Algorithmic Analysis of Medieval Arabic Biographical Collections

By Maxim Romanov

> Biographical dictionaries seem to be, for the researchers in the Islamic Arabic library, both a blessing and a curse.
>
> Wadād al-Qāḍī[1]

Introduction

With at least forty thousand unique titles identifiable for the period before 1900 CE (see below), the Arabic written tradition is one of the greatest treasuries of knowledge in human history. Covering practically every aspect of Islamic culture, this tradition is particularly rich in extensive historical sources such as chronicles and biographical collections. Numbered in the hundreds, these multivolume texts cover practically every aspect of Islamic history and culture: from conquests, dynastic vicissitudes, and urban unrest, to food prices, long-distance trade, plagues, and natural disasters, as well as practically anything imaginable in between. The overall volume of individual titles is often equally astonishing. One of the largest surviving texts, "The History of Islam" (*Taʾrīḫ al-islām*) of al-Ḏahabī, a fourteenth-century Damascene scholar, is a mammoth of Arabic biographical literature that covers seven centuries of Islamic history (c. 600–1300 CE) through over thirty thousand biog-

I would like to express my gratitude to Sarah Bowen Savant and Matthew Thomas Miller, my dear colleagues and collaborators within the Open Islamicate Texts Initiative, who read the draft of the article and made a great number of valuable suggestions; to three anonymous reviewers, whose comments helped to further improve my article; to Gregory Ralph Crane, who has been supporting digital studies of Arabic written tradition first at Tufts University (2013–15) and then at Leipzig University (2015–2017).

Note on transliteration: The article uses a somewhat unconventional transliteration system, which was developed to facilitate computational analysis. Unlike more traditional transliteration schemes the current one uses strict one-to-one representation, with every Arabic letter and short vowel transcribed distinctively, which allows for an automatic conversion between transliteration and the Arabic script. The overall scheme should be easily recognizable to Arabists (new letters are as follows: *ŧ* for *tāʾ marbūṭat*; *ã* for dagger *alif*; and *á* for *alif maqṣūrat*). Additionally, all attached conjunctions, prepositions, and pronominal suffixes are separated with "-". Whenever applicable, toponyms are given in their current American spelling. Bibliographical references and quotations preserve their original transliteration schemes.

[1] Wadād al-Qāḍī, "Biographical Dictionaries: Inner Structure and Cultural Significance," in *The Book in the Islamic World: The Written Word and Communication in the Middle East*, ed. George N. Atieh (Albany, 1995), 93.

Speculum 92/S1 (October 2017).
DOI: 10.1086/693970, 0038-7134/2017/92S1-0009$10.00.

raphies and about ten thousand descriptions of historical events—spanning over fifty volumes (approximately 3.4 million words) in one of its modern editions.[2] The overall number of biographical records in texts that are currently available digitally already exceeds four hundred thousand. Together these narrative sources make up the richest gold mine of information on Islamic history and culture, and they are particularly important for the period prior to the fifteenth century, for which very few primary documents and archives are available.

For decades, scholars of Islamic history have recognized the value of these sources. The potential of the quantitative approach to these sources has been conceptualized and demonstrated by several scholars of Islamic history who worked independently in different countries in the 1970s and 1980s.[3] However, the excessive volume of even individual titles posed a formidable challenge. Previous attempts to study Arabic historical sources relied on the use of mechanically sortable index cards, punch cards, early computers that stored data on magnetic tapes, and, most recently, relational databases. None of these approaches, however, allowed one to surpass the bottleneck of data extraction and processing, and the methods remained extremely time-consuming. The number of data-driven studies is still small, and the potential of the approach has not been realized.[4] The unfathomable Arabic biographical and historical texts became "both a blessing and a curse."

This remains largely the status quo, but the recent digital turn offers new opportunities. In the course of the past decade thousands of premodern Islamic texts have become available in full-text digital formats through a number of online open-access libraries, while at the same time rapid development of computational methods of text analysis has provided a number of novel approaches for studying large textual corpora. These two developments help to overcome the main limitations of conventional quantitative historical studies and take full advantage of the information trapped in these voluminous texts. First, modern computers make data extraction fast, scalable, and flexible; second, computational methods of text analysis allow one to maintain solid connections between collected quantifiable data and full-text passages that deal with specific instances, and thus bring together quantitative and qualitative dimensions of analysis—or, in other words, distant and close reading.

The focus of the current article is on a method of algorithmic analysis of Arabic biographical collections. What is understood here by algorithmic analysis is

[2] Al-Ḏahabī, *Taʾrīḫ al-islām wa-wafayāt al-mašāhīr wa-al-aʿlām*, ed. ʿUmar Tadmurī, 2nd ed., 52 vols. (Beirut, 1990).

[3] See Richard W. Bulliet, "A Quantitative Approach to Medieval Muslim Biographical Dictionaries," *Journal of the Economic and Social History of the Orient* 13/2 (1 April 1970): 195–211; Stanislav M. Prozorov and Maxim G. Romanov, "Metodika izvlecheniya i obrabotki informatsii iz arabskih istochnikov (na materiale istoriko-biograficheskoi literaturi)" [Principles and procedures of extracting and processing the data from Arabic sources (based on materials of historical-cum-biographical literature)], *Oriens/Vostok* 4 (2003): 117–27.

[4] For a detailed discussion of previous approaches to biographical data in Arabic and Islamic studies, see Maxim G. Romanov, "Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunnī World (661–1300 CE)" (PhD diss., University of Michigan, 2013), 51–58.

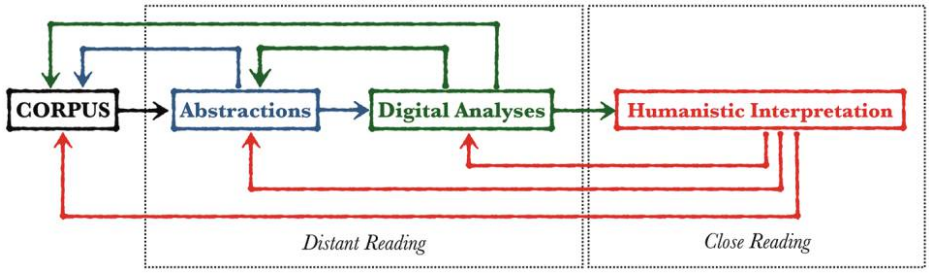

Fig. 1. The iterative nature of algorithmic analysis

a step-by-step reduction of a text in a natural language to a machine-readable abstraction, which is then followed by the analysis of shapes, relations, and structures.[5] More generally, algorithmic analysis can be viewed—as shown in Fig. 1—as an iterative process of engagement with texts, their abstractions, and their interpretations, where preliminary results of later steps of the loop can suggest how one can improve earlier steps to attain better results. Although my focus here is on this particular genre, similar approaches can be applied to any type of text in any language as long as it displays some internal regularity that can be identified and exploited for the reduction of the initial text to a machine-readable abstraction. It should be noted, however, that the method will be most effective for extensive collections of information units with similar internal structures. In the context of the Arabic written tradition, this method can be used most effectively with lexicographical dictionaries, collections of legal decisions (sing. *fatwá*), gazetteers, comprehensive geographies, interpretations of the Qurʾān, collections of the sayings of the Prophet (Ḥadīṯ), bibliographies, and other dictionary-like texts.

The Workflow

The process of algorithmic analysis involves a series of steps, which can be summed up as follows: (1) finding the machine-readable text of a book; (2) tagging the logical structure of the book; (3) tagging—manually and semiautomatically—relevant data in the structured text (alternatively, extracting relevant data automatically); (4) extracting and modeling tagged data; (5) visualizing and analyzing results. All these

[5] I found Stephen Ramsay's notion of "algorithmic criticism" particularly inspiring for my thinking about algorithmic analysis. See Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (Urbana, 2011). Franco Morretti's and Matthew Jockers's work has been equally thought provoking and inspiring: Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London, 2007); Franco Moretti, *Distant Reading* (London, 2013); Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana, 2013). The method was first presented several years ago when it was at the very early stage of development as part of a dissertation project, where it was then first implemented: see Romanov, "Computational Reading." This article offers an overview of the method in its most recent form and describes relevant ongoing work aimed at scaling up the approach. See Maxim Romanov, "Toward the Digital History of the Pre-Modern Muslim World: Developing Text-Mining Techniques for the Study of Arabic Biographical Collections," in *Analysis of Ancient and Medieval Texts and Manuscripts: Digital Approaches*, ed. Tara Andrews and Carolin Macé (Leuven, 2014), 229–44. This publication summarizes the status of the method in 2012.

steps occur in this order only procedurally, but not necessarily on the conceptual level, where one finds oneself constantly thinking about what kind of data can be extracted from a given source—and how exactly—and what kind of processes it can help to model.

Step 1

Finding an electronic text of a medieval Arabic book has become rather easy over the past decades as a number of open-access electronic libraries have appeared in the Middle East.[6] Most of these libraries are repositories of text files in a variety of formats—usually HTML, TXT, or MS Word—and offer no analytical tools, except for basic search capabilities. Before proceeding any further, one must collate the found text with the printed edition on which it is based in order to establish its overall adequacy. In most cases, these electronic texts are high-quality reproductions of printed editions (they seem to be produced with the double-keying method)[7] and, for this reason, inherit all the potential and real issues of critical editions of the printed era. It is also worth stressing here that most of these digital texts are based on printed editions that are widely used in the field of Arabic and Islamic studies.

Step 2

The tagged logical structure offers one an ability to work with every logical unit of a book on the machine level. To provide this structural tagging, I am using a lightweight scheme of my own design, whose current version is named Open-Arabic mARkdown (more on it in the final section). Built on regular expressions[8] and implemented in EditPad Pro (https://www.editpadpro.com/), the scheme offers the dynamic highlighting of tagging patterns and the folding of a tagged text into a table of contents. In a nutshell, this tagging task can be described as (1) collating the electronic text with the relevant printed edition, and (2) ensuring that all words of chapter headers are on the same line and prepended with relevant mARkdown tags. For example, chapter headers of the first level receive the tag "### |"; those of the second, "### ||"; those of the third, "### |||", and so on. Logical units of specific types have their own patterns.

[6] Altogether, the libraries that I was able to survey include over 30,000 texts. The largest online libraries are *al-Maktabat al-šāmilat* (http://www.shamela.ws, 6,300 texts); *al-Mishkāt* (http://www.almeshkat.net, 7,300 texts); *Ṣayd al-fawāʾid* (http://www.saaid.net, 10,000 texts); *al-Warrāq* (http://www.alwaraq.com, 860 texts); and *al-Maktabat al-šīʿat* (http://www.shiaonlinelibrary.com, 1,970 texts). Other digital libraries are available on CDs, DVDs (for example, *al-Muʿjam al-fiqhī*, Qom, Iran, 1,130 texts), and even external HDD (*al-Jāmiʿ al-kabīr*, ʿAmmān, Jordan, 2,400 texts).

[7] The double-keying transcription method is confirmed to be the most accurate digitization approach: see Susanne Haaf, Frank Wiegand, and Alexander Geyken, "Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text," *Journal of the Text Encoding Initiative* 4 (8 March 2013), doi:10.4000/jtei.739.

[8] An integral part of most programming languages, "regular expressions" is a minilanguage for describing search patterns. For more details, see http://www.regular-expressions.info/.

Fig. 2. Automatically tagged entities in a biography: "@SOC01" for "descriptive names" (*nisbats*) that behave as social markers; "@TOP01" for place names (toponyms); "@YY167" for dates (year statements). You can see a mild problem with what happens when right-to-left and left-to-right languages appear in the same document: the order of symbols in tags appears different ("SOC01@", "TOP01@", "YY167@"), but the logical order remains correct.

Step 3

After the structure is tagged, one can either design a data-extraction routine or manually tag needed information. A combination of automatic tagging (using entity lists) and manual disambiguation offers perhaps the optimal solution. Fig. 2 shows an example of an automatically tagged biography using entities lists for toponyms and "descriptive names"; year statements are rather regular in classical Arabic and can be identified with regular expressions and converted into numbers.[9] The morphology of tags is as follows: the tag starts with @, which is followed by SOC or TOP, which introduces the category of an entity, and concludes with two numbers—the first one marks the length of any prefix that should be dropped[10] and the second the length of an entity in words. When the tag is properly entered in front of the necessary word or the word group (up to three words), it is dynamically highlighted. Automatically inserted tags are highlighted in black.

To avoid false positives, automatic tags can be disambiguated in the manner shown on Fig. 3: "@SOC01" becomes "@S01"; "@TOC01" becomes "@T01"; "@YY167" becomes the year of death "@YD167". These are manual variations of automatic tags for the same categories—they are shorter (to make manual tagging easier and faster) and are highlighted with different colors for the ease of visual recognition. To make things more accessible, Fig. 4 shows the same information in English.

Note that automatic tagging of toponyms can be enhanced by inserting a relevant URI, a Uniform Resource Identifier, of a relevant toponym from a gazetteer of Islamic places. In our particular case, this will be al-Ṭurayyā Gazetteer (https://althurayya.github.io/), which we have developed with this purpose in mind. The URIs in the gazetteer are designed to be human readable and aid in disambiguation of complicated cases, such as, for example, Ṭarābulus/Aṭrābulus—the toponym that may refer either to the city of Tripoli in North Africa (sometimes appearing in the sources as Ṭarābulus al-Ġarb, "Tripoli of the West") or to Tripoli in Levant (sometimes appearing in the sources as Ṭarābulus al-Šarq, "Tripoli of the East"); in such an ambiguous case, the URIs of both places can automatically

[9] Date statements can offer a valuable insight into a large Arabic corpus as well as specific books: see my blog post "Chronological Coverage of an Arabic Corpus: An Experiment with Date Statements," https://alraqmiyyat.github.io/2016/03-29.html.

[10] Certain prepositions (*wa-*, *fa-*, "and") and conjunctions (*li-*, "for"; *bi-*, "in, with," etc.) are attached to words in Arabic.

Fig. 3. Manually disambiguated tagged entities in a biography: "@S01" for "descriptive names" (*nisbats*); "@T01" for place names (toponyms); "@YD167" for the date of death (year statements).

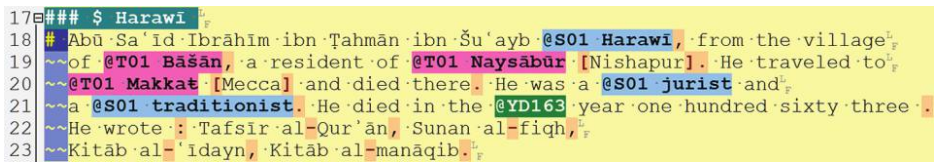

Fig. 4. Manually disambiguated tagged entities in the translated version of the Arabic biography given above.

be inserted and flagged for disambiguation. A quick glance at the text is usually enough to determine which of two Tripolis is referred to, and the coordinates encoded into the URIs—"ATRABULUS_131E328N_S" and "ATRABULUS_358E344N_S"—help one to decide which of the URIs must be removed (the first URI is for the city in North Africa; the second, for that in Levant). The use of URIs in the process of tagging allows one to pull out all available information on tagged places from the gazetteer, such as transliteration of the place name, its coordinates, settlement type categorization, and regional classification; for example, "ATRABULUS_358E344N_S" can be transformed into *Aṭrābulus*, a *town* in the region of *al-Šām (Greater Syria)* with the coordinates *34.4 LAT*, *35.8 LON*.

Step 4

Now that we have our data tagged automatically and, ideally, disambiguated, one can proceed to extracting, enriching, and modeling this data. In terms of modeling some explanations are required, particularly to those not familiar with Islamic history.[11] Traditional Arabic biographies usually include three major markers—chronological, geographical, and onomastic/social—which can be used in a variety of distant-reading modes of analysis. In the example above we have all three of them: (1) dates—in our case, the year of death, 163/780 CE ("@YD163"); (2) locations with which this person is associated—the village of Bāšān ("@T01 Bāšān"), the city of Herat ("@S01 Harawī"), the city of Nishapur ("@T01 Naysābūr"), and the city of Mecca ("@T01 Makkat"); and (3) "descriptive names" (sing. *nisbat*)—a "jurist" ("@S01 faqīh [jurist]"), a "traditionist" ("@S01 muḥaddiṯ [traditionist]", a specialist in the study and transmission of "the words of the Prophet"), and, again, a Herati

[11] One should think of modeling, to quote Willard McCarty, as "a continual process of coming to know by manipulating representations." See his "Modeling: A Study in Words and Meanings," in Susan Schreibman, Ray Siemens, and John Unsworth, *A New Companion to Digital Humanities*, 2nd ed. (Chichester, UK, 2016), http://www.digitalhumanities.org/companion/; and, more extensively, Willard McCarty, *Humanities Computing* (Basingstoke, UK, 2014), 20–72.

("@S01 Harawī"), a person who is strongly associated with the city of Herat (in this particular case, this person got the name "Herati" because he comes from the village of Bāšān in the district of Herat/Harāt). From this profile we get this person's terminus ante quem; we can construct his geographical network (on the level of settlements, and—through the gazetteer—on the level of regions); and we also know what kind of religious specialization he had and to what geographical community he belonged (onomastic data often also provide social, professional, occupational, communal, and ethnic markers). Combining thousands of such biographical profiles and subsetting them with different parameters, we can get detailed insights into chronological and geographical patterns of a variety of social, religious, and professional groups that can be identified in a specific biographical collection.

It should be noted that the tagging does not have to be limited to these three types of markers and can be extended in a similar manner to any other relevant category, especially around linguistic patterns. For example, one can tag specific phrases that describe biographees' religious training, tag people they studied under, or, as immediately relevant to our example, tag books that they composed.

In terms of extraction, with a relatively simple script (in my case, written in Python), one can automatically extract tagged data from all biographies and convert them into a format suitable for further analysis and visualization. The script performs the following: (1) it numbers all biographies sequentially (using biographical tags, "### $", as anchors); (2) it splits the entire text of the book into individual biographies; (3) it extracts all tagged items with regular expressions and reformats everything into a CSV-format file, where the abstraction of our biography will look as follows (assuming we used URIs from the gazetteer to tag and disambiguate toponyms):[12]

```
======================
id, item, category
======================
000006, 163, year_of_death
000006, BASHAN_623E342N_S, toponym
000006, NAYSABUR_587E361N_S, toponym
000006, MAKKA_398E213N_S, toponym
000006, harawī, descriptive_name
000006, faqīh, descriptive_name
000006, muḥaddiṯ, descriptive_name
======================
```

Converted into such a format, our data can now be enriched, reshaped, and reorganized into subsets for a variety of research questions. (Further data manipulations and visualizations are performed in R, https://www.r-project.org/.) The easiest example of the enrichment of our data can be given on geographical information: now that we have our geographical data in the form of URIs, we can add

[12] Additional lists of aliases should also be constructed and used in order to unify different forms of the same words or different names of the same entities. For example, such lists allow us to unify different names commonly employed for Baġdād (Madīnaṯ al-salām, Baġḏād, and Baġdād); the same approach, amplified with some scripting, can also be used to unify various morphological forms of the same words. Arabic morphology is particularly challenging because of a plethora of attached prefixes and suffixes, which in different, often stackable, combinations can multiply words into over fifty variations; existing morphological analyzers do not yet offer a reliable solution, especially for classical Arabic.

additional geographical information from the gazetteer, such as coordinates and regions; regions are particularly relevant, since they will allow one to move between local and regional levels of data analysis. In a similar way, a detailed onomastic table can provide broader categories for conducting analysis on a higher level: for example, descriptive names like *faqīh*, "jurist"; *qāḍī*, "judge"; and *muftī*, "jurisconsult" can be thus combined into a broader category of "legal professions," and one can then graph and map both specific legal professions as well as all legal professionals together as a broader category.

The filtering and subsetting of the enriched data then can be performed in the following manner: (1) One first identifies a type (let's say "jurisconsults") or a broader category (in this case, the corresponding category will be "legal professions") and filters the data set using the selected value. (2) The filtered results will have the IDs of all the biographees associated with the selected type or the broader category, and this list of IDs can be used to resubset the main data set to get all relevant chronological, geographical, and onomastic markers. (4) Aggregating chronological markers, we can now build a graph of the temporal distribution of jurisconsults or legal professionals more generally. (5) Aggregating geographical markers, we can build a cartogram of their spatial distribution. (6) Combining chronological and geographical markers, we can also build cartograms of spatial distribution for different chronological periods that trace the spatial dynamic of their distribution over time. (7) Combining geographical data further, we can also build cartograms of interregional connections, and also trace how the configuration and density of these connections was changing over time.

Keeping in mind that all biographies now have IDs, one can easily go back and forth between distant and close reading of relevant biographies, thus improving the outcome of both.

Analysis

With several proposed exploratory visualizations, we can now take a closer look at the source of our short biography in its entirety. The text in question—the *Hadiyyat al-ʿārifīn* (The Gift to the Knowledgeable)—is a biobibliographical collection written by Ismāʿīl Bāšā al-Baġdādī (d. 1338/1919 CE). Although de facto the text is modern, it follows very closely in the footsteps of medieval texts of this kind and is effectively part of the tradition; additionally, chronologically, we get the most extensive coverage from this collection, as it covers the period from the beginning of Islam in the seventh century CE up to the end of the nineteenth century.

From the very little that we know about him,[13] Ismāʿīl Bāšā wrote two extensive bibliographical texts: the first one, *Īḍāḥ al-maknūn fī Ḏayl ʿalá Kašf al-ẓunūn*, is the continuation of the famous *Kašf al-ẓunūn* of Ḥāǧī Ḫalīfaṯ (d. 1067/1656 CE),[14]

[13] See J. J. Witkam, "Ismāʿīl Pasha Baghdādlî," in *Encyclopaedia of Islam (EI2-Online)*, 2nd ed., ed. P. Bearman, Th. Bianquis, C. E. Bosworth, E. van Donzel, and W. P. Heinrichs (Malden, MA, 2016), available online at http://referenceworks.brillonline.com/browse/encyclopaedia-of-islam-2. For the edition of this text, see Ismāʿīl Bāšā al-Baġdādī, *Hadīyat al-ʿārifīn asmāʾ al-muʾallifīn wa-aṯār al-muṣannifīn*, 6 vols. (Beirut, 1992).

[14] He is also known as Kātib Čelebi; see Orhan Şaik Gökyay, "Kātib Čelebi", in Bearman et al., *EI2–Online*.

and mirrors its structure—a dictionary of book titles, organized alphabetically; the second one is the *Hadiyyaṯ al-ʿārifīn*, which contains essentially the same information, but grouped into biographical records, where all works attributed to a given author are listed after a short biography.[15] The *Hadiyyaṯ al-ʿārifīn* is organized alphabetically, and then chronologically within each letter.[16]

Although one cannot possibly expect such a collection to be comprehensive and exhaustive, this is the largest bibliography of books written in the Islamic world that we have available. So, we can still hope to get valuable insights into cultural production—the appearance of new works—in the Islamic world up until the beginning of the twentieth century. For the sake of space and the mere fact that the analysis of this collection deserves a separate study, I will focus on broad spatial and chronological patterns that can be discerned in the data.

Insight 1: Cultural Production over Time

First of all, our algorithmic analysis allows us to get a better understanding of the overall coverage of the collection itself: it includes almost 8,800 authors and over 40,000 book titles—with most authors being attributed 1 to 4 titles (interquartile range). The overall chronological distribution of authors (Fig. 5) displays a steady upward trend up to 1200/1785 CE, reflecting the general historical situation: as the Islamic world keeps expanding geographically and the Muslim population growing, we find more individuals getting involved in the process of cultural production.

Displaying the same trend for the period up to 1200/1785 CE, the graph of books (Fig. 6) makes the prominent early period (200–450 / 815–1058 CE) more noticeable. Although this period is usually strongly associated with the translation movement from Greek into Arabic,[17] it is probably even more important for the formation of Islam as a religious system—particularly for the development of the Ḥadīṯ canon[18] and the crystallization of theological views.[19] Spikes are also due to a few very prominent polymaths: al-Suyīṭī (d. 911/1505 CE)—585 works; Ibn ʿArabī

[15] The majority of works listed in the *Hadiyyaṯ al-ʿārifīn* are in Arabic, "the Latin of the Islamic world," although about 10 percent of books are written in Persian and Turkish (either the language is explicitly mentioned, or the title of the work includes a Persian or Turkic word—most commonly, *nāmah*, Pers./Turk. "book"); Persian and Turkish works are not excluded from the analysis.

[16] It is worth pointing out here that, when it comes to biographical material, alphabetical organization is secondary in Islamic culture; the primary form of organization would be chronological, divided into "generations" or "cohorts" (sing. *ṭabaqaṯ*)—authors of later generations would often take this information, and edit, supplement, and reorganize it alphabetically. See Franz Rosenthal, *A History of Muslim Historiography* (Leiden, 1952): al-Saḫāwī's *al-Iʿlān bi-l-tawbīḫ*, translated in Rosenthal's book, is particularly rich on notes about who updated and reorganized whose work.

[17] Dimitri Gutas, *Greek Thought, Arabic Culture: The Graeco-Arabic Translation Movement in Baghdad and Early ʿAbbāsid Society (2nd–4th/8th–10th Centuries)* (London, 1998).

[18] See, for example, "Phase 3: The Age of 'Six Books' (c. 200–400/912–1009)," in Scott C. Lucas, *Constructive Critics, Ḥadīth Literature, and the Articulation of Sunnī Islam: The Legacy of the Generation of Ibn Saʿd, Ibn Maʿīn, and Ibn Ḥanbal* (Leiden, 2004), 73–86.

[19] According to the *Hadiyyaṯ al-ʿārifīn*, about 90 percent of almost five hundred "refutations" (Ar. *radd*) of different groups and specific beliefs were written during this period (peaking 250–450/864–1058).

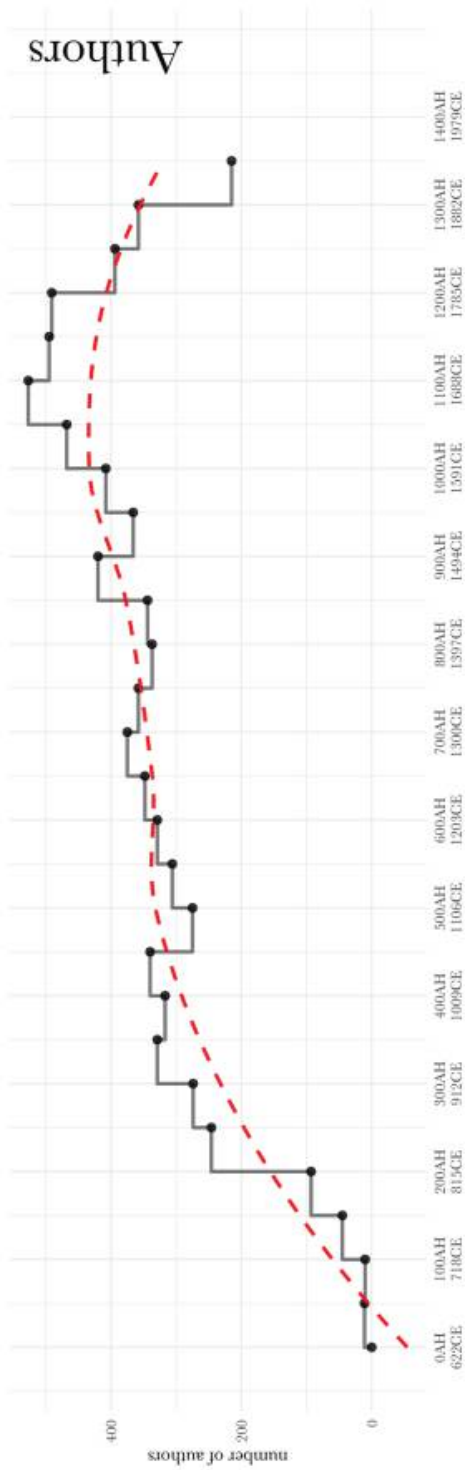

Fig. 5. Chronological distribution of authors

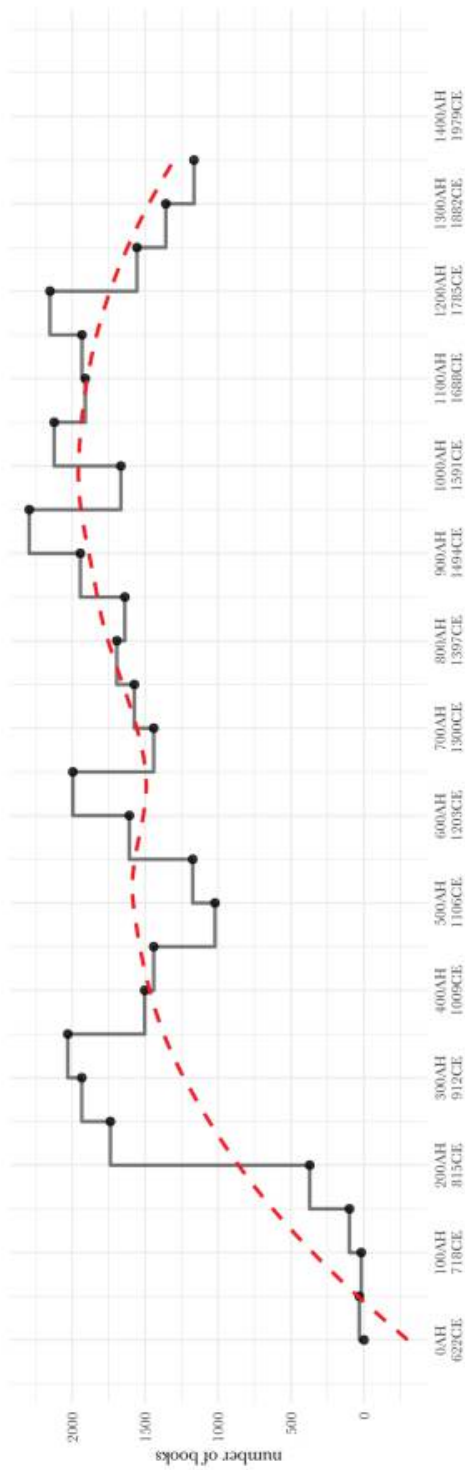

Fig.6. Chronological distribution of books

(d. 638/1240 CE)—425 works; al-Kindī (d. 256/870 CE)—256 works; al-Madāʾinī (d. 225/840 CE)—223 works; al-Nābulusī (d. 1143/1730 CE)—204 works; Ibn al-Jawzī (d. 597/1200 CE)—201 works, and quite a few other prolific authors.

The decline of both graphs after 1200/1785 CE most likely indicates the unavailability of bibliographical information for our author. The geographical coverage of the collection also starts shrinking at roughly the same period. It should be noted that all chronological data sets tend to exhibit this trend. For example, the trend can be observed in al-Ḏahabī's own continuation, *Ḏayl*, to his massive "The History of Islam" (*Taʾrīḫ al-islām*), where the number of biographies per period drops dramatically. One can equally see this in Brill's bibliographical database *Index Islamicus* as well as in *Harvard Open Metadata* on the 12 million books that Harvard libraries hold. The only difference is that the lag gets shorter as we get closer to our time—for premodern Arabic sources this lag is 100 to 150 years; in modern data sets, 10 to 20 years.

Splitting our data geographically—see Fig. 7—we can also discover which regions played leading roles in cultural production. What we discover from the results is that, as we suspected, the collection does not cover all the regions of the Islamic world, particularly regions that became part of the Islamic world in the later periods and in geographical terms remained peripheral to the core: sub-Saharan Africa, the Indonesian archipelago, the Volga region, and Eastern Europe. At the same time, all core regions—the historical heartlands—of the Islamic world are covered quite well.

It should be pointed out that the bar chart here shows the *presence* of authors in those regions, as many of them traveled (sometimes extensively) and composed their books at different locations. In other words, our biographee—who lived in Nishapur, but died in Mecca—appears both in the column of Iran (Īrān) and that of Arabia (Jazīrat al-ʿarab). Such treatment of data is also justified because regions in their prime tend to attract people from less prosperous ones.

We can get a better understanding of regional contributions by graphing regional data chronologically. Fig. 8 shows the top five contributing regions: Anatolia (Rūm), Iraq (al-ʿIrāq), Iran (Īrān), Syria (al-Šām), and Egypt (Miṣr) are homes to the highest number of individuals engaged in cultural production across the Islamic world. The chronological distribution of authors in those regions (as well as in the regions that are not graphed here) display a rather distinct pattern: cultural production is on the rise during economic and political stability, usually marked by the early rule of strong dynasties: the ʿAbbāsids in Iraq; the dynasties of the "Iranian intermezzo," followed by the Tīmūrids and the Ṣafawids, in Iran; the Mamlūks in Syria and Egypt;[20] the Ottomans in Anatolia. It should be noted, however, that the increase in cultural production in these cases is not necessarily due to rulers' patronage but rather due to the stability and predictability of social and economic life that their rule brings about. Although many rulers did act as patrons

[20] The rule of the Fāṭimids in Egypt marked the shift in the ideology—from Sunnism to Ismāʿīlī Shiʿism—which featured the rise in numbers of Ismāʿīlī writings; however, these numbers are overshadowed by the decline in Sunnī writings—as well as of Sunnī communities in general—in Egypt. On Ismāʿīlī authors, see Ismail K. Poonawala and Teresa Joseph, *Biobibliography of Ismāʿīlī Literature*, Studies in Near Eastern Culture and Society (Malibu, 1977), 467–69.

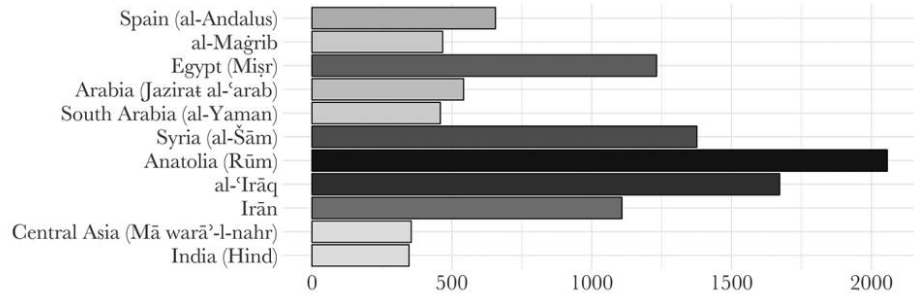

Fig. 7. Regional Contributions

of "fine literature," most books in the *Hadiyyat al-ʿārifīn* deal with religious subjects—Qurʾānic exegesis, "words of the Prophet" (*Ḥadīṯ*), Islamic law, and so on—and they were composed in the framework of the development of local religious communities, whose florescence depended more on overall political and economic stability. In this regard, the example of Iraq may be quite telling: the early period of ʿAbbāsid rule is marked by a very significant rise, which comes to a halt when the ʿAbbāsids lose their sovereignty and become the puppets, first, of their generals, then of the Būyids, and then of the Saljūqs, regaining their power only briefly at the end of their rule, which is ended dramatically by the Mongol invasion. Needless to say, the real historical picture is always more complicated than the space of this article allows.

Insight 2: Cultural Connections

Our biobibliographical data also offers a significant amount of geographical information with which one can model geographical networks of connections. A network of an individual can be represented by connecting all places mentioned in that individual's biography. Fig. 9 shows the geographical network from our sample biography, where possible paths are generated from the route network of that period using the shortest path (the Dijkstra algorithm) and the optimal path (a modified Dijkstra algorithm that avoids stretches with a small number of settlements along the way). For our purposes, however, a somewhat more simplified approach for modeling the network will work better. First of all, we want to move from the level of settlements to the level of regions: they become the nodes, which are connected with each other directly—as the crow flies—without using route networks.[21] In the case of our sample biography, the network is thus simplified to a single arc between Iran and Arabia. One can then combine route networks of a particular group of individuals in order to see a broader pattern. Arguably, by combining individual networks from specific periods—with every

[21] The problem with route networks is that they change over time, and it is very difficult to recreate route networks for all the periods covered in our collection; more importantly, however, route networks will forefront the most traveled sections of the network, rather than the density of connections among the regions.

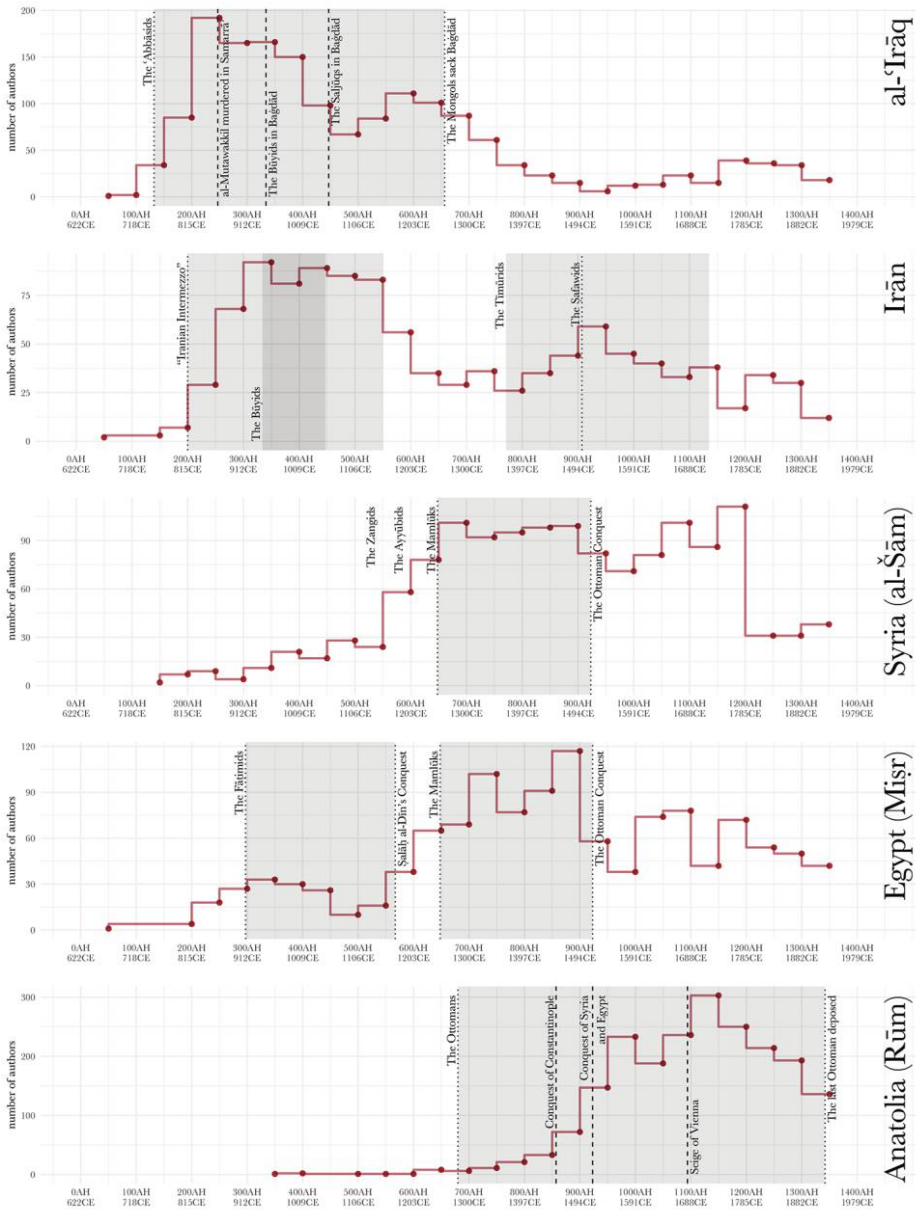

Fig. 8. Most prominent Islamic regions over time

shared node becoming bigger, and every shared edge thicker—one can get an idea of how the Islamic world was connected in that particular period, and more interestingly, what constituted its core: namely, the constellation of the most prominent and interconnected regions.

Fig. 9. Geographical network of the biographee from the sample biography (using our al-Ṭurayyā Gazetteer, https://althurayya.github.io/).

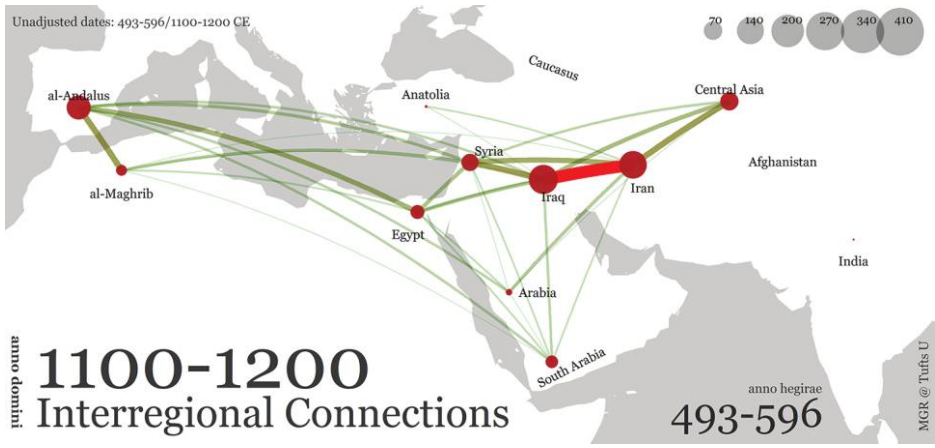

Fig. 10. The Iraqi-Iranian core in the twelfth century CE

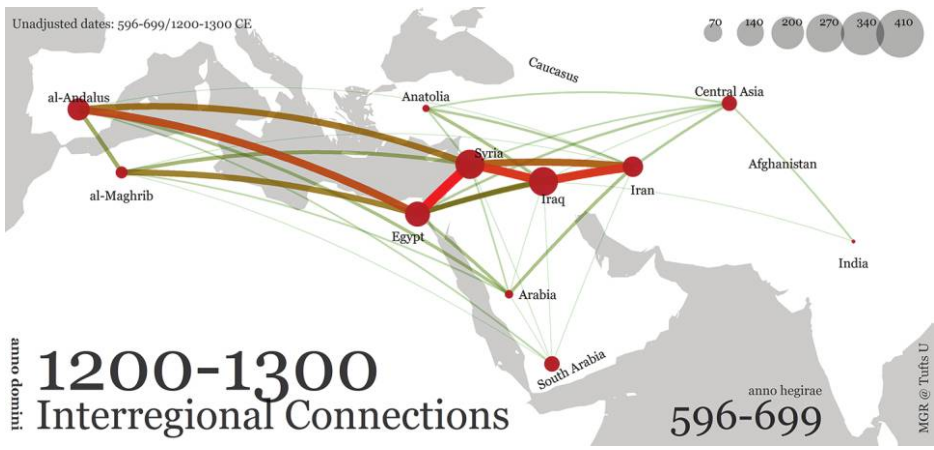

Fig. 11. Massive migrations of the thirteenth century CE

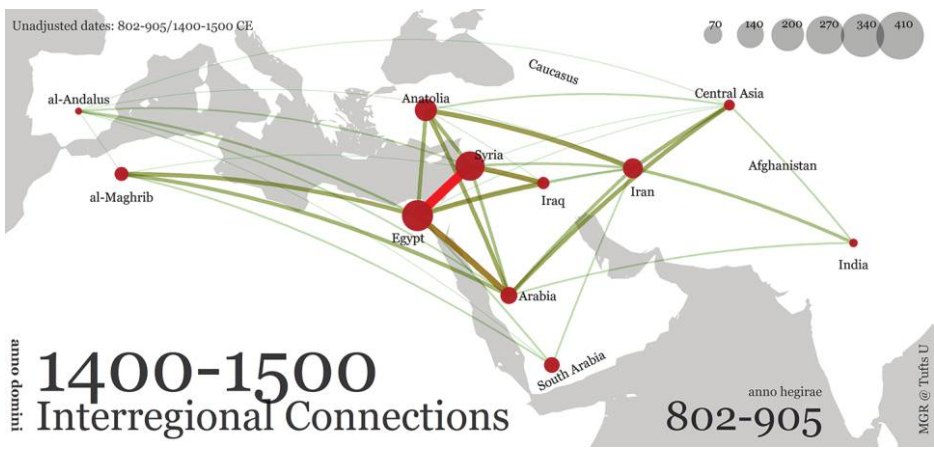

Fig. 12. New Mamlūk core of the fourteenth and fifteenth centuries CE

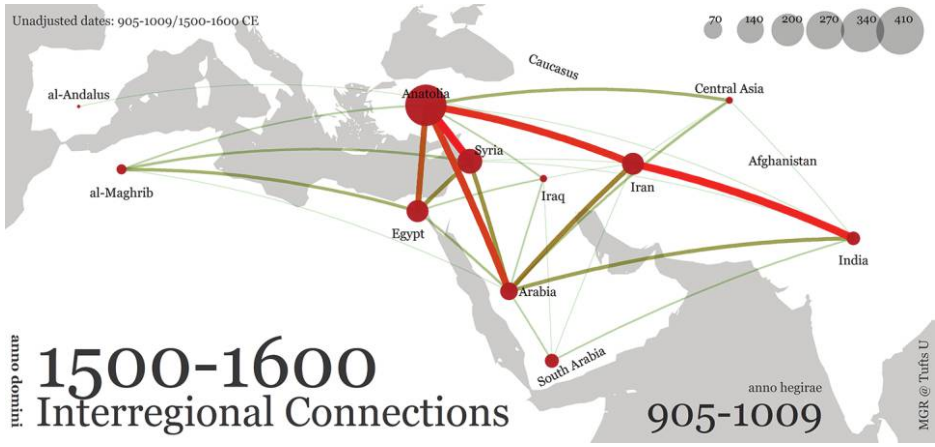

Fig. 13. Reconfiguration of the sixteenth century CE

Practically up until 1200 CE (Fig. 10), Iraq and Iran remain the core of the Islamic world:[22] they are strongly connected with one another—a very significant number of the men of letters (mostly religious scholars who write predominantly in Arabic) come from Iran during this period. Spain (al-Andalus), which, based on our data, thrives during 900–1300 CE, forms its own core with North Africa (al-Maġrib). The West and the East are too far from each other to maintain strong connections.

During the thirteenth century CE (Fig. 11), we find the strongest connections among the Eastern and Western regions of the Islamic world. Although one might expect this to indicate a certain tranquility that permitted travel, what we see is in fact the result of the crises both in the East and the West of the Islamic world. In Spain, Muslims are losing their ground, and a significant number of scholars start moving east to North Africa, Egypt, and Syria; Iran and Iraq are suffering from their own crises, most notably "the big chill" of the eleventh to early twelfth centuries CE, which destroys the economic prosperity of the Iranian regions and pushes nomads from the Turco-Mongolian steppe further and further into the Iranian plateau.[23] The Mongols usually take the blame for the destruction of the great cities of Iran and Iraq (most notably, Baġdād); however, judging by the data from biographical collections, by the time they show up and deliver the finishing blow, all the previously prominent urban centers are long in decline. It is during this period that we find Iranians and Iraqis leaving their homes, relocating to Syria

[22] As I show elsewhere, on data from a significantly larger biographical collection, the core for this period is more complex, particularly since what we come to understand as "Iran" in that period includes several major provinces, almost each one of them similar in size to Iraq. See Maxim Romanov, "After the Classical World: The Social Geography of Islam (c. 600–1300 CE)," in *Ars Islamica: Festschrift in Honor of Stanislav Mikhailovich Prozorov* and ed. Mikhail Piotrovsky and Alikber Alikberov (Moscow, 2016), 247–77.

[23] See, most notably, Richard W. Bulliet, *Cotton, Climate, and Camels in Early Islamic Iran: A Moment in World History* (New York, 2009).

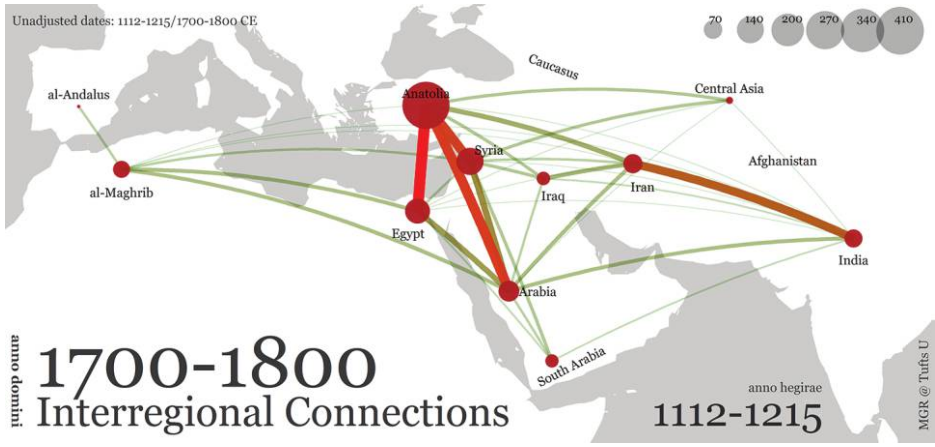

Fig. 14. The Turco-Arabic and Indo-Iranian cores in the eighteenth century

and Egypt, which in the two centuries to follow form a new core under Mamlūk rule (Fig. 12).

The sixteenth century marks a significant reconfiguration of the Islamic world, most notably with the rise of the "gunpowder empires": the Ottomans in Anatolia (Rūm) and the regions conquered by them, Mamlūk Syria and Egypt, which had formerly been the core; the Ṣafawids in Iran; and the Mughals in India (not graphed here). Fig. 13 displays this reconfiguration, marked by the rise of the Ottoman Empire and the reorientation of Iran, when significant numbers of Iranian scholars begin moving to Anatolia, but even more so to India.[24]

The last map—Fig. 14—shows the split of the Islamic world into two distinct cores of the Ottoman Empire, which gains control over almost the entire Arab world and the Indo-Iranian core. This split begins in the seventeenth century and remains equally distinct in our data up to the end of the nineteenth century.

Scaling Things Up

These graphs and maps show only a fraction of what can be done with the data extracted with the proposed approach even from a single biographical collection.[25] The next logical step is to study data from *all* available biographical collections—this step, however, requires even further formalization and infrastructural development.

A series of activities to this end are at the core of the Open Arabic Project, which has been ongoing for the past three years, first at Tufts University (2013–15) and

[24] See, for example, Masashi Haneda, "Emigration of Iranian Elites to India during the 16–18th Centuries," *Cahiers d'Asie Centrale* 3 (1997): 129–43, https://asiecentrale.revues.org/480.

[25] For more examples of such analysis of data from a different collection, see Maxim Romanov, "Toward Abstract Models for Islamic History," in *The Digital Humanities and Islamic and Middle East Studies*, ed. Elias Muhanna (Berlin, 2016), 117–49, http://www.degruyter.com/view/books/9783110376517/9783110376517-007/9783110376517-007.xml.

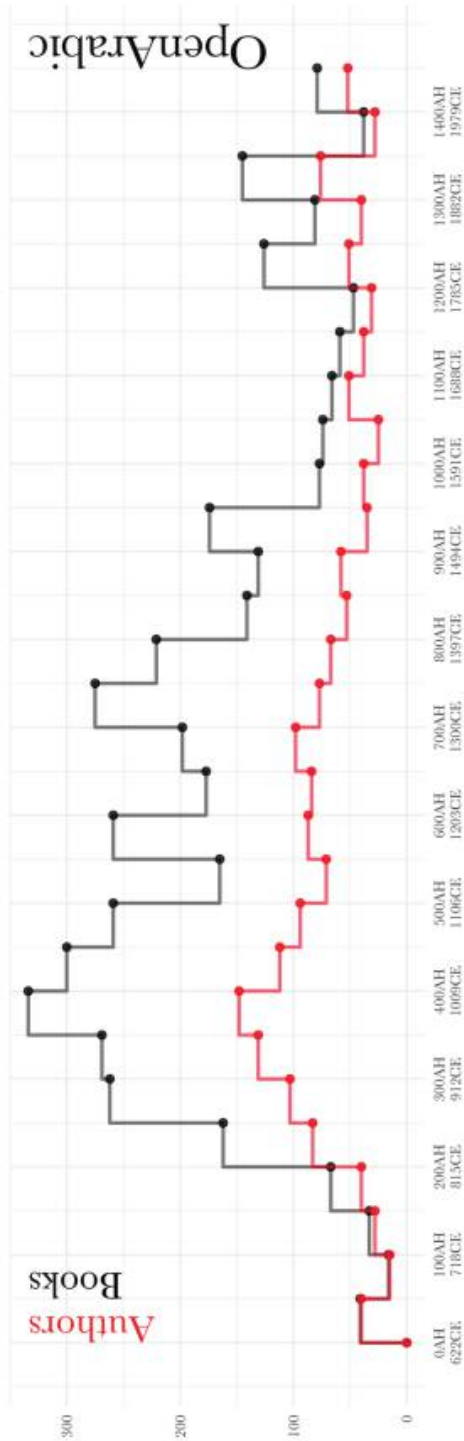

Fig. 15. Chronological distribution of authors and books in the Open Arabic Corpus

then at Leipzig University (2015–2017), within the broader vision of the Open Philology and Global Philology projects led by Gregory Crane, professor of classics at Tufts University and the holder of the Humboldt Chair of Digital Humanities at Leipzig University. One of the major efforts of these projects is the creation of machine-actionable corpora in historical languages (e.g., Open Greek and Latin, Open Persian, Open Arabic) and the development of tools and methods facilitating their analysis.[26]

Open Arabic is currently merging with a larger collaborative effort, Open Islamicate Texts Initiative (OpenITI),[27] that brings together scholars from Leipzig University, the University of Maryland (College Park), Aga Khan University (London), and the University of Vienna and aims to construct the first machine-actionable scholarly corpus of premodern Islamicate texts—first in Arabic and Persian, and later in other languages of the Islamic world.[28] Currently, Open Arabic includes several major components (at varying stages of development) that are meant to facilitate not only the large-scale analysis of Arabic biographical literature, but also of Arabic written tradition more generally.

At the heart of *Open Arabic* is the first instantiation of the corpus of premodern and early modern Arabic texts based on materials collected from several online open-access collections. The corpus now includes 1,850 authors and 4,280 unique titles (740 million words; with multiple editions, 1.34 billion words). However, when we compare our corpus with the data from the *Hadiyyat al-ʿārifīn*, it becomes clear that despite its considerable size, it still covers only a fraction of the Arabic written legacy—21 percent of authors and about 10 percent of book titles (very provisionally, of course).[29] The chronological distribution of authors and books (Fig. 15) also makes it clear that its coverage is heavily skewed toward the earlier period. The main goals for the further development of the corpus are (1) to provide detailed machine-actionable metadata suitable for research purposes;[30] (2) to vet collected texts for quality and tag their structure; (3) to expand the corpus by incorporating new digital texts from online collections that have not been covered so far, and also by using OCR to digitize published editions that are out of copyright;[31] (4) to create additional instantiations of the corpus that will facilitate specific

[26] See the website of the Humboldt Chair at http://www.dh.uni-leipzig.de/.

[27] The term "Islamicate" was introduced by Marshall Hodgson to refer to all things Islamic and non-Islamic, religious and nonreligious, that have been produced in the part of the world that we now know as the Islamic world. See Marshall G. S. Hodgson, *The Venture of Islam: Conscience and History in a World Civilization*, vol. 1, *The Classical Age of Islam* (Chicago, 1974), 57–60.

[28] For more details on the initiative, see http://iti-corpus.github.io/. OpenITI resources are available at https://github.com/OpenITI.

[29] How the data from the *Hadiyyat al-ʿārifīn* correlates with what had been published is impossible to say since to date no one has conducted a study of how many books written in the Islamic world have been published.

[30] Unfortunately, metadata created by librarians (for example, from http://www.worldcat.org/) is not suitable for research purposes mainly because of the complexities of the traditional Arab/Islamic name, where its six major components are used interchangeably without any consistent logic; book titles pose similar issues. Although designed to solve issues of this kind, the Virtual International Authority File (https://viaf.org/) is of no help here.

[31] Building on the foundational open-source OCR work of the Alexander von Humboldt Chair for Digital Humanities at Leipzig University (LU), the OpenITI team has achieved accuracy rates for clas-

forms of computational analysis, such as, for example, text reuse detection (https://github.com/dasmiq/passim), topic modeling (https://github.com/ThomasK81/ToPan), and stylometric analyses (https://github.com/computationalstylistics/stylo).

The entire corpus is available at https://github.com/OpenArabic; it is organized in compliance with the standards of canonical text services (CTS) as implemented in the CapiTainS Suite. The CapiTainS Suite has been developed for the maintenance of textual data at the Perseus Digital Library (Tufts University) as an important step toward Linked Open Data. These standards also make our corpus easy to expand.[32]

We have developed a lightweight tagging scheme—OpenArabic mARkdown—to facilitate the conversion of raw texts into machine-actionable formats as well as to facilitate data collection and extraction. Two main issues prompted the development of the scheme. First, we sought to avoid problems that one faces when paired symbols (such as angle brackets), left-to-right and right-to-left languages, and connected scripts[33] occur in the same document, making even a simple editing task overly complicated. Second, a lightweight and easy-to-use tagging scheme is of utmost necessity when one has to work with multivolume texts that make up the core of the Arabic written tradition.[34] Currently, OpenArabic mARkdown offers an easy-to-use scheme for structural tagging (3–6 symbols per tag) and a limited number of tags for semantic patterns and entities. The detailed description of the scheme can be found at https://alraqmiyyat.github.io/mARkdown/. It can be downloaded and used in EditPad Pro (https://www.editpadpro.com/).

Additionally, we are developing a Python library that will facilitate algorithmic analysis and conversion routines (for example, from OpenArabic mARkdown to TEI XML), as well as work with Open Arabic more generally. Last but not least, we are working on an exploratorium (in D3) that will allow users to explore abstractions of biographical collections through a series of interactive data visualizations (including graphs, maps, networks, tables, and so forth).

---

sical Arabic-script texts in the high nineties. On the results, see our working paper: Benjamin Kiessling, Matthew Thomas Miller, Sarah Bowen Savant, and Maxim Romanov, "Important New Developments in Arabographic Optical Character Recognition (OCR)," https://www.academia.edu/28923960/. We are currently working on a web interface for our OCR software.

[32] CapiTainS Suite was originally developed by Thibault Clérice (Leipzig University) and Bridget Almas (Tufts University). For more information, see http://capitains.github.io/.

[33] In comparison with Hebrew, the issues with Arabic are further aggravated by the fact that the computer dynamically changes the shape of each letter depending on its place in a word—and does that for all the letters. This creates a lot of issues on all operating systems and finding an editor that can properly handle this dynamic letter form selection and display is quite a challenging task. For example, none of the major text editors for the Mac offers proper support for Arabic script (which affects most of the languages of the Islamic world—Arabic, Persian, Urdu, pre-reform Turkic languages, etc.).

[34] For example, the longest biographical collection, "The History of Damascus" (*Taʾrīḫ [madīnat] Dimašq*) of Ibn ʿAsākir (d. 571/1175 CE), is a seventy-volume book of ten million words; currently, there almost two hundred books in OpenITI that are over the one-million-word threshold.

Maxim Romanov, University of Vienna (maxim.romanov@univie.at.ac)