read me: data_sakhrit

Till Grallert[1,2,]

2026-11-10

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

Correspondence: Till Grallert <till.grallert@hu-berlin.de>

This repository contains machine-actionable 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 The data set is based on the website ارشيف الشارخ للمجلات الادبية والثقافية العربية.(صخر n.d. provides outdated numbers) Prior to 2019, this website was available at https://archive.sakhrit.co/.

about this repository

Release v0.1 contains the full data set as present on ارشيف الشارخ in autumn 2026. v1.0 will require some data cleaning and warant more and better documentation. However, I consider v0.1 to be fully usable for reliable and significant computational research and exploring the full data set for close reading.

folder structure

```
├── data            # folder for the main data set as CSV, RDA, and TEI/XML
│   ├── derivatives  # variants of the main data set in various serialisations, such as BibTeX or MODS/XML
│   ├── subsets      # computationally generated subsets of the main data set
│   │   ├── journals  # one file per journal named after its QID (Wikidata)
│   │   └── topics    # by topic
│   └── v1           # original data set from 2018
├── docs            # documentation for this repository
└── src             # all scripts (R, XSLT, bash) for scraping and processing the data
    └── deprecated
```

File names are based on the old name of the data source (“sakhrit”).

documentation

Detailled documentation for this repository and the data set can be found in the docs/ folder.

- data.md provides detailed documentation of the data set as the main delivery of this repository. Please note the comments on the data quality.
- workflow.md provides a description of the workflow, processes, and decisions for scraping and processing the data

data models, serialisations

The main serialisation in this data set is CSV. We provide BibTeX serilisation for import into Zotero or similar reference managers. The journal metadata is also serialised as TEI/XML as we had pipelines for transforming TEI/XMl into QuickStatements for contributing data to Wikidata.

meta information

background

Periodicals are compound texts and there are, to my knowledge, no concerted efforts to provide searchable metadata for Arabic periodicals below the bound copy, such as issues, rubrics, and articles. The only exceptions are a few editions of individual periodicals, back-issues of newspapers in their publishers' archives, and ارشيف الشارخ للمجلات الادبية والثقافية العربية. This website provides metadata including titles, authors, issue number, and publication date for more than 300.000 articles in 309 cultural and literary magazines published between 1870 and 2023, with the vast majority dating to the second half of the twentieth century. ارشيف الشارخ, thus, enables an unprecedented search for authors and their articles across multiple journals.

ارشيف الشارخ also provides digital images of all magazines in its collections. However, it must be noted that the proclaimed scans of original copies cannot be relied upon without additional checks. They have rendered at least one entire volume of المقتبس (*al-Muqtabas*, Q12193626) based on the defective digital text from المكتبة الشاملة in a original-looking layout and serve them as "fakesimiles" (Grallert 2022, sec. 14). Other facsimiles are of such a low resolution to render them absolutely illegible.

ارشيف الشارخ was set up by and, with the move to the new URL, finally named after Muhammad al-Sharikh (Q6892907), a Kuwaiti businessman and founder of the Sakhrit software company that developed the first Arabic OCR software صخر. Unlike what many websites make you to believe (c.f. أرشيف الشارخ" للمجلات الأدبية والثقافية العربية في حلة جديدة" 2021"; Charafeddine 2020; الكويتية 2020), an earlier iteration of ارشيف المجلات الادبية والثقافية العربية (note the absence of الشارخ from the name) was online at archive.sakhrit.co between 2016 and 2019 (Grallert 2016; Salamon 2016). The earlierst snapshot of a beta version on the Internet Archive is dated to January 2015 and already displays the صخر logo in its head. During the year in beta, the number of magazines increased from 122 to 187.

purpose of this data set

Unfortunately, ارشيف الشارخ is a silo disconnected from the larger knowledge ecologies. The website neither makes the bibliographic data available beyond their graphical web interface, nor do they link to external authority files.

The current data set aims at mitigating both points. The main purpose is to make the bibliographic metadata collected by ارشيف الشارخ available in open and well-established serialisations for **computational analysis** and **integration into bibliographic information systems**. To this end, we provide the entire data set and various subsets as CSV, BibTeX, and TEI/XML. Other serialisations can be easily generated.

The initial goal in creating this data set was the diachronical analysis of co-citational networks of authors who published in the same periodical issue.

Large as it might be, the ارشيف الشارخ data set covers only a sliver of the Arabic periodical press. In order to integrate these data into large information ecologies and the semantic web, we aim at reconciling as many entities as possible with existing authority files and knowledge graphs, most notably Wikidata. This work has been completed on the journal level. Here we have reconciled all publications with Wikidata items created hundreds of new items for periodicals, which had not yet been found there. Reconciling all of the 61885 distinct authors in this data set is beyond the scope and capability of the current endeavour. I have, however, started reconciling the most important authors (determined by *betweenness centrality* and *degree* in the aforementioned social network analysis) for the period between 1870 and 1930.

responsibilities

All responsibility for scraping and processing the data rests with the authors mentioned in this file's YAML block.

The original data were and are continuously gathered by the website ارشيف الشارخ, which does not provide information on their data collection processes.

licence

It is our conviction that bibliographic metadata cannot be copyrighted. We conceive of all data in the `data/` folder as being in the public domain (CC0).

The documentation is published as CC BY 4.0 and code is covered by the MIT license.

references

Charafeddine, Fatme. 2020. "New OA Resource: Al-Sharekh Archive." Access to Mideast; Islamic Resources (AMIR). November 2, 2020. https://amirmideast.blogspot.com/2020/11/new-oa-resource-al-sharekh-archive.html.

Grallert, Till. 2016. "Archive.sakhrit.co's Failure as a Source for Digitsed Imagery of Arabic Journals." Open Arabic Periodical Editions. April 22, 2016. https://openarabicpe.github.io/2016-04-22-review-archive-sakhrit/.

———. 2022. "Open Arabic Periodical Editions: A Framework for Bootstrapped Scholarly Editions Outside the Global North." Edited by Roopika Risam and Alex Gil. *Digital Humanities Quarterly* 16 (June). https://doi.org/10.63744/w9dk2dm7m92p.

Salamon, Anaïs. 2016. "أرشيف المجلات الأدبية والثقافية العربية." Aldébaran - Lire et repérer les sciences humaines sur Internet. June 21, 2016. https://journals.openedition.org/aldebaran/7233.

"""2021 ". أرشيف الشارخ" للمجلات الأدبية والثقافية العربية في حلة جديدة. MEO. May 25, 2021. https://middle-east-online.com/%D8%A3%D8%B1%D8%B4%D9%8A%D9%81-%D8%A7%D9%84%D8%B4%D8%A7%D8%B1%D8%AE-%D9%84%D9%84%D9%85%D8%AC%D9%84%D8%A7%D8%AA-%D8%A7%D9%84%D8%A3%D8%AF%D8%A8%D9%8A%

D8%A9-%D9%88%D8%A7%D9%84%D8%AB%D9%82%D8%A7%D9%81%D9%8A%D8%A9-%D8%A7%D9%84%D8%B9%D8%B1%D8%A8%D9%8A%D8%A9-%D9%81%D9%8A-%D8%AD%D9%84%D8%A9-%D8%AC%D8%AF%D9%8A%D8%AF%D8%A9.

الكويتية, جريدة الجريدة. 2020. "حصاد : الشارخ يطلق نسخة جديدة من «أرشيف المجلات الثقافية والأدبية»." جريدة الجريدة الكويتية. January 13, 2020. https://www.aljarida.com/articles/1578838864459493000.

صخر, محمد الشارخ-. n.d. "حول أرشيف الشارخ." أرشيف الشارخ للمجلات الأدبية والثقافية العربية. Accessed November 10, 2026. https://archive.alsharekh.org/About.