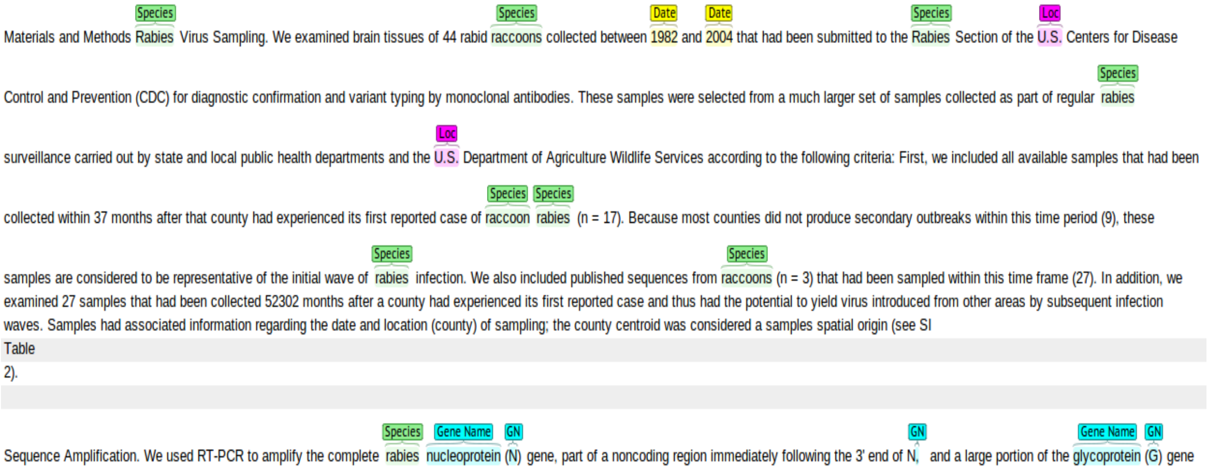


Figure 3. Sufficiency Calculation Example

In cases where no sufficient location data was found from the "country" field of a GenBank record, the Java parser searched through its "strain" field. This was done because some viral strains such as influenza include their location of origin integrated into their names. For example, the influenza strain "A/duck/Alberta/35/76" indicates that the geographic origin of the strain is Alberta. The different sections of a strain field are separated by either forward slash, parenthesis, comma, colon, hyphen or underscore and so we used a set of delimiters consisting of these characters to split this field. Each string thus retrieved was queried as before on the ISO country code table and the GeoNames table. GeoNames often returned matches for strings like 'raccoons' and 'chicken' which were actually meant to be names of host species within the "strain" field, and so a list of



Materials and Methods Rabies Virus Sampling. We examined brain tissues of 44 rabid raccoons collected between 1982 and 2004 that had been submitted to the Rabies Section of the U.S. Centers for Disease Control and Prevention (CDC) for diagnostic confirmation and variant typing by monoclonal antibodies. These samples were selected from a much larger set of samples collected as part of regular rabies surveillance carried out by state and local public health departments and the U.S. Department of Agriculture Wildlife Services according to the following criteria: First, we included all available samples that had been collected within 37 months after that county had experienced its first reported case of raccoon rabies (n = 17). Because most counties did not produce secondary outbreaks within this time period (9), these samples are considered to be representative of the initial wave of rabies infection. We also included published sequences from raccoons (n = 3) that had been sampled within this time frame (27). In addition, we examined 27 samples that had been collected 52302 months after a county had experienced its first reported case and thus had the potential to yield virus introduced from other areas by subsequent infection waves. Samples had associated information regarding the date and location (county) of sampling; the county centroid was considered a samples spatial origin (see SI

Table

2).

Sequence Amplification. We used RT-PCR to amplify the complete rabies nucleoprotein (N) gene, part of a noncoding region immediately following the 3' end of N, and a large portion of the glycoprotein (G) gene

Figure 4. Example of annotation including all four entities

some of the most frequently seen host name mentions in these records was manually created and filtered out before querying GeoNames.

Some of the place mentions contained very specific location information which resulted in GeoNames not finding a match for them. A list was created for strings like ‘north’, ‘south-east’, ‘governorate’ etc. which when removed from a place mention may produce a match. In cases of potential place mentions which contained any one of these strings and for which GeoNames returned no matching result, a second query was performed after removal of the string.

**Evaluation of Sufficiency Computation:** We manually annotated 10% of all influenza records in GenBank which reference at least one PubMed Central article as sufficient or insufficient based on our sufficiency criteria (5731 records). We then ran our program on these records and compared system results with annotated results.

### 2.3 Development/Evaluation of NER systems

**Creation of Gold Standard Corpus:** We created a gold standard corpus consisting of twenty-seven manually-annotated full-text PubMed Central articles in order to evaluate the performance of NER systems for tagging location, gene, species and date mentions in text. The articles corresponded to over 8,500 GenBank records and were randomly sampled using the subset of extracted GenBank records which contained a link to PubMed Central articles and had insufficient geographic metadata.

Three annotators tagged the following four entities in each article using the freely available annotation tool, BRAT (Stenetorp et al., 2012): gene names, locations, dates and species. Figure 4 provides an example of the manual annotation in BRAT. We annotated all mentions of each entity type, not only those relevant to zoonotic viruses, in order to evaluate system performance. A total of over 19,000 entities were annotated within this corpus. The number of tokens annotated was about 24,000. A set of annotation guidelines was created for this process (available upon request). Before creating the guidelines, each annotator individually annotated six common articles and compared and discussed their results to devise a reasonable set of rules for annotating each entity. After discussion, the annotators re-annotated the common articles based on the guidelines and divided the remaining articles amongst themselves. The inter-annotator agreement was calculated for each pair of annotators. The annotated corpus will be made available at diego.asu.edu/downloads.

**Development of Automated Location Tagger:** We developed a dictionary-based NER system using the GeoNames database for automated identification of location mentions in text. The dictionary used by this system, which we will hereby refer to as GeoNamer, was created by retrieving distinct place names from the GeoNames table and filtering out commonly used words from the retrieved set. Words filtered out include stop words such as ‘are’ and ‘the’, generic place names such as ‘cave’ and ‘hill’, numbers like ‘one’ and ‘two’, domain specific words such as ‘biology’ and ‘DNA’, most commonly used surnames like ‘Garcia’, commonly used animal names such as ‘chicken’ and ‘fox’ and other miscellaneous words such as ‘central’. This was a crucial step since the GeoNames database contains a wide array of commonly used English words which may cause a large volume of false positives if not removed. The final dictionary consists of 5,396,503 entries. In order to recognize place mentions in a given set of text files, GeoNamer first builds a Lucene index on the contents of the files. It then constructs a phrase query for every entry in the Geonames dictionary and runs each query on the Lucene index. The document id, query text, start offset and end offset for every match found is written to an output file. We chose this approach because of its simplicity and efficiency.

**Evaluation of NER Systems:** Four different NER systems for identifying species, gene, date and location mentions in text were evaluated using the created gold standard. The evaluated systems include LINNEAUS (Gerner et al., 2010), BANNER, Stanford SUTime (Chang and Manning, 2012) and GeoNamer. LINNEAUS, BANNER and Stanford SUTime are widely-used, state-of-the-art open source NER systems for recognition of species, gene and temporal expressions respectively. GeoNamer is the system we developed in this work for the purpose of tagging locations, as described earlier.

## 3 Results

### 3.1 Sufficiency Analysis

The system for classifying records as sufficient or insufficient was found to have an accuracy of 72% as compared to manual annotation. 98% of the errors was due to insufficient records being called sufficient. The results of the sufficiency analysis are given in Table 1. 64% of all GenBank records extracted for this project contained insufficient geographic information. Amongst the seven studied viruses, WEE had the highest and EEE had the lowest percentage of insufficient records.

| **Virus Type** | **Number of Entries** | **% Insufficient** |
|---|---|---|
| WEE | 67 | 90 |
| Rabies | 4450 | 85 |
| WNV | 1084 | 79 |
| SLE | 141 | 74 |
| Hanta | 1745 | 66 |
| Influenza | 51734 | 62 |
| EEE | 374 | 51 |
| All | 59595 | 64 |

Table 1. Percentage of GenBank records with insufficient geographic information for each zoonotic virus studied in this project

### 3.2 Gold Standard Corpus

The results for the comparison of the annotations performed by our three annotators on 6 common papers can be found in Table 2. We used the F-score between each pair of annotators as a measure of inter-rater agreement and had over 90% agreement with overlap matching and over 86% agreement with exact matching in all cases. The final gold standard corpus contained approximately 19,000 entities corresponding to approximately 24,000 tokens.

| **Entity** | **F-score (*A*,*B*) (Exact; Overlap)** | **F-score (*A*, *C*) (Exact; Overlap)** | **F-score (*B*, *C*) (Exact; Overlap)** |
|---|---|---|---|
| Date | .975; .978 | .979; .987 | .962; .973 |
| Gene | .914; .926 | .913; .932 | .911; .954 |
| Location | .945; .961 | .907; .931 | .914; .935 |
| Species | .909; .956 | .874; .940 | .915; .959 |
| Virus | .952; .958 | .947; .966 | .947; .955 |
| **Mean** | **.939; .956** | **.924; .951** | **.930; .955** |

Table 2. Frequency of Annotated Entities for 6 common annotated papers

### 3.3 Performance Analysis of NER Systems

The performance metrics for the NER systems at tagging the desired entities in the test set are listed in Table 3. The highest performance was achieved by Stanford SUTime for date tagging. Tagging of genes had the lowest performance.

| **Entity** | **Precision (Exact; Overlap)** | **Recall (Exact; Overlap)** | **F-score (Exact; Overlap)** |
|---|---|---|---|
| BANNER | 0.070; 0.239 | 0.114; 0.395 | 0.087; 0.297 |
| GeoNamer | 0.452; 0.626 | 0.658; 0.783 | 0.536; 0.696 |
| LINNEAUS | 0.853; 0.962 | 0.563; 0.658 | 0.678; 0.781 |
| Stanford SUTime | 0.800; 0853 | 0.681; 0.727 | 0.736; 0.785 |

Table 3. Performance Statistics of NER

## 4 Discussion

Based on our analysis, at least half of the GenBank records for each of the studied zoonotic viruses lack sufficient geographic information, and the proportion of insufficient records can be as high as 90%. Our automated system for classifying records as insufficient or sufficient was found to have an accuracy of 72% with 98% of the errors being a result of insufficient records being called sufficient. Therefore, our computed estimate of insufficiency is very likely to be an underestimation of the actual problem. The virus with the highest level of sufficiency, EEE, had a large number of records with county level information in the "country" field. However, the insufficient records for this virus typically contained no place mention, not even at the country level. A key reason for our calculated percentage of sufficient GenBank records being higher for these seven viruses than what has been previously computed by Scotch et al. (2011) was the inclusion of the "strain" field. The "strain" field often contained specific location information which, when combined with place mentions present within the "country" field, made the record geographically sufficient. The virus for which the inclusion of "strain" field had the greatest impact on boosting the sufficiency percentage was influenza. Most of the GenBank records associated with this virus had structured "strain" fields from which the parser could easily separate place mentions using GeoNames.

Although the sufficiency classifications produced by our system were correct most of the time, there were a few cases where a record got incorrectly labeled as insufficient even when it contained detailed geographic information. This typically happened because GeoNames failed to return matching results for these places. For instance, the country field "India: Majiara,WB" was not found to be sufficient even though Majiara is a city in India because GeoNames has no entry for it. In some cases the lack of matching result was due to spelling variations of the place name. For instance the country field "Indonesia: Yogjakarta" was called insufficient since "Yogjakarta" is spelled as "Yogyakarta" in GeoNames. Sometimes the database simply did not contain the exact string present in the GenBank record. For instance, it does not have any entry for the place "south Kalimantan" but it contains the place name "kalimantan". The number of sufficient records which were called insufficient by our system due to inexact matching were greatly mitigated by removing strings such as "south" from the place mention, as described in the "Methods" section.

Most of the NER systems performed significantly better with overlap measures than with exact-match measures. This is because our annotation guidelines typically involved tagging the longest possible match for each entity and the automated systems frequently missed portions of each annotation. Stanford SUTime had the best overlap f-measure of 0.785, closely followed by LINNEAUS with an overlap f-measure of 0.781. Although Stanford SUTime was fairly effective at finding date mentions in text, it tagged all four-digit-numbers such as "1012" and "2339" as years, leading to a number of false positives. The poor recall of LINNEAUS was mostly caused because the dictionary used by LINNEAUS tagged only species mentions in text while we tagged genus and family mentions as well. It also missed a lot of commonly used animal names such as monkey, bat, badger and wolf. GeoNamer was the third best performer with the highest recall but second lowest precision. This is because the GeoNames dictionary contains an extensively large list of location names, many of which are commonly used words such as "central". Even though we filtered out a vast majority of these words, it still produced false positives such as "wizard". However, its performance was considerably better than that of the Stanford location tagger used by Scotch et al. (2011) which was found to have a recall, precision and f-score of 0.26, 0.81 and 0.39 respectively. The improved performance was achieved because of the higher recall of our system. The GeoNames dictionary provides an extensive coverage of all location mentions in the world and the Stanford NER system, which is a CRF classifier trained on a different dataset, was not able to recognize many of the place mentions present in full-text PMC articles related to GenBank records.

BANNER showed the poorest performance amongst all the entity taggers evaluated in this paper. In fact, the f-score we achieved for BANNER in this study was much lower that its past f-score of 0.42 within the domain of articles related to GenBank records for viral isolates (Scotch et al., 2011). As mentioned by Scotch et al. (2011), a key reason for BANNER's poor performance in this domain is the difference between the data set used to train the BANNER model and the annotation corpus used to test this system. The version of BANNER used in these two studies was trained on the training set for the BioCreative 2 Gene

Mention task, which comprised of 15,000 sentences from PubMed abstracts. These abstracts often contained the full names for gene and protein mentions while the full-text articles we used mostly contained the abbreviated forms of gene names, which BANNER tended to miss. The articles also contained abbreviated forms of several entities such as viral strain name (e.g. H1N1) and species name (e.g. VEEV) which look similar to abbreviated gene names. Therefore, BANNER often misclassified these entities as gene mentions. A possible reason for BANNER having a much lower performance in this study than in the previous study conducted by Scotch et al (2011) is the presence of a large number of tables in the journal articles we selected. BANNER is a machine learning system based on conditional random fields which uses orthographic, morphological and shallow syntax features extracted from sentences to identify gene mentions in text. Such features do not help greatly for extraction from tables. Therefore, BANNER was often not able to identify the gene mentions in the tables present within our corpus, thereby producing false negatives. Moreover, it tagged several entries within the table as a single gene name, thereby producing false positives as well. This reduced both the recall and precision of BANNER.

Although this study explores the problem of insufficient geographic information in GenBank more thoroughly than past studies, the number of papers annotated as the gold standard is still limited. Thus, the performance of the taggers reported can be construed as a preliminary estimate at best. The set of taggers and their performance seem to be adequate for a large-scale application, with the exception of BANNER. However, we did not make any changes to the BANNER system (specifically, re-training) since changes to it are not possible until sufficient data is annotated for retraining.

## 5 Conclusions and Future Work

It can be concluded that the majority of GenBank records for zoonotic viruses do not contain sufficient geographic information concerning their origin. In order to enable phylogeographic analysis of these viruses and thereby monitor their spread, it is essential to develop an efficient mechanism for extracting this information from published articles. Automated NER systems may help accelerate this process significantly. Our results indicate that the NER systems LINNEAUS, Stanford SUTime and GeoNamer produce satisfactory performance in this domain and thus can be used in the future for linking GenBank records with their corresponding geographic information. However, the current version of BANNER is not well-suited for this task. We will need to train BANNER specifically for this purpose before incorporating it within our system.

We are currently altering the component of our program which classifies records as sufficient or insufficient in order to reduce the number of errors due to insufficient records being called sufficient. We are also manually looking through GenBank records for zoonotic viruses with insufficient geographic metadata and linking them to the location mentions in related articles which we deem to be the most likely location of collection for the given viral isolate. The resulting annotated corpus will be used to train and evaluate an automated system for populating GenBank geographic metadata. We have already covered all GenBank records related to Encephalitis viruses and close to 10% of all records related to Influenza which are linked to PubMed Central articles. The annotation process has revealed that a large proportion of the information allowing linkage of GenBank records to geographic metadata is often present in tables within the articles in addition to textual sentences. Therefore, we have developed a Python parser for automatically linking GenBank records to location mentions using tables from the HTML version of the PubMed Central articles. Future work will include further expansion of this annotation corpus and the development of an integrated system for enhancing GenBank geographic metadata for phylogeographic analysis of zoonotic viruses.

## Acknowledgement

Research reported in this publication was supported by the NIAID of the NIH under Award Number R56AI102559 to MS and GG. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

## References

Avise, John C. (2000). Phylogeography : the history and formation of species Cambridge, Mass.: Harvard University Press.

Chang, Angel X., and Christopher Manning. "SUTime: A library for recognizing and normalizing time expressions." LREC. 2012.

Ciccozzi M, et al. Epidemiological history and phylogeography of West Nile virus lineage 2. Infection, Genetics and Evolution. 2013:17;46-50.

Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL 2005); 2005. p. 363–70.

Gerner M, Nenadic G, and Bergman CM. LINNAEUS: A species name identification system for biomedical literature. BMC Bioinformatics. 2010;11(85).

Gray RR, and Salemi M. Integrative molecular phylogeography in the context of infectious diseases on the human-animal interface. Parasitology-Cambridge. 2012;139:1939-1951

Krauss, H. (2003). Zoonoses: infectious diseases transmissible from animals to humans (3rd ed.). Washington, D.C.: ASM Press.

Leaman R and Gonzalez G. BANNER: An executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing. 2008;13:652-663.

Scotch, Matthew, et al. Enhancing phylogeography by improving geographical information from GenBank. Journal of biomedical informatics. 2011;44:S44-S47.

Stenetorp P, et al. BRAT: A Web-based Tool for NLP-Assisted Text Annotation. EACL '12 Proceedings

Wallace, R.G. and W.M. Fitch, Influenza A H5N1 immigration is filtered out at some international borders. PLoS One, 2008. 3(2): p. e1697.

Weidmann M, et al. Molecular phylogeography of tick-borne encephalitis virus in Central Europe. Journal of General Virology. 2013;94:2129-2139.