

# Analysis of ProMED-mail information network

Author: Nathan Breit (breit@ecohealthalliance.org)

## Synopsis

This analysis is intended to provide actionable insights to ProMED contributors about the news sources and organizations the content posted on ProMED-mail is derived from. The key question it seeks to answer is "How can news gathering efforts be focused to improve ProMED-mail's efficiency?" Toward this end, news sources with long reporting delays are identified, ProMED's historical reporting delay is measured and compared year by year and a method for determining the best sources to use to gather information about specific organizations is presented. Depending on interest from ProMED editors, more detailed analysis can be performed. For example, there are many ways the relationships between organizations and news sources could be further explored. If the data presented here is useful, we could develop interactive visualizations around it that allow users to drill into their results further, and we could develop GRITS features to do similar types of analysis in a more generalized manner.

## General statistics about the posts this analysis was derived from

The posts all came from the main ProMED-mail feed. In future work, this analysis could be extended to the regional feeds (e.g. ProMED-RUS, ProMED-FRA...).

Date range: 08/20/94 to 07/02/15

Total posts: 50165

Total articles posted: 69233

## Which news sources have the most articles posted on ProMED?

Sources are identified using the "Source:" headers at the top of some articles. If the source is "None" it means that no header was identified for the article. Some source name are resolved to generalized canonical names based on the keywords they contain. For example, "USA CDC, Division of Vector-Borne Infectious Division" is resolved to CDC because it contains the keyword "CDC". Only the most prolific sources are resolved at the moment.

Total sources: 26436

Total sources with only 1 article: 22783

Out[364]:

|    | source                              | number of articles |
|----|-------------------------------------|--------------------|
| 0  | None                                | 14095              |
| 1  | WHO                                 | 2461               |
| 2  | Reuters                             | 1994               |
| 3  | OIE                                 | 1990               |
| 4  | Associated Press                    | 1166               |
| 5  | Xinhua News Agency                  | 949                |
| 6  | BBC                                 | 947                |
| 7  | CDC                                 | 837                |
| 8  | ABC                                 | 501                |
| 9  | Times of India                      | 479                |
| 10 | New York Times                      | 399                |
| 11 | Eurosurveillance Weekly             | 367                |
| 12 | Office International des Epizooties | 329                |
| 13 | UN                                  | 328                |
| 14 | CIDRAP News                         | 328                |
| 15 | Saudi Arabia Ministry of Health     | 253                |
| 16 | CNN                                 | 230                |
| 17 | Washington Post                     | 186                |
| 18 | The Guardian                        | 164                |
| 19 | Outbreak News Today                 | 125                |

### Take-aways

- ProMED relies on many news companies based in English speaking countries (BBC, ABC, CNN, New York Times...). This may mean that news sources based in other countries are underutilized.
- Most sources have only a single article associated with them. It may be worth periodically examining some of these to see if they continue to produce valuable content.
- Webscraping services that do automated notifications of new content would do well to start with the sources at the top of this list.

# Most frequently mentioned organizations

Organizations are identified by using NLP techniques to scan the body of articles on ProMED. In the future, similar techniques could be used to find mentions of other things such as diseases and locations. However, this type of analysis is error prone. There are many false positives and some of the terms identified are ambiguous. For example, the term "Ministry of Health" does not make it clear which Ministry of Health is being mentioned. Use of contextual information, such as the locale the article was published in, may help disambiguate some organizations. Countries and political constructs (European Union) also appear frequently in the organization list. It is unclear as to whether they should be included as organizations.

Total organizations: 82295

Ambiguous organizations:

Out[296]:

|           | <b>_id</b>                                 | <b>count</b> |
|-----------|--|--------------|
| <b>1</b>  | Center for Disease Control                 | 10089        |
| <b>2</b>  | Ministry of Health                         | 3422         |
| <b>5</b>  | Food and Drug Administration               | 2981         |
| <b>9</b>  | Health Ministry                            | 1700         |
| <b>10</b> | Department of Health                       | 1695         |
| <b>11</b> | Ministry of Agriculture                    | 1604         |
| <b>12</b> | Centers for Disease Control and Prevention | 1317         |
| <b>13</b> | Health Department                          | 1242         |
| <b>30</b> | National Institute of Health               | 479          |
| <b>32</b> | Agriculture Ministry                       | 472          |
| <b>33</b> | Department of Agriculture                  | 471          |
| <b>40</b> | Department of Natural Resources            | 420          |
| <b>41</b> | The Ministry of Health                     | 413          |
| <b>59</b> | Department of Public Health                | 259          |
| <b>77</b> | National Institutes of Health              | 217          |
| <b>84</b> | Ministry of Public Health                  | 204          |
| <b>88</b> | Agriculture Department                     | 197          |

Unambiguous organizations:

Out[297]:

|           | <b>_id</b>                              | <b>count</b> |
|-----------|---|--------------|
| <b>3</b>  | World Health Organization               | 3346         |
| <b>4</b>  | European Union                          | 3099         |
| <b>6</b>  | USA                                     | 2877         |
| <b>7</b>  | United States Department of Agriculture | 2725         |
| <b>8</b>  | United Kingdom                          | 2175         |
| <b>14</b> | Office International des Epizooties     | 1235         |
| <b>15</b> | Food and Agriculture Organization       | 1153         |
| <b>16</b> | Health Protection Agency                | 1116         |
| <b>17</b> | Centers for Disease Control             | 899          |
| <b>18</b> | Reuters                                 | 842          |
| <b>19</b> | Canadian Food Inspection Agency         | 798          |
| <b>20</b> | Médecins Sans Frontières                | 788          |
| <b>21</b> | UNICEF                                  | 766          |
| <b>22</b> | European Commission                     | 716          |
| <b>23</b> | Federal Bureau of Investigation         | 707          |
| <b>24</b> | World Health Organisation               | 673          |
| <b>25</b> | Sociedade Brasileira de Virologia       | 613          |
| <b>26</b> | Sierra Leone                            | 605          |
| <b>27</b> | Agence France Presse                    | 602          |
| <b>28</b> | Democratic Republic of Congo            | 601          |

## Which sources report on which organizations the most?

This data is intended to show which news sources report on which organizations the most. The table below shows the top twenty pairings for a select set of organizations. This information could be presented visually using something akin to a Sankey diagram (see the mockup [here](https://docs.google.com/presentation/d/1SEgH6in0YrbgymOVc8ggQQ88v-SEntXLSNMPC790EuU/edit?usp=sharing) (<https://docs.google.com/presentation/d/1SEgH6in0YrbgymOVc8ggQQ88v-SEntXLSNMPC790EuU/edit?usp=sharing>)).

Out[370]:

|    | count | org                                     | src   |
|----|-------|---|---|
| 7  | 311   | European Union                          | Reuters   |
| 8  | 268   | World Health Organization               | Reuters   |
| 10 | 234   | Health Protection Agency                | BBC   |
| 18 | 170   | United States Department of Agriculture | Reuters   |
| 22 | 146   | World Health Organization               | CDC   |
| 25 | 140   | World Health Organization               | WHO   |
| 30 | 117   | UNICEF                                  | WHO   |
| 31 | 116   | World Health Organization               | Associated Press                                  |
| 33 | 112   | Russian Federation                      | WHO   |
| 34 | 110   | World Health Organization               | CIDRAP News                                       |
| 37 | 105   | Berger SA                               | GIDEON (Global Infectious Disease & Epidemiolo... |
| 43 | 99    | World Health Organization               | New York Times                                    |
| 45 | 95    | European Union                          | BBC   |
| 50 | 88    | United States Department of Agriculture | OIE   |
| 51 | 88    | National Institute of Virology          | Times of India                                    |
| 54 | 87    | United States Department of Agriculture | Associated Press                                  |
| 61 | 80    | World Health Organization               | Xinhua News Agency                                |
| 62 | 80    | EAEC                                    | Eurosurveillance Edition 2013, 18(37)             |
| 65 | 78    | European Union                          | Eurosurveillance Weekly                           |
| 72 | 74    | International Federation of Red Cross   | WHO   |

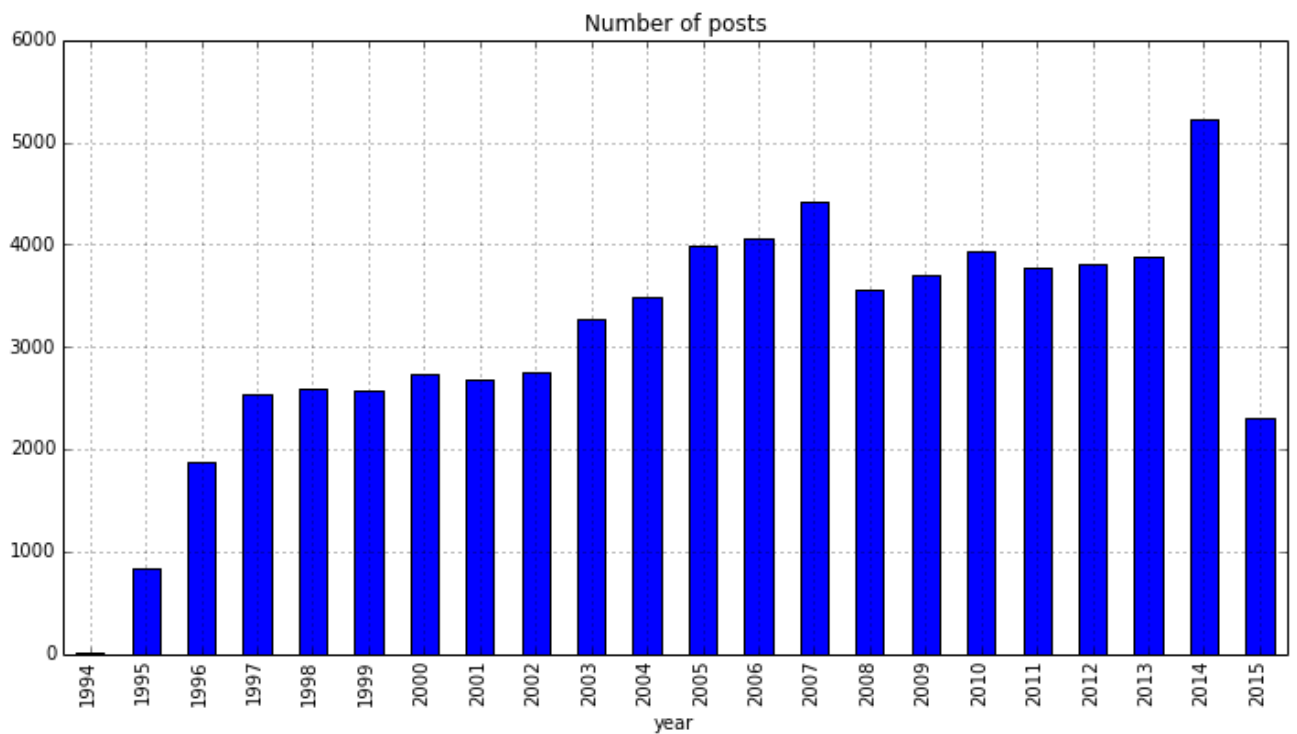
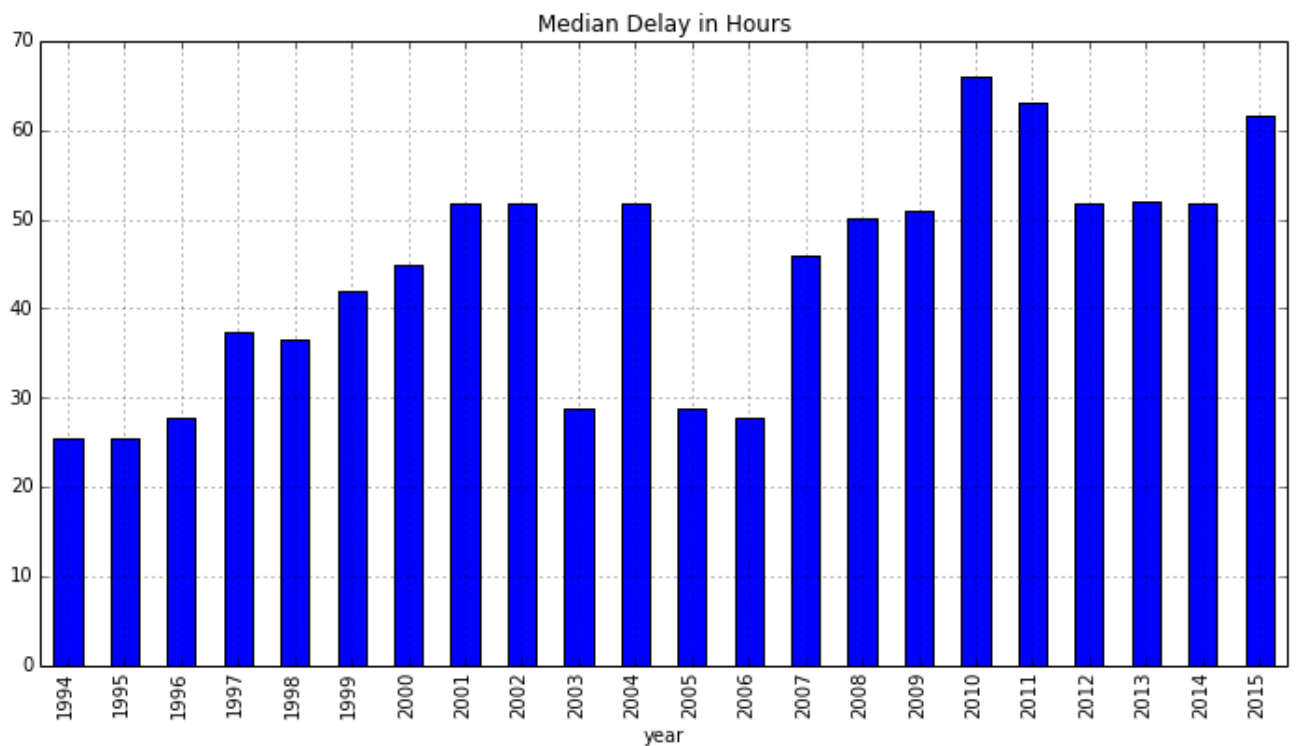
## Take-aways

- If there are organizations you would like to keep an eye on, this type of analysis can be used to select news sources that report on them most frequently. For example, Reuters articles mention the European Union, the WHO, and the United States Department of Agriculture more than any other source, so monitoring Reuters reports is likely to reveal information about those organizations.
- This type of analysis could be targeted at specific organizations that are underreported on. It may be beneficial to find sources that report on organizations that are reported on by relatively few major sources. In this list "UNICEF" is an example of such an organization.
- Sources could be compared by the number of organizations they mention as a proxy for their breadth of reporting.
- News sources tend to mention themselves which results in some pairings that are not so useful. For instance, the WHO frequently mentions itself in its publications. These types of results could be filtered out, although it can be interesting to compare the number of times an organization mentions itself in relation to the amount that others mention it.

## How long does it take articles to appear on ProMED from their original publication date depending on the source?

Delay is measured as the difference between the time of article publication and the time it was posted on ProMED-mail. The delays are not all exact as the article publication timestamps might only include a date and might not specify a time zone. Furthermore, the publication date of many articles is unknown.

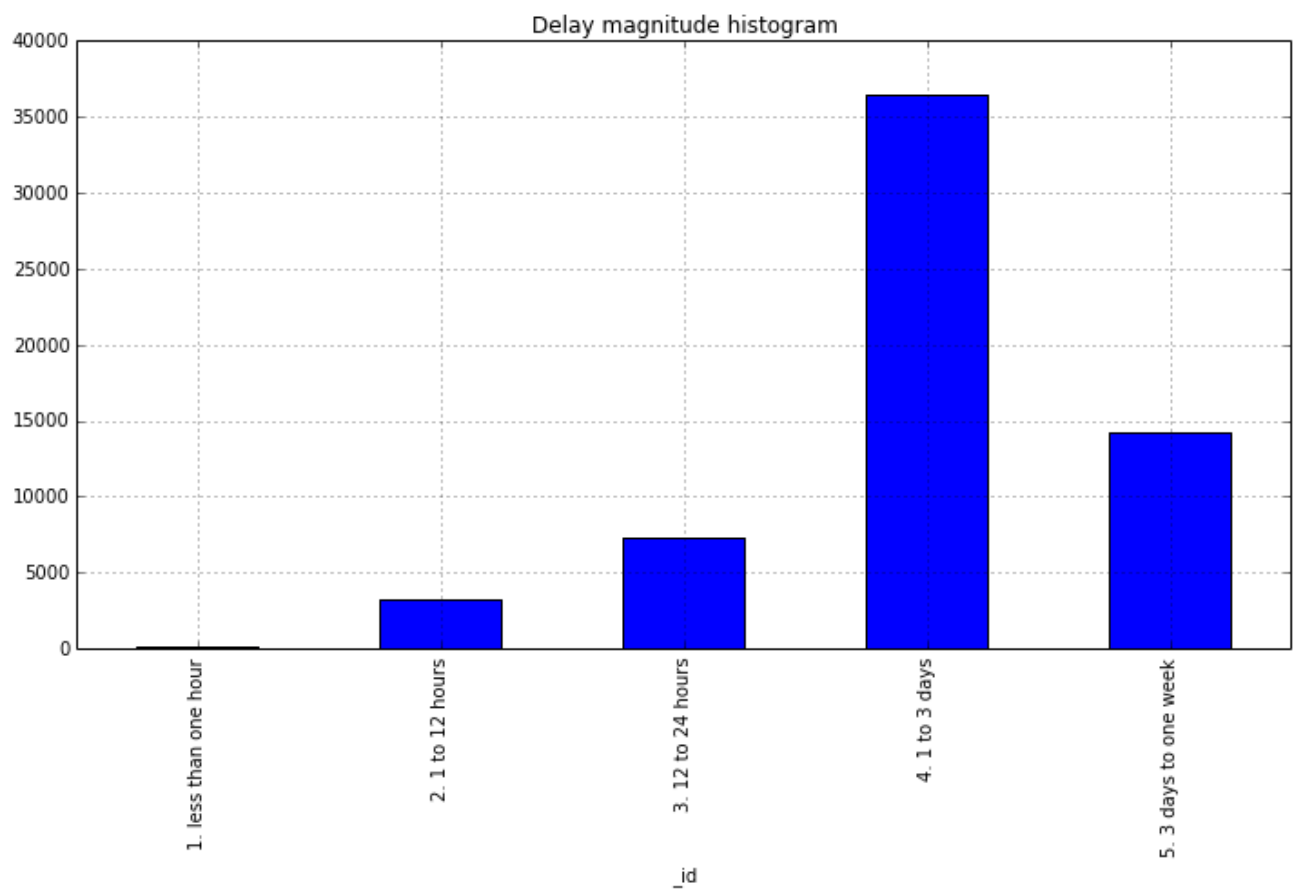
Articles where a publication date was found: 68038 / 69233



### Take aways:

- The median delay can be viewed as a performance metric.
- The general trend seems to be rising delay times. This could be due to having to sort through a larger number of spam posts. However, 2003, 2005, and 2006 all have usually low delay times even though the number of posts is high than previous years.

In future work, we would like to present this data in greater detail, for example by using box and whiskers plots to present the range of delays and by providing ways to drill into the data and explore the causes of delays.



## How long does it take articles to appear on ProMED from their original publication date depending on the source?

Mean delay is heavily skewed for most sources by a few articles with typos in their publication year, or some unusual circumstance that causes the delay to be years long. This makes the median delay a better indication of the typical delay for a source.

The chart below shows only sources with 50 or more articles posted on ProMED-mail.

Out[365]:

|  | median delay in hours | mean       | number of articles |
|--|-----------------------|------------|--------------------|
| <b>source</b>  |                       |            |                    |
| <b>News media</b>  | 27.559722             | 67.859529  | 95                 |
| <b>Saudi Arabia Ministry of Health</b>                               | 27.674167             | 187.605080 | 244                |
| <b>WHO</b>   | 28.175556             | 130.189569 | 2389               |
| <b>Reuters</b>   | 28.833333             | 82.175462  | 1956               |
| <b>CNN</b>   | 29.296806             | 132.105511 | 226                |
| <b>Eurosurveillance Weekly</b>                                       | 30.925556             | 90.610129  | 361                |
| <b>GIDEON (Global Infectious Disease &amp; Epidemiology Network)</b> | 32.075278             | 96.880469  | 87                 |
| <b>Nando Net</b>   | 34.552222             | 42.733362  | 58                 |
| <b>New York Times</b>  | 38.870556             | 165.328102 | 391                |
| <b>Xinhua News Agency</b>  | 40.000694             | 156.752588 | 936                |
| ...  | ...                   | ...        | ...                |
| <b>The Global Dispatch</b>   | 61.479444             | 82.573354  | 53                 |
| <b>The Horse</b>   | 62.129444             | 361.186676 | 61                 |
| <b>UN</b>  | 64.004722             | 131.420316 | 323                |
| <b>Los Angeles Times</b>   | 64.371250             | 80.341314  | 52                 |
| <b>The Hindu</b>   | 64.769583             | 122.453762 | 116                |
| <b>Radio Dabanga</b>   | 67.045556             | 88.195146  | 57                 |
| <b>Outbreak News Today</b>   | 68.581111             | 403.804639 | 123                |
| <b>Angola Press</b>  | 77.000000             | 112.460374 | 72                 |
| <b>Thanh Nien News</b>   | 85.002500             | 120.168620 | 66                 |
| <b>Prensa Latina</b>   | 97.470278             | 110.632739 | 65                 |

45 rows × 3 columns

Posts from the source with the greatest median delay (Prensa Latina).

Out[367]:

|     | link  | delayHours |
|-----|---|------------|
| 0   | <a href="http://www.promedmail.org/direct.php?id=14438">http://www.promedmail.org/direct.php?id=14438</a>     | 432.000556 |
| 1   | <a href="http://www.promedmail.org/direct.php?id=316957">http://www.promedmail.org/direct.php?id=316957</a>   | 335.001389 |
| 2   | <a href="http://www.promedmail.org/direct.php?id=2746">http://www.promedmail.org/direct.php?id=2746</a>       | 316.000000 |
| 3   | <a href="http://www.promedmail.org/direct.php?id=1182070">http://www.promedmail.org/direct.php?id=1182070</a> | 252.247222 |
| 4   | <a href="http://www.promedmail.org/direct.php?id=3249609">http://www.promedmail.org/direct.php?id=3249609</a> | 195.056667 |
| 5   | <a href="http://www.promedmail.org/direct.php?id=616798">http://www.promedmail.org/direct.php?id=616798</a>   | 194.001111 |
| 6   | <a href="http://www.promedmail.org/direct.php?id=316957">http://www.promedmail.org/direct.php?id=316957</a>   | 191.001389 |
| 7   | <a href="http://www.promedmail.org/direct.php?id=2499421">http://www.promedmail.org/direct.php?id=2499421</a> | 171.671111 |
| 8   | <a href="http://www.promedmail.org/direct.php?id=30982">http://www.promedmail.org/direct.php?id=30982</a>     | 166.002500 |
| 9   | <a href="http://www.promedmail.org/direct.php?id=27838">http://www.promedmail.org/direct.php?id=27838</a>     | 164.002500 |
| ... | ...   | ...        |
| 55  | <a href="http://www.promedmail.org/direct.php?id=2969">http://www.promedmail.org/direct.php?id=2969</a>       | 52.000000  |
| 56  | <a href="http://www.promedmail.org/direct.php?id=429324">http://www.promedmail.org/direct.php?id=429324</a>   | 51.833333  |
| 57  | <a href="http://www.promedmail.org/direct.php?id=1553832">http://www.promedmail.org/direct.php?id=1553832</a> | 45.740833  |
| 58  | <a href="http://www.promedmail.org/direct.php?id=1142303">http://www.promedmail.org/direct.php?id=1142303</a> | 43.775000  |
| 59  | <a href="http://www.promedmail.org/direct.php?id=1251636">http://www.promedmail.org/direct.php?id=1251636</a> | 41.705556  |
| 60  | <a href="http://www.promedmail.org/direct.php?id=2746">http://www.promedmail.org/direct.php?id=2746</a>       | 28.000000  |
| 61  | <a href="http://www.promedmail.org/direct.php?id=2207525">http://www.promedmail.org/direct.php?id=2207525</a> | 27.833333  |
| 62  | <a href="http://www.promedmail.org/direct.php?id=2207527">http://www.promedmail.org/direct.php?id=2207527</a> | 27.833333  |
| 63  | <a href="http://www.promedmail.org/direct.php?id=555233">http://www.promedmail.org/direct.php?id=555233</a>   | 26.000833  |
| 64  | <a href="http://www.promedmail.org/direct.php?id=1606529">http://www.promedmail.org/direct.php?id=1606529</a> | 19.837500  |

65 rows × 2 columns

## Take-aways

- Checking for reports from the sources at the bottom of the list more frequently would improve the timeliness of the reports on ProMED.
- Many of the sources with longer median delays have relatively fewer reports published on ProMED. It could be that their articles are not posted as often as others because they aren't checked as frequently, or it could be that they aren't checked as frequently because they don't have as many articles worthy of posting. An experiment where a set of sources is checked every day could determine the degree to which it is one case or the other.
- Many of the sources with longer median delays are based in non-English speaking countries. Focusing on news sources in this category may improve ProMED's balance of reporting (see the top sources at the top) and decrease its overall reporting delay.

## **Visualization mockups**

**(<https://docs.google.com/presentation/d/1SEgH6in0YrbgymOVc8qgQQ88v-SEntXLSNMPC790EuU/edit?usp=sharing>)**