**Analysis of ProMED­mail information network**

Author: Nathan Breit (breit@ecohealthalliance.org)

## Synopsis

This is an analysis of the news sources and organizations in content posted to ProMED­mail. The key question it seeks to answer is "How can news gathering efforts be focused to improve ProMED­mail's efficiency?" We hope to provide actionable insights to ProMED contributors. We are interested in finding out whether our suggestions are valid and useful, and learning how they could be improved.

Toward this end, we identify news sources with long reporting delays, measure ProMED's historical reporting delay and compare it year by year, and present a method for determining the best sources for gathering information from specific organizations. Depending on interest from ProMED editors, we can perform a more detailed analysis. For example, there are many ways to further explore the relationships between organizations and news sources. If the data presented here are useful, we could develop interactive visualizations that allow users to drill into the results, and we could develop GRITS features to do similar types of analysis in a more generalized manner.

## General statistics about the posts

The posts used in this analysis all came from the main ProMED­mail feed. In future work, it could be extended to the regional feeds (e.g., ProMED­RUS, ProMED­FRA.).

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?

## Method

We identify sources using the ‘source’ headers above articles in a ProMED report. If the source is ‘none’ it means that we didn't find a header for the article. Some source names are resolved to generalized canonical names based on keywords they contain. For example, "USA CDC, Division of Vector­Borne Infectious Division" matches CDC because it contains "CDC". We only resolve the most prolific sources at the moment.

## Analysis

Total sources: 26435

Total sources with only 1 article in a ProMED report: 22783 Top 20 sources:

Out[23]:

|  |  |  |
| --- | --- | --- |
|  | **source** | **number of articles** |
| **0** | None | 14095 |
| **1** | WHO | 2461 |
| **2** | OIE | 2319 |
| **3** | Reuters | 1994 |
| **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** | UN | 328 |
| **13** | CIDRAP News | 328 |
| **14** | Saudi Arabia Ministry of Health | 253 |
| **15** | CNN | 230 |
| **16** | Washington Post | 186 |
| **17** | The Guardian | 164 |
| **18** | Outbreak News Today | 125 |
| **19** | The Hindu | 116 |

## Take­aways

ProMED relies on many news companies based in English speaking countries (e.g., BBC, ABC, CNN, New York Times). This may mean that news sources based in other countries are underutilized, or just that English news sources are higher quality. We're interested in ProMED's thoughts on this issue.

Most sources have only a single article used by ProMED. It may be worth periodically examining some of these to see if they continue to produce valuable content.

Web-scraping services that provide automated notifications of new content would do well to start with the sources at the top of this list.

# Most frequently mentioned organizations

## Method

We use natural language processing to identify organizations mentioned in news articles in ProMED reports. In the future, we could use similar techniques 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, "Ministry of Health" does not make it clear which Ministry of Health is mentioned. Contextual information, such as where the article was published, may help disambiguate some organizations. Countries and political constructs (e.g., European Union) also appear frequently in the organization list. It is unclear whether they should be included as organizations.

At this point, only limited name resolution is done on acronyms, so some organizations appear under multiple names (e.g. World Health Organisation and World Health Organization).

## Analysis

Total organizations: 82295

Top 20 ambiguous organizations:

|  |  |  |
| --- | --- | --- |
|  | **name** | **number of mentions** |
| **1** | Center for Disease Control | 10089 |
| **2** | Ministry of Health | 3422 |
| **5** | Food and Drug Administration | 2981 |
| **9** | Health Ministry | 1700 |
| **10** | Department of Health | 1695 |
| **11** | Ministry of Agriculture | 1604 |
| **12** | Centers for Disease Control and Prevention | 1317 |
| **13** | Health Department | 1242 |
| **17** | Centers for Disease Control | 899 |
| **30** | National Institute of Health | 479 |
| **32** | Agriculture Ministry | 472 |
| **33** | Department of Agriculture | 471 |
| **40** | Department of Natural Resources | 420 |
| **41** | The Ministry of Health | 413 |
| **59** | Department of Public Health | 259 |
| **77** | National Institutes of Health | 217 |
| **84** | Ministry of Public Health | 204 |
| **88** | Agriculture Department | 197 |

Top 20 unambiguous organizations:

|  |  |  |
| --- | --- | --- |
|  | **name** | **number of mentions** |
| **3** | World Health Organization | 3346 |
| **4** | European Union | 3099 |
| **6** | USA | 2877 |
| **7** | United States Department of Agriculture | 2725 |
| **8** | United Kingdom | 2175 |
| **14** | Office International des Epizooties | 1235 |
| **15** | Food and Agriculture Organization | 1153 |
| **16** | Health Protection Agency | 1116 |
| **18** | Reuters | 842 |
| **19** | Canadian Food Inspection Agency | 798 |
| **20** | Médecins Sans Frontières | 788 |
| **21** | UNICEF | 766 |
| **22** | European Commission | 716 |
| **23** | Federal Bureau of Investigation | 707 |
| **24** | World Health Organisation | 673 |
| **25** | Sociedade Brasileira de Virologia | 613 |
| **26** | Sierra Leone | 605 |
| **27** | Agence France Presse | 602 |
| **28** | Democratic Republic of Congo | 601 |
| **29** | Centre for Health Protection | 563 |

# Which news sources report on an organization the most?

This data intended to show which news sources report on certain organizations the most. The table below demonstrates 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/1SEgH6in0YrbgymOVc8qgQQ88v­SEntXLSNMPC790EuU/edit? usp=sharing)).](https://docs.google.com/presentation/d/1SEgH6in0YrbgymOVc8qgQQ88v-SEntXLSNMPC790EuU/edit?usp=sharing)

## Analysis

|  |  |  |  |
| --- | --- | --- | --- |
|  | **mentions** | **organization** | **news source** |
| **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 |
| **24** | 140 | World Health Organization | WHO |
| **28** | 132 | United States Department of Agriculture | OIE |
| **33** | 117 | UNICEF | WHO |
| **34** | 116 | World Health Organization | Associated Press |
| **36** | 112 | Russian Federation | WHO |
| **37** | 110 | World Health Organization | CIDRAP News |
| **40** | 105 | Berger SA | GIDEON (Global Infectious Disease & Epidemiolo... |
| **44** | 99 | World Health Organization | New York Times |
| **47** | 95 | European Union | BBC |
| **53** | 88 | National Institute of Virology | Times of India |
| **54** | 87 | United States Department of Agriculture | Associated Press |
| **62** | 80 | World Health Organization | Xinhua News Agency |
| **63** | 80 | EAEC | Eurosurveillance Edition 2013, 18(37) |
| **66** | 78 | European Union | Eurosurveillance Weekly |
| **75** | 74 | International Federation of Red Cross | WHO |

**Take­aways**

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?

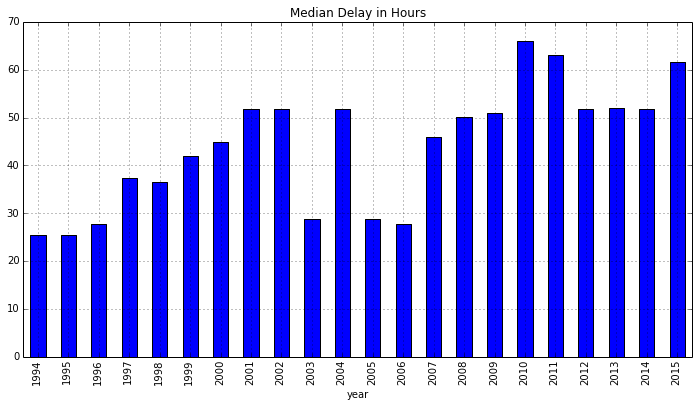
## Method

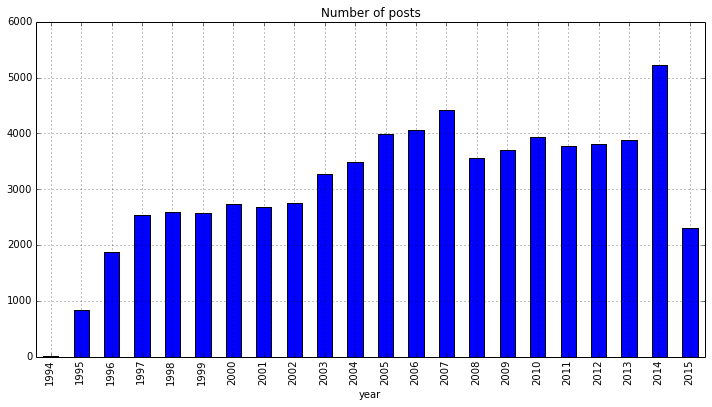
Delay is measured as the difference between the time of article publication and the time it was posted on ProMED­mail. The delays are not all exact as the article publication timestamps might only include a date and might not specify a time zone. Furthermore, the publication date of many articles is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

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.

## Analysis

Articles where a publication date was found: 68038 / 69233



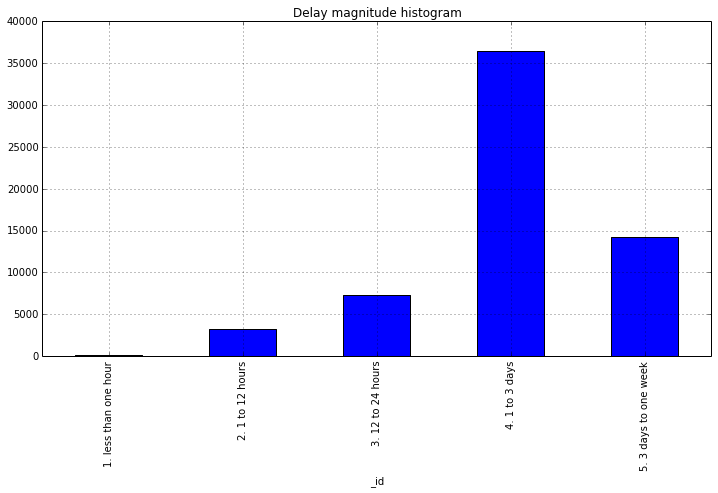


## Take aways:

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.



The chart below illustrates only sources with 50 or more articles posted on ProMED­mail.

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

44 rows × 3 columns

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

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

65 rows × 2 columns

## Take­aways

Checking for reports from the sources at the bottom of the median delay table 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 non­English 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)](https://docs.google.com/presentation/d/1SEgH6in0YrbgymOVc8qgQQ88v-SEntXLSNMPC790EuU/edit?usp=sharing)