Documentation of the workflow for scraping and processin the data

Till Grallert[1,2,]

2026-11-03

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

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

Process, workflow

1. Scrape the entire website with wget. Done in autumn of 2018
2. Scrape all journal landing pages with Curl script and extract the JSON-LD with regex
   - convert the JSON-LD to <tei:bibl> nodes with regex
3. Convert and process the scraped data in R. This is done with code from https://github.com/OpenArabicPE/convert_sakhrit-to-tei
   - processing
     - extract structured information from the HTML tables
     - group by journal
     - filter by publication date
   - Save data frames to file
     - as CSV and RDA: full data set
     - as generic XML: subset data set for 1870–1920
4. Convert and process the XML with XSLT. This is done with code from https://github.com/OpenArabicPE/convert_sakhrit-to-tei
   - Convert generic XML to TEI/XML
     - Data cleaning
     - ☒ Entity linking (journals)
   - Convert TEI/XML to CSV
5. scrape missing issues with R. Done in 2026.
   - including data extraction etc.
   - ☒ 1–21200 done by 2026-09-27
   - ☒ 21200-22100: 21206 is missing!
     - ☒ 21200-21700
     - ☒ 21700-22000, 2026-10-04, 2026-10-06
   - ☒ 22100-23000: 2026-10-12, six issues are missing
   - ☒ 23000-23300, 2026-10-04
   - ☒ 23300-23500:
   - ☒ 23500-24000: comprising many of the magazines published before 1930

- ☒ 24000-24500: 2026-10-15
- ☒ 24500–: it seems that no issues exist in this range
- ☒ there are no issues for journal 397. Manually check the website for the first issue ID: 24308
- ☐ some IDs in a sequence of issue IDs for a journal are missing. This can be checked against issue numbers. If there is a corresponding gap in issue numbers, the issue most likely was wrongfully recorded as non existing in sakhrit_issues-status.csv

6. Upload data for magazines to Wikidata
   - ☒ Convert TEI/XML to QuickStatements
   - ☒ Upload to Wikidata with QuickStatements 3.0
     - Batches: 53407, 53407, 53429

scrape

with the relaunch of the website at https://archive.alsharekh.org/, most, if not all HTML pages also incapsulate JSON-LD, which helps immensely with scraping the metadata.

journal / magazine pages

Although journal IDs appear to be sequential, there are quite a few, which do not lead to a magazine. But I noticed in 2026 that I had not indeed scraped all journals in 2018, when I (or rather the machine) stoped at 278. Testing IDs 1–500, it appears that in 2026 the highest assigned ID is 397.

The magazines' main landing pages contain some JSON-LD:

```
<script type="application/ld+json">
{
"@context": "https://schema.org",
"@type": "Periodical",
"name": "مجلة عالم الفكر ",
"issn" : "34",
"description": "مجلة عالم الفكر  رئيس التحرير أحمد مشاري العدواني تاريخ الإصدار 1970 نوعية الإصدار فصلية بلد الإصدار الكويت",
"publisher": {
  "@type": "Organization",
  "name": "عالم الفكر "
},
"startDate": "1970","endDate": "2010",
  "inLanguage": "Arabic",
  "url": "https://archive.alsharekh.org/magazineYears/34"
}
</script>
```

This does not include the name of the editor-in-chief or the publication place as structured data. Instead, the description is rather formulaic and one could assume everything between رئيس التحرير and تاريخ الإصدار to be the name of the editor-in-chief and the remainder of the string after بلد الإصدار to the country of publication.

I scraped these pages with CURL, extracted the JSON-LD and converted it to TEI/XML with regex. The resulting file is sakhrit_journals.TEIP5.xml.

- Regex

- to find all JSON-LD: <script type="application/ld\+json">(\s+.+$)+?\s+</script>
- to mark-up content of description
  - (مجلة\s+)(.+?)(\s+رئيس التحرير\s+)(.+?)(\s+تاريخ الإصدار\s+.+?\s+نوعية الإصدار\s+.+?\s+بلد الإصدار\s+)(.+?)"
  - $1<title level="j">$2</title>$3<editor> <persName>$4</persName> </editor>$5<pubPlace> <placeName>$6</placeName> </pubPlace>"
- dates
  - ("startDate": ")(\d{4})
    - $1<date type="onset" when="$2">$2</date>
  - (\s+إلى\s+)(\d{4})
    - $1<date type="terminus" when="$2">$2</date>
  - ("endDate": ")(\d{4})
    - $1<date type="terminus" when="$2">$2</date>
- IDs
  - ("issn": ")(\d+)
  - $1<idno type="jid">$2</idno>
- language
  - ("inLanguage": ")(Arabic)
  - $1<textLang mainLang="ar">$2</textLang>

issue pages (https://archive.alsharekh.org/contents/JID/IID)

It is important to note from the outset that, unfortunately, there is no way of guessing the issue IDs for any particular journal. Journal pages link to annual subpages through some Javascript and only from there one will get the actual issue page. In addition, every existing issue ID will always provide the table of contents for this issue, no matter which journal ID it is combined with. However, only the correct combination (1 in 300 journals) will also provide the necessary metadata for the issue, including the journal it belongs too. This means that in the worst case, one has to test every possible combination of issue and journal IDs.

With the correct combination of journal and issue ID, metadata on the issue level is provided in the <meta> tag and as JSON-LD

```
<meta name="og:description" property="og:description" content="مجلة الرؤيا - العدد 3, رئيس التحرير أحمد حمدي , تاريخ الإصدار 1 يناير 1983 , نوعية الإصدار فصلية , بلد الإصدار الجزائر">

<script type="application/ld+json">
{
 "@context": "https://schema.org",
 "@type": "Periodical",
 "name": "مجلة الرؤيا - العدد 3",
 "issn" : "146 - 14864",
 "description": "مجلة الرؤيا - العدد 3, رئيس التحرير أحمد حمدي , تاريخ الإصدار 1 يناير 1983 , نوعية الإصدار فصلية , بلد الإصدار الجزائر",
 "publisher": {
  "@type": "Organization",
  "name": "الرؤيا"
 },
 "startDate": "1 يناير 1983",
 "inLanguage": "Arabic",
 "url": "https://archive.alsharekh.org/contents/146/14864"
```

```
}
</script>
```

- metadata on the journal
  - title:
    - in description
    - in publisher/name
  - issue number: description
  - editor: description
  - place: description
  - language: inLanguage
  - id: issn
- metadata on the issue:
  - date:
    - in description
    - in startDate
  - url: url
  - id: issn

The page then contains a HTML table of contents with sufficient information for every article. The article pages then do not provide any additional information beyond the link to the facsimiles!

```
<tr _ngcontent-serverapp-c85="" class="ng-star-inserted">
   <th _ngcontent-serverapp-c85="" scope="row">محمد الشريف مساعدية</th>
   <td _ngcontent-serverapp-c85=""><a _ngcontent-serverapp-c85="" href="/Articles/146/14864/334689">:كلمة العدد
بين الواقع والتزييف</a></td>
   <td _ngcontent-serverapp-c85="">4</td>
</tr>
```

issues missing from the 2018 data set

I saw in 2026 (based on the divergence in authors), that the original scrape of the entire website did not catch all article pages. It would make sense to try and get all issue IDs and then the article IDs from the issue pages.

We can establish the journals with missing issues with the help of sakhrit_journals-summary.csv, which provides differences between the date of the latest scraped issue and the terminus of the journal provided by the journal pages.

```
df.journals.summary %>%
  tidyr::drop_na(terminus) %>%
  dplyr::mutate(terminus.diff = as.numeric(terminus) - year(terminus.collection)) %>%
  dplyr::filter(terminus.diff != 0) %>%
  dplyr::select(journal.id, id.wd, journal.title, onset, terminus, terminus.diff) %>%
  dplyr::arrange(onset) %>%
  knitr::kable(digits = 0, format = "pipe", caption = "missing data")
```

note that this only covers missing issues from the end of the distribution. The code results in

journal.id	id.wd	journal.title	onset	terminus	terminus.diff
j107	Q6033538	المقتطف	1876	1952	11
j278	Q12238943	العرفان	1909	1996	57
j65	Q141431890	المجمع العلمي العراقي	1950	2009	8
j255	Q86547607	الآداب	1953	1996	35
j64	Q141431896	المعرفة	1962	2020	6
j263	Q141431811	التراث الشعبي	1963	1983	19
j128	Q141431918	حوليات الجامعة التونسية	1964	2013	8
j129	Q12187413	الدوحة	1969	2013	27
j234	Q141431906	الموقف الأدبي	1977	2023	8
j171	Q141431812	التراث العربي	1979	2023	10
j235	Q120357476	آفاق الثقافة والتراث	1993	2019	4
j190	Q141431909	النص الجديد	1993	2000	2
j180	Q141431946	قوافل	1994	2020	5
j232	Q141431957	نزوى	1994	2019	5
j55	Q141431899	الملتقى	1997	2011	1
j250	Q141431830	الجسرة الثقافية	1998	2021	20
j170	Q141431798	الآداب العالمية	2004	2022	6
j236	Q141431933	طنجة الأدبية	2010	2019	5
j238	Q141431793	إضاءات نقدية	2011	2015	2
j194	Q141431950	كيكا	2013	2017	3

- observations
  - there might be huge gaps in the missing range! in the case of https://archive.alsharekh.org/magazineYears/278 the gap spans the years between 1946 and 1979
  - there are subpages for years for each magazine: https://archive.alsharekh.org/magazineYear/278/2819. but the URLs cannot be guessed!
  - there are subpages for issues and their contents: https://archive.alsharekh.org/contents/278/20041
    - with #cURL, one has to provide the explicit URL and thus either harvest them, guess them, or iterate over the entire range that reaches at least 30000. Wrong issue IDs do not cause an error, but a partially empty page!
    - Issue IDs might be consecutive for journals, but this would mean that journals are not conceptualised to receive updates and Q141431896 and others were updated since our initial scrape of the data set
  - it is interesting to see that the journal ID in the URL is completely irrelevant!!!

  - https://archive.alsharekh.org/contents/278/300 is identical to https://archive.alsharekh.org/contents/98/300 and https://archive.alsharekh.org/contents/5/300, which is the correct one!
  - Importantly, the metadata for the issue is missing from the wrong URLs. Additionally, while the article IDs in these pages are correct, the URLs will contain the wrong journal IDs and it is impossible to establish the correct journal!
  - However, the results of wrong URLs won't include the JSON-LD content and we can look for that
- looking
- how to download the missing data (i.e. how to get the missing issue and article IDs)? Neither cURL nor wget was fit for the job, which caused me to write R scripts for testing journal and issue combinations

new scrape

Using the script download_sakhrit-issue-pages.sh. Due to the issue of finding the correct combination of journal and issue IDs, this script is rather expensive to run, even though I factored in some assumptions about more similar issues being closer together (i.e. issue IDs being at least partially sequential for each journal).

Note that running the script in base R dramatically improves the success rates!

- problems
  - for some issue IDs all journal IDs seem to be wrong even though the issue does exist …
    - e.g. 24133 and the following IDs
    - data cleaning in sakhrit_issues-status:
      - remove all FALSE combinations for the issue IDs which have a known journal ID
      - check for issue IDs that have only FALSE combinations despite all journal IDs having been tested
    - it seems that journal 389 is not part of my data set, even though https://archive.alsharekh.org/contents/389/24133 clearly exists!

original scrape (2018-01-31)

Scraped contents.aspx?CID= with wget

- provides information: author (if known), article title, number of first page
- ☐ The CSV contains unclosed quotes for some articles, investigate!
  - i.e. 92185

differences between 2018 and 2026 data sets

Apparently, some issue ID have been deprecated / superseeded since 2018. This is true for a least some issues of "العربي" (Q4702220, jid:142) from the 1980s, such as iid:17568 and iid:17573.

- iid:14510 had a faulty publication date and is superseeded by iid:21516

article pages

All information is provided by the issue pages.

original scrape (2018-01-31)

Scraped ArticlePages.aspx?ArticleID= with wget

- provides information: journal title, issue, date, place, author, facsimiles, issue URL, author URL
- missing information: article title

authors' pages

I used a ZSH script to iterate over the alphabetical index and download the HTML pages. From there, I extracted the names and IDs from the HTML table:

<a _ngcontent-sc28="" href="/AuthorArticles/49401">1. رأس الماء عيسى

- regex: href="/AuthorArticles/(\d+)"\s*>\d+\.(.+?)
- result: 49401, رأس الماء عيسى

somehow, it appears that there is only a limited overlap between the index of authors and the author names in the bibliographic data of individual articles. There are more than 10.000 name strings from the article metadata, which are not found in the authors index. About a similar number of authors from the index are not found in the article data. The latter indicates that **the original scrape of article metadata was incomplete**!!

original scrape (2018-01-31)

scraped the author detail pages with wget: authorsArticles.aspx?AID=

- Problems:
  - the content of the page is limited to 30 entries. A JS provides links to further pages but they were not scraped with wget.
  - after running some 16000 of these pages through R, 1005 had 30 (and possibly more) articles. The remaining 15092 authors can be safely used

data quality

- Not all authors in the data set actually have an index page on the website. non-Arab authors, for instance, seem to be systematically excluded.
- Authorship information includes some impossible identifications:
  - عائشة عبد الرحمن, known as بنت الشاطئ, cannot have been the author of an article published in 1910, as she was only born 3 years later
- author names have only been cleaned to some extend (most likely depending on the knowledge of the cataloguers and based on the batches of issues, people were working on)
  - Louis Cheikho
    - "لويس شيخو","Q1871590","1898-01-01"
    - "ل . ش","Q1871590",NA
  - Jamīl Sidqī al-Zahāwī
    - "جميل صدقي الزهاوي","Q4115227","1909-07-01"
    - "زهاوي زادة جميل صدقي",Q4115227,"1896-09-15"
    - "جميل زهاوي",Q4115227,"1920-02-01"
    - "جميل الزهاوي",Q4115227,"1919-08-01"