Documentation of the data set in this repository

Till Grallert

[1] Humboldt-Universität zu Berlin
[2] NFDI4Memory

2026-11-03

This data set comprises bibliographic metadata and summary statistics for 334127 articles by 61885 distinct authors published in 17602 issues of 309 Arabic-speaking journals between 1870 and 2023.

The source of this data set is the website ارشيف الشارخ للمجلات الادبية والثقافية العربية. For more information on the background and purpose of this data set see this repository's read me.

For more insights into the data set see the section characteristics of the data set below.

main file / single source of truth

The main file for metadata on the article level is an R dataframe sakhrit.rda. Most other files can be derived from this main file using sakhrit_generate-subsets.R. The association between entities in this data set and Wikidata as our main external authority file are kept in sakhrit_journals.csv and sakhrit_authors.csv, which are never computationally over-written by any scripts in this repo. Metadata on the journal level is not kept in this file as this bloats file size with redundant information for each of the more than 1000 articles per journal on average. For convenience, we have, however, provided this bloated file as sakhrit_full.rda.

The columns in the dataframe sakhrit.rda are

- journal.id: numerical ID for a periodical as provided by ارشيف الشارخ and prefixed with j.
- issue.id: numerical ID for an issue as provided by ارشيف الشارخ and prefixed with i.
- issue.no: issue number or nummerical range. extracted from the information provided by ارشيف الشارخ.
- date.publication.iso: ISO data, computationally converted from the Arabic text strings provided by ارشيف الشارخ.
- article.id: numerical ID for an article as provided by ارشيف الشارخ and prefixed with a.
- article.title: article title. Arabic string.
- author.id: our own numerical ID for a person, prefixed with p.
- author.name: single author name. Arabic string.
- page.from: starting page of an article. Integer.

Leading and trailing whitespaces have been removed from all strings. All other whitespace has been collapsed.

IDs

ارشيف الشارخ uses numerical IDs for all entities. This creates duplicate IDs for different types of entities. To avoid such confusion and ease computational analysis, we have prefixed the IDs in our data set (see the subsections on IDs for each type of entity below).

journals

sakhrit_journals-summary.csv provides basic bibliographic metadata for every journal as scraped from the website.

- journal.id: the ID of the periodical as provided by the website.
- journal.title: the title of the periodical as provided by the website. Obvious mistakes have been quietly corrected (see below)
- editor: name of the editor as provided by the website
- place: place of publication as provided by the website

To this, we computationally added some basic descriptive statistics:

- onset.collection: the date of the first issue scraped from the website
- terminus.collection: the date of the last issue scraped from the website
- issues: the number of issues scraped from the website
- articles: the number of articles scraped from the website
- authors: the number of individual authors published in this journal

We also semi-automatically added some qualitative information. This is maintained in and pulled from sakhrit_journals.csv

- id.wd: Wikidata QIDs. All journals were reconciled with Wikidata and we contributed missing information to Wikidata.
- colour: a hex colour code, which can be used for consistent plotting.
- place.id.wd: Wikidata QID of the publication place. This was provided through automated reconciliation with authority files.

IDs

ارشيف الشارخ provides a nummerical ID for journals (JID, prefix j), which can be resolved with https://archive.alsharekh.org/magazineYears/JID

data quality

Some of the publication data provided by the website was evidently false. For titles already found in authority files, such as Wikidata, this can be easily corrected by pulling publication data from their APIs.

- Some titles were not exactly as printed on the magazines' title pages and have been corrected.
- Some publication places are wrong, namely for Jacob Sanua's publications, which sharikh records as published in Egypt, when they were actually published in Paris

issues

sakhrit_issues-summary.csv provides basic bibliographic metadata for every issues as scraped from the website.

- issue.id: the ID of the periodical issue as provided by the website.
- issue.no: numerical information provided by the website but converted from the Arabic strings. It seems that they have kept the original issue numbers as evidenced by the restart of the issue count in the middle of any given (Gregorian) year.
- journal.id: ID linking this issue to a journal in our data set
- date.publication.iso: the website provides Egyptian Arabic strings according to the Gregorian calendar. These have been converted to ISO notation with regex. A few issues carried obviously faulty publication years. These have been corrected as documented in sakhrit_clean-data.R.

In addition, we provide some computed information

- articles: the number of articles in this issue scraped from the website
- pages.min: the minimum number of pages in this issue. This is computed as the difference between the smallest and the largest starting page number of the articles in this issue plus one.

IDs

ارشيف الشارخ provides a nummerical ID for issues (IID, prefix i), which can be resolved with https://archive.alsharekh.org/contents/JID/IID. These sites provide bibliographic metadata for an issue (number, publication date), and a table of contents with links to all articles in this issue. Note that correct metadata depends on the correct combination of JID and IID.

data quality: publication dates

ارشيف الشارخ provides publication dates according to the Gregorian calender in Egyptian Arabic. As such, these dates will most likely not match the information provided by the front inspices / mastheads particularly of magazines published before the mid-20th century and should thus be treated very causiously. Take, for example, المقتبس (al-Muqtabas, Q12193626). ارشيف الشارخ assumes a publication schedule in which volumes correspond to Gregorian years and issues to Gregorian months (i.e. according to this data set (1)1 المقتبس was published on 1 January 1906). This is despite the fact that المقتبس clearly stated its publication schedule on the front page of every volume as adhering to the Muslim hijrī calendar for both volumes and issues, as one can easily see on ارشيف الشارخ facsimile of that issue's first page (on the issue of dating periodicals see Grallert 2021, 66–69). }

articles

sakhrit_articles.csv provides bibliographic metadata on the article level based on the tables of contents for each issue scraped from the website. To some extent, this is the core of the data set.

- article.id: the ID of an article as provided by the website
- issue.id: the ID of the issue, the article was published in, linking to more information in sakhrit_issues-summary.csv.
- author.id: the ID of the article's author, linking to more information in sakhrit_people-summary.csv.

- article.title: the article title as scraped from the website. Minor corrections for orphaned quotation marks and commata were silently made.
- page.from: the page number of an article's innitial page. Information on the length of an article is not provided but the starting page of the next article could be a proxy.

IDs

ارشيف الشارخ provides a nummerical ID for articles (AID, prefix a), which can be resolved with https://archive.alsharekh.org/Articles/JID/IID/AID. The resulting page will provide a link to the facsimile of the article, no matter which JID and IID are used. The correct bibliographic metadata is only provided for the correct combination of JID, IID, and AID.

data quality

The quality of the data is currently unkown beyond some cursory checks:

- missing data: absolutely unkown. The practices of recording the tables of contents is not generally documented and might have differed between journals and over time.
  - rubrics:
    - I suspect with relatively high confidence that shorter articles in common rubrics were not recorded.
    - it appears that rubric titles were generally not recorded, as we do not have large clusters of articles, whose titles commence with the same words.
    - take for example a246792. The facsimile clearly shows the rubric "صدر حديثا" or recently published. But this is not recorded in the article's metadata.
- authors:
  - not a single article carries more than one author name. First, it is unlikely to not a single article in this huge data set was originally attributed to more than one person. Second, the practice of listing translators vs. original authors for pieces translated into Arabic is unclear. Given the large number of non-Arab authors in the data set, it seems that, at least in later years, solely the original was recorded.

One journal, آمال (*Āmāl*, Q141431781), was recorded twice and the original issue and article IDs have been deprecated between the 2018 and 2026 scrapes. It would be worthwhile to check the differences between those two sets to learn something about changing cataloguing practices.

means to test data quality

The first step to systematically test the data quality, is to compare the current data set to other data sets with overlapping coverage. The first choice for such a comparison would be my digital scholarly editions of المقتبس (Grallert 2020) and لغة العرب (*Lughat al-ʿArab*, Q12237314) (Grallert 2022).

people

sakhrit_authors-summary.csv provides information based on the tables of contents of every issue. Note that, while the majority of rows refer to people, quite few provide other information on an article's source, such as "… عن مجلة"

- author.id: a unique ID prefixed p and assigned by sakhrit_process-data.R to account for author strings without an ID on the website
- person.id: the numerical ID of a name string as scraped from the index of authors on the website. Note that many rows do not have an official ID from the website (see section on data quality below).
- id.wd: the Wikidata QID of a person. The association between people and Wikidata are maintained in sakhrit_people-wikidata.csv.
- author.name: string as found on the website.

In addition we computed some basic descriptive statistics:

- journals: the number of periodicals in our data set, an author published in.
- issues: the number of periodical issues in our data set, an author published in.
- articles: the number of articles in our data set published by that author
- onset: earliest publication date of any article by this author in our data set.
- terminus: latest publication date of any article by this author in our data set. These dates can be used to identify authors as they provide some indication of life dates.

IDs

ارشيف الشارخ provides a nummerical ID for authors (PID), which can be resolved with https://archive.alsharekh.org/AuthorArticles/PID. These sites provide a list of articles for each author. However, the PID is not used in any other context.

Note that explorative analysis of the dataset revealed that **not all authors have their own author page** and thus an ID! This is particularly true for foreign authors published in Arabic translation. Also note that the thousands of anonymous articles cannot possibly be linked to from an author page. We have therefore **created our own ID for authors**, which uses the prefix p. This should probably be changed in a release.

In addition, some people with identical name strings have two author IDs, i.e. طه السويفي, is recorded as both PID:16045 and PID:31623.

data quality

Sharikh conducted some basic normalisation in recording personal names. Thus, almost all titles have been removed (only a handful of names carry the Ottoman honorary addresses رفعتلو, عزتلو etc.) and a substantial number of authors have received their own author ID (but not all by a large margin).

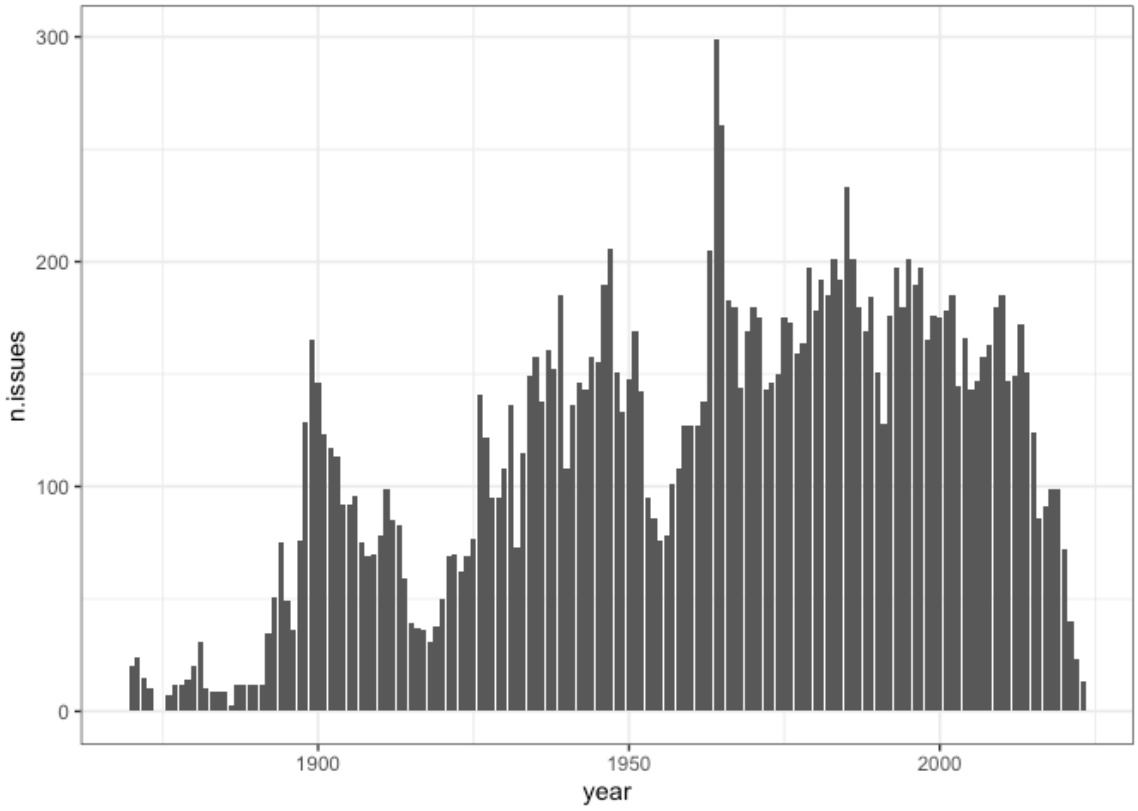
Spelling and name forms have not been normalised. Thus, **non-Arab authors** are frequently recorded with multiple spellings (consider, for instance, آرت بوتشوالد, آرت بوكوالد, and آرت بونشفالد). Different regional preferences for final ي and ة have not been normalised and names ending in these letters are frequently duplicated across the data set, e.g. البغدادى-البغدادي, أناتولى-أناتولي.

Arab names are not as stable for the earlier period in this data set as they are later on. Thus, Jamal Sidqi al-Zahawi (Q4115227) is recorded as جميل صدقي الزهاوي, زهاوي زادة جميل صدقي, جميل زهاوي, and جميل الزهاوي.

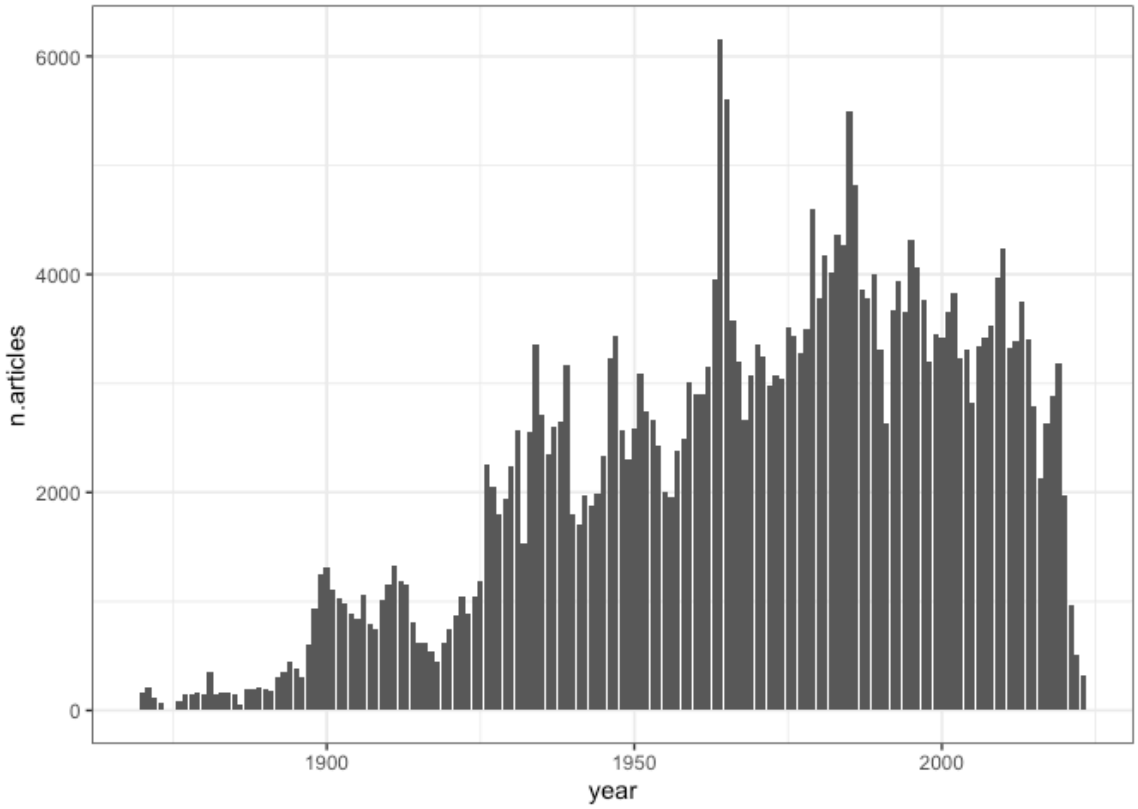
characteristics of the data set

temporal distribution of the data set

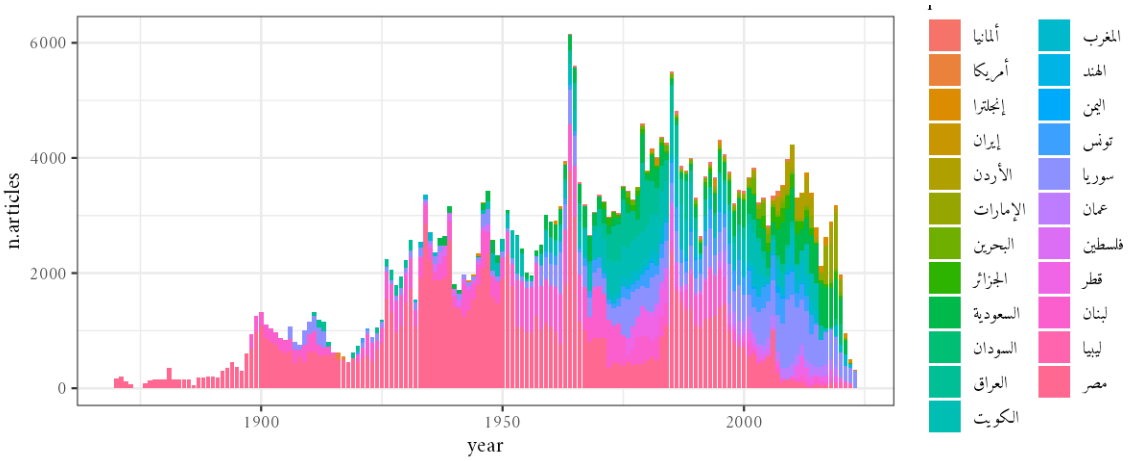
```
# set a general theme for all ggplots
theme_set(theme_bw())
# a bit of pre-processing
df.issues <- df.issues %>%
  dplyr::mutate(year = year(date.publication.iso))
# plot temporal distribution of issues
df.issues %>%
  dplyr::group_by(year) %>%
  summarise(n.issues = n()) %>%
  ggplot() +
   geom_col(aes(x = year, y = n.issues))
```




```
# plot temporal distribution of articles
df.issues %>%
  dplyr::group_by(year) %>%
  summarise(n.articles = sum(articles)) %>%
  ggplot() +
    geom_col(aes(x = year, y = n.articles))
```




```
knitr::opts_chunk$set(dev = "ragg_png")
# articles and publication place
df.data <- df.issues %>%
  dplyr::left_join(df.journals.summary, by = join_by(journal.id)) %>%
  dplyr::rename(articles = articles.x) %>%
  dplyr::select(year, place, articles) %>%
  dplyr::group_by(place, year) %>%
  dplyr::summarise(n.articles = sum(articles), .groups = "drop_last") %>%
  dplyr::ungroup()
plot.data <- df.data %>%
  ggplot(aes(x = year, y = n.articles, fill = place)) +
    geom_col() +
    theme(
      text = element_text(family = "Amiri", face = "plain")
    )
# convoluted way of plotting in order to correctly display Arabic script
file <- knitr::fig_path('.png')
agg_png(file, width = 1200, res = 150)
plot.data
invisible(dev.off())
knitr::include_graphics(file)
```



Grallert, Till. 2020. "An open, collaborative, and scholarly digital edition of Muḥammad Kurd ʿAlī's monthly journal 'al-Muqtabas' (Cairo and Damascus, 1906-1917/18)." https://doi.org/10.5281/zenodo.597319.

———. 2021. "Catch Me If You Can! Approaching the Arabic Press of the Late Ottoman Eastern Mediterranean Through Digital History." Edited by Simone Lässig. *Geschichte Und Gesellschaft* 47 (June): 58–89. https://doi.org/10.13109/gege.2021.47.1.58.

———. 2022. "An open, collaborative, and scholarly digital edition of Anastās Mārī al-Karmalī's monthly journal "Lughat al-ʿArab" (Baghdad, 1911–14)." https://doi.org/10.5281/zenodo.3514384.