DHQ: Digital Humanities Quarterly
2022
Volume 16 Number 2

Open Arabic Periodical Editions: A Framework for Bootstrapped Scholarly Editions Outside the Global North

Till Grallert <till_dot_grallert_at_fu-berlin_dot_de>, Orient-Institut Beirut https://orcid.org/0000-0002-5739-8094

Abstract

This paper introduces and evaluates the project Open Arabic Periodical Editions (OpenArabicPE) as a case study of minimal computing. It confronts hyperbolic promises of mass digitization and computational methods for the exploration of digitized cultural heritage as a hegemonic episteme rooted in 20th-century, English-speaking, neoliberal capitalism from the margins. OpenArabicPE is a framework for open, collaborative, and scholarly digital editions of early Arabic periodicals from the late Ottoman Eastern Mediterranean. It addresses the specific affordances of a historical multilingual society, whose material heritage continues to be looted, destroyed, and neglected; whose material heritage resists digitization efforts by being dependent on non-Latin scripts and, for instance, non-Gregorian calendars; and whose contemporary heirs cannot draw on the vast resources in wealth and socio-technical infrastructures of the Global North. Centered around generosity and minimal computing, OpenArabicPE is run by volunteers and currently hosts six editions with some 630 journal issues and more than 7 million words, without any funding, by re-purposing data, software, and infrastructures.

Introduction

Early Arabic periodicals published across the Eastern Mediterranean from the mid-19th to the early 20th centuries were at the core of social developments and formative discourses of modern(izing) societies that carry continued prominence across the region and beyond, such as the Arabic (cultural) renaissance (*nahḍa*), Arab nationalism, and the Islamic reform movement (*ṣalafiyya*). Yet, the vast majority of journals and newspapers remain obscure and understudied beyond a few well-known titles from Beirut and Cairo. As physical artifacts — frequently printed with cheap ink on cheap paper and handled without care in transit — they are vulnerable to neglect and active destruction. Neoliberal defunding of cultural heritage institutions as a global phenomenon is compounded by an onslaught of iconoclasts, failing institutions, and wars ravaging Syria, Yemen, and Iraq. 1

Turning to Alex Gil and Élika Ortega's question of "What do we need?" [Gil and Ortega 2016, 29], the answer, therefore, is simple: preservation and access. Societies of the Global South have a right to unhampered access to their own cultural record — a cultural record that is frequently held by institutions of the Global North. Access would also allow the scholarly communities concerned with the histories of Middle Eastern societies to study them through their own cultural production. 2

Taking up this issue, the following paper consists of two parts. I begin with challenging the equation of "digitization = access" by outlining the layers of inaccessibility inherent in existing digitization efforts of textual cultural artifacts from the late Ottoman Eastern Mediterranean. The Eastern Mediterranean, between Mesopotamia in the East and the Libyan desert in the West, the mountains of Anatolia in the north, and the Arabian Peninsula in the south, is home to historically multilingual and multiscriptural societies. Predominantly Arabic speaking, the area was part of the Ottoman Empire for four centuries until the end of World War I. The material cultural heritage of these societies has been and continues to be looted, destroyed, and neglected, as exemplified by the recent scandal surrounding the Museum of the Bible in Washington or the destruction of Palmyra by the so-called Islamic State of Iraq and Syria. Their textual cultural heritage resists digitization by being dependent on scripts other than Latin (such as Arabic, Armenian, Syriac, or Hebrew scripts) and calendars, as well as corresponding conceptions of time other than the Gregorian (such as the Ottoman fiscal or *mālī* calendar, the reformed Julian calendar, or the Islamic *hijrī* calendar). This is further exacerbated by their contemporary heirs' relative lack of resources in wealth and socio-technical infrastructures when compared to the Global North. The second part introduces the project Open Arabic Periodical Editions (OpenArabicPE) as a practical critique of these layers of inaccessibility that creatively bootstraps existing data, tools, and infrastructures into a framework to produce and distribute open scholarly periodical editions drawing on the affordances of the Global South.[1] I argue that OpenArabicPE demonstrates the feasibility of bootstrapping approaches by having published six full-text editions of Arabic journals originally published in Baghdad, Cairo, and Damascus between 1892 and 1918 with a total of 41 volumes comprising 645 issues, 12,830 pages that link to digital facsimiles,[2] and more than six million words (Table 2), without any resources beyond voluntary labor and private laptops. 3

Inaccessibility of Digitized Arabic Periodicals

The first layer of inaccessibility is a knowledge gap rooted in the physical artifact and its history. Scholars have written on the late Ottoman periodical press of the Eastern Mediterranean since the early twentieth century [Ṭarrāzī, Fīlīb dī. 1914] [Yalman 1914] and there is a plethora 4

of encyclopedic work on the Arabic press [Muruwwa 1961] [Rifāʿī 1969] [Dāghir 1978] [Ilyās 1982] [Khūriyya 1976] [ʿAbduh 1948] [Tikrītī 1969] [Ḥasanī 1957], listing titles and dates of first publication. Yet, the history of the vast majority of individual titles remains unknown, and only a handful of periodicals have been the subject of systematic studies [Glaß 2004] [Cioeta 1979] [Dierauff 2018] [Beška 2017]. This is particularly true for all places beyond the traditional scholarly focus on Beirut and Cairo.[3] If we were interested in the series of food riots that shook cities across the predominantly Arabic speaking provinces of the Ottoman Empire in summer 1910 [Grallert 2019], we would not know which contemporary newspapers and journals from Homs, Hama, or Gaza could be consulted for reports on these events.

5 The second layer of inaccessibility is a direct consequence of the knowledge gap and can be summarized as a digitization bias rooted in collection bias and survival bias [Gooding 2018, 52–61]. Publishers sought to minimize the cost of publishing, and, as a result, the durability of printed newspaper and periodical copies were not necessarily their primary concerns. Surviving copies are scattered across libraries and private collections around the globe. In the absence of a survey of collection histories, I deduce from the copies I have seen that many collections came about by chance and reflect local, regional, and global distribution networks more than specific collection policies. We must also take into account individual publishers' agency in influencing collections through donations of library copies, even long after their publication. The Ottoman-Arab intellectual, journalist, founder of the Arab Scientific Academy, and Minister of Education, Muḥammad Kurd ʿAlī (1876–1953), for example, established the monthly journal *al-Muqtabas* (The Digest) in Cairo in 1906. Kurd ʿAlī moved to his hometown of Damascus after the Young Turk Revolution of 1908, where he continued to publish *al-Muqtabas* until the end of World War I [Seikaly 1981]. In 1914, six years after its publication in Cairo, Kurd ʿAlī gifted a library copy of the second year of *al-Muqtabas* to his friend Jūrj Fākhūri. Two years later, he gifted another library copy of the same volume to the library of the Ṣalāḥiyya College in Jerusalem. The former is now held by the University of Minnesota and a digital facsimile can be accessed through HathiTrust. The latter ended up in the al-Aqṣā Mosque's library in Jerusalem and was digitized by the British Library's Endangered Archives Programme.[4]

6 Preservation of these brittle materials is costly, and frequent, prolonged wars, regime changes, and economic crises continue to have a devastating effect on surviving holdings. The Lebanese National Library, for instance, was shut down and its collections, including extensive periodical collections, were hastily stuffed into boxes and stored in the port of Beirut in 1975, where they remained for the next forty-odd years. A rehabilitation project has been underway since 2003, and reading rooms were opened to the public in 2018, but the status of the periodical collections remains opaque.

7 Locating a specific title and issue or all titles published in a particular place is a tedious endeavor exacerbated by the state of online catalogs. Existing catalogs are not necessarily published or digitized beyond onsite records. The website of the Lebanese National Library, for instance, still advertises its catalog as forthcoming. Union catalogs, on the other hand, have fallen out of fashion and should probably be read as historical sources rather than finding aids [El-Hadi 1965] [Hopwood 1970] [Aman 1979] [De Jong 1979] [Iḥdādan 1984] [Khūrī 1985] [Höpp 1994] [Atabaki and Rustămova-Tohidi 1995]. Known and confirmable collections are predominantly located in the Global North and are frequently incomplete. Figure 1 illustrates the combined survival, collection, and cataloging biases with the example of *al-Muqtabas*. Based on data from WorldCat and the Arabic Union Catalogue (ArUC), we can locate only nine complete collections worldwide: five in the US, two in Lebanon, and one each in Egypt and Saudi Arabia, including a reprint published by Dār Ṣādir in Beirut in 1992.[5]

8 Digitizing and hosting digital artifacts are expensive, and the cost scales almost linearly. In order to justify their expenses, curators will almost always turn to the rare and beautiful. Hundreds of thousands of periodical pages in foreign languages are frequently not considered important enough to warrant this investment for institutions in the Global North. Funds, infrastructures, and access to the physical artifacts, on the other hand, are commonly not available in the Global South. In consequence, we witness a neo-colonial absence of the Global South from the digital cultural record [Risam 2018] [Gooding 2018, 149–57] [Thylstrup 2018, 79–100]. In the absence of aggregators or an index, specific titles are incredibly hard to find in the patchwork of existing collections of digitized periodicals. Figure 2 shows the distribution of all Arabic periodical titles published across the Middle East and North Africa between 1789 and 1929, based on Project Jarāʾid, a scholarly crowdsourcing effort we have been running since 2012 to gather publication and holdings data. Based on this data set, only 504 or 15.59% of the 3232 titles could be located in collections and only 148 or 4.58% have been at least partially digitized [Mestyan and Grallert 2020] [Mestyan and Grallert et al. 2020].

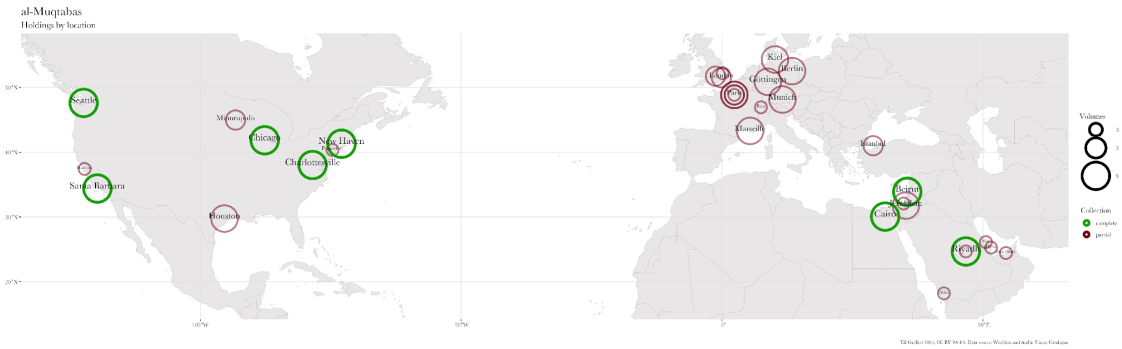

Figure 1. Geographic distribution of library holdings of al-Muqtabas (Cairo and Damascus, 1906–18) as recorded by WorldCat and Arabic Union Catalog (ArUC). The size of circles corresponds to the number of volumes in a collection. The color indicates whether collections are complete (green) or incomplete (red). The map is available online at https://doi.org/10.5281/zenodo.4154171.

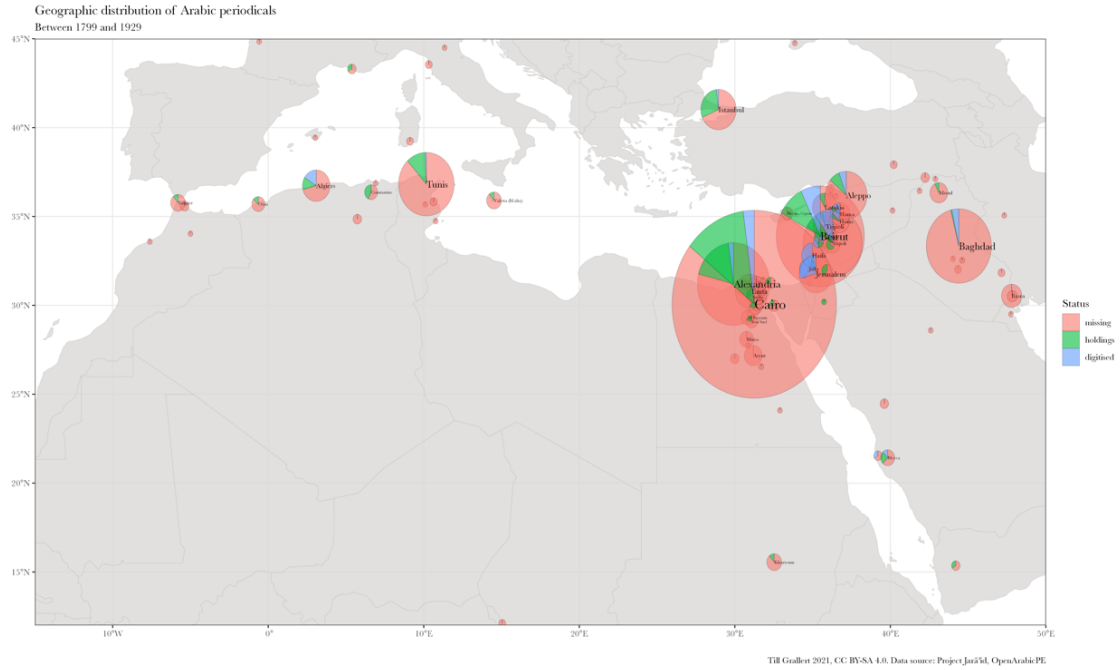

Figure 2. Geographic distribution of Arabic periodical titles published across the Middle East and North Africa between 1789–1929. The size of the pie charts corresponds to the total number of titles published at a location. Slices show the percentage of known holdings and digitized collections. This and other maps as well as the same data set are available at https://doi.org/10.5281/zenodo.4815717.

The third layer of inaccessibility is one of socio-technical infrastructures: digital infrastructures, despite all promises towards the opposite, are rooted in the epistemic hegemony of late 20th-century, Anglo-American, neoliberal capitalism. Most digitized periodicals are kept in data silos without any means for interchange or interoperability in the form of application programming interfaces (APIs) or the option to download data in machine-actionable, standardized, open file formats. Reading access to these silos is frequently restricted by paywalls and geo-fencing, a method to restrict access based on the geographic location of a user's IP address. Table 1 provides an overview of the accessibility of the ten largest online platforms serving digitized Arabic periodicals. *al-Muqtabas*, for example, is commonly deemed in the public domain in the United States. Copies from the University of Minnesota and Princeton University are openly available online at HathiTrust — for scholars at member institutions and the general public in the U.S. as determined by a user's IP address.[6] Everyone else will see blank pages. Downloading content in order to circumvent ill-suited interfaces is often limited to individually identifiable users. Automated downloads frequently violate terms of use, and most vendors try to prevent this on the technical level — at least in one case, detected attempts of automated downloads will result in the vendor punishing subscribing institutions for violations from within their IP range with a blanket block.

platform	vendor	vendor type	UI language	paywall	log-in required	geo-fencing	APIs	download	download formats	machine-readable metadata
al-maktaba al-shāmila		private, non-profit	ar	no	no	no	no	yes	EPUB, PDF (journal)	no
arshīf al-majallāt al-adabiyya wa-l-thaqāfiyya al-ʿarabiyya		private, non-profit	ar	no	no	no	no	no	NA	no
Endangered Archives Programme	British Library	public, non-profit	en	no	no	no	IIIF	no	NA	no
Early Arabic Printed Books	Cengage Gale	commercial	en	yes	yes	no	no	yes	PDF	no
Global Press Archive	East View	commercial	en	partial	yes	no	no	yes	PDF (page)	no
HathiTrust	HathiTrust	ppp, non-profit	en	no	yes	yes	yes	yes	PDF, txt (volume)	no
Institut du Monde Arabe		public, non-profit	ar, en, fr	no	no	no	no	no	NA	no
Arab American Newspapers Project (beta)	NC State University	public, non-profit	en	no	yes	no	no	yes	PDF[7] (issue) *currently disabled*	no
Translatio	University of Bonn	public, non-profit	de, en, fr	no	no	no	IIIF	yes	PDF (issue)	METS/ MODS
WikiSource	Wikimedia foundation	private, non-profit	ar, en, fr, …	no	no	no	yes	yes	EPUB, MOBI, PDF	no

Table 1. Accessibility of platforms serving digitized Arabic periodicals.

Proprietary web-interfaces are commonly neither tailored to the display of Arabic material nor themselves available in Arabic [Mansour and Gadallah 2020] [Wrisley and Jarkas 2019] [Wrisley and Jarkas 2016], while bibliographic metadata is regularly limited to the issue level and is woefully inadequate. This is due to ambiguity of data found in the physical artifacts, such as ambiguous references to calendars in mastheads, limited knowledge about these artifacts among librarians and the contractors who did the actual digitizing work, and software stacks incapable of handling anything but Western scripts and hegemonic Western concepts of dates and names. In consequence, often incorrect bibliographic metadata are recorded and shared in transcription to Latin script — of which there are as many varieties as there are languages written in Latin script (see Figure 3) [Grallert 2021]. Take for example, the Ottoman Turkish newspaper *Yenī Taṣvīr-i Efkār* from Istanbul. Holdings at the al-Aqṣā Mosque's library were digitized through the Endangered Archives Programme. Ottoman Turkish was written in Arabic script until 1924. The catalogers transcribed a truncated title into Latin script but did so according to the rules for Arabic, which resulted in the grammatically wrong title *Taşwīr Afkār*. In addition, they provided Jerusalem as place of publication instead of Istanbul.[8] Although practices have been improving in recent years and many catalogs can now display Arabic script, re-cataloging the original script is costly and not a priority for vendors.

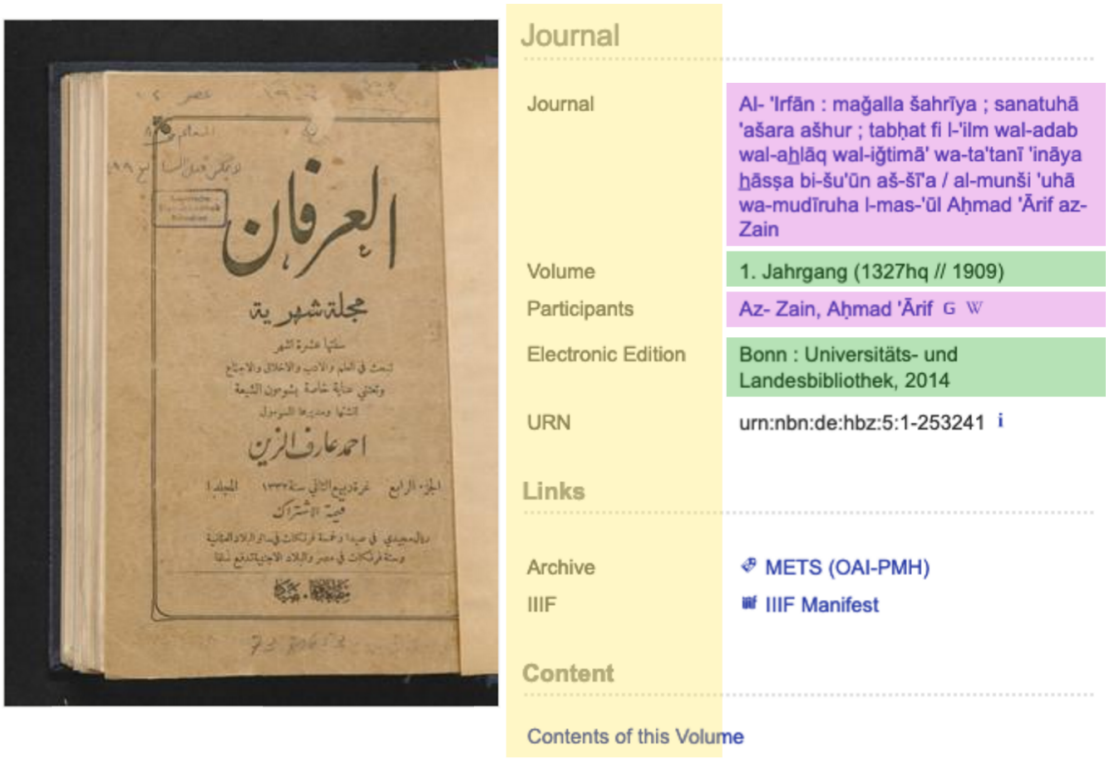

Figure 3. Annotated interface of the Translatio platform (University of Bonn) for a digitized copy of the Arabic journal *al-ʿIrfān*. Next to a facsimile of the front cover, they provide an English interface (yellow) for the bibliographic information, some of which is presented in German (green) and some in Arabic (purple) in Latin transcription transcribed into Latin script (purple) according to the system of the "Deutsche Morgenländische Gesellschaft" (DMG).

The second aspect of socio-technological inaccessibility is the hegemonic approach of computational systems to script through the metaphors of discrete characters and typesetting and by conceptually equating scripts with (national) languages [Fiormonte et al. 2015]. Yet, Arabic script is a writing system for many Asian and African languages [Mumin and Versteegh 2014]. It runs from right-to-left, with most letters connecting within a word. Letters have up to four forms depending on their position within a string, while multiple letters share the same basic shape (*rasm*). Diacritical signs (dots) above and below the *rasm* have been introduced to unambiguously identify a letter (see Figure. 4). There are only a few exceptions to these script-specific writing rules across languages.[9] However, dots are not strictly necessary for reading, and their use is subject to taste. Some regional cultural practices almost consistently omit them for specific letters, particularly at the end of words. Unicode — as a character-encoding standard that enables the vendor-independent exchange and interoperability of multilingual textual material of the networked age — does not adequately encode this cultural preference for ambiguity and forces transcribers to select a specific interpretation of the string in front of them. Transcribers have to either normalize orthographic variance or pick visually matching but "wrong" glyphs. On the other hand, because it is organized into (national) languages, Unicode provides multiple code points for the same glyph, introducing a potentially large variance in possible encodings for the same string of letters depending on the input keyboard of the transcribers. Egyptian Arabic speakers, for instance, would virtually never write the two dots underneath a final *yāʾ* (U+064A: ي). To mirror their cultural preference, they can either select the Arabic *alif maqṣūra* (U+0649: ى) or the Persian *ye* (U+06CC: [10].(ى] Unfortunately, search algorithms built into modern operating systems are not aware of this variance, and any application software relying on them will return skewed results without additional efforts at regularization or a reduction to the *rasm* [Milo and González Martínez 2019].

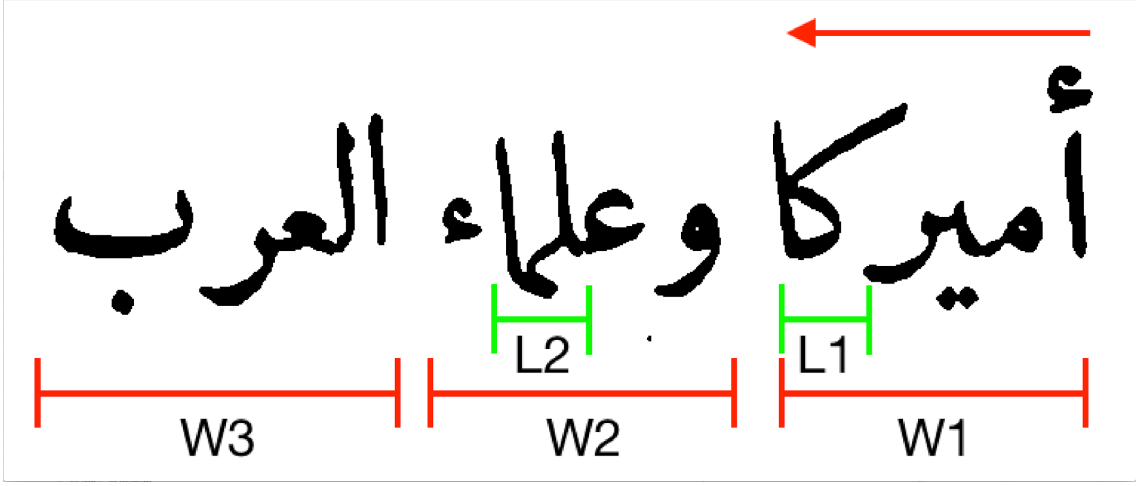

Figure 4. Example of basic Arabic typesetting from [Zakham 1907]: "America and the Arab scholars." The red arrow indicates the reading direction. Words are underlined in red. Ligatures of multiple letters are underlined in green.

12 The fourth layer of inaccessibility is the digital artifact itself: "Digitized" periodicals commonly mean scanned images (facsimiles) with limited or no text layer due to the particular challenges of Arabic script for optical character recognition (OCR) and the difficulty in automated layout recognition for dense multi-column page layouts. In addition to the characteristics of Arabic script outlined above, and depending on font and writing style, letters will form ligatures. They will not necessarily sit on a single base line, and base lines can be tilted (Figure 4). Segmentation into letters as the traditional approach to OCR has consequently proven extremely complicated and error prone. Accuracy rates for leading Arabic OCR solutions are well below 75% on the word level [Alghamdi and Teahan 2017] [Alkhateeb, Abu Doush, and Albsoul 2017] [Märgner and El Abed 2012] [Habash 2010], which causes the Internet Archive to state that the "language [is] not currently OCRable" (e.g., [Kurd ʿAlī 1923]). This is about to change with recent developments in machine-learning based approaches to OCR and hand-written text recognition (HTR), from Kraken, to Transkribus and Tesseract, which generally have shown to reliably produce high levels of accuracy independent of input language and script, including Arabic [Kiessling et al. 2017]. The Open Islamicate Texts Initiative Arabic-script OCR Catalyst Project (OpenITI ACOP) will train recognition models for the most frequent fonts and types [Open Islamicate Texts Initiative 2019], and their technology will eventually find its way into HathiTrust [HathiTrust Research Center Awards Three ACS Projects for 2020]. However, there is an important catch beyond the still unsolved layout recognition: machine-learning requires enormous resources — from labor time and the skills needed to use specialist software, to necessary hardware requirements and power consumption [Alkaoud and Syed 2020] [Strubell, Ganesh, and McCallum 2019] [Baillot et al.].

13 Commercial vendors and even some public institutions claim opaque break-throughs for in-house OCR technologies, but none share their software, evaluation reports, or even full-text layers (e.g., [Early Arabic Printed Books from the British Library]). Instead, their interfaces foreground full-text "search" capabilities with hits superimposed onto the digital facsimiles. Exploratory searches often reveal a high number of false positives among the search results while the extent of false negatives, strings that should have been found but weren't, and the veracity of full-text operations remains inherently unknown.

14 Finally, informal online libraries of Arabic literature, most prominent and popular among them *al-Maktaba al-shāmila* (The Comprehensive Library, 2005–) [Verkinderen 2020], provide access to a small number of full-text editions of Arabic periodicals based on the work of anonymous human transcribers. Unfortunately, they do not provide information on editorial principles, the quality of the transcription, or the material artifacts they were based on. In addition, such informal "editions" lack information linking the digital remediation to the original artifact, namely bibliographic metadata and page breaks, which makes them almost impossible to employ for scholarly research. In a final twist, I found that the informal editions from *al-Maktaba al-shāmila*, with all their gaps and omissions, were rendered as images with a pseudo-original layout to provide "fakesimiles" and served through *Arshīf al-majallāt al-adabiyya wa-l-thaqāfiyya al-ʿarabiyya* (Archive of Literary and Cultural Arabic Journals), the largest Arabic online platform for historical periodicals.[11] While this reveals the cultural significance of seemingly faithful reproductions, the existence of such "fakesimiles" requires any scholar to exert additional scrutiny before using material from these informal platforms.

15 These four layers of inaccessibility have an already detrimental effect on any Arabic speaker from Syria, Jordan, or Palestine with only limited knowledge of foreign languages, particularly English, and no affiliation with wealthy institutions from the Global North. They are, however, further compounded by the infrastructural dependencies of digital remediations on access to devices, Internet connections, and electricity. Neither of these can be taken for granted for the societies of the Global South. In the — all too frequent — worst case, devices are old, Internet connections are slow, traffic is prohibitively expensive, and electricity is perilously scarce [Aiyegbusi 2019].

16 Because OpenArabicPE was set-up in and run from Beirut, Lebanon, I illustrate this multi-layered "digital divide"[12] with the Lebanese example: only 15% of inhabitants have access to a landline and less than 1% have a broadband subscription, while 73% have a mobile phone

[Central Intelligence Agency 2020]. Many people rely on mobile Internet connections because wired infrastructures are insufficiently deployed or because they cannot sign up for a landline due to an unofficial rent agreement. They might also have no access to a personal computer or the necessary private space to use one due to cramped living conditions. Mobile internet connections, while by and large reliable and fast (4G), are prohibitively expensive. 20GB traffic or 30 days of service, whichever is less, cost about 40 USD, while at least 30% of the population lived on less than 120 USD a month in 2018 [Chadi 2018] [Fadel 2018]. Consequently, plans to levy a daily one-dollar tax on the usage of widely popular WhatsApp ignited wide-spread popular protest against the government and the entire political class in October 2019. Electricity has been another major issue. For the last decades, rising demand could not be matched by constantly decreasing supply. The single, state-run utility relies on insufficient power plants and imported, subsidized fuel. Consequently, the capital city of Beirut has been suffering from regular daily power cuts of three hours, even before the accelerating economic collapse that commenced in 2019. Already at this point, electricity had been expensive and easily came in at about 120 USD per month per household despite having been heavily subsidized. Generator subscriptions to offset power cuts were equally expensive.[13] The situation has deteriorated since. At the point of writing, the local currency lost 94% of its purchasing power for buying the necessary fuel for generating power on the international market, and Beirut had about one hour of electricity per day.

17 The impact of these infrastructural realities is not only limited to individual users of digital cultural artifacts, but also severely impacts institutions planning to host them. At my home institution, a well-funded foreign research institute, for instance, we share a single connection of 24Mbps. Even under perfect conditions, simply loading the landing page for a single volume of *al-Muqtabas* at the Endangered Archives Programme requires three seconds. With 20 colleagues and another 20 library users all trying to access online services and resources, load time quickly multiplies tenfold and more. Browsing through a large number of scanned images behind such a bottleneck is a daunting task. Uploading multiple gigabytes of high-resolution scans to a cloud computing service for machine-learning based OCR, for instance, is practically impossible, and we prefer to ship hard drives and wait for months for the results.

Bootstrapping for Access

18 I set up OpenArabicPE in August 2015 in order to address these socio-technical layers of inaccessibility of existing digitized Arabic periodicals with the affordances of the Global South and the tools at hand. Building on the guiding principles of simplicity and credibility for the sake of accessibility and sustainability, OpenArabicPE was inspired by Alex Gil's talk at Digital Humanities Institute-Beirut, during which he introduced the audience to the ideas of minimal computing, and by discussions and workshops at the 2015 Digital Humanities Summer Institute. A second strand of inspiration came from ideas about crowdsourcing and democratic public domain editions based on distributed version control systems that had emerged from the open-source communities in the early 2010s [Wittern 2013] [Forster 2012] [Reeve 2015] [Shaffer 2013b] [Terras 2016]. The idea is simple: re-use — sometimes creatively — openly available data, tools, and infrastructures to produce and share open, collaborative, scholarly digital editions of early Arabic periodicals. OpenArabicPE's resources were and remain extremely constrained: we have no funds, no staff beyond volunteering interns, and no equipment beyond our own computers. OpenArabicPE, therefore, also represents (and reflects the changing boundaries of) our own ability "to produce, disseminate, and preserve digital scholarship ourselves, without the help we can't get, even as we fight to build the infrastructures we need at the intersection of, with, and beyond institutional libraries and schools" [Gil and Ortega 2016, 29]. As such, OpenArabicPE constitutes a contribution to the digital commons [Wittel 2013] [Hall 2016] and provides a practical critique of neoliberal capitalism along the lines of what Valeria Graziano, Marcell Mars, and Tomislav Meda have recently called "pirate care" [Graziano, Mars, and Medak 2019] [Mars and Medak 2019].

The Data Layer

19 On the data level, OpenArabicPE combines the virtues of immensely popular, but non-academic, informal online libraries of volunteers with academic and institutional scanning efforts and editorial expertise. We transformed the digital text of six Arabic periodicals (see Table 2) — published in Baghdad, Cairo, and Damascus between 1892 and 1918 from *al-Maktaba al-shāmila* — from HTML (packaged as EPUB) into an open, standardized plain-text file format (XML) following the Text Encoding Initiative's (TEI) guidelines [TEI Consortium 2020] [DFG-Praxisregeln 2016]. Additionally, we linked each page to digital facsimiles from various sources, namely the Endangered Archives Project, Translatio, HathiTrust, and *Arshīf al-majallāt al-adabiyya wa-l-thaqāfiyya al-ʿarabiyya*. Each periodical issue is modeled as a single file to align with the original organizational principle of these compound texts. We then model each issue with light structural mark-up for articles, sections, mastheads, and bylines, as well as other bibliographic information.[14] As far as possible, authors and locations are identified and linked to local and international authority files. Bibliographic metadata on every article and in common standardized formats such as BibTeX and Metadata Object Description Schema (MODS) is then automatically generated from the modelled TEI XML source.[15] Thus, we provide the advantages of truly digital editions and a means to verify the text layer and our mark-up against facsimiles of the original artifact.

20 *Al-Maktaba al-shāmila* has been repeatedly evaluated for building scholarly corpora with a focus on distant reading approaches to classical texts [Belinkov et al. 2016] [Alrabiah, Al-Salman, and Atwell 2013]. OpenArabicPE was the first to focus on modern genres and to expand this basis into reliable and citable digital scholarly editions. Going through two journals in our corpus for the purpose of locating page breaks also revealed some of *al-Maktaba al-shāmila*'s editing practices, such as the omission of all footnotes and terms in scripts other than Arabic from the transcription. Most of the journals made extensive use of footnotes and articles covering recent developments in technology and sciences at the turn of the 20th century, often providing French, English, and Latin terms for clarification, which makes such systematic omissions relevant for evaluating distant reading approaches to the texts. Other scholars already noted that *al-Maktaba al-shāmila*'s editors omitted

Rashīd Riḍā's commentary on the Qurʾān (*tafsīr*) from their transcription of the monthly journal *al-Manār* (The Lighthouse) even though it accounted for more than one-fifth of the journal's content [Zemmin 2016, 232]. A combination of algorithmic searches for uncharacteristically short pages with close reading of these pages further substantiated the hypothesis of human transcribers without quality control strategies such as double keying. Most omissions can be plausibly explained by common human errors: skipping a few words on a long line, jumping a small number of lines, or turning two pages at once. In other instances, transcribers left completely unmarked comments in the transcription itself, stating, for example, that they couldn't read the following lines in the copy in front of them.[16] From these notes and common normalization of spelling variants, we can safely deduce that the transcribers were Arabic speakers. We are still looking for ways to adequately acknowledge their work beyond a generic reference in the metadata section of our TEI files.

Linking page breaks to facsimiles, although trivial, proved extremely labor-intensive because page breaks seemingly did not matter enough to the anonymous transcribers to be consistently recorded. Each of the ~8000 page breaks in the journals *al-Muqtabas* and *al-Ḥaqāʾiq* needed to be manually marked up by volunteers.[17] 21

Periodical	Place	Dates[18]	DOI	Volumes	Issues	Words	Size (MB) [19]
al-Ḥaqāʾiq	Damascus	1910–13	10.5281/zenodo.1232016	3	35	298090	15
al-Manār	Cairo	1898–1918		20	387	c.3000000	NA
al-Muqtabas	Cairo, Damascus	1906–18	10.5281/zenodo.597319	9	96	1981081	107.6
al-Ustādh	Cairo	1892–93	10.5281/zenodo.3581028	1	42	221447	13.1
al-Zuhūr	Cairo	1910–13	10.5281/zenodo.3580606	4	39	292333	25.6
Lughat al-ʿArab	Baghdad	1911–14	10.5281/zenodo.3514384	3	34	373832	35
total				41	645	c.6166783	

Table 2. OpenArabicPE's corpus of periodical editions.

The Presentation Layer

TEI XML is certainly not simple and comes with a steep learning curve. Bidirectional XML with left-to-right tags and right-to-left text is not particularly accessible to readers and editors either, even if specialized XML-editing software would better support it (Figure 5). However, TEI is both sustainable and credible as an underlying format, and the direction of a script is only relevant if rendered on a two-dimensional surface for human readers and editors. 22

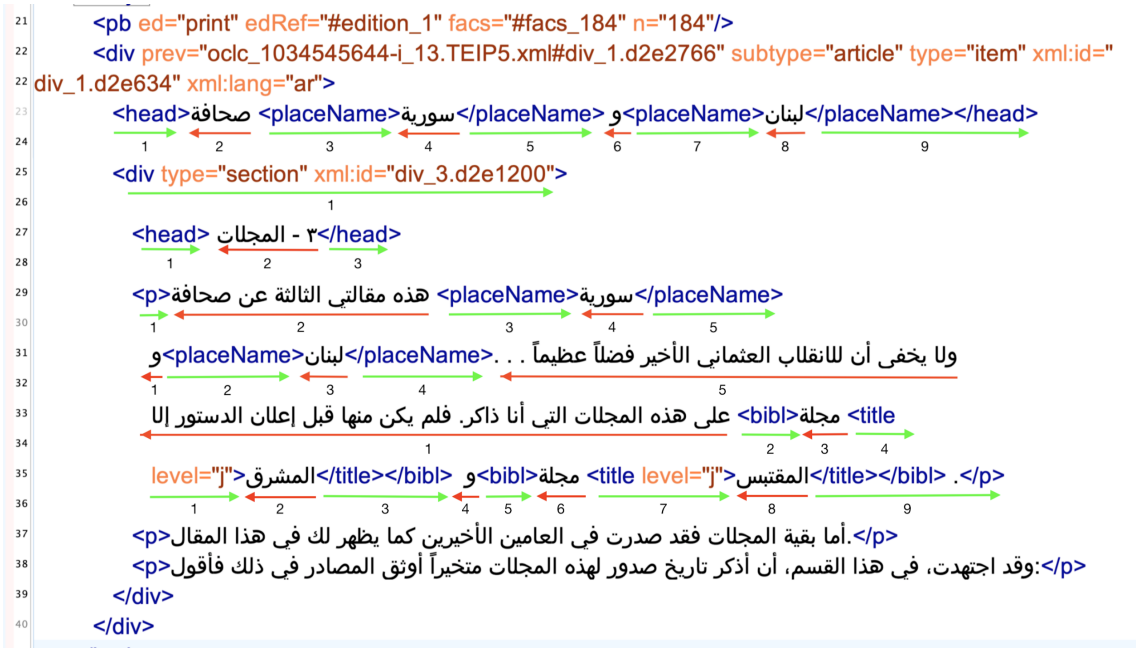

Figure 5. Example of bidirectional XML from the beginning of [Dammūs 1911]. The colored arrows indicate reading direction. The reading order is indicated by the numbers below the errors.

Presenting facsimiles and text side-by-side for the validation of the latter with the help of the former is core to OpenArabicPE's claim of 23

credibility. Therefore — and in order to make the editions accessible to readers with as little overhead as possible — we adapted the TEI Boilerplate to our needs (Figure 6).[20] TEI Boilerplate is based on the idea of directly rendering XML files in a user's web browser by using the built-in support for XSLT (eXtensible Stylesheet Language Transformations). This application of XSLT allows users to transform TEI XML into HTML on-the-fly and to tailor the structure and content of this HTML to the specific project needs. The HTML representation can then be styled with CSS and interacted with through Javascript, just like any other webpage [Walsh et al. 2016] [Walsh and Simpson 2013]. Our heavily modified adaptation of the TEI Boilerplate stylesheets adds support for right-to-left scripts; the side-by-side view of page images and text; a table of contents; browsing to neighboring issues; stable links to articles, sections, and paragraphs; and machine-actionable bibliographic metadata on the article level, and the necessary support for the specific structural requirements of periodicals. Finally, we added a simple system to provide multiple localizations, and most parts of the (simple) interface are available in Arabic [Grallert and Walsh 2020].

Figure 6. Webview of [Dammūs 1911], which renders the XML from Figure. 5.

The most obvious advantage of rendering XML files locally on the reader's computer is the removal of dependencies on backend servers and Internet connections. Editions can be downloaded, distributed through USB keys, and run locally. Generating ad hoc and on-the-fly reading copies also removes the need to actively generate, maintain, and upload multiple derivative versions of our editions and guarantees that human readers will always see the latest version. 24

Linking to digital facsimiles already available on the Internet is the preferred option from our project's point of view, as scanning, maintaining, and hosting large files is prohibitively expensive.[21] But these facsimiles tax the readers as linked online facsimiles require stable, fast, and affordable Internet connections; loading our webview for a single periodical issue of 56 pages requires 8MB even though IIIF allows us to limit the quality of images to grayscale and a width of 800dpi. A simple parameter setting, therefore, allows users to switch off the loading of online facsimiles altogether. Downloading images once and serving them locally would be the preferred option, but it would most likely violate user agreements, licenses, and copyright depending on vendors and jurisdictions. Nevertheless, such downloads are rather trivial as links to the digital facsimiles are an inherent part of the TEI XML files' `<facsimile>` node. 25

The Infrastructure

Editions and tools are maintained on GitHub, a staple of the open-source software development and digital humanities communities since 2008 [Lawson 2013] [Massey 2013] [Shaffer 2013a]. Built upon the open-source version control system .git, originally developed for the Linux kernel, GitHub provides an ever-evolving infrastructure for collaborative editing, including issue trackers and wikis for documentation. In this environment, each edit to our editions is transparently documented and credited to an individual contributor with their email address and a timestamp. GitHub also hosts the webview of our editions directly from the code repositories through GitHub Pages. Most importantly, GitHub 26

is free to use as long as all code repositories are openly accessible to the public.[22] This means that any reader spotting an error or omission can contribute their corrections or comments. The only threshold to participation is registering a free GitHub account and continued support for GitHub by Microsoft, which bought the platform in 2018 (see below).

27 Full-text search across the entire corpus is a core feature most users would expect to see implemented. There are at least three distinct aspects to this feature, none of which are easy to implement in a static environment. First, users want any search to be aware of our corpus' structural mark-up. That is, any search should be able to return hits on a specific structural level, such as a volume, issue, section, or article. Second, users will want to see keywords in context (KWIC) to evaluate a potentially large number of hits and because we are used to hegemonic search interfaces. Third, we expect search engines to be "smart" in the sense that they are aware of linguistic peculiarities and allow for some orthographic and morphologic variance.

28 Distributing digital corpora of hundreds of periodical issues without a backend severely restricts the ability to search and browse across issues. String search on plain-text files, such as XML, is, of course, a basic functionality of any personal computer's operating system and there are plenty of free and open solutions to leverage additional power for fine tuning queries through XPath or regular expressions (RegEx). For an expert user, this solution will satisfy the first two demands once they download the entire corpus. We will also gradually change our encoding scheme to one TEI XML file per article for each periodical issue once it is completely marked-up, in order to set a more useful default item level for file based search operations (these files will then be compiled into issues upon rendering by the webviewer). However, most users will not readily have the skills and tools for taking advantage of this approach. As a result, we have also turned to Google's Programmable Search Engine and Search Console. This requires a Google account linked to the editions' URLs, which, for various political and legal reasons, might not be an option for other projects.[23] Once set up, the Programmable Search Engine causes Google to crawl and index all editions and provides a basic Javascript snippet to integrate a project-specific Google search box into any website. Google's spiders are not aware of the TEI XML or its structure as they crawl the HTML page generated by the webview, but the Programmable Search Engine provides KWIC and all the bells and whistles of Google's algorithmic power. Moving to one file per article will again improve the granularity of this approach.

29 Browsing and searching bibliographic metadata, on the other hand, can be easily implemented. The free and open reference manager Zotero is another staple of the digital humanities community and beyond. Among its features, Zotero allows users to share an unlimited amount of bibliographic references for free through a feature called Groups. Zotero also provides a well-documented API for uploading, updating, and accessing all data beyond the user interface. This allows us to build and maintain a bibliographic index by hosting references to all articles in our corpus in a project specific Zotero Group. Each reference links to the relevant section in the periodical editions through a stable URL.[24] Readers can, therefore, use both Zotero's web interface and the standalone client as a port of entry to the entire OpenArabicPE corpus.

Facilitating Re-use Through Licenses

30 Within the academic framework in which we are operating, we depend on explicit licenses to facilitate the use and re-use of our data and tools.[25] We share all scripts with permissible MIT licenses whenever possible, and we assume that the content of periodicals published across the Eastern Mediterranean before 1920 is in the public domain even under the most restrictive definitions (i.e., U.S. copyright law) — an interpretation that is explicitly shared by U.S.-based vendors serving digital facsimiles. The enormous amount of human labor required for digitizing cultural artifacts, on the other hand, tends to be actively hidden by vendors and users. Within academic historiography, for instance, extensive use of digital surrogates instead of the physical artifact and reliance on their affordances remains an open secret: everybody does it but only few acknowledge this practice. This is evidenced by a search on JSTOR for references to the URLs of works from *al-Maktaba al-shāmila* in academic texts, which returned a grand total of only 16 journal articles and book chapters, some of which are shared with the 11 hits of a search for "*al-Maktaba al-shāmila*" (c.f., [Miller, Romanov, and Bowen 2018, 104]). Reasons for this omission range from a distrust of the digital, vaguely inspired by Walter Benjamin's ideas on the loss of aura through mechanical reproduction of an artifact [Benjamin 2005 [1968]], to perseverance of the established order of things and, ultimately, deception. Licenses will not solve dishonest citation practices, but retaining copyright of our own editorial contributions in the form of a Creative Commons Attribution-ShareAlike 4.0 International license is at least a reminder that all contributors need to be transparently credited for their work.

The Challenges of Bootstrapping

31 As has become clear above, the main advantage of bootstrapping is its reliance on open and free-to-use external services, tools, and data. This is also its greatest weakness, as all these services, tools, and data are beyond our control. Every one of these dependencies will eventually break. The only question is when they will do so. Our main mitigation strategies are open and widely supported standards to prevent lock-ins and loss of data, and a minimal software stack to keep the number of potential breaking points as small as possible. Mitigation strategies also depend on the evaluation of the relative importance of threats and components for the overarching goals of our project.

32 The data layer is at the core of the project. Once modeling and editing of a periodical issue in TEI is complete, this part of the data layer will remain fairly stable. This stability is documented by versioned releases of our editions at GitHub. Such versioned editions protect users from the inevitable content drift of digital editions and ascertains that a referenced edition hasn't been changed in the meantime [Broyles 2020]. GitHub, however, is a commercial platform owned by Microsoft since 2018. Like any other commercial platform, they might change their

business model or go out of business altogether at any point in the future.

GitHub and competing platforms for distributed version control are, therefore, no solution for sustainable, long-term access. Long-term preservation and accessibility require costly continuous maintenance, which is out-of-reach for projects, such as ours, which cannot afford to pay running costs for years to come. Publicly funded data repositories, on the other hand, guarantee the integrity and availability of data for the foreseeable future. Zenodo is an open and free-to-use, E.U.-funded research data repository originally developed by the European Organization for Nuclear Research (CERN) in Geneva. In addition to its core service of providing long-term archiving of digital data, Zenodo registers versioned Digital Object Identifiers (DOI) and integrates with various Open Science infrastructures, such as scholarly aggregators of metadata or Open Researcher and Contributor ID (ORCID) for unambiguously identifying contributors. Zenodo also integrates with GitHub by automatically archiving each release of data and tools and recording all contributors as found in the .git repository. This makes it a perfect choice for our GitHub-centered workflow. 33

The combination of platforms for distributed version-control with public repositories provides the necessary accessibility and sustainability of our data layer. But our editions' claim to credibility builds on the linkage between textual editions and digital facsimiles. These links to externally hosted facsimiles are the most volatile component of the data layer, and we already encountered three major instances of link rot. Two were caused by vendors moving servers and changing protocols: The British Library moved the Endangered Archives Programme to IIIF in 2017 and *Arshīf al-majallāt al-adabiyya wa-l-thaqāfiyya al-ʿarabiyya* moved to a new domain in 2019, while HathiTrust removed the facsimiles of Princeton's copy of *al-Haqāʾiq* (vol.1, 1910) from the public domain without any explanation in 2016.[26] Such link rot inevitably requires a lot of manual labor to figure out the patterns in new URLs (if any) and to write the necessary scripts to update all TEI files. Derivative formats, such as bibliographic metadata, although part of the data layer and preserved on Zenodo, can easily be regenerated from the main TEI files whenever changes to the infrastructure or updated standards necessitate it. 34

The World Wide Web and web browsers are fluid environments that will gradually break the presentation layer. As this presentation layer depends on specific versions of software, standards, and infrastructures, its functionality and looks cannot readily be preserved and archived. TEI Boilerplate was the only viable option to render TEI files in the browser without a backend in 2015, and it still works. All major browsers continue to support XSLT transformation but with two limitations. First, JavaScript Object Notation (JSON) has, in many cases, superseded XML as an exchange format for serialized (structured) data, and XSLT support is limited to version 1, which was superseded by XSLT 2 in 2007 and XSLT 3 in 2017. If browsers continue to support XSLT, it won't be further developed. Second, browser vendors tighten the screws on security. Running XSLT transformations on local files is considered attempted cross-site scripting and, consequently, is aborted by Google's Chrome browser. The Firefox browser will transform local XML files but only with remotely hosted XSLT. At the time of writing, only Apple's Safari allows for rendering our editions without any Internet connection. In the meantime, the TEI community has spawned new and exciting infrastructures for presenting TEI-based editions to human readers. Based on the TEI Processing Model [Turska, Cummings, and Rahtz 2016], the TEI Publisher provides a flexible publication and editing framework as long as one has access to hosted server space. The Javascript-based CETEIcean, on the other hand, has matured into an alternative to TEI Boilerplate with the full support of the TEI consortium [Cayless and Viglianti 2018] [Cayless and Viglianti 2020]. All future developments of OpenArabicPE's webview will therefore move to CETEIcean. 35

Hosting the webview is not a challenge in itself. There will always be services that provide hosting for the small amounts of data that our digital editions contain (Table 2) for free. They are, however, of varying convenience. Data might need to be copied from the code repository to the hosting provider. In this regard, GitHub pages is certainly the most convenient solution. It allows us to directly serve the TEI files from any branch of the code repositories, ensuring that the presentation layer always renders the latest version of the data layer. Switching providers will introduce a major instance of link rot for users of our editions. At a bare minimum, all URLs pointing from the bibliographic metadata to the editions will need to be updated. Renting a project specific domain, such as openarabicpe.org, will protect against such basic link rot for a small monthly fee, but it might require a credit card with the ability to transfer USD abroad, which is not necessarily available to projects across the Eastern Mediterranean. We have therefore decided to stick with the default domain provided by GitHub (openarabicpe.github.io). 36

Conclusion

This paper introduced our project, Open Arabic Periodical Editions, as a concrete response to the multiple layers of inaccessibility for the textual cultural heritage of the predominantly Arabic speaking societies of the Eastern Mediterranean. We demonstrate the possibility of bootstrapping existing data, tools, and infrastructures into a framework for producing and distributing open scholarly periodical editions. This framework addresses the constraints of material that resists digitization and operates within the socio-technical affordances of a particular region of the Global South. The problems and practical approaches outlined above are, to a large extent, not specific to digital editions of Arabic periodicals or the particular environment in Lebanon. All tools, workflows, and infrastructures can be readily adapted to other literary genres in different scripts and languages with very little effort. Documenting our approaches and workflows is a major, ongoing contribution to this effort. 37

With regards to project specific outcomes, OpenArabicPE has the potential to radically alter and influence an emerging field of Arab Periodical Studies by making this material accessible for computational methods [Grallert 2021]. The editions have been used as the necessary ground truth in initial efforts to train genre-specific models for machine-learning approaches to OCR with very promising results, in collaboration with Sinai Rusinek. Planned collaborations for improving layout-recognition algorithms will help to eventually remove the dependency on human 38

transcribers and bring the digitization of the rich Arabic textual heritage on par with at least the smaller Western languages.

Beyond building infrastructures “without the help we can’t get” [Gil and Ortega 2016, 29], OpenArabicPE highlights how minimal effort is frequently needed to improve existing infrastructures. Often, adding basic support for the `@xml:lang` and `@lang` attributes to XML and HTML environments with a little CSS snippet significantly improves their usability. Lobbying browser vendors to include a few lines to this end in their default CSS seems to be a worthwhile and achievable goal. In less optimistic terms, OpenArabicPE contributes to laying open the brutal neocolonial character of our contemporary knowledge infrastructures and digital economies. Character and text encoding standards, natural language processing and named entity recognition, ontologies, thesauri, and gazetteers (whether part of the semantic web or not) have all been developed without consideration of or input from the majority of human societies. OpenArabicPE is part of larger communities of practitioners that gravitate towards, but are in no way limited to, the Right To Left conference at the Digital Humanities Summer Institute (DHSI), the Digital Humanities Institute-Beirut (DHIB), the Islamicate Digital Humanities Network (IDHN), the Digital Orientalist, or the Historical Middle East Data Alliance. Together, we are raising attention for the needs of Arabic-centered scholarship and right-to-left scripts more generally within the larger digital humanities community.[27]

Notes

[1] Open Arabic Periodical Editions is available at https://openarabicpe.github.io/.

[2] The linking has only been completed for five periodicals, which is reflected by this page count.

[3] For a more detailed overview of the state of Arab Periodical Studies see [Grallert 2021].

[4] Handwritten note on al-Muqtabas 2.1, p.1. The digital facsimile is available at: https://babel.hathitrust.org/cgi/imgsrv/image?id=umn.319510029968616;seq=12. For more information about the Endangered Archives Programme, visit http://eap.bl.uk

[5] The Arabic Union Catalogue (ArUC) is available at https://www.aruc.org/.

[6] The University of Minnesota copy is available at https://catalog.hathitrust.org/Record/100658549 and the Princeton University copy is available at https://catalog.hathitrust.org/Record/008882293.

[7] Claims to provide a “proprietary” PDF version, whose text layer is currently incompatible with Adobe’s PDF reader. Download has, therefore, been disabled.

[8] See, for example, https://eap.bl.uk/archive-file/EAP119-1-18-1.

[9] For an introduction to the particularities of Arabic script see [Nemeth 2017, 14–22].

[10] On the background of Unicode and its application to Arabic, see [Nemeth 2017, 400–406]. Nemeth also provides the most concise overview of the work of Thomas Milo, the most profound critic of digital approaches to Arabic script and the founder of DecoType [Nemeth 2017, 400–406].

[11] There is a huge and unmarked gap in *al-Maktaba al-shāmila*’s transcription of *al-Muqtabas* 5.7 between p.463 and p.466, which is reproduced in this “fakesimile.”

[12] For a recent conceptualization of the term see [Ragnedda 2019]. Unfortunately, while acknowledging the multiple socio-technological layers, they focus solely on the Internet and do not mention the hegemony of English. This is also true for [Townsend 2013, 168–93].

[13] It must be noted that corruption and waste are not the only culprits to blame. Israel deliberately damaged and destroyed power plants in the 2006 war.

[14] Our TEI customization for Arabic periodicals is openly available at https://github.com/OpenArabicPE/OpenArabicPE_ODD.

[15] Our code is openly available at https://github.com/OpenArabicPE/convert_tei-to-bibliographic-data. It uses `<tei:biblStruct>` as an intermediary format and can also generate YAML and Zotero RDF.

[16] See [Sharīf 1911, 422] for an example, stating, “Four lines are unclear on p.422.”

[17] In other instances, such as the journals Lughat al-ʿArab and al-Ustādh, *al-Maktaba al-shāmila* did provide page breaks that correspond to a printed edition. My gratitude goes to Dimitar Dragnev, Talha Güzel, Dilan Hatun, Hans Magne Jaatun, Xaver Kretzschmar, Daniel Lloyd, Klara Mayer, Tobias Sick, Manzi Tanna-Händel, and Layla Youssef, who contributed their time to this task.

[18] The current cut-off date is 1918.

[19] Size includes TEI files and bibliographic metadata (BibTeX, MODS).

[20] TEI Boilerplate is available at http://dcl.slis.indiana.edu/teibp/.

[21] If we were to produce our own facsimiles, I would recommend uploading them to the Internet Archive, which also provides basic IIIF support since 2015 [IIIF Documentation].

[22] Students and teachers have access to private repositories and additional features otherwise reserved for paying users through GitHub’s education program. Proof of eligibility is based on institutional email addresses.

[23] There are at least two issues here. The first is practical access to Google services. Some countries, such as China, ban their citizens from using Google or prevent access to its services for anyone within their IP range, while others, such as Iran, are blocked from accessing Google due to, for instance, U.S. sanctions. There are also perfectly good political reasons for objecting to using a platform whose business-model is built upon creating and monetizing highly specific behavioural profiles of individual users.

[24] This includes stable BibTeX cite keys for referencing the periodicals in academic writing. See our Zotero Group at https://www.zotero.org/groups/OpenArabicPE.

[25] For an introduction to the conceptual critique of copyrights see [Hall 2016, 1–9]. For a specific application to TEI XML files see [Hannesschläger 2020].

[26] My inquiries about the legal basis for this decision have not been replied to.

[27] For more information, see the Right To Left conference at https://dhsi.org/dhsi-2021-online-edition/dhsi-2021-online-edition-aligned-conferences-and-events/dhsi-2021-right-to-left/; Digital Humanities Institute-Beirut (DHIB) https://dhibeirut.wordpress.com/; the Islamicate Digital Humanities Network (IDHN) https://idhn.org/; Digital Orientalist https://digitalorientalist.com/; and the Historical Middle East Data Alliance https://github.com/Hist-ME.

Works Cited

ʿAbduh 1948 ʿAbduh, Ibrāhīm. *Aʿlām al-ṣiḥāfa al-ʿArabiyya* [Eminent personalities of the Arabic press]. al-Qāhira: Maktabat al-Ādāb, 1948. http://hdl.handle.net/2333.1/v9s4n21b.

Aiyegbusi 2019 Aiyegbusi, Babalola Titilola. “Decolonizing Digital Humanities: Africa in Perspective.” In *Bodies of Information: Intersectional Feminism and Digital Humanities*, edited by Elizabeth Losh and Jacqueline Wernimont, 434–46. Minneapolis: University of Minnesota Press, 2019. https://dhdebates.gc.cuny.edu/read/d02c3ed5-0c55-4de9-88de-5f543fecd130/section/c9862793-ef00-4d6b-a30c-2f3d354e4e94#ch23.

Alghamdi and Teahan 2017 Alghamdi, Mansoor and William Teahan.“Experimental Evaluation of Arabic OCR Systems.” *PSU Research Review* vol. 1.3 (2017): 229–41. https://doi.org/10/gh4457.

Alkaoud and Syed 2020 Alkaoud, Mohamed and Mairaj Syed. “On the Importance of Tokenization in Arabic Embedding Models.” In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 119–29. Barcelona: Association for Computational Linguistics, 2020. https://aclanthology.org/2020.wanlp-1.11.

Alkhateeb, Abu Doush, and Albsoul 2017 Alkhateeb, Faisal, Iyad Abu Doush, and Abdelraoaf Albsoul. “Arabic Optical Character Recognition Software: A Review.” *Pattern Recognition and Image Analysis* vol. 27.4 (2017): 763–76. https://doi.org/10/gh445n.

Alrabiah, Al-Salman, and Atwell 2013 Alrabiah, Maha, A Al-Salman, and E S Atwell. *The Design and Construction of the 50 Million Words KSUCCA*. The University of Leeds, 2013.

Aman 1979 Aman, Mohammed M. *Arab Periodicals and Serials: A Subject Bibliography*. New York: Garland, 1979.

Atabaki and Rustămova-Tohidi 1995 Atabaki, Touraj and Solmaz Rustămova-Tohidi. *Baku Documents: Union Catalogue of Persian, Azerbaijani, Ottoman Turkish and Arabic Serials and Newspapers in the Libraries of the Republic of Azerbaijan*. London: Tauris Academic Studies, 1995.

Baillot et al. Baillot, Anne, James Baker, Madiha Zahrah Choksi, Alex Gil, Ana Lam, Alicia Peaker, Walter Scholger, Torsten Roeder, and Jo Lindsay Walton. “Digital Humanities and the Climate Crisis: A Manifesto.” Accessed October 12, 2021. https://web.archive.org/web/20210923032455/https://dhc-barnard.github.io/envdh/.

Belinkov et al. 2016 Belinkov, Yonatan, Alexander Magidow, Maxim Romanov, Avi Shmidman, and Moshe Koppel. “Shamela: A Large-Scale Historical Arabic Corpus.” *arXiv Preprint*, 2016. arXiv:1612.08989.

Benjamin 2005 [1968] Benjamin, Walter. “The Work of Art in the Age of Mechanical Reproduction.” In *Illuminations*, edited by Hannah Arendt, translated by Harry Zohn. New York: Schocken Books, 2005 [1968].

Beška 2017 Beška, Emanuel. *From Ambivalence to Hostility: The Arabic Newspaper Filasṭīn and Zionism, 1911-1914*. Bratislava: Slovak Academic Press, 2017.

Broyles 2020 Broyles, Paul A. “Digital Editions and Version Numbering.” *Digital Humanities Quarterly* vol. 14.2 (2020). https://doi.org/10.17613/bde5-rp28.

Cayless and Viglianti 2018 Cayless, Hugh and Raffaele Viglianti. “CETEIcean: TEI in the Browser.” In *Proceedings of Balisage: The Markup Conference* 21 (2018). https://www.balisage.net/Proceedings/vol21/html/Cayless01/BalisageVol21-Cayless01.html.

Cayless and Viglianti 2020 Cayless, Hugh and Raffaele Viglianti. *CETEIcean* (version 1.2.1). JavaScript. *Text Encoding Initiative Consortium*, 2020. https://github.com/TEIC/CETEIcean.

Central Intelligence Agency 2020 Central Intelligence Agency. “Lebanon.” *The World Factbook*, 2020. https://web.archive.org/web/20201012153026/https://www.cia.gov/library/publications/the-world-factbook/geos/le.html.

Chadi 2018 Chadi. “UNDP Latest Poverty Assessment Report: 30% of Lebanese Are Poor.” *Blog Baladi,* February 17, 2018. https://web.archive.org/web/20210121081411/https://blogbaladi.com/undp-latest-poverty-assessment-report-30-of-lebanese-are-poor/.

Cioeta 1979 Cioeta, Donald J. “Thamarāt Al-Funūn, Syria’s First Islamic Newspaper, 1875-1908”. PhD dissertation, 1979. University of Chicago.

DFG-Praxisregeln 2016 DFG-Praxisregeln. “Digitalisierung.” Bonn: Deutsche Forschungsgemeinschaft, 2016. http://www.dfg.de/formulare/12_151/12_151_de.pdf.

Dammūs 1911 Dammūs, Ḥalīm Ibrāhīm.“Ṣiḥāfat Sūriyya wa-Lubnān” [The Press of Syria and Lebanon]. *al-Zuhūr* 2.4 (June 1, 1911).

https://openarabicpe.github.io/journal_al-zuhur/tei/oclc_1034545644-i_15.TEIP5.xml#div_1.d2e634.

De Jong 1979 De Jong, Fred. “Arabic Periodicals Published in Syria Before 1946: The Holdings of Zahiriyya Library in Damascus.” *Bibliotheca Orientalis* 36 (1979): 292–300.

Dierauff 2018 Dierauff, Evelin. “Negotiating Ethno-Confessional Relations in Late Ottoman Palestine: Debates in the Arab Palestinian Newspaper Filasṭīn (1911-1914).” PhD dissertation, 2018. Universität Tübingen.

Dāghir 1978 Dāghir, Yūsuf Aḥmad. *Qāmūs al-ṣiḥāfa al-Lubnāniyya 1858-1974* [Encyclopaedia of the Lebanese Press, 1858-1974]. Bayrūt: al-Maktaba al-Sharqiyya al-Kubrā, 1978.

Early Arabic Printed Books from the British Library Early Arabic Printed Books from the British Library. Accessed September 16, 2019. https://web.archive.org/web/20190916173504/https://p-www.gale.com/primary-sources/early-arabic-printed-books-from-the-british-library.

El-Hadi 1965 El-Hadi, Mohamed M. *Union List of Arabic Serials in the United States: The Arabic Serial Holdings of Seventeen Libraries*. *Occasional Papers* 75 (1965). Urbana: University of Illinois, Graduate School of Library and Information Science.

Fadel 2018 Fadel, Rosette. “Third of Lebanese Live in Poverty, Experts Say.” *An-Nahar*, September 21, 2018. https://web.archive.org/web/20210515205913/https://www.annahar.com/english/article/865485-third-of-lebanese-live-in-poverty-experts-say.

Fiormonte et al. 2015 Fiormonte, Domenico, Desmond Schmidt, Paolo Monella, and Paolo Sordi. “The Politics of Code. How Digital Representations and Languages Shape Culture.” In *ISIS Summit Vienna 2015 — the Information Society at the Crossroads*. Vienna: MDPI AG, 2015. https://doi.org/10/gkzc7v.

Forster 2012 Forster, Chris. “Public Domain Editions.” *Chris Forster*, June 21, 2012. https://web.archive.org/web/20210907150427/http://cforster.com/2012/06/drill-baby-drill/.

Gil and Ortega 2016 Gil, Alex and Élika Ortega. “Global Outlooks in Digital Humanities: Multilingual Practices and Minimal Computing.” In *Doing Digital Humanities: Practice, Training, Research*, edited by Constance Crompton, Richard J. Lane, and Ray Siemens, 22–34. NY: Routledge, 2016.

Glaß 2004 Glaß, Dagmar. *Der Muqtaṭaf und seine Öffentlichkeit. Aufklärung, Räsonnement und Meinungsstreit in der frühen arabischen Zeitschriftenkommunikation*. 2 vols. Würzburg: Ergon Verlag, 2004.

Gooding 2018 Gooding, Paul. *Historic Newspapers in the Digital Age: “Search All About It.”* London: Routledge, 2018.

Grallert 2019 Grallert, Till. “Urban Food Riots in Late Ottoman Bilād Al-Shām as a 'Repertoire of Contention.'” In *Crime, Poverty and Survival in the Middle East and North Africa: The “Dangerous Classes” Since 1800*, edited by Stephanie Cronin, 157–76. London: I.B. Tauris, 2019. https://doi.org/10.5040/9781838605902.ch-010.

Grallert 2021 Grallert, Till. “Catch Me If You Can! Approaching the Arabic Press of the Late Ottoman Eastern Mediterranean Through Digital History” *Geschichte und Gesellschaft* vol. 47.1 (2021): 58–89. https://doi.org/10/gkhrjr.

Grallert and Walsh 2020 Grallert, Till and John Walsh. *TEI Boilerplate for Arabic Editions* (version 0.8). XSLT. Open Arabic Periodical Editions, 2020. https://doi.org/10.5281/zenodo.597307.

Graziano, Mars, and Medak 2019 Graziano, Valeria, Marcell Mars, and Tomislav Medak. “The Pirate Care Project.” *Centre for Postdigital Cultures, Coventry University*, 2019. https://web.archive.org/web/20210926194036/https://pirate.care/.

Habash 2010 Habash, Nizar Y. “Introduction to Arabic Natural Language Processing.” Synthesis Lectures on Human Language Technologies 10. Williston, VT: Morgan & Claypool, 2010. https://doi.org/10/ffr8nh.

Hall 2016 Hall, Gary. *Pirate Philosophy: For a Digital Posthumanities*. Cambridge, MA: The MIT Press, 2016.

Hannesschläger 2020 Hannesschläger, Vanessa. “Common Creativity International: CC-Licensing and Other Options for TEI-Based Digital Editions in an International Context.” *Journal of the Text Encoding Initiative* 11 (June 2020). https://doi.org/10/gg3wts.

Ḥasanī 1957 Ḥasanī, ʿAbd al-Razzāq. *Tārīkh al-ṣiḥāfa al-ʿIrāqiyya* [History of the Iraqi Press]. Baghdād: Maṭbaʿa al-Zahrāʾ, 1957. http://hdl.handle.net/2333.1/m63xspj1.

HathiTrust Research Center Awards Three ACS Projects for 2020 “HathiTrust Research Center Awards Three ACS Projects for 2020.” *HathiTrust Digital Library,* July 7, 2020. https://web.archive.org/web/20210414053200/https://www.hathitrust.org/htrc-awards-three-acs-projects.

Hopwood 1970 Hopwood, Derek. *Arabic Periodicals in Oxford: A Union List*. Oxford: St. Antony’s College, 1970.

Höpp 1994 Höpp, Gerhard. *Arabische und islamische Periodika in Berlin und Brandenburg 1915 - 1945*. Berlin: Verlag Das Arabische Buch, 1994.

IIIF Documentation “IIIF Documentation.” *Internet Archive*. Accessed July 1, 2020. https://iiif.archivelab.org/iiif/documentation.

Ilyās 1982 Ilyās, Jūzīf. *Taṭawwur al-ṣiḥāfa al-Sūriyya fī miʾat ʿām: 1865-1965* [Development of the Press in Syria During a Century, 1865-1965]. Bayrūt: Dār al-Niḍāl, 1982.

Iḥdādan 1984 Iḥdādan, Zāhir. *Bībliyūghrāfiyā al-ṣiḥāfa al-Jazāʾiriyya* [Bibliography of the Algerian Press]. al-Jazāʾir: al-Muʾassasat al-Waṭaniyya li-l-Kitāb, 1984.

Khūriyya 1976 Khūriyya, Yūsif Q. *al-Ṣiḥāfa al-ʿArabiyya fī Filasṭīn 1876-1948* [The Arabic Press in Palestine 1876-1948]. Bayrūt: Muʾassasat al-Dirāsāt al-Filasṭīniyya, 1976.

Khūrī 1985 Khūrī, Yūsuf Quzmā. *Mudawwanat al-ṣiḥāfa al-ʿArabiya* [A Record of the Arabic Press]. Edited by ʿAlī Dhū al-Fiqār Shākir. Vol. 1: Miṣr. Bayrūt: Maʿhad al-Inmāʾ al-ʿArabī, 1985.

Kiessling et al. 2017 Kiessling, Benjamin, Matthew Thomas Miller, Maxim Romanov, and Sarah Bowen Savant. "Important New Developments in Arabographic Optical Character Recognition (OCR)". *Al-ʿUṣūr Al-Wuṣṭā* 25 (2017): 1–13. https://www.middleeastmedievalists.com/wp-content/uploads/2017/11/UW-25-Savant-et-al.pdf.

Kurd ʿAlī 1923 Kurd ʿAlī, Muḥammad. *Gharāʾib al-Gharb* [The Oddities of the West]. 2nd ed. Vol. 1. Miṣr: al-Maṭbaʿa al-Raḥmāniyya, 1923. http://archive.org/details/1_20191109_20191109_1843.

Lawson 2013 Lawson, Konrad M. "GitHub101." *ProfHacker*, 2013. https://web.archive.org/web/20131208102848/http://chronicle.com/blogs/profhacker/tag/github101.

Mansour and Gadallah 2020 Mansour, Nadirah and Marwa Gadallah. "al-Iḥtiyāj li-l-wājihāt al-iliktrūniyya bi-l-lughat al-ʿArabiyya" [The Need for Electronic Interfaces in Arabic]. Presented at the Digital Orientalisms Twitter Conference 2020 (#DOsTC2020), June 20, 2020. https://twitter.com/NAMansour26/status/1274361436215574529.

Mars and Medak 2019 Mars, Marcell and Tomislav Medak. "System of a Takedown: Control and de-Commodification in the Circuits of Academic Publishing." In *Archives: In Search of Media*, edited by Andrew Lison, Marcell Mars, Tomislav Medak, and Rick Prelinger, 47–68. Lüneburg / Minnesota: meson press / University of Minnesota Press, 2019. https://doi.org/10.14619/1501.

Massey 2013 Massey, Harrison. "GitHub, Academia, and Collaborative Writing." *HASTAC*, October 12, 2013. https://web.archive.org/web/20210126160642/https://www.hastac.org/blogs/harrisonm/2013/10/12/github-academia-and-collaborative-writing.

Mestyan and Grallert 2015 Mestyan, Adam, and Till Grallert. "A Chronology of Nineteenth-Century Periodicals in Arabic (1800-1900): A Research Tool." Accessed April 22, 2016. https://web.archive.org/web/20160422071133/https://www.zmo.de/jaraid/.

Mestyan and Grallert 2020 Mestyan, Adam and Till Grallert. "Jara'id: A Chronology of Arabic Periodicals (1800-1929)." 2020 Edition, Accessed October 14, 2021. https://projectjaraid.github.io/.

Mestyan and Grallert et al. 2020 Mestyan, Adam, Till Grallert et al. "Jarāʾid: A Chronology of Arabic Periodicals (1800-1929)." *Zenodo*. Accessed October 14, 2021. https://doi.org/10.5281/zenodo.4399240.

Miller, Romanov, and Bowen 2018 Miller, Matthew Thomas, Maxim G. Romanov, and Sarah Bowen Savant. "Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans". *International Journal of Middle East Studies* vol. 50.1 (2018): 103 –9. https://doi.org/10/gg865d.

Milo and González Martínez 2019 Milo, Thomas and Alicia González Martínez. "A New Strategy for Arabic OCR: Archigraphemes, Letter Blocks, Script Grammar, and Shape Synthesis." In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, 93–96. DATeCH2019. NY: Association for Computing Machinery, 2019. https://doi.org/10/gmscgz.

Mumin and Versteegh 2014 Mumin, Meikal and Kees Versteegh. *The Arabic Script in Africa: Studies in the Use of a Writing System*. Leiden: Brill, 2014.

Muruwwa 1961 Muruwwa, Adīb. *al-Ṣiḥāfa al-ʿArabiyya: nashʾatuhā wa taṭawwuruhā* [The Arabic Press: its Spread and Development]. Bayrūt: Dār Maktabat al-Ḥayyāt, 1961.

Märgner and El Abed 2012 Märgner, Volker and Haikal El Abed, eds. *Guide to OCR for Arabic Scripts*. London: Springer, 2012. https://doi.org/10.1007/978-1-4471-4072-6.

Nemeth 2017 Nemeth, Titus. *Arabic Type-Making in the Machine Age: The Influence of Technology on the Form of Arabic Type, 1908-1993*. Leiden: Brill, 2017. https://doi.org/10.1163/9789004349308.

Open Islamicate Texts Initiative 2019 Open Islamicate Texts Initiative (OpenITI). "The Open Islamicate Texts Initiative Arabic-Script OCR Catalyst Project (OpenITI AOCP)." *Medium*, August 29, 2019. https://medium.com/@openiti/openiti-aocp-9802865a6586.

Ragnedda 2019 Ragnedda, Massimo. "Conceptualising the Digital Divide." In *Mapping the Digital Divide in Africa: A Mediated Analysis*, edited by Bruce Mutsvairo and Massimo Ragnedda, 27–44. Amsterdam: Amsterdam University Press, 2019. https://doi.org/10.2307/j.ctvh4zj72.6.

Reeve 2015 Reeve, Jonathan. "Introducing Git-Lit." *Jonathan Reeve: Computational Literary Analysis*, September 8, 2015. https://web.archive.org/web/20151021061552/http://jonreeve.com/2015/09/introducing-git-lit/.

Rifāʿī 1969 Rifāʿī, Shams al-Dīn al-. *Tārīkh al-ṣiḥāfa al-Sūriyya* [History of the Syrian Press]. 2 vols. al-Qāhira: Dār al-Maʿārif bi-Miṣr, 1969.

Risam 2018 Risam, Roopika. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston: Northwestern University Press, 2018. https://doi.org/10.2307/j.ctv7tq4hg.

Seikaly 1981 Seikaly, Samir. "Damascene Intellectual Life in the Opening Years of the 20th Century: Muhammad Kurd ʿAli and Al-Muqtabas." In *Intellectual Life in the Arab East, 1890-1939*, edited by Marwan Rafat Buheiry, 125–53. Beirut: American University of Beirut, 1981.

Shaffer 2013a Shaffer, Kris. "Push, Pull, Fork: GitHub for Academics." *Hybrid Pedagogy* /, May 26, 2013. https://web.archive.org/web/20170425194229/http://www.digitalpedagogylab.com/hybridped/push-pull-fork-github-for-academics.

Shaffer 2013b Shaffer, Kris. "GitHub for Academics: The Open-Source Way to Host, Create and Curate Knowledge." *Impact of Social Sciences*, June 4, 2013. https://web.archive.org/web/20160625163656/http://blogs.lse.ac.uk/impactofsocialsciences/2013/06/04/github-for-academics/.

Sharīf 1911 Sharīf, Ṣāliḥ al-."Nuṣīḥa li-l-Yamāniyīn" [Advice to the Yemenites]. *al-Ḥaqāʾiq* 1.11 (May 30, 1911). https://OpenArabicPE.github.io/journal_al-haqaiq/tei/oclc_644997575-i_11.TEIP5.xml#div_4.d1e704.

Strubell, Ganesh, and McCallum 2019 Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and Policy Considerations for Deep

Learning in NLP." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–50. Florence, Italy: Association for Computational Linguistics, 2019. https://doi.org/10/ggbgzx.

TEI Consortium 2020 TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (version 4.0.0). XML. *Zenodo*. Accessed October 14, 2021. https://doi.org/10.5281/zenodo.3413524.

Ṭarrāzī, Fīlīb dī. 1914 Ṭarrāzī, Fīlīb dī. *Tārīkh al-ṣiḥāfa al-ʿArabiyya* [History of the Arabic Press]. 3 vols. Bayrūt: al-Maṭbaʿa al-Adabiyya, 1914.

Terras 2016 Terras, Melissa."Crowdsourcing in the Digital Humanities." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 420–38. Chicester: Wiley, 2016. https://doi.org/10.1002/9781118680605.ch29.

Thylstrup 2018 Thylstrup, Nanna Bonde. *The Politics of Mass Digitization*. Cambridge, MA: The MIT Press, 2018.

Tikrītī 1969 Tikrītī, Munīr Bakr. *al-Ṣiḥāfa al-ʿIrāqiyyqa wa-ittijāhātuhā al-siyāsiyya wa-l-ijtimāʿiyya wa-l-thaqāfiyya min 1869-1921* [The Iraqi Press and its Political, Social, and Cultural Trends between 1869 and 1921]. Baghdād: Maṭbaʿat al-Irshād, 1969. http://hdl.handle.net/2333.1/gf1vhp1w.

Townsend 2013 Townsend, Anthony M. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. New York: Norton, 2013.

Turska, Cummings, and Rahtz 2016 Turska, Magdalena, James Cummings, and Sebastian Rahtz. "Challenging the Myth of Presentation in Digital Editions" *Journal of the Text Encoding Initiative* 9 (September 2016). https://doi.org/10/ghh5sx.

Verkinderen 2020 Verkinderen, Peter."Al-Maktaba Al-Shāmila: A Short History" *KITAB*, December 3, 2020. http://kitab-project.org/2020/12/03/al-maktaba-al-shamila-a-short-history/.

Walsh and Simpson 2013 Walsh, John and Grant Leyton Simpson. "TEI Boilerplate" *Journal of Digital Humanities* vol. 2.3 (2013). http://journalofdigitalhumanities.org/2-3/tei-boilerplate/.

Walsh et al. 2016 Walsh, John, Grant Simpson, Saeed Moaddeli, Till Grallert, and Gioele Barabucci. *TEI Boilerplate* (version 1.1.0). XSLT. Accessed October 14, 2021. https://github.com/TEI-Boilerplate/TEI-Boilerplate.

Wittel 2013 Wittel, Andreas. "Counter-Commodification: The Economy of Contribution in the Digital Commons" *Culture and Organization* vol. 19.4 (2013): 314–31. https://doi.org/10/gmqgqq.

Wittern 2013 Wittern, Christian. "Beyond TEI: Returning the Text to the Reader." *Journal of the Text Encoding Initiative* 4 (2013). http://jtei.revues.org/691.

Wrisley and Jarkas 2016 Wrisley, David Joseph and Najla Jarkas. "On Translating Voyant Tools into Arabic." *David Joseph Wrisley*, June 9, 2016. https://web.archive.org/web/20190704164337/http://djwrisley.com/on-translating-voyant-tools-into-arabic/.

Wrisley and Jarkas 2019 Wrisley, David Jospeh and Najla Jarkas. "RTL Software Localization and Digital Humanities: The Case Study of Translating Voyant Tools into Arabic." Presented at the Digital Humanities Summer Institute: Right2Left Workshop, June 8, 2019.

Yalman 1914 Yalman, Ahmet Emin. "The Development of Modern Turkey as Measured by Its Press." PhD dissertation, 1914. Columbia University.

Zakham 1907 Zakham, Yūsuf."Amīrkā wa-ʿulamāʾ al-ʿArab." [America and Arab Scholars]. *al-Muqtabas* vol. 2.11 (February 14, 1907). https://OpenArabicPE.github.io/journal_al-muqtabas/tei/oclc_4770057679-i_13.TEIP5.xml#div_8.d1e1249.

Zemmin 2016 Zemmin, Florian. "Modernity Without Society? Observations on the Term Mujtamaʿ in the Islamic Journal Al-Manār (Cairo, 1898–1940)." *Die Welt Des Islams* vol. 56.2 (2016): 223–47. https://doi.org/10/ggwwhh.