Meta-research: International authorship and collaboration across bioRxiv preprints
RJ Abdill, EM Adamowicz, R Blekhman

Supplement: Database schema

The information used for this paper is organized into seven tables in a publicly available PostgreSQL database. It is important to note that, in contrast to our previous bioRxiv dataset, this one will not be continually updated—there were extensive manual corrections to this data, particularly the "affiliation_institutions" table, which would not be practical to maintain long-term.

- **article_authors**: Each entry represents one author of one preprint.
    - **id** (integer, primary key): A unique identifier for each entry. IDs for entries for the same article reflect the author order.
    - **name** (text): The author name, as scraped from bioRxiv.
    - **affiliation** (text): For most authors, this is an exact copy of the "original_affiliation" field described below. For authors whose country-level affiliation was corrected using their email address, this field was modified to read "EMAIL INFERENCE: {country name}" to indicate the inferred association.
    - **orcid** (text): The ORCID iD string displayed next to a single author on bioRxiv. These are formatted as web URLs. It is important to note that while these are ostensibly unique identifiers for individual researchers, there are hundreds of examples in this dataset where multiple authors on a single paper reported the same ORCID.
    - **email** (text): The email address listed for the author, as scraped from bioRxiv.
    - **article** (integer): A unique identifier for a bioRxiv preprint. These reflect the "article" ID assigned in the Rxivist dataset (https://doi.org/10.5281/zenodo.3758736).
    - **observed** (date): The datestamp on the most recent version of the specified preprint, or the date on which the specified author was "observed" to be associated with the specified institution.
    - **name_nopunc** (text): The "name" field stripped of punctuation and capitalization.
    - **name_nomi** (text): The "name_nopunc" field stripped of any middle initials using regular expressions.
    - **original_affiliation** (text): The author's institutional affiliation string, as scraped from bioRxiv. Note that while bioRxiv requests institution name and departmental information in separate fields at time of submission, these are concatenated together when displayed on the bioRxiv website. They are recorded here as presented online.
- **institutions:** Stores basic data about each institution referenced in the ROR results.
    - **id** (integer, primary key): A unique identifier for each entry. ID values have no semantic meaning.
    - **name** (text): The name of the institution, as reported by the ROR API.
    - **ror** (text): A web URL to the institution's entry on the ROR.org website.

    - **grid** (text): The unique identifier assigned to the institution in the Global Research Identifier Database (GRID), as reported by the ROR API.
    - **country** (text): The alpha-2 abbreviation of the country in which the institution is located.
- **countries:** Basic data about each country referenced in the ROR results. Each entry represents a single country.
    - **alpha2** (text, primary key): The "alpha-2" country code for the country as defined in ISO 3166-1.
    - **name** (text): The country name.
    - **continent** (text): Currently unused.
- **affiliation_institutions:** Links affiliation strings to a canonical institution.
    - **affiliation** (text, primary key): A value from the "affiliation" field of the "article_authors" table. Each unique affiliation string has an entry in this table.
    - **institution** (integer): A value from the "id" field of the "institutions" table.
- **baseline_affiliation_institutions:** Links affiliation strings to a canonical institution. Stores the same data as the *affiliation_institutions* table, but this table reflects the entries generated from the ROR data processing *before any manual corrections*.
- **article_traffic:** Stores monthly download counts for each preprint. Each entry represents traffic data for a single preprint in a single month.
    - **id** (integer, primary key): A unique identifier for the entry. This field does not have a reliable semantic meaning.
    - **article** (integer): The identifier of a single preprint from the "article" field of the "article_authors" table. One article can have many entries in this table.
    - **month** (integer): The 1-indexed identifier of the month for which the traffic data was recorded. 1=January, 2=February, and so on.
    - **year** (integer): The 4-digit year for which the traffic data was recorded.
    - **abstract** (integer): The number of views of the preprint abstract on the bioRxiv website in the specified month, as scraped from bioRxiv.
    - **pdf** (integer): The number of downloads of the preprint PDF in the specified month, as scraped from bioRxiv.
- **publications:** Links each published preprint to the journal that published it and the DOI of its published version.
    - **article** (integer, primary key): The identifier of a single preprint from the "article" field of the "article_authors" table. Each preprint has at most one record in this table.
    - **doi** (text): The DOI of the published version of the preprint, as scraped from bioRxiv.
    - **journal** (text): The title of the journal that published the preprint, as scraped from bioRxiv. Several manual corrections were made to this field to accommodate

journals that appear under multiple titles; see the “manual_edits.sql” file for all details.