Harmonizing publisher names

Harmonizing publisher names

Last data file 08-10-2024, last updated processing 16-01-2025.

One of the most challenging aspects of data harmonization lies in publishers' information. Although it holds significant research potential, publisher data is often noisy, with a high degree of variability. Publishers can appear as individuals, organizations, or publishing houses, each with multiple name variations. To harmonize this information, we employed two methods – rule-based and vector similarity-based approaches – and kept the results as separate columns.

In the rule-based approach, we crafted over 300 regular expressions to standardize publisher names. These patterns handled a range of tasks, such as removing common suffixes (e.g., 'printing,' '& co.') and standardizing variants of prominent publishers. This application of rules successfully reduced variance in the publishers column by 8.8%, from 34,397 unique entries to 31,356.

For the second approach, we leveraged the rule-harmonized publisher names to create semantic embeddings. Using OpenAI's text-embedding-3-small model (OpenAI 2023), we generated 1000-dimensional vectors for each name form. We then clustered publishers within each location based on cosine similarity, using harmonized place names as an anchor. Operating under the assumption that many publishers, particularly those contributing to data variability, are active in a single location, we set a relatively low cosine similarity threshold (0.7) to encourage more inclusive clustering. The vector-based grouping brought together similar publisher names in each location, regardless of length or language. This approach further reduced the variance in publishers by an additional 22.9%, leaving 23,484 unique names or 682% of the original number of unique entries.

To evaluate the workflow's quality, we created a test set focused on one city in Estonia, Viljandi. The rule-based approach proved conservative but accurate, yielding 31 correct links and no errors across 516 publisher names. The vector-based method contributed an additional 128 links: 72 were clearly correct, while 56 were ambiguous or incorrect. The ambiguous links typically connected publishers with related meanings but distinct identities (e.g., societies from different professions within the same city). Table 1 shows examples of the links made in either approach. Whether these links are useful will depend on the analyst's goals; we recommend that users manually review these connections to assess their relevance.