# Diseases across the Top Five Languages of the PubMed Database: 1961-2012

**The ACM Web Science 2014 Conference Data Visualization Challenge**


Angela Zoss[1,2]
angela.zoss@duke.edu

Trevor Edelblute[2]
tedelblu@indiana.edu

Inna Kouper[2]
inkouper@indiana.edu

[1]Duke University, [2]Indiana University


This visualization focuses on diseases in the biomedical literature in English, French, German, Japanese and Russian between 1961 and 2012. We mapped and visualized the titles of the articles from MEDLINE/PubMed, a database maintained by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH), to the categories of diseases from the International Classification of Diseases (ICD-10-CM, 2014).

The initial dataset (Light, Polley and Borner, 2013; Rowe, Ambre, Burgoon, Ke, and Börner, 2009) had 21,788,173 records, covering the time period from 1809 to 2013 and containing year of publication, title, and author names. About 2 million of records in this dataset (1,701,298) were missing year information and 2,379 records contained no data due to erroneous identifiers (PMID). To fix the errors and complete the dataset, we retrieved all the data again from the PubMed API, using the NCBI eBot utility (2013) and processed it using the BioPython toolkit (2013). The resulting dataset contained the correct publication dates as well as data about the language of publication. Records with erroneous PMIDs were removed after consulting PubMed staff, who confirmed that the erroneous PMIDs were the result of publications that were ingested into PubMed but subsequently removed because they were reviews, announcements, or erratum notices.

The publications were first limited to a date range with sufficiently good coverage, 1946 to 2012, and to the top five most frequent languages of publications, namely, English, Russian, German, French, and Japanese. To match publication titles to disease categories, we created a Python classifier that matched N-grams from the ICD-10 to N-grams generated from the titles (N = [1, 2, 3]). The classifier produces several matches that are weighed based on the frequencies of N-grams associated with each category as well as on the length of the phrases matched. To improve the results, we first removed stop words using a standard list of English stop-words and then manually cleaned N-grams that generated high-frequency false positive results.

Significant scale differences between the English language publications and those in the other languages necessitated additional data processing. The data were normalized by splitting the papers by language and year and calculating the percentage of papers classified into each category. Due to sparseness of publications in languages other than English before 1961, we focused on the literature that was published between 1961 and 2012, which resulted in 4.24 million publications being matched to disease categories.

The visualization consists of both large alluvial diagrams that show the distributions of publications in the top five languages over disease categories and also small streamgraphs that show trends in publications for each disease category across the five languages. The streamgraphs use the normalized percentage publication data. Both alluvial diagrams and streamgraphs were produced with the RAW visualization suite (RAW, 2014). Additionally, the wordle.net toolkit was used to generate word clouds of the top 20 most frequent N-grams matched to each disease category (Feinberg, 2013). All visualizations components were recolored, styled, and arranged in Adobe Illustrator CS6 (2012).

## References


Adobe Illustrator CS6 (Version 16.0) [Software]. (2012). Adobe Systems Inc. Available from http://www.adobe.com/products/illustrator.html?promoid=KAUCB

BioPython (Version 1.63) [Software]. (2013). Python tools for biological computations. Available from http://biopython.org/wiki/Biopython

Feinberg, J. (2013). Wordle [Software]. Available from http://www.wordle.net/

International Classification of Diseases, Tenth Revision, Clinical Modification - ICD-10-CM. (2014). Retrieved from http://www.cdc.gov/nchs/icd/icd10cm.htm#icd2014

Light, R. P., Polley, D. E., & Börner, K. (2013). Open Data and Open Code for Big Science of Science Studies. In *Proceedings of International Society of Scientometrics and Informetrics Conference*, pp. 1342—1356. Retrieved from http://cns.iu.edu/docs/publications/2013-light-sdb-sci2-issi.pdf

NCBI eBot – Perl Script Generator to Implement an NCBI E-utility Pipeline [Software]. (2013). Retrieved from http://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi

RAW Visualization Suite [Software]. (2014). DensityDesign Lab. Available from http://raw.densitydesign.org/.

Rowe, G. L., Ambre, S. A., Burgoon, J. W., Ke, W. & Börner, K. (2009). The Scholarly Database and its Utility for Scientometrics Research. *Scientometrics, 79*(2). Retrieved from http://cns.iu.edu/docs/publications/2009-larowe-sdb.pdf.


## Acknowledgements


This work used Data Capacitor II at Indiana University, which is supported by the National Science Foundation under Grant No. CNS-0521433.