Evidence of Open Access in Google Scholar: a large-scale analysis

Description of data files

Raw data files

The “raw_data” folder contains CSV files that were used to compute OA levels:

- **oa_gs_articles.csv**: 2,269,022 rows, 4 columns. Each row provides information about one of the documents selected for analysis. The columns contain the following information:
    - doi (string). DOI (Digital Object Identifier) of a document.
    - doc_type (string). Type of document. Possible values: “Article”, “Review”.
    - journal (string). Title of the journal where the document was published.
    - pub_year (integer). Year of publication of the document. Possible values: 2009, 2014.

- **oa_gs_fa_urls.csv**: 3,549,300 rows, 7 columns. Each rows provides information about a link to a freely available (FA) link to a document from **oa_gs_articles.csv**, as found by Google Scholar. The columns contain the following information:
    - doi (string). DOI (Digital Object Identifier) of a document in **oa_gs_articles.csv**.
    - fa_url (string). URL pointing to a freely available version of the document, as found in Google Scholar.
    - host (string). The fa_url was parsed to extract the domain in which the document was hosted.
    - prim_fa_url (logic). TRUE if fa_url was the link that Google Scholar displayed as the primary FA URL. FALSE if this fa_url was a secondary full text version.
    - host_type (string). Category of host. Possible values:
        - social_network
        - publisher
        - repository
        - institution
        - unknown
        - harvester
        - other
    - oa_type (string). Type of Open Access. Possible values:
        - freely_available
        - green
        - bronze
        - gold
        - delayed
        - hybrid

    - only source (logic). TRUE if fa_url was the only FA version of the article that Google Scholar could find. FALSE otherwise.

- **oa_gs_gold_journals.csv**: 1151 rows, 1 column. Each row is the title of a journal included in DOAJ (Directory of Open Access Journals). The columns contain the following information:
    - gold_oa_journal (string). Title of the journal, as recorded by the Web of Science.

- **oa_gs_host_types.csv**: 116,295 rows, 2 columns. Each row contains a website address, and its categorisation.
    - host (string): a website address.
    - host_type (string): Category of host. Possible values:
        - social_network
        - publisher
        - repository
        - institution
        - unknown
        - harvester
        - other

- **oa_gs_licenses.csv**: 2,328,276 rows, 5 columns. Each row contains information on the license of a document in **oa_gs_articles.csv**. The columns contain the following information:
    - doi (string). DOI (Digital Object Identifier) of a document in **oa_gs_articles.csv**.
    - license (string). A declaration of a license, or the string “no license found”, if it was not possible to find the license of the document.
    - delay (integer). Number of days since the date of publication of the article for the license to come into effect, as recorded in CrossRef.
    - license_provenance (string): origin of the license data. Possible values: “crossref”, “publisher meta tags”.
    - open_access_license (logic): TRUE if license is an Open Access License. FALSE otherwise.

- **os_gs_wos_sc_mapping.csv**: 252 rows, 4 columns. Each row displays the correspondence between one of the 252 Web of Science subject category, and the three levels (high, middle, and low) of the NOWT classification scheme.
    - wos_sc (string). Name of one of the 252 Web of Science subject categories.
    - low_nowt (string). Name of one of the 35 low-level subject categories of the NOWT classification scheme.
    - middle_nowt (string). Name of one of the 14 middle-level subject categories of the NOWT classification scheme.
    - high_nowt (string). Name of one of the 7 high-level subject categories of the NOWT classification scheme.

Results

The “results” folder contains the summary tables that display the specific percentages of Open Access and Free Availability of scientific publications, aggregated at various levels, as well as the distribution of links by host:

- **oa_gs_summary_countries.csv**: 237 rows, 16 columns. Open Access (OA) and Free Availability (FA) levels aggregated at the level of country of affiliation of the authors of the documents.

- **oa_gs_summary_low_nowt.csv**: 35 rows, 16 columns. OA and FA levels aggregated at the level of high-level categories of the NOWT classification scheme (those found in **os_gs_wos_sc_mapping.csv**).

- **oa_gs_summary_high_level.csv**: 7 rows, 16 columns. OA and FA levels aggregated at the level of high-level categories of the NOWT classification scheme (those found in **os_gs_wos_sc_mapping.csv**).

- **oa_gs_summary_journals.csv**: 11,236 rows, 16 columns. OA and FA levels aggregated at the level of journals.

- **oa_gs_summary_journals_countries.csv**: 302,790 rows, 17 columns. OA and FA levels aggregated at the level of journals and countries of affiliation.

- **oa_gs_summary_journals_pubyear.csv**: 19,238 rows, 17 columns. OA and FA levels aggregated at the level of journals and publication years.

- **oa_gs_summary_journals_pubyear_countries.csv**: 426,528 rows, 18 columns. OA and FA levels aggregated at the level of journals, publication years, and countries of affiliation.

- **oa_gs_summary_journals_pubyear_wos_sc.csv**: 31,837 rows, 18 columns. OA and FA levels aggregated at the level of journals, publication year, and WoS subject categories.

- **oa_gs_summary_journals_pubyear_wos_sc_countries.csv**: 735,901 rows, 19 columns. OA and FA levels aggregated at the level of journals, publication year, WoS subject categories, and countries of affiliation.

- **oa_gs_summary_journals_wos_sc.csv**: 18,494 rows, 17 columns. OA and FA levels aggregated at the level of journals, and WoS subject categories.

- **oa_gs_summary_journals_wos_sc_countries.csv**: 519,173 rows, 18 columns. OA and FA levels aggregated at the level of journals, WoS subject categories, and countries of affiliation.

- **oa_gs_summary_pubyear.csv**: 2 rows, 16 columns. OA and FA levels aggregated at the level of publication years.

- **oa_gs_summary_pubyear_countries.csv**: 436 rows, 17 columns. OA and FA levels aggregated at the level of publication years and countries of affiliation.

- **oa_gs_summary_pubyear_wos_sc**: 504 rows, 17 columns. OA and FA levels aggregated at the level of publication years, and WoS subject categories.

- **oa_gs_summary_pubyear_wos_sc_countries**: 47,191 rows, 18 columns. OA and FA levels aggregated at the level of publication years, WoS subject categories, and countries of affiliation.

- **oa_gs_summary_wos_sc.csv**: 252 rows, 16 columns. OA and FA levels aggregated at the level of WoS subject categories.

- **oa_gs_summary_wos_sc_countries.csv**: 27,485 rows, 17 columns. OA and FA levels aggregated at the level of countries of affiliation.

- **oa_gs_host_distr.csv**: 116,271 rows, 6 columns. Number of FA links provided by each host. Columns:
    - host (string). Address of the host.
    - host_type (string): Category of host. Possible values:
        - social_network
        - publisher
        - repository
        - institution
        - unknown
        - harvester
        - other
    - n_documents (integer). Number of FA documents provided by the host.
    - perc_only_source (numeric). Percentage of documents (relative to n_documents) for which host is the only FA source (the document is not freely available from any other source).
    - prim_version (integer). Number of times that FA documents in this host are selected as the primary free full text source by Google Scholar.
    - perc_prim_version (numeric). Percentage of prim_version, relative to n_documents (prim_version / n_documents * 100).

All the oa_gs_summary_* data files contain, apart from the appropriate grouping columns (journal, pub_year, wos_category, country), the following columns:

- n_documents (integer). Number of documents in the group.
- oa_fa_all (numeric). Percentage of overall availability considering all sources, relative to n_documents.
- publisher_all (numeric). Percentage of Open Access provided by the publisher (combination of Gold, Hybrid, Delayed, and Bronze), relative to n_documents.
- gold (numeric). Percentage of documents in the group published in Gold OA journals, relative to n_documents.

- hybrid (numeric). Percentage of documents in the group published as Hybrid OA, relative to n_documents.
- delayed (numeric). Percentage of documents in the group published as Delayed OA, relative to n_documents.
- bronze (numeric). Percentage of documents in the group published as Bronze OA, relative to n_documents.
- green (numeric). Percentage of documents in the group that are freely accessible from repositories, relative to n_documents.
- green_only (numeric). Percentage of documents in the group that are freely accessible from repositories, excluding those that are also freely accessible from the publisher, relative to n_documents.
- fa_all (numeric). Percentage of documents in the group that are freely available from sources other than the publisher or repositories (combination of institution, social_network, harvester, and other), relative to n_documents.
- fa_only (numeric). Percentage of documents in the group that are freely available from sources other than the publisher or repositories (combination of institution, social_network, harvester, and other), excluding those that are also available from the publisher or from repositories. Relative to n_documents.
- institution (numeric). Percentage of documents in the group that are freely available from the website of an academic institution (excluding institutional repositories), relative to n_documents.
- social_network (numeric). Percentage of documents in the group that are freely available from academic social networks (ResearchGate or Academia.edu), relative to n_documents.
- harvester (numeric). Percentage of documents in the group that are freely available from harvester websites (CiteSeerX, Semantic Scholar, CORE), relative to n_documents.
- other (numeric). Percentage of documents in the group that are freely available from sources other than the ones described above, relative to n_documents.