Algorithmic Analysis of Medieval Arabic Biographical Collections

Maxim Romanov
Leipzig University

March 3, 2017

Abstract

Arabic biographical collections constitute one of the most voluminous and unexplored genres in the Arabic literary tradition. They are particularly valuable as a source for the social history of the Islamic world, especially up until 1500 CE, before which we are often poorly served by documentary evidence. Numbered in the hundreds, biographical collections include hundreds to tens of thousands of biographies and thus are ideal for prosopographical research of any kind. Scholars have recognized the value of these texts for decades, but their sheer volume has posed a formidable challenge, so their potential has remained untapped. This paper offers an efficient method for studying these texts through algorithmic analysis, which is here understood as a step-by-step reduction of texts written in a natural language to machine-readable data, and exploratory techniques that rely heavily on the use of graphs, cartograms and networks to identify and interpret chronological, geographical and social patterns from these texts. While the method so far has been applied only to a small number of our texts, I will also lay out current work in progress aimed at the development of maintainable infrastructure that will facilitate the analysis of not only all surviving texts, but also all of them taken together.

The article is to be submitted to the special issue of *Speculum*.

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

1 Introduction[2]

With at least 40,000 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 14th-century Damascene scholar, is a 50-volume mammoth of Arabic biographical literature (~3,4 million words) that covers seven centuries of Islamic history (*c.* 600–1300 CE) through over 30,000 biographies and about 10,000 descriptions of historical events.[3] The overall number of biographical records in texts that are currently available digitally already exceeds 400,000. 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 15th 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.[4] However, the excessive

[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 : [Washington, D.C.]: State University of New York Press; Library of Congress, 1995), 93.

[2] *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 one-to-one letter representation, with every Arabic letter 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 marbūṭaŧ*; *ã* for dagger *alif*; and *á* for *alif maqṣūraŧ*). Additionally, all attached conjunctions, prepositions, pronominal suffixes are separated with "-". The final version will adopt the transliteration system of the edition where the paper will be published. Whenever applicable, toponyms are given in their current American spelling. Bibliographical references and quotations preserve their original transliteration schemes.

[3] Al-Ḏahabī, *Taʾrīḫ al-islām wa-wafayāt al-mašāhīr wa-al-aʿlām*, ed. ʿUmar Tadmurī, 2nd ed., 52 vols. (Bayrūt: Dār al-Kitāb al-ʿArabī, 1990).

[4] See, Richard W. Bulliet, "A Quantitative Approach to Medieval Muslim Biographical Dictionaries," *Journal of the Economic and Social History of the Orient* 13, no. 2 (April 1, 1970): 195–211, doi:10.2307/3596086, Stanislav M. Prozorov and Maxim G. Romanov, "Principles and Procedures of Extracting and Processing the Data from Arabic Sources (Based on Materials of Historical-cum-biographical Literature) / Original Title: Metodika Izvlecheniya I Obrabotki Informatsii Iz Arabskih

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. 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.

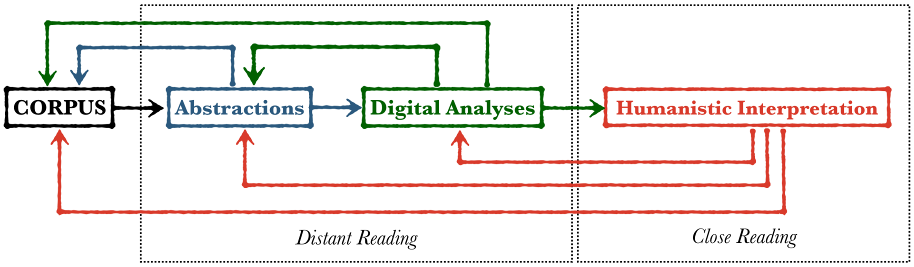


Figure 1: An iterative nature of algorithmic analysis

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

Istochnikov (Na Materiale Istoriko-biograficheskoi Literaturi)," *Oriens/Vostok* 4 (2003): 117–27.

[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*, 1st Edition (Urbana, Chicago & Springfield: University of Illinois Press, 2011). Franco Morretti's and Matthew Jockers's work has been equally thought provoking and inspiring: Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History* (London - New York: Verso, 2007), Franco Moretti, *Distant Reading*, 1st ed. (Verso, 2013), Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*, 1st

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 one how to 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 it will be most effective for extensive collections of information units with similar internal structure. 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.

2 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 semi-automatically—relevant data in the structured text (alternatively, extracting relevant data automatically); 4) extracting and modeling tagged data; 5) visualizing and analyzing results. All these 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 a printed edition on which it is based in order to establish its

Edition (University of Illinois Press, 2013). The method was first presented several years ago when it was at the very early stage of development as a part of a dissertation project, where it was then first implemented: see, Maxim G. Romanov, "Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunnī World (661-1300 CE)" (PhD thesis, University of Michigan, 2013). This article offers an overview of the method in its most recent form and describes relevant ongoing work aimed at scaling up the approach.

[6] Altogether, the libraries that I was able to survey include over 30,000 texts. The largest online libraries are: *al-Maktabaŧ al-šāmilaŧ* (www.shamela.ws; 6,300 texts); *al-Mishkāŧ* (www.almeshkat.net; 7,300 texts); *Ṣayd al-fawāʾid* (*www.saaid.net*; 10,000 texts); *al-Warrāq* (*www.alwaraq.com*; 860 texts); *al-Maktabaŧ al-šīʿaŧ* (www.shiaonlinelibrary.com; 1,970 texts); other libraries come 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).

overall adequacy. In most cases, these electronic texts are high-quality reproductions of printed editions (they seem to be produced with double-keying method),[7] and, for this reason, inherit all of 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 `OpenArabic 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 tag '`### |`', those of the second—'`### ||`', those of the third—'`### |||`', and so on. Logical units of specific types have their own patterns.

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. Figure 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 a prefix that should be dropped[10] and the second the length of an entity in words. When the tag is properly entered

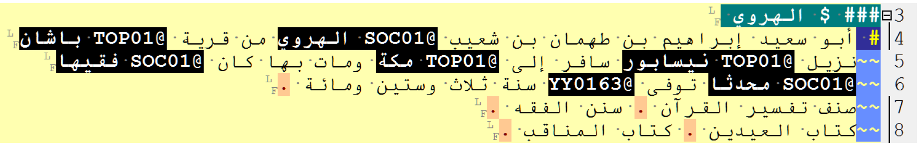
[7] The double-keying transcription method is confirmed to be the most accurate digitization approach, see Susanne Haaf, Frank Wiegand, and Alexander Geyken, "Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text," *Journal of the Text Encoding Initiative*, no. Issue 4 (March 8, 2013), doi:10.4000/jtei.739.

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

[9] Date statements can offer a valuable insight into a large Arabic corpus as well as specific books, see my blogpost "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.

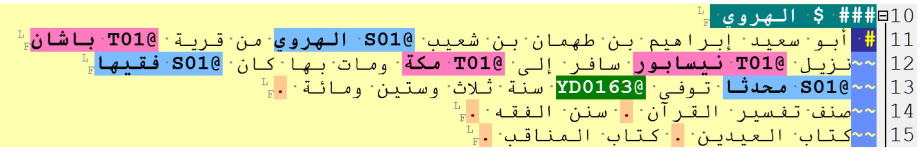
in front of the necessary word or the word group (up to 3 words), it is dynamically highlighted. Automatically inserted tags are highlighted in black.



```
### $ الهروي
# أبو سعيد إبراهيم بن طهمان بن شعيب @SOC01 الهروي من قرية @TOP01 باشان
~~نزيل @TOP01 نيسابور سافر إلى @TOP01 مكة ومات بها كان @SOC01 فقيها
~~@SOC01 محدثا توفي @YY0163 سنة ثلاث وستين ومائة .
~~صنف تفسير القرآن . سنن الفقه .
~~كتاب العيدين . كتاب المناقب .
```

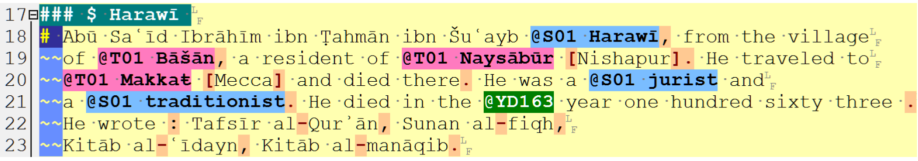
Figure 2: Automatically tagged entities in a biography: '`@SOC01`' for "descriptive names" (*nisbat*s) that behave as social markers; '`@TOP01`'—for place names (toponyms); '`@YY167`'—for dates (year statements). You can see a mild issue of what happens when right-to-left and left-to-right languages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

To avoid false positives, automatic tags can be disambiguated in the manner shown on Figure 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, Figure 4 shows the same information in English.



```
### $ الهروي
# أبو سعيد إبراهيم بن طهمان بن شعيب @S01 الهروي من قرية @T01 باشان
~~نزيل @T01 نيسابور سافر إلى @T01 مكة ومات بها كان @S01 فقيها
~~@S01 محدثا توفي @YD0163 سنة ثلاث وستين ومائة .
~~صنف تفسير القرآن . سنن الفقه .
~~كتاب العيدين . كتاب المناقب .
```

Figure 3: Manually disambiguated tagged entities in a biography: '`@S01`' for "descriptive names" (*nisbat*s); '`@T01`'—for place names (toponyms); '`@YD167`'—for the date of death (year statements).



```
### $ Harawī
# Abū Saʿīd Ibrāhīm ibn Ṭahmān ibn Šuʿayb @S01 Harawī, from the village
~~of @T01 Bāšān, a resident of @T01 Naysābūr [Nishapur]. He traveled to
~~@T01 Makkat [Mecca] and died there. He was a @S01 jurist and
~~a @S01 traditionist. He died in the @YD163 year one hundred sixty three .
~~He wrote : Tafsīr al-Qurʾān, Sunan al-fiqh,
~~Kitāb al-ʿīdayn, Kitāb al-manāqib.
```

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

NB: Automatic tagging of toponyms can be enhanced by inserting URIs of respective toponyms from a gazetteer of Islamic places. In our particular case, this will be al-Ṯurayyā Gazetteer (`https://althurayya.github.io/`), which was developed with this purpose in mind. The URIs in the gazetteer are designed to be human readable and aid in with disambiguation of complicated cases, such as, for example Ṭarābulus/Aṭrābulus—the toponym that may refer to the city of Tripoli in North Africa (sometimes appearing in the sources as Ṭarābulus al-Ġarb, 'Tripoli of the West') and 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 be automatically 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 of *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. 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`') and the city of Herat ('`@S01 Harawī`'), the city of Nishapur ('`@T01 Naysābūr`'), the city of Mecca ('`@T01 Makkaẗ`'); 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 ('`@S01 Harawī`'), a person who is strongly associated with the city of Herat (in this particular case, this person got the name of 'Herati' because he comes from the village of Bāšān in the district of Herat/Harāẗ). 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 provides social, professional, occupational, communal and ethnic markers). Combining thousands of such biographical profiles together 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.

NB: 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 specific linguistic patterns. For example, one can tag specific phrases that describe biographees' religious training, tag people one studied under, or, as immediately relevant to our example, books that one 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 it into a format suitable for further analysis and visualization. The script performs the following: 1) numbers all biographies sequentially (using biographical tags, '`### $`', as anchors); 2) splits the entire text of the book into individual biographies; 3) 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):[11]

```
======================
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 subset 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 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', *muftī*, 'jurisconsult' can be thus combined into a broader category of 'legal professions' and one can then

[11] 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 (Madīnaŧ 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 50 variations; existing morphological analyzers do not yet offer a reliable solution, especially for classical Arabic.

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 re-subset the main data set to get all relevant chronological, geographical and onomastic markers;
3) aggregating chronological markers, we can now build a graph of the temporal distribution of jurisconsults or legal professionals more generally;
4) aggregating geographical markers, we can build a cartogram of their spatial distribution;
5) combining chronological and geographical markers, we can also build cartograms of spatial distribution for different chronological periods and with that trace their spatial dynamic of their distribution over time;
6) 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.

3 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 *Hadiyyaŧ al-ʿārifīn* ("The Gift to the Knowledgeable")—is a bio-bibliographical 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 the 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 7th century CE up to the end of the 19th century CE.

From the very little that we know about him,[12] Ismāʿīl Bāšā wrote two extensive bibliographical texts—the first one, *Īḍāḥ al-maknūn fī Ḏayl ʿalá Kašf al-ẓunūn*, is the

[12] See, Witkam, J.J., "Ismāʿīl Pas̲h̲a Bag̲h̲dādli̊", in *EI2–Online*. For the edition of this text, see: Ismāʿīl Bāšā al-Baġdādī, *Hadīyaŧ al-ʿārifīn asmāʾ al-muʾallifīn wa-aṯār al-muṣannifīn*, 6 vols. (Bayrūt: Dār al-kutub al-ʿilmīyaŧ, 1992).

continuation of the famous *Kašf al-ẓunūn* of Ḥāǧī Ḫalīfat (d. 1067/1656 CE),[13] which mirrors its structure with the main unit being the book and them all organized alphabetically; the second one is the *Hadiyyat al-ʿārifīn*, which contains essentially the same information, but grouped into biographical records, which are organized alphabetically (and then chronologically within each letter).[14]

Although one cannot possibly expect for 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 in the Islamic world up until the beginning of the 20th 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.

3.1 Insight 1: cultural production over time

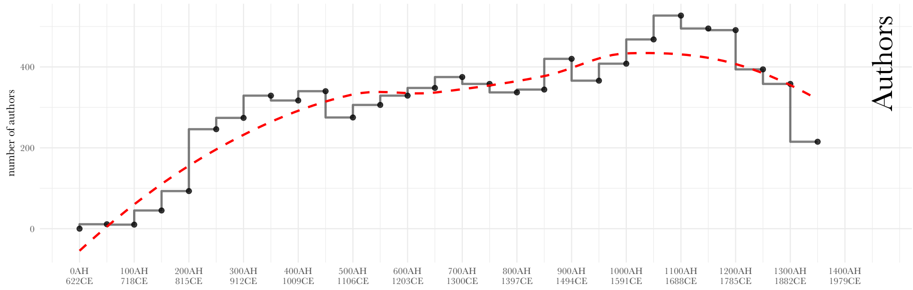


Figure 5: Chronological distribution of authors

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 (Figure 5) displays a steady upward trend up until 1200/1785 CE, reflecting the general historical situation: as the Islamic world keeps expanding geographically and the Muslim population growing, we find more individual getting involved in the process of cultural production.

Displaying the same trend for the period up 1200/1785 CE, the graph of books (Figure 6) makes the prominent early period (200–450 AH / 815–1058 CE) more

[13] He is also known as Kātib Čelebi, see: Şaik Gökyay, Orhan, "Kātib Čelebi", in *EI2–Online*.

[14] It is worth pointing 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. *ṭabaqat*)—authors of later generations would often take this information, edit, supplement and reorganize alphabetically. See, Franz Rosenthal, *A History of Muslim Historiography* (Leiden: E. J. Brill, 1952), passim: 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.

noticeable. Although this period is usually strongly associated with the translation movement from Greek into Arabic,[15] it is probably even more important for the formation of Islam as a religious system: particularly for the development of the Ḥadīṯ canon[16] and the crystallization of theological views.[17] Spikes are also due to a few very prominent polymaths: al-Suyīṭī (d. 911/1505 CE)—585 works; Ibn ʿArabī (d. 638/1240 CE)—425 works; al-Kindī (d. 256/870 CE)—256; al-Madā’inī (d. 225/840 CE)—223 works; al-Nābulusī (d. 1143/1730 CE)—204; Ibn al-Jawzī (d. 597/1200 CE)—201 works, and quite a few other prolific authors.

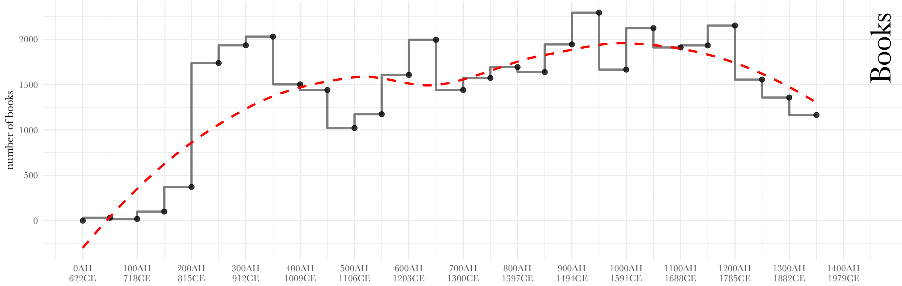


Figure 6: Chronological distribution of books

The decline of both graphs after 1200/1785 CE most likely indicates the unavailability of bibliographical information to our author. The geographical coverage of the collection also starts shrinking roughly at the same period. It should be noted that all chronological datasets 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 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 datasets—10 to 20 years.

Splitting our data geographically—Figure 7—we can also discover which regions played the leading role 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: Subsaharan Africa, the Indonesian Archipelago, the Volga region, and Eastern Europe. At

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

[16] 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 ; Boston: Brill, 2004), 73–86.

[17] According to the *Hadiyyat al-ʿārifīn*, about 90% of almost 500 "refutations" (Ar. *radd*) of different groups and specific beliefs were written during this period (peaking 250–450 AH / 864–1058 CE).

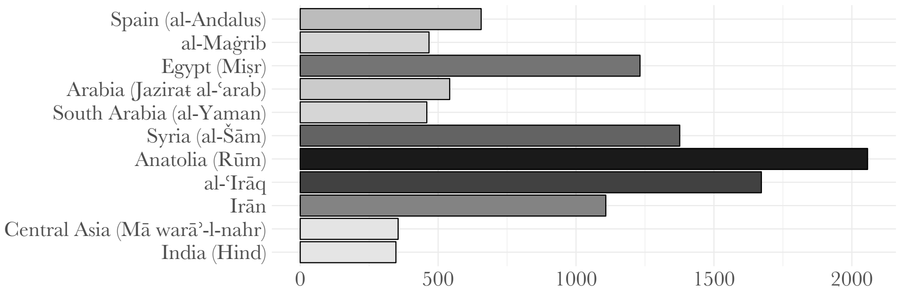


Figure 7: Regional Contributions.

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 (Jazīraŧ 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—Figure 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; dynasties of the "Iranian intermezzo", followed by the Tīmūrids and the Ṣafawids in Iran; the Mamlūks in Syria and Egypt;[18] 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 of "fine literature," most books in the *Hadiyyat al-ʿārifīn* deal with religious subjects—Qurʾānic exegesis, "words of the Prophet" (Ḥadīṯ), Islamic law, etc.—and they were composed more in the framework of the development of local religious communities, whose florescence depended on the overall political and economic stability. In this regard, the example of Iraq might be quite telling: the early period of ʿAbbāsid rule is marked by a very significant rise, which comes

[18] 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 in 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, Calif.: Undena Publications, 1977), 467–69.

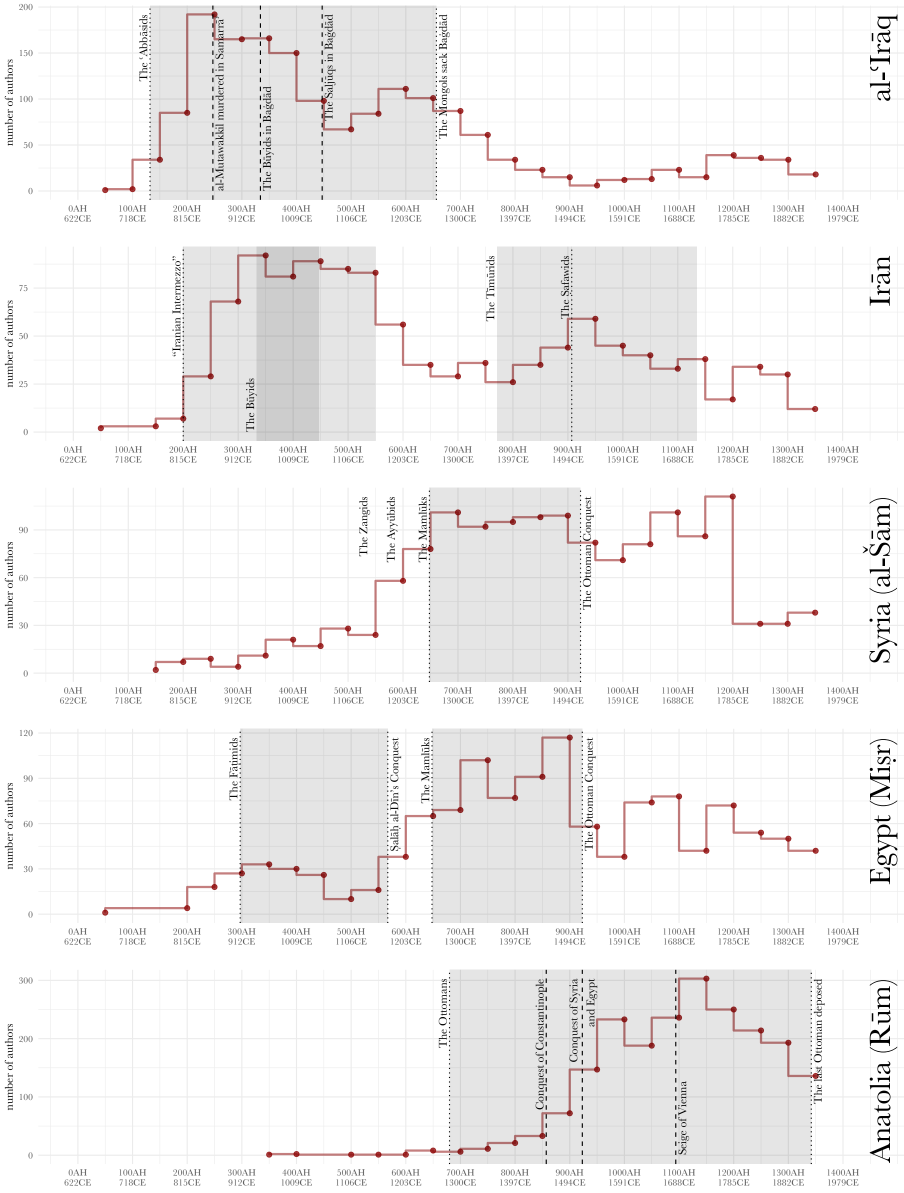
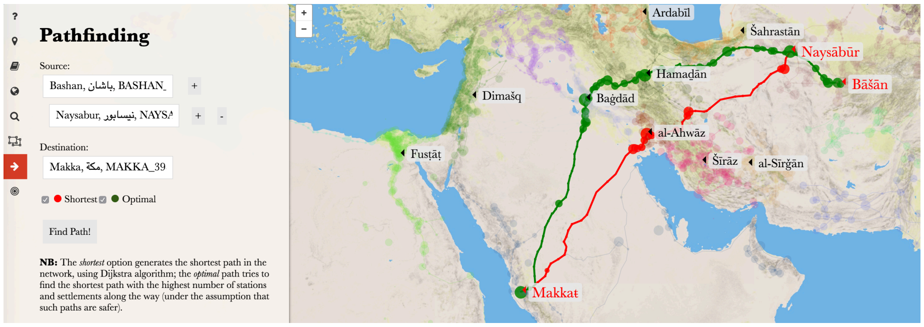


Figure 8: Most prominent Islamic regions over time.

to a halt when the ʿAbbāsids lose their sovereignty and become the puppets first, of their generals, then—the Būyids, and then—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 space of this article allows.

3.2 Insight 2: Cultural Connections

Our bio-bibliographical 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—Figure 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 (Dijkstra algorithm) and the optimal path (modified Dijkstra algorithm that avoids stretches with a small number of settlements along the way).



Pathfinding

Source:

Bashan, باشان, BASHAN_

Naysabur, نيسابور, NAYSA

Destination:

Makka, مكة, MAKKA_39

☑ ● Shortest ☑ ● Optimal

NB: The *shortest* option generates the shortest path in the network, using Dijkstra algorithm; the *optimal* path tries to find the shortest path with the highest number of stations and settlements along the way (under the assumption that such paths are safer).

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

For our purposes, however, a bit 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.[19] 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 period—with every 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 most prominent and inter-connected regions.

[19] The problem with the route network 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.

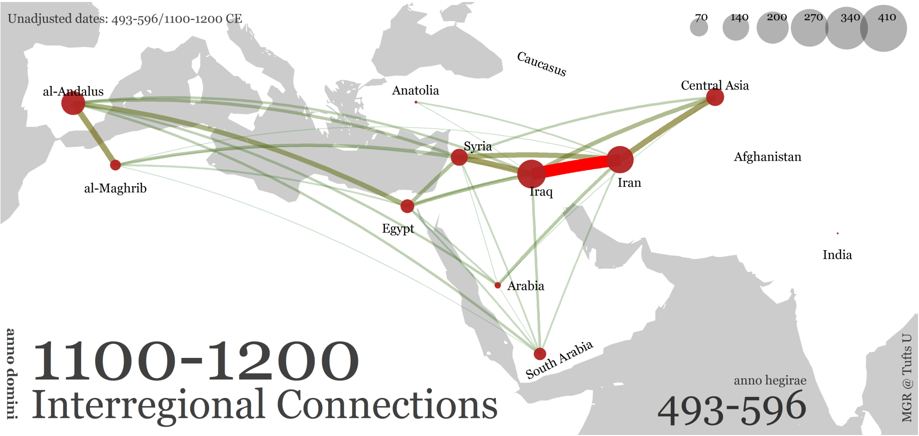


Figure 10: The Iraqi-Iranian core in the 12th century CE.

Practically up until 1200 CE (Figure 10), Iraq and Iran remain the core of the Islamic world:[20] they are strongly connected with each other—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 the 10–13th centuries, forms more of 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 13th century CE (Figure 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 thier own crises, most notably—"The Big Chill" of the 11th–early 12th 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.[21] 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

[20] 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 is several major provinces, with almost each one of them being 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*, ed. Mikhail Piotrovsky Alikber Alikberov (Moscow: Russian Academy of Sciences (Institute of Oriental Studies) & "Vostochnaya Literatura", 2016), 247–77.

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

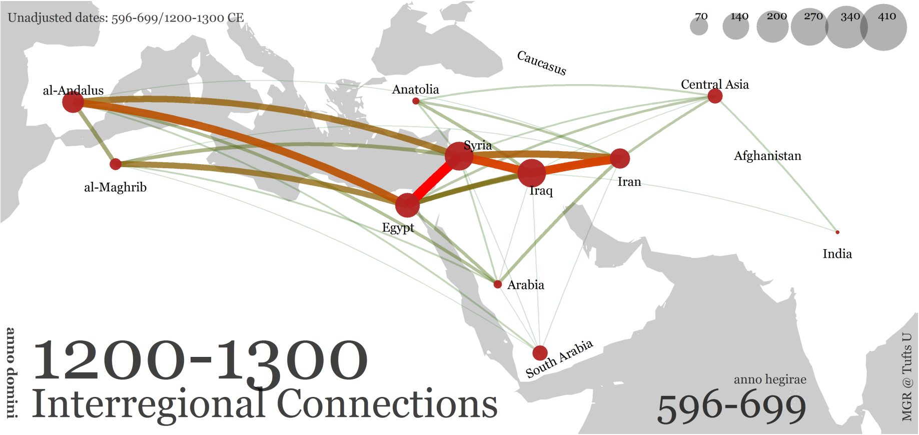


Figure 11: Massive migrations of the 13th century CE.

period that we find Iranians and Iraqis leaving their homes, relocating to Syria and Egypt, which in the two centuries to follow form a new core under Mamlūk rule (Figure 12).

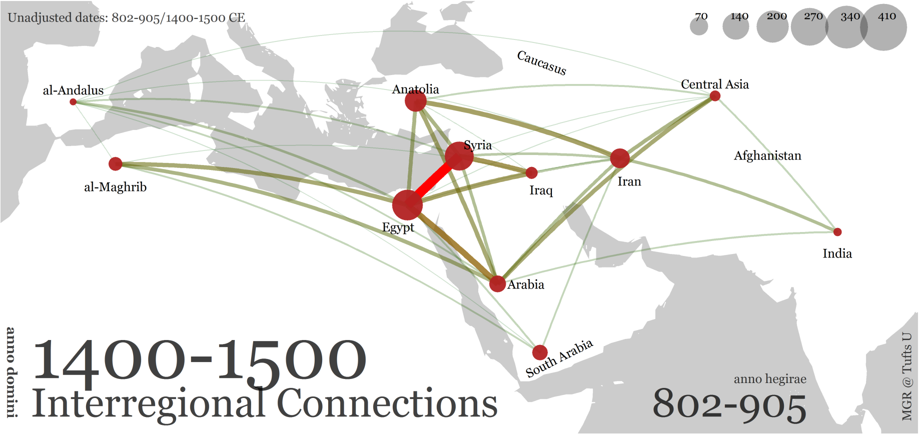


Figure 12: New Mamlūk core of the 14th and 15th centuries CE.

The 16th century marks a significant reconfiguration of the Islamic world: most notably with the rise of the "gunpowder empires"—the Ottomans in Anatolia (Rūm) and their successful conquests of the former core—Mamlūk Syria and Egypt; the Ṣafawids in Iran, and the Mughals in India (*not graphed here*). Figure 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.[22]

[22] See, for example, Masashi Haneda, "Emigration of Iranian Elites to India During the 16-18th

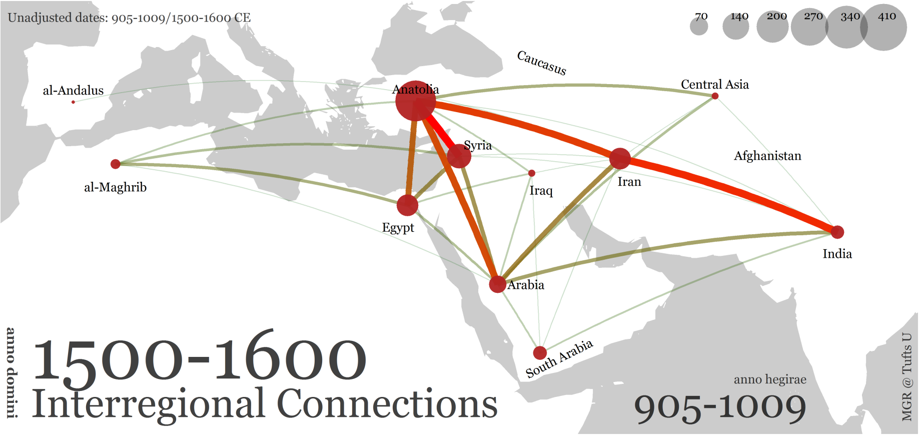


Figure 13: Reconfiguration of the 16th century CE.

The last map—Figure 14—shows the split of the Islamic world into two distinct cores of the Ottoman Empire which gains control over almost entire Arab world and the Indo-Iranian core. This split begins in the 17th century and remains equally distinct in our data up until the end of the 19th century.

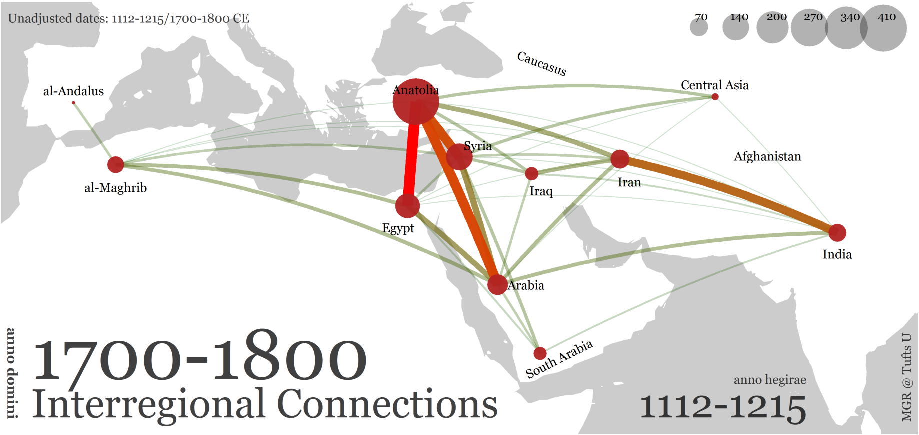


Figure 14: The Turco-Arabic and Indo-Iranian cores in the 18th century.

Centuries," *Cahiers d'Asie Centrale*, no. 3 (October 1, 1997): 129–43, https://asiecentrale.revues.org/480.

4 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.[23] 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 (2013–2015) and now—at Leipzig University (2015–) within the broader vision of the Open Philology and Global Philology projects led by Gregory Crane, the 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.[24]

Open Arabic is currently merging with a larger collaborative effort—*Open Islamicate Texts Initiative* (OpenITI)[25]—that brings together scholars from Leipzig University, the University of Maryland (College Park), and Aga Khan University (London), and aims to construct the first scholarly machine-actionable scholarly corpus of premodern Islamicate texts—first in Arabic and Persian, and later in other languages of the Islamic world.[26] 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 texts 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 Arabic written legacy—21% of authors and about 10% of book titles (*very provisionally, of course*).[27] The chronological distribution of authors and books

[23] 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 & Middle East Studies* (Berlin, Boston: De Gruyter, 2016), 117–49, http://www.degruyter.com/view/books/9783110376517/9783110376517-007/9783110376517-007.xml.

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

[25] The term 'Islamicate' was introduced by Marshall Hodgson to refer to all things Islamic and non-Islamic, religious and non-religious 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.*, vol. 1 (Chicago: University of Chicago Press, 1974), 57–60.

[26] For more details on the initiative, see: `http://iti-corpus.github.io/`. OpenITI resources will be soon available at `https://github.com/OpenITI`.

[27] How the data from the *Hadiyyat al-ʿārifīn* correlates with what had been published is impossible to

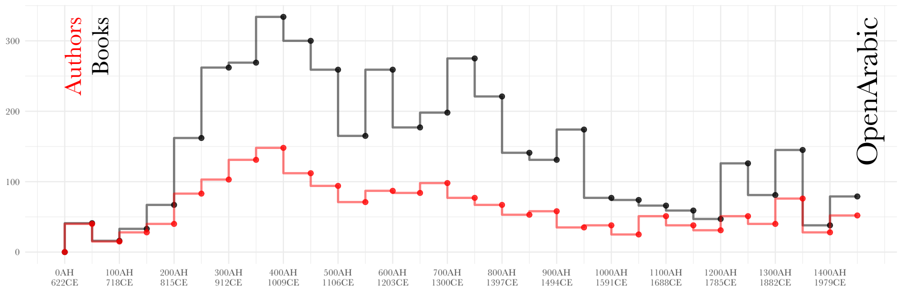


Figure 15: Chronological distribution of authors and books in the OpenArabic Corpus.

(Figure 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;[28] 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 OCR-ing published editions that are out of copyright;[29] 4) to create additional instantiations of the corpus that will facilitate specific 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.[30]

We have developed a lightweight tagging scheme—`OpenArabic mARkdown`—to facilitate the conversion of raw texts into machine-actionable formats as

say since to date no one has conducted a study of how many books written in the Islamic world have been published so far.

[28] 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, *Virtual International Authority File* (`https://viaf.org/`) is of no help here.

[29] Building on the foundational open-source OCR work of the Leipzig University's (LU) Alexander von Humboldt Chair for Digital Humanities, the OpenITI team has achieved accuracy rates for classical Arabic-script texts in the high nineties. On the results, see our working paper: Benjamin Kiessling, Matthew Thomas Miller, Sarah Bowen Savant, Maxim Romanov. "Important New Developments in Arabographic Optical Character Recognition (OCR)" at https://www.academia.edu/28923960/. We are currently working on a web-interface for our OCR software.

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

well as to facilitate data collection and extraction. Two main issues prompted the development of the scheme. First, to avoid problems that one faces when paired symbols (such as angle brackets), left-to-right and right-to-left languages, and connected scripts[31] 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.[32] 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 (`pyoa`) that will facilitate algorithmic analysis and conversion routines (for example, from `OpenArabic mARkdown` to TEI XML), as well as the 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, etc.).

References

al-Baġdādī, Ismāʿīl Bāšā. *Hadīyat al-ʿārifīn asmāʾ al-muʾallifīn wa-aṯār al-muṣannifīn*. 6 vols. Bayrūt: Dār al-kutub al-ʿilmīyat, 1992.

al-Ḏahabī. *Taʾrīḫ al-islām wa-wafayāt al-mašāhīr wa-al-aʿlām*. Edited by ʿUmar Tadmurī. 2nd ed. 52 vols. Bayrūt: Dār al-Kitāb al-ʿArabī, 1990.

al-Qāḍī, Wadād. "Biographical Dictionaries: Inner Structure and Cultural Significance." In *The book in the Islamic world: the written word and communication in the Middle East*, edited by George N. Atieh, 93–122. Albany : [Washington, D.C.]: State University of New York Press; Library of Congress, 1995.

Bearman, P., Th. Bianquis, C.E. Bosworth, E. van Donzel, and W.P. Heinrichs, eds. *Encyclopaedia of Islam (EI2-Online)*. Second edition. Malden, MA: Brill Online, 2016. http://referenceworks.brillonline.com/.

Bulliet, Richard W. "A Quantitative Approach to Medieval Muslim Biographical

[31] 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 Mac offer proper support for Arabic script (which affects most of the languages of the Islamic world—Arabic, Persian, Urdu, pre-reform Turkic languages, etc.).

[32] For example, the longest biographical collection, "The History of Damascus" (*Taʾrīḫ [madīnat] Dimašq*) of Ibn ʿAsākir's (d. 571/1175 CE), is a 70-volume book of 10 million words; there almost two hundred books over 1-million word threshold.

Dictionaries.” *Journal of the Economic and Social History of the Orient* 13, no. 2 (April 1, 1970): 195–211. doi:10.2307/3596086.

———. *Cotton, Climate, and Camels in Early Islamic Iran: A Moment in World History*. New York: Columbia University Press, 2009.

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

Haaf, Susanne, 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*, no. Issue 4 (March 8, 2013). doi:10.4000/jtei.739.

Haneda, Masashi. “Emigration of Iranian Elites to India During the 16-18th Centuries.” *Cahiers d’Asie Centrale*, no. 3 (October 1, 1997): 129–43. https://asiecentrale.revues.org/480.

Hodgson, Marshall G. S. *The Venture of Islam: Conscience and History in a World Civilization. Vol. 1. the Classical Age of Islam.* Vol. 1. 3 vols. Chicago: University of Chicago Press, 1974.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. 1st Edition. University of Illinois Press, 2013.

Lucas, Scott C. *Constructive Critics, Ḥadīth Literature, and the Articulation of Sunnī Islam: The Legacy of the Generation of Ibn Saʿd, Ibn Maʿīn, and Ibn Ḥanbal*. Leiden ; Boston: Brill, 2004.

Moretti, Franco. *Distant Reading*. 1st ed. Verso, 2013.

———. *Graphs, Maps, Trees: Abstract Models for Literary History*. London - New York: Verso, 2007.

Poonawala, Ismail K., and Teresa Joseph. *Biobibliography of Ismāʿīlī Literature*. Studies in Near Eastern Culture and Society. Malibu, Calif.: Undena Publications, 1977.

Prozorov, Stanislav M., and Maxim G. Romanov. “Principles and Procedures of Extracting and Processing the Data from Arabic Sources (Based on Materials of Historical-cum-biographical Literature) / Original Title: Metodika Izvlecheniya I Obrabotki Informatsii Iz Arabskih Istochnikov (Na Materiale Istoriko-biograficheskoi Literaturi).” *Oriens/Vostok* 4 (2003): 117–27.

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. 1st Edition. Urbana, Chicago & Springfield: University of Illinois Press, 2011.

Romanov, Maxim. “After the Classical World: The Social Geography of Islam (c. 600—1300 CE).” In *ARS ISLAMICA: Festschrift in Honor of Stanislav Mikhailovich Prozorov*, edited by Mikhail Piotrovsky Alikber Alikberov, 247–77. Moscow: Russian

Academy of Sciences (Institute of Oriental Studies) & “Vostochnaya Literatura”, 2016.

———. “Toward Abstract Models for Islamic History.” In *The Digital Humanities and Islamic & Middle East Studies*, 117–49. Berlin, Boston: De Gruyter, 2016. http://www.degruyter.com/view/books/9783110376517/9783110376517-007/9783110376517-007.xml.

Romanov, Maxim G. “Computational Reading of Arabic Biographical Collections with Special Reference to Preaching in the Sunnī World (661-1300 CE).” PhD thesis, University of Michigan, 2013.

Rosenthal, Franz. *A History of Muslim Historiography*. Leiden: E. J. Brill, 1952.