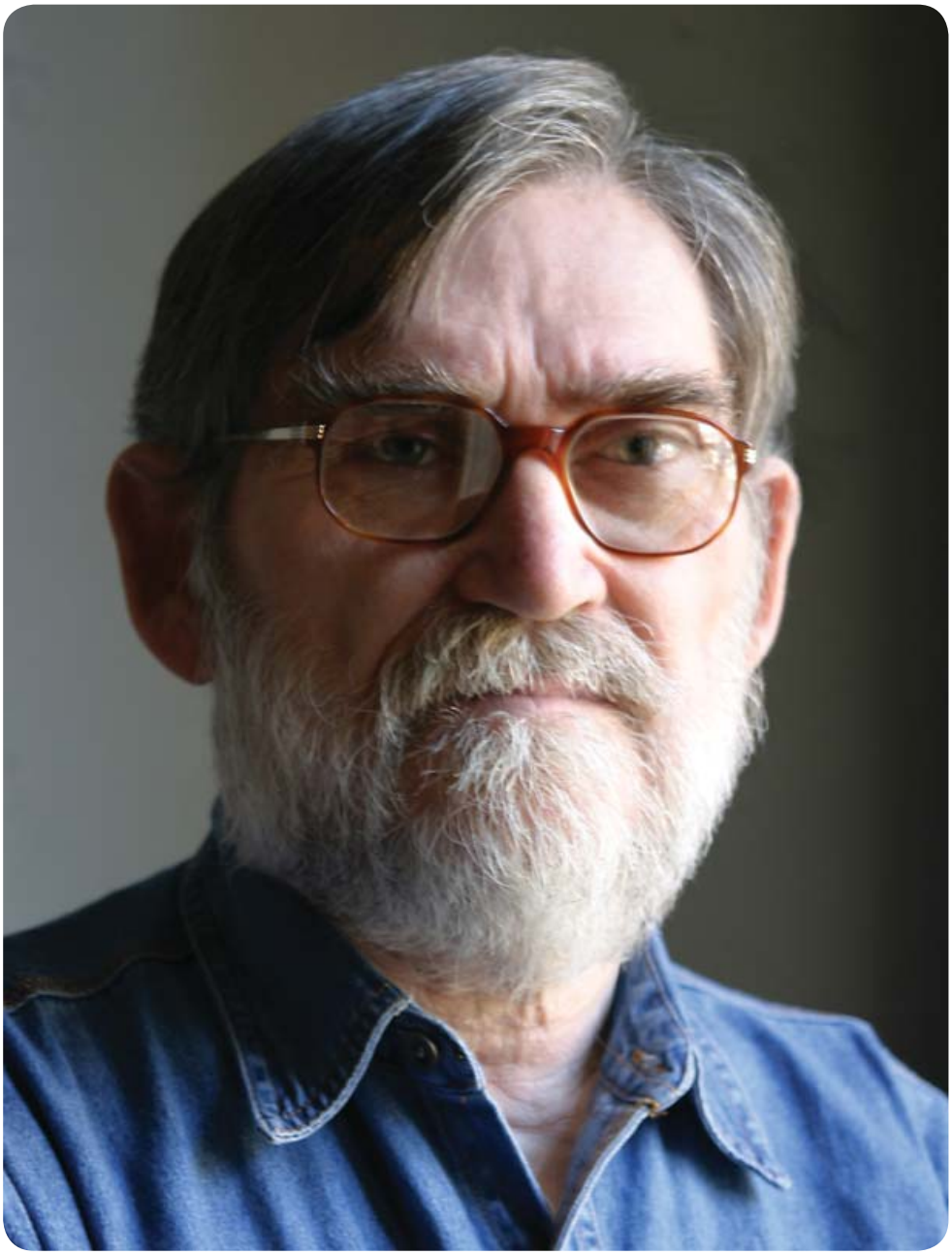
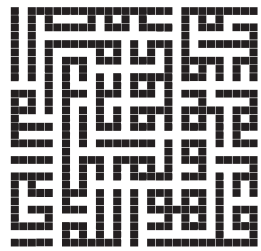


Станислав Михайлович ПРОЗОРОВ

THE RUSSIAN ACADEMY OF SCIENCES
Institute of Oriental Studies



ARS ISLAMICA

In Honor of Stanislav Mikhailovich PROZOROV

Edited by Mikhail B. PIOTROVSKY
and Alikber K. ALIKBEROV

Moscow
«Vostochnaya literatura» publishers
2016

Maxim Romanov

Digital Age, Digital Methods

> [A] canon of two hundred novels, for instance, sounds very large for nineteenth-century Britain ..., but it is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows — and close reading won't help here, a novel a day every day of the year would take a century or so... And it's not even a matter of time, but of method: a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it *isn't* a sum of individual cases: it's a collective system, that should be grasped as such, as a whole — and the graphs that follow are one way to begin this.
>
> Franco Moretti. *Graphs, Maps, Trees*[1].

Although Franko Moretti wrote this about literary history, this perfectly applies to any field of humanities. The availability of great volumes of information, on one hand, made our life easier, but on the other, made it incredibly more difficult since such abundance of information forces us into the

[1] *Moretti F.* Graphs, Maps, Trees: Abstract Models for Literary History // Verso. 2007. No 4.

stack overflow[1] mode — we have to deal with much more information than we can possibly process, let alone effectively analyze. According to Brill's Index Islamicus, the largest bibliographical database on the Islamic world, the number of publications practically doubled over the past two decades (the number of published books grew almost four times!) if compared to almost the entire preceding history of Islamic studies as a field of scholarly inquiry: approximately 284,000 publications (with about 47,000 books) for the period of 1990–2009 against approximately 145,000 publications (with about 13,000 books) for the period of 1900–1989 (See, Figure 1).

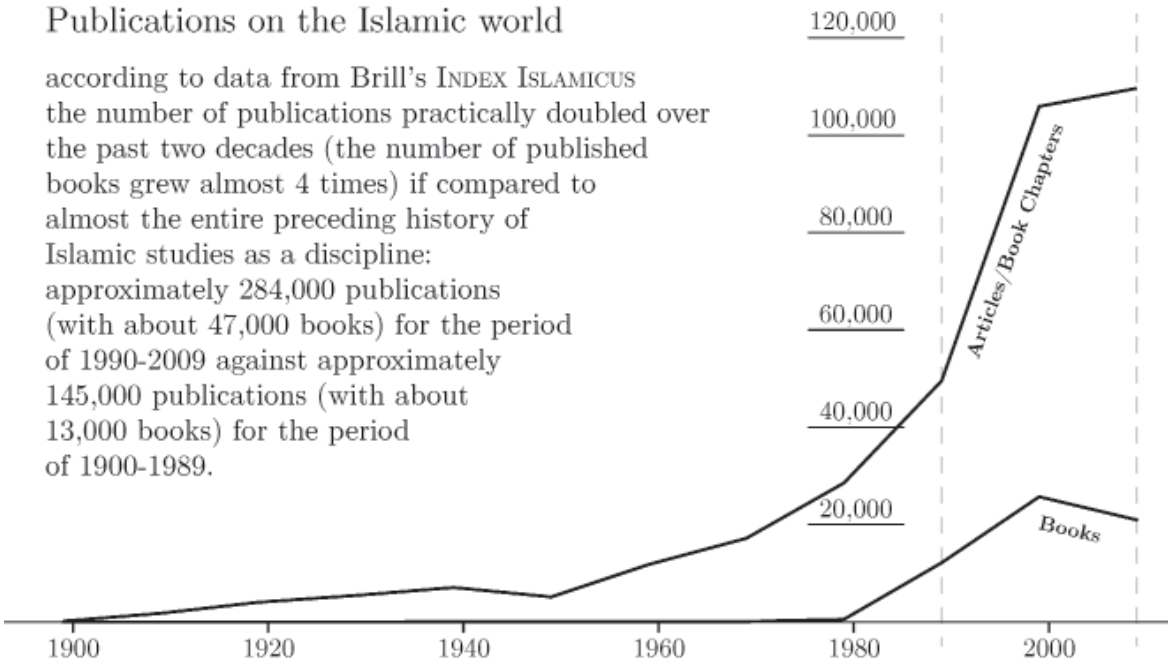


Figure 1: *Publications on the Islamic world. Source: Brill's Index Islamicus, accessed via http://web.ebscohost.com.proxy.lib.umich.edu/, October, 2012*

[1] In programming, stack overflow is an error condition that occurs when there is no room in the stack for a new item. Often, this will overwrite the adjacent memory locations causing hard-to-trace bugs.

Developments in the digital world affected the area of primary research as well. Historians now have access to digital forms of primary sources: scanned images of printed books, high-resolution images of manuscripts, fully-searchable versions of these texts etc. The pace of different fields in this area varies, but digital libraries of primary sources are now available for classical Greek, Latin, Arabic, Persian, Hebrew, Chinese as well as for early forms of European languages, such as English, German, Italian, Old Church Slavonic etc.

The area of classical Arabic has undergone a very significant change: thousands of primary sources have been scanned and uploaded to a great number of websites all over the world. Among the largest resources of books in classical Arabic are: *al-Maktaba al-waqfiyya* (www.waqfeya.org), *Multaqá ahl al-ḥadīth* (www.ahlalhdeeth.com), the Internet Archive (www.archive.org), *Wīkī maṣdar* (ar.wikisource.org), Hathi Trust (www.hathitrust.org) etc. More important, however, is that a great number of these texts have been converted into a fully-searchable format (henceforth, eTexts).

Classical Arabic Corpus

At the moment it is hardly possible to give even an approximate estimate to the overall volume of the classical Arabic corpus[1], but one can get some glimpses by looking into available digital libraries that include classical Arabic texts. While these libraries are huge, most of them suffer from strong ideological bias and include only Sunnī texts that passed the Salafī test on orthodoxy. However, even this fraction of the entire classical Arabic legacy is quite impressive. The commercially available software, *al-Jāmiʿ al-kabīr* (al-Turāth, Jordan), includes 2,400 titles (approximately 5,500 volumes) and adds up to almost 400 million words; another commercially available library, *al-Muʿjam al-fiqhī* (Qom), includes over 1,100 titles[2]. The largest online

[1] For now I can say that in his voluminous *Hadīyat al-ʿārifīn* ("The Gift of the Knowledgeable"), Ismāʿīl Bāshā al-Baghdādī listed over 8,700 men of letters, who contributed to the treasury of Islamic written culture up to the early 20th century. See: Ismāʿīl Bāshā al-Baghdādī, *Hadīyat al-ʿārifīn Asmāʾ al-muʾallifīn Wa-athār al-muṣannifīn*. 6 vols. Bayrūt, 1992).

[2] *Al-Muʿjam al-fiqhī* was developed under the patronage of *āyatullāh* al-Gulpāyagānī and includes both Sunnī and Shīʿī texts (some in Persian). Reviews/manuals for *al-Muʿjam al-fiqhī* and selected libraries by al-Turāth will soon be available at URL: http://maximromanov.github.io (written by Michael Bonner, Stanislav Prozorov and Maxim Romanov).

libraries of Arabic texts surpass *al-Jāmiʿ al-kabīr* and *al-Muʿjam al-fiqhī* combined: as of September 2012, *al-Maktaba al-shāmila* included over 5,800 titles (over 800 million words); *al-Mishkāt* — over 7,300; *Ṣayd al-fawāʾid* — over 10,000; while *al-Warrāq*, perhaps the first online library to appear, had only 860 titles (Figure 2).

Digital Library	Media	Titles	Vols	Words	uTokens†
al-Jāmiʿ al-kabīr	HDD, Win.95	**2,400**	**5,552**	**394,5 mln**	**0.524%**
al-Maktaba al-shamīla	www.shamela.ws	**5,869**	no data	**821,5 mln**	no data
al-Mishkāt	www.almeshkat.net	**7,375**	no data	no data	no data
Ṣayd al-fawāʾid	www.saaid.net	**10,159**	no data	no data	no data
al-Muʿjam al-fiqhī ††	DVD, Windows	**1,131**	**3,000‡†**	no data	no data
al-Warrāq	www.alwaraq.net	**862**	no data	no data	no data

Figure 2: *Major Digital Libraries of Classical Arabic. † uTokens show the number of unique word forms in the library; †† al-Muʿjam al-fiqhī includes both Sunnī and Shīʿī texts; ‡† the number of volumes is approximate*

Al-Jāmiʿ al-kabīr and *al-Maktaba al-shāmila* have a very significant number of books written by authors who died before 1900 CE, with most of the titles distributed in the period between 800 CE and 1700 CE: at least 2,000 titles in *al-Jāmiʿ al-kabīr* and 3,200 in *al-Maktaba al-shāmila* (Figure 3). Just to get an idea of how big these libraries are, one can compare them with other digital libraries of books in various historical languages. For example, the Perseus Project at Tufts University, the largest digital library of classical texts in Greek and Latin, contains about 14 million words[1]. For

[1] As of March 2011, the corpus contains 7.5 million words of Greek and 6.5 million words of Latin. In all other aspects, however, the Perseus Project surpasses all digital resources available to the scholars of the Islamic world.

comparison, three largest books in *al-Jāmi‘ al-kabīr* amount to 19 million words[1].

Of all other digital corpora of historical languages in general, it seems that only the corpus of classical Chinese is comparable to that of classical Arabic. Funded by the Hong Kong government, the Chant library includes most of the pre-Qin and Han corpus amounting to about 60 million characters. A Taiwanese database at Academia Sinica has a selection of texts from all the periods of Chinese history, totaling about 150 million characters. The largest digital library of Chinese texts is probably Siku Quanshu, the 18th century imperial library, which contains about 800 million characters[2].

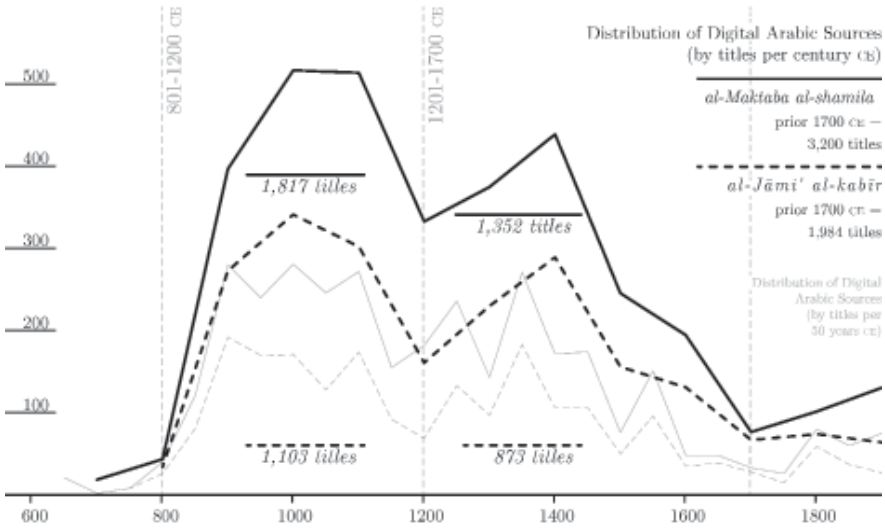


Figure 3: *Distribution of Digitized Arabic Sources (by Titles). The graphs of both libraries are strikingly similar, which strongly suggests that that al-Maktaba al-shāmila has been built upon al-Jāmi‘ al-kabīr. Additionally, two periods vividly stand out: 801 –1200 CE and 1201 –1700 CE*

[1] These texts are: *Taʾrīkh madīnat Dimashq* of Ibn ‘Asākir (d. 571/1176), which is 8,1 million words; *‘Umdat al-qārī fī sharḥ Ṣaḥīḥ al-Bukhārī* of al-‘Aynī (d. 855/1451), which is 4,6 million words; and *Tāj al-‘arūs* of al-Zabīdī (d. 1205/1791), which is 3,9 million words.

[2] I am deeply grateful to Donald Sturgeon, a PhD student at the University of Hong Kong, for providing me with this information on Chinese digital libraries. Donald Sturgeon develops his own digital library of classical Chinese texts — the Chinese Text Project (URL: http//www.ctext.org). Although still relatively small in comparison to the above-listed giants, his web-library of about 17 million characters is equipped with a variety of digital research tools, which will make any Arabist green with envy.

Most of the electronic texts of classical Arabic sources are based on specific paper editions that are widely used by the Western scholars of Islam, and, by and large, reproduced exceptionally well. Such abundance of digital texts calls for new methods of text analysis. Equipped only with the traditional methods of historical inquiry, we will not be able to make sense of this constantly growing digital corpus of primary sources in classical Arabic. In this paper I propose a method that relies on a number of digital techniques of text analysis that emerged at the intersection of statistics, corpus linguistics and computer science. In doing so, I will focus on the genre directly relevant to my research: traditional Islamic biographical collections. Although currently this method is most effective for the analysis of sources that are composed of structurally similar blocks of information — collections of Prophetic reports (*ḥadīth*), interpretations of the al-Qurʾān (*tafsīr*), collections of legal decisions (sing. *fatwá*), chronicles etc. –it can be adopted to the study of any kinds of texts in any oriental or occidental language.

Biographical Collections

Biographical collections is one of the most prominent genres of Islamic pre-modern literature. Paul Auchterlonie's reference lists about 230 biographical collections, published completely or partially up to the 1980s[1]. The total number of titles in this genre is somewhere between three and four hundred. In addition to their numbers, one must definitely stress their size: most biographical collections are multivolume, they usually cover long periods — in most cases measured in centuries — and include up to tens of thousands of biographical records. The text of the largest[2] biographical collection –"The History of Islam" (*Taʾrīkh al-islām*) of al-Dhahabī (d. 748/1347 CE) — takes up 52 volumes, covers 700 years of

[1] See: *Auchterlonie P.* Arabic Biographical Dictionaries: a Summary Guide and Bibliography (Durham [Durham]: Middle East Libraries Committee, 1987).

[2] It is the largest in terms of coverage. The palm for being the longest one, however, belongs to Ibn 'Asākir (d. 571/1176) and his 70 volumes of *Taʾrīkh madīnat Dimashq* ("The History of the City of Damascus"), the longest book in *al-Jāmi' al-kabīr*.

Islamic history and includes over 30,000 biographies[1]. The overall number of biographies in these sources reaches hundreds of thousands, but he more precise figure still remains a mystery. Some of these biographical collections include only members of particular social and/or religious groups (*ṭabaqāt* — prosopographical dictionaries, in the strict sense), others cover specific locations and/or time periods (biographical dictionaries), yet others are all-inclusive — at least they are imagined as such by their authors — and also cover historical events ("obituary chronicles").

A great number of biographical records in these sources are rather short notices — often just the name of a person with dates of his (more rarely — her) life, whether precise or approximate. However, even onomastic data alone provides historians with a lot of valuable information, mainly thanks to the part of Muslim name known as *nisba*, "descriptive name." In strict grammatical terms, *nisba* is an adjective formed from a noun by means of adding suffix "ī" and thus denoting a relation to the noun from which it was formed, i.e. Baghdad + ī = Baghdādī, meaning something or someone related to the city of Baghdad. In historical terms, however, *nisba* is not necessarily limited to this particular morphological pattern, but is rather used for any word that can meaningfully describe a person, including but not limited to such morphological patterns as *fāʿil*, *faʿīl*, *faʿʿāl*[*a*], *dhū shay*[*ayn*], *mufaʿʿil*, *mufāʿil*, etc. This is particularly true in the case of al-Samʿānī who included them all into his *Kitāb al-ansāb*.

Let's take a close look at one of such names: Abū l-Faraj ʿAbd al-Raḥmān, the son of (ibn) ʿAlī, the son of (ibn) Muḥammad, ..., the son of (ibn) [so-and-so], ..., the son of (ibn) Muḥammad, the son of (ibn) Abī Bakr al-Ṣiddīq, al-Jawzī, al-Qurashī, al-Taymī, al-Bakrī, al-Baghdādī, al-Ḥāfiẓ, al-Mufassir, al-Ḥanbalī, al-Wāʿiẓ, al-Ṣaffār. This name includes nine meaningful "descriptive names," which tell us that this particular person belonged to the clan of Taym (al-Taymī) of the tribe of Quraysh (al-Qurashī) and was a descendant of Abū Bakr al-Ṣiddīq (al-Bakrī), the first of the four Rightly-guided caliphs of the Islamic community; a native of Baghdad (al-Baghdādī) and a jurist of the Ḥanbalī school of law (al-Ḥanbalī), he distinguished himself as a knowledgeable transmitter of Islamic tradition (al-Ḥāfiẓ), an exegete of the Qurʾān (al-Mufassir) and

[1] Muḥammad ibn Aḥmad al-Dhahabī. *Taʾrīkh al-islām wa-wafayāt al-mashāhīr wa-al-aʿlām* / ed. ʿUmar Tadmurī. 52 vols. 2nd ed. Bayrūt, 1990.

a public preacher (al-Wā'iẓ); the last *nisba* (al-Ṣaffār) also tells us that he comes from a family that earned its living selling copper utensils (*nuḥās*). Thus, the onomastic information alone is tantamount to the social profile of a person. Studied as a whole, such social profiles have a capacity of transforming into a unique looking glass through which the historian can study different aspect of Islamic history, which are otherwise indiscernible. Additionally, such profiles form a unique body of data, which is ideally suitable for different forms of sociological and spatial analyses.

People who became the subjects of these biographical records were not simple commoners. By and large, they were representatives of religious, administrative, military and literary élites. Nonetheless, the lives of these notables often presented with so many details that studying them as a whole will also shed light on the life of rank-and-file believers.

Prior Studies: Database Approach

Previous studies of biographical collections relied on quantitative methods and relational databases, both analog and digital. Clearly conceptualized by Richard Bulliet in 1970[1], a new quantitative approach was very promising, but also extremely laborious and time-consuming, even with the help of early computers, which were anything but user-friendly back in the day. The main problem with the approach was posed by the very advantage it was supposed to exploit: the limitless quantifiable data available for analysis. Anyone attempting to implement this approach had to set very strict limits in order to accomplish the research project. One had to clearly define research goals, select a limited amount of sources and carefully consider kinds of data required for the research. After careful planning one had to peruse the selected sources, manually extract required information and then to record it either on a paper media, or to code it for transferring to "the memory bank of a computer". These technologically imposed limita-

[1] *Bulliet Richard W.* A Quantitative Approach to Medieval Muslim Biographical Dictionaries // Journal of the Economic and Social History of the Orient *13*. 1970. No. 2. April 1. P. 195–211, doi:10.2307/3596086.

tions affected the usability of the extracted information and in most cases the potential of created databanks was exhausted by the end of research projects for which they were created.

In the late 90s, advancements of computer technologies and availability of personal computers stimulated a few more attempts. Although it was expected that digital relational databases open up a new range of questions that can be asked that would hitherto have been unthinkable "without 500 monks at hand"[1], in real life they offered no significant improvements for the most tedious process of entering data: their creation remained equally time consuming and many projects were never finished. Overall, the potential of the approach is still far from being realized.

Novel Approach: Text-mining

Text-mining approach is cardinally different. Not a method with clearly defines boundaries, it is rather an open-ended set of computational techniques that allow extracting meaningful information from unstructured texts. This approach capitalizes on a number of developments of our digital age. The wide support of the Unicode standard made it possible to apply text-mining methods to oriental languages. Scripting languages commonly used in text-mining, such as Python or Perl, allow one to work with an unlimited number of texts and design complex analytical tasks. Combined with statistical and linguistic methods, these tools offer limitless ways for studying voluminous historical texts. Unlike traditional database approach, text-mining allows to expand the databank of primary texts and incorporate new digital methods and techniques as they became available. While the structure of relational databases imposes strict limitations on what research questions can be asked of data, the limitations of text-mining approach seem to lurk only in the ability of researchers to

[1] The quote is from: *Mathisen Ralph W.* Where Are All the PDBs?: The Creation of Prosopographical Databases for the Ancient and Medieval Worlds // *Prosopography Approaches and Applications: A Handbook*. University of Oxford, Linacre College Unit for Prosopographical Research, 2007. P. 95. The article is an interesting overview of the use of databases — both analog and digital — in history (prosopography, to be more precise). The main issue with them is eloquently expressed in the very title of Mathisen's article.

transform complex research questions into working algorithms and then translate them into the scripting language of choice. At the hands-on level, text-mining approach offers a more efficient work flow and allows one to begin the analysis of data almost instantaneously, gradually increasing the complexity of research tasks.

Text-mining relies on patterns. As repetitive, quantifiable and morphologically similar structures, they permeate the entire approach. Designing any text-mining task one first samples sources to collect textual patterns that encapsulate required information[1]; then one translates these textual patterns into the language of regular expressions — another kind of patterns — which the machine can interpret and use to locate and extract all instances of data that fit the initial textual patterns. Extracted data are then repeatedly reorganized, segmented and analyzed in order to discover patterns of historical significance.

Although text-mining approach is quantitative in its nature, its main advantage is that it enhances qualitative analysis. Because of its nature, text-mining encourages — even demands — seamless connection between the texts of primary sources and research data extracted from them, allowing one to switch easily between quantitative and qualitative tasks. Thus, text-mining techniques allow one to explore vast corpora of primary sources and discover all thematically relevant passages. Such exhaustive coverage will make any close reading thoroughly contextualized and more reliable[2].

In the preliminary quantitative study of extracted data I rely on the principles and techniques of exploratory data analysis. Pioneered in the 70s by John Tukey[3], this statistical approach is based on the underlying assumption that "the more one knows about the data, the

---

[1] Scholars of Islamic history discussed some of these patterns although in a quite different context. See, for example, Albrecht Noth's discussion of "transitional formulae" in Noth Albrecht, Conrad Lawrence I. *The Early Arabic Historical Tradition: a Source-critical Study*. Princeton, N.J., 1994, particularly Chapter 4.

[2] My analysis of dream-tales from two Ḥanbalī biographical collections was a test run for this approach. See: *Romanov M.* Dreaming Ḥanbalites: Dream-Tales in Prosopographical Dictionaries // *Dreams and Visions in Islamic Societies,* ed. Özgen Felek and Alexander Knysh. N.Y., 2012. 31–50.

[3] *Tukey John W.* Exploratory Data Analysis. Addison-Wesley Series in Behavioral Science. 1977. Even though written in a pre-digital age, this book is considered foundational for anyone seriously interested in data analysis. For a quick-fix, one can opt for a much more concise

more effectively data can be used to develop, test and refine theory"[1]. Exploratory data analysis is based on the principles of *openness* and *skepticism*: they call for the analysis of data without preformed expectations or theoretical assumptions while remaining open to new ways of looking at the same data. Sociologists are well aware of the questionnaire problem when questions may frame responders' answers. Questioning their sources, historians are not immune from this problem either. For this reason it is important to explore sources to make sure that research questions are relevant. This is particularly important in the case of multivolume biographical collections, which, as the inside joke goes, are easier to write than to read. Exploratory data analysis emphasizes visual representation of data, specifically over summary statistics. Arguably, carefully designed visualizations can convey the complexity of data without fracturing the subtlety of intricate connections into a linear narrative, which readers may never be able to fully reassemble. In many cases visual display of data is the only way to bring together complex combinations of variables. Thus, visualizations are used for "visual evidence, visual reasoning, and visual understanding"[2].

Unlike in the database approach where one extracts all kinds of information processing one biography at a time, repeating this process until a specific source is exhausted, in the text-mining approach one extracts one kind of information but processing all biographies of a specific source at once. Such a change of perspective has two advantages. First, dealing with only one type of information streamlines the entire process which is particularly important since regular expressions and text-mining scripts must be continuously readjusted for better performance. Second, one can begin analysis right after a specific kind of data has been extracted. Such step-by-step exploratory evaluation of data is helpful for guiding further research. Moreover, it encourages one to continue with the study, while with conven-

representation of this approach for social sciences: *Hartwig F., Dearing Brian E.* Exploratory Data Analysis. A Sage University Paper. 1979.

[1] *Hartwig F., Dearing Brian E.* Exploratory Data Analysis, 9.

[2] This area of exploratory data analysis was largely developed by Edward Tufte. The quote is from: *Tufte Edward R.* The Visual Display of Quantitative Information. Graphics Press. 1992. 8.

tional relational databases one usually gets a depressing feeling that this will never end.

In the examples that will follow I use *Taʾrīkh al-islām* of al-Dhahabī, volumes 4–52 that cover the period of 41–700 AH/661–1300 CE. I had to exclude the first three volumes, since structurally they are very different and thus unsuitable for the method in its current form.

Preparing eTexts

First of all, eTexts must be prepared for text-mining scripts. The whole process can be described as "normalization", which in this context refers to two different tasks. In terms of computational linguistics and natural language processing (NLP) "normalization" refers to the disambiguation of spelling, formatting and encoding irregularities of Arabic text[1]. In terms of text-mining "normalization" refers to providing meta-data — i.e. data that describes data — for every unit of text. In this case, "normalization" informs us about the context from which specific information was extracted, thus making queries meaningful. Consider the following example: if one searches for the word "preacher" (*wāʿiẓ*) in any biographical dictionary, it will result in hundreds or even thousands of hits; browsing through all the hits will not necessarily lead to anything meaningful. However, if the very same dictionary has been split into biographies, the search will tell the number of people who were either preachers themselves or had some connections with preachers. Furthermore, if the same search is applied specifically to the onomastic section of each biography — a section that occurs in the beginning of each biography and ends right before the details on religious education — it will actually give us the number of preachers in the dictionary. This will narrow down the number of biographies for close reading, which is particularly important when dealing with four or five digit figures.

[1] For more details on "normalization", see: *Habash Nizar Y*. Introduction to Arabic Natural Language Processing. [San Rafael, Calif.], 2010. P. 21–23.

In practical terms, the task of "normalization" refers to tagging the structure of each eText. In the case of a biographical dictionary, one marks titles of chapters (preserving hierarchy) and starting points of biographical records. Tagging the structure is a rather tiresome task, however, it allows preserving the entire eText of the source and has to be done only once. The processed eText becomes machine-readable and can be used until its research value is completely exhausted. This process can also be streamlined, first, by using short tags for the markup of all structural elements and, second, by using highlighting schemes, which help to avoid typos and to make structural elements easily visible.

After the tagging is complete, one can take a closer look at first statistics from the source in question. To begin with, one now knows the number of meaningful blocks of information: volumes 4–52 of *Taʾrīkh al-islām* that cover the period of 41–700 AH include biographical records on 28,927 individuals and descriptions of 5,288 events[1].

Another statistic that is worth looking at is the lengths of these biographies. It would be logical to assume that the length of a biography reflect the importance of that person for the author, and, arguably, the significance of the person for Islamic history in general. If we accept that the length of biography is proportional to protagonist's status, then, using basic methods from exploratory data analysis, we can split all the biographical records from *Taʾrīkh al-islām* into "modal" biographies and biographies-outliers.

The boxplot in Figure 4 presents the basic statistics about the distribution of lengths, while the histogram displays the shape of the distribution. Most importantly, 89% of biographies — "modal" biographies — are less than 144 words (with 50% in the range from 29 to 75 words, with the median at 44), while the rest 11% are statistical outliers. A derivative from the statistical term "mode" that refers to the most frequent value in the population, the word "modal" is used here to qualify biographies of the most typical, "average" members of élites. Biographies-outliers, on the other hand, should represent cultural outliers — the most extraordinary members of élites[2]. Thus, "modal" biographies are most likely to contain

[1] The actual number of events is most likely higher. Since my main goal was to tag biographies, I was not as rigorous tagging events as I was with biographies. Tagging can be corrected at any moment without disrupting the research process.

[2] On cultural outliers, see Malcolm Gladwell. *Outliers: The Story of Success*. 2011.

representative information on larger social processes in the Islamic world. At the same time, patterns discovered in biographies-outliers — the remaining 11% — are least likely to represent élites in general and therefore should not be used for broad extrapolations.

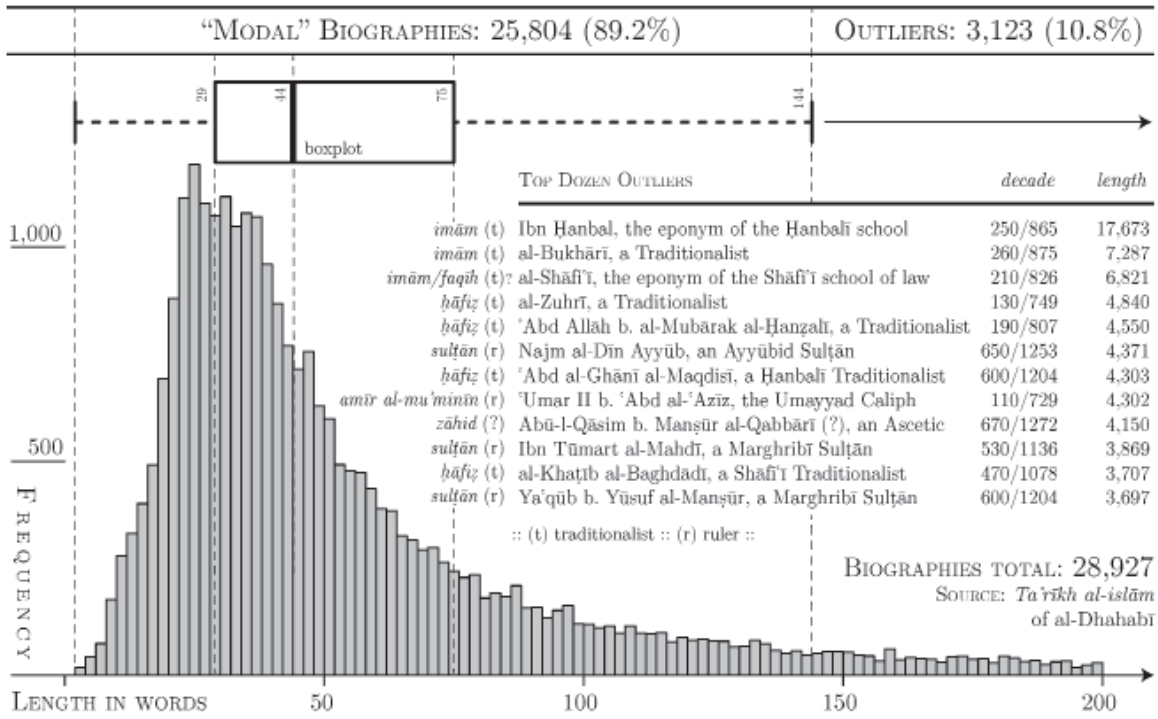


Figure 4: *"Modal" Biographies and Biographies-Outliers in Taʾrīkh al-islām*

Additionally, close look at biographies-outliers may help to get a better idea of al-Dhahabī's biases in *Taʾrīkh al-islām*. Arguably, longest are the biographies of individuals toward whom the author had strong emotional feelings — either positive or negative — which drove him to spend more time and effort to write long, detailed accounts. The "top dozen" from al-Dhahabī's *Taʾrīkh al-islām* looks quite interesting (Figure 4). A dozen is, by no means, a statistically significant sample, but it should suffice to entertain my point.

Everybody in this dozen is a famous person, with an exception of Abū-l-Qāsim b. Manṣūr al-Qabbārī, an ascetic from Alexandria who is kind of a "dark horse" in this list. In terms of socio-religious roles, the list is clearly dominated by the Traditionalists (7 out of 12). On the very top we have Aḥmad b. Ḥanbal, followed by al-Bukhārī, and only then by Muḥammad al-Shāfi'ī, which might be quite unexpected for a Shāfi'ī author, who is expected to favor his own kind[1]. (All three are qualified primarily as *imām*s; Muḥammad al-Shāfi'ī is also qualified as *faqīh*, "jurist"). The only other Shāfi'ī on this list is al-Khaṭīb al-Baghdādī, however, al-Dhahabī goes on for about three pages writing about him as an outstanding Traditionalist before even mentioning that he was one of the most prominent Shāfi'ī jurists. Thus, if al-Dhahabī favored any socio-religious group — however feeble this provisional conclusion may be — he favored the preservers of the Prophetic tradition. Even more interesting — in the light of al-Dhahabī's criticism of the jurists and their inability to serve the *umma*[2] — is the presence of two Ḥanbalīs in this list — Aḥmad b. Ḥanbal himself and 'Abd al-Ghanī al-Maqdisī, who were prominent community leaders[3].

Exploratory analysis: Dates

With the eText converted into a machine-readable format, one can start the extraction of data that can be used for more meaningful exploratory data analysis. The complexity of algorithms for the extraction of dates will differ depending on how dates are recorded in a given eTexts and what level of details one wants to preserve. The easiest way to begin, however, is to use the chronological division of a source in question, if it is available. Luckily, al-Dhahabī divided his "History" into decades. Figure 5 shows the histogram of this distribution of biographies and several LOESS curves that smooth out the "noise" of data allowing to see larger trends. This curve

[1] See, *de Somogyi J.* The Taʾrīkh Al-islām of adh-Dhahabī // Journal of the Royal Asiatic Society of Great Britain and Ireland. 1932. No. 4. October 1. P. 847.

[2] *Kevin Jaques R.* Authority, Conflict, and the Transmission of Diversity in Medieval Islamic Law. Leiden, 2006. P. 4.

[3] On the communal role of Aḥmad b. Ḥanbal, see: *Nimrod Hurvitz.* The Formation of Ḥanbalism: Piety into Power. London, New York, 2002.

may reflect either the actual historical changes that took place in the Islamic world, or al-Dhahabī's inability to consult other sources on earlier periods, or his bias towards certain groups which he decided to ignore for some reason, or some combination of the three. For now, let's take a closer look at the curve itself.

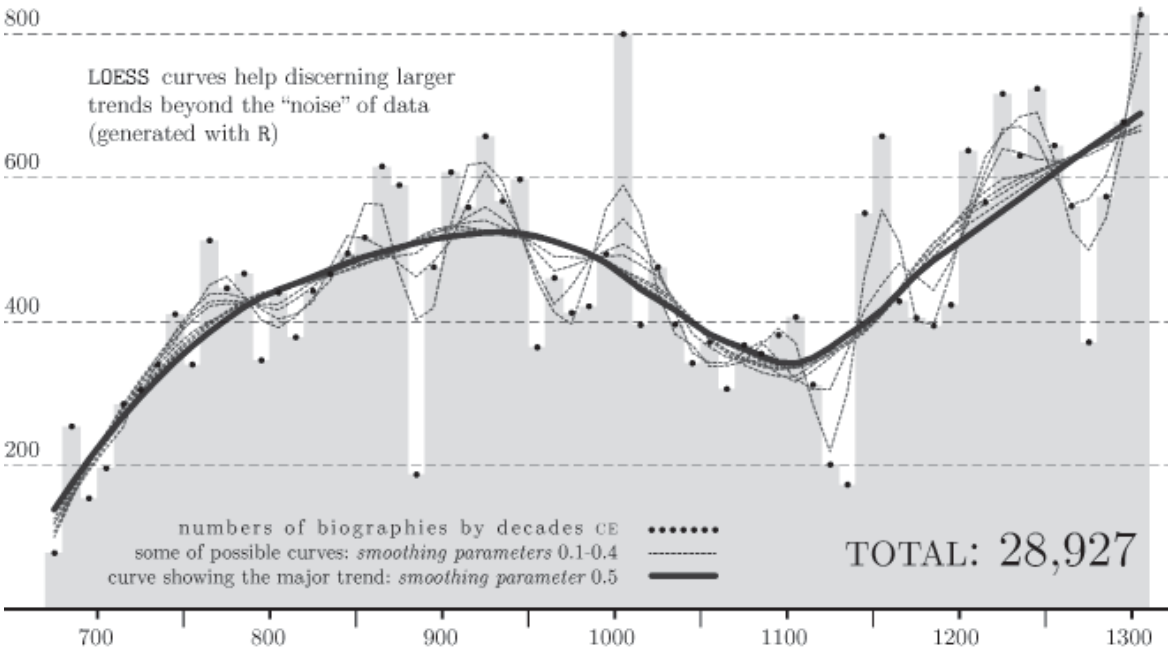


Figure 5: *Distribution of Biographies by Decades CE*

The major curve reflects the death dates of individuals. For this reason, it makes sense to adjust the curve to the left — 30 or 40 years back in time, so to speak. Thus adjusted, it will reflect the early years of those individuals who — young and daring — would have been on the lookout to seize opportunities offered by political, religious, economic, cultural and social circumstances of their time. The adjusted biographical curve then should also reflect the number of these opportunities: the higher the number of opportunities, the higher the number of the individuals who could use them.

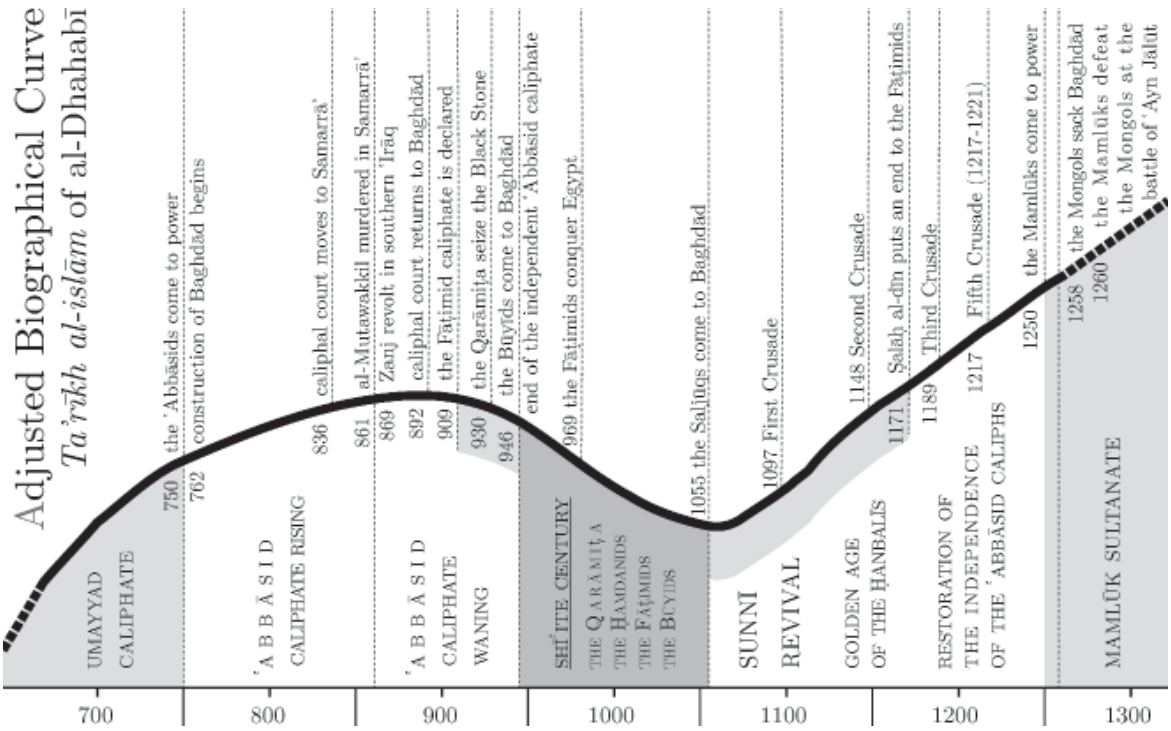


Figure 6: *Adjusted Biographical Curve*

Figure 6 shows that the adjusted curve follows almost perfectly the major developments in the Muslim world: the flourishing of the ʻAbbāsid caliphate, its crisis and the end of the independence of the ʻAbbāsid caliphs; the Shīʻīte century — the Qarāmiṭa who caused great unrest and even dared to seize the Black Stone from al-Kaʻba; the Buyīds who took the power from the ʻAbbāsids; the Fāṭimids who remained the main ideological rival of the ʻAbbāsids — both real and rhetorical — until Ṣalāḥ al-dīn ended their caliphate; the Sunnī revival — the Saljūqs, the saviors of Sunnism, who took the ʻAbbāsids under their wing; the golden age of Ḥanbalism; brief independence of the ʻAbbāsid caliphs, followed by the final fall of the dynasty and the rise of the Mamlūks. Interestingly, the Shīʻīte century occupies almost the entire declining segment of the curve, while the end of the Fāṭimid dynasty marks the point where the curve returns to its highest point before the decline[1]. Ideally, of course, one further

[1] Generated in the same manner, the Shīʻīte curve spikes where the main curve drops.

needs to split this cumulative curve into regional ones and look into how they correlate with historical developments in their respective regions as well as in the Caliphate in general.

Exploratory analysis: Locations

With each biography contextualized chronologically, one can contextualize them geographically. The most complicated issue is how one can computationally extract all meaningful toponyms from an unstructured text. This issue can be solved with methods and techniques from computational linguistics — mainly frequency lists and n-gram. While the concept of frequency list is quite self explanatory, that of n-gram requires some clarifications. In general, n-gram refers to a contiguous sequence of items extracted from a specific text, where n stands for the number of these items: bigram for two items, trigram for three, etc. N-grams can be used for a variety of purposes: tracing word usage, language identification, machine translation, speech recognition, spelling correction, entity detection, information extraction, etc.[1] Entity detection is of particular importance for current research purposes. The main idea behind such usage of n-grams is that preceding words narrow down the possibilities of what the following words can be. For example, if Arabic/Persian *y-z-d* is preceded by *s-k-n*, it is a geographical entity — "he lived in [the city of] Yazd"; while preceded by *l-m*, it is a verbal form — "he/it did not increase". This basic principle can be used to identify toponyms in classical Arabic sources heuristically.

With the toponymic data extracted one can trace how the importance of a specific place changed over time. The most efficient — and logical — way to study this kind of information is to put it on a series of geographical maps[2]. If a biography mentions some place, it is implied that the subject of

[1] N-grams are also actively used in protein and DNA sequencing. With the appearance of Google Ngram Viewer humanists began to use n-grams in their research and teaching.

[2] I relied on R, a free and open-source statistical software with GIS capabilities. I could not have done the coding for this part without Benjamin Schmidt's post "Wide World of Physics" at his blog "Sapping Attention" (URL: http://sappingattention.blogspot.com) and the help of Missy Plegue at the Center for Statistical Consultation and Research (CSCAR), University of Michigan, who helped me to make sense of the code written by Benjamin Schmidt.

the biography is somehow affiliated with that place; some, in fact many, biographies mention more than one place. The maps that will follow consider all of them. It is logical to presume that prominent places are mentioned more often, since they offered more opportunities be they of economic, political or religious nature. Evenly spaced across the entire period of 41–700 AH/661–1300 CE, the following four maps visualize the frequencies of top 100 toponyms mentioned in biographies of al-Dhahabī's *Taʾrīkh al-islām*: almost 13,000 biographies (44,4%) mention top 100 toponyms, altogether slightly over 25,000 times.

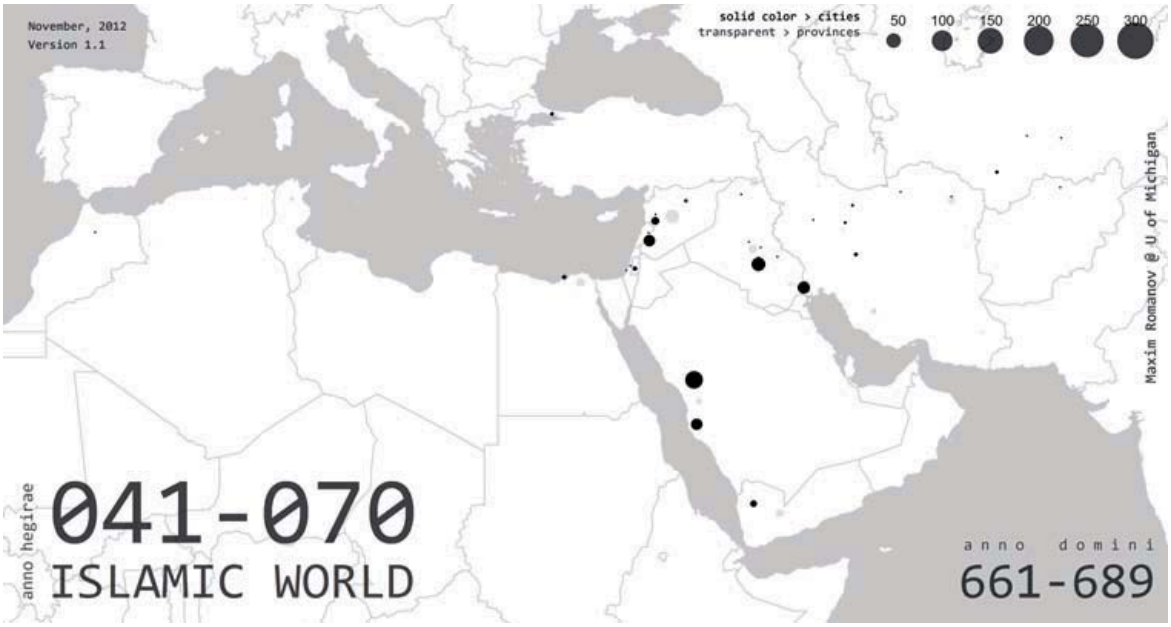


Figure 6: *Top 100 Toponyms: 41–70 AH/661–689 CE*

Figure 6 shows the period of 41–70 AH/661–689 CE, the earliest three decades from the considered section of *Taʾrīkh al-islām*. One can clearly see the major centers: Mecca, Medina, Basra, Kufa and, to a lesser extent, Damascus. The empire grows eastward toward northern Iran and Central Asia, but cities in those regions are not prominent yet. This will change soon.

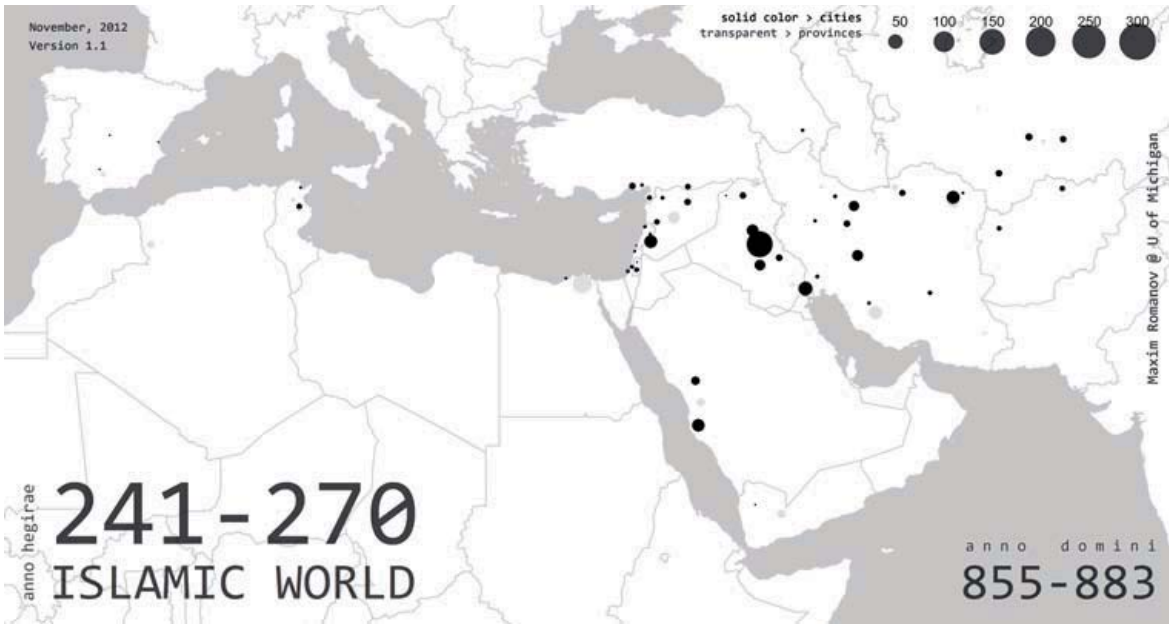


Figure 7: *Top 100 Toponyms: 241–270 AH/855–883 CE*

Two centuries later, 241–270 AH/855–883 CE (Figure 7), both northern Iran and Central Asia gain prominence. Greater Syria, al-Shām, is now more visible on the map of the Islamic world. Nothing significant seems to be going on in the Iberian Peninsula, the Maghreb and Egypt (Miṣr). Former centers — Medina, Basra and Kufa — begin to lose their prominence; Mecca will suffer least because of its sacred status for the entire Islamic community. Yet, all the cities are dwarfed in comparison to Baghdad, the capital of the ʻAbbāsid caliphate which is about to plunge into the turmoil of disintegration. This disintegration seems to have significantly affected the numbers of notable people in al-Dhahabī's "History" — few decades later the number of biographies drops significantly and returns to its previous highest point only about two centuries later (See, Figure 6 above).

Two more centuries later, 441–470 AH/1049–1077 CE (Figure 7), Islamic Spain (al-Andalus) is clearly visible on the map. The western and northern parts of al-ʻAwāṣim[1] are reconquered by the Byzantine Empire in the middle of the 10th century CE. The temporary capital of the ʻAbbāsid empire,

[1] A region at the western part of the modern border between Turkey and Syria.

Sāmarrāʾ practically disappears after the court is moved back to Baghdad. The rule of the Buyīds negatively affects Iraq, while most part of Iran, Central Asia and Afghanistan continue flourishing during this period of the rise of the Saljūqs.

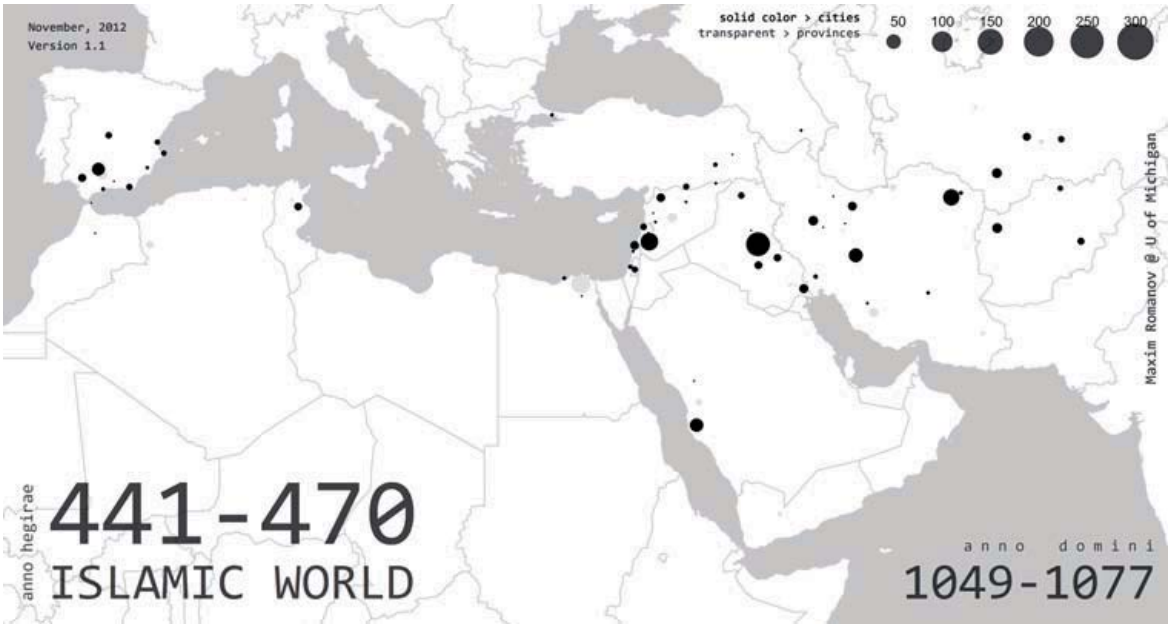


Figure 7: *Top 100 Toponyms: 441–470 AH/1049–1077 CE*

The end of the period, 671–700 AH/1272–1300 CE (Figure 8), is drastically different. Spain is being erased from the map of the Islamic world by the ongoing Reconquista. Having flourished in the earlier periods, the eastern lands of the Muslim world — Iran, Central Asia and Afghanistan — fade away in the aftermath of the Mongol invasions. Still a prominent spot on the map, Baghdad will never recover from the Mongol sack of 1258 CE and by the last decade covered in *Taʾrīkh al-islām* will shrink into a tiny dot. Consolidated under the Ayyūbids, Syria is now the seat of a new power — the Mamlūks, the saviors of the world from the Mongol menace. Under the Mamlūk rule, Egypt appears prominently on the map — with Cairo ready to become the next major center of the Islamic world.

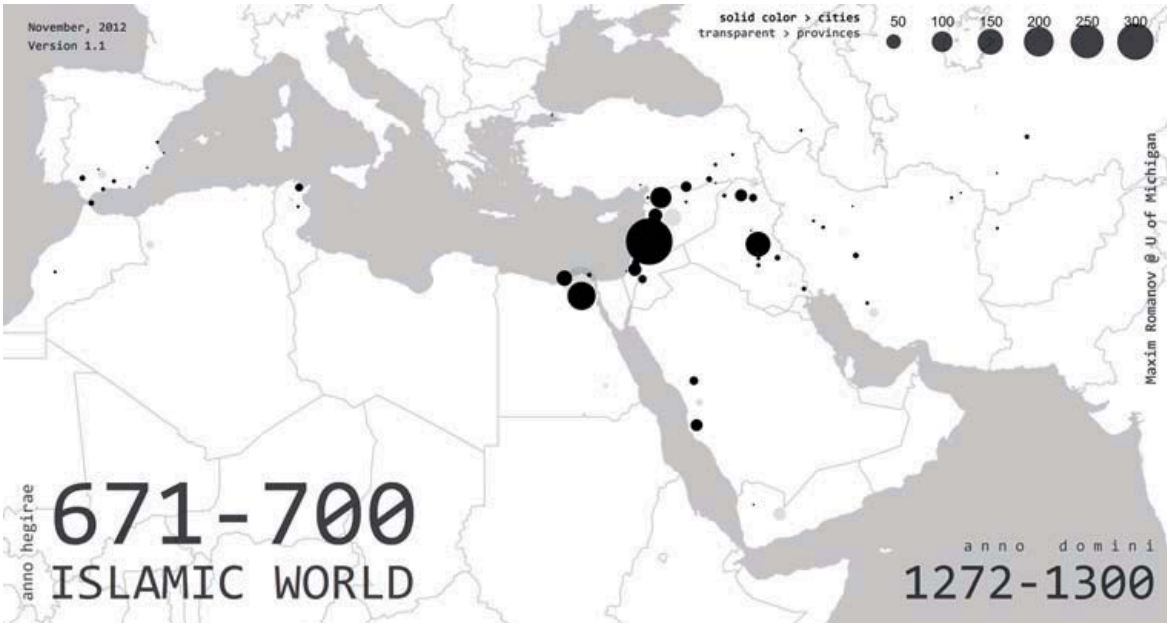


Figure 8: *Top 100 Toponyms: 671–700 AH/1272–1300 CE*

These are just 4 maps and they hardly do the justice to the toponymic data from *Taʾrīkh al-islām*. The best way to visualize these maps is to animate them — this will allow to bring out the dynamics of change. The animated map of these data — **50 seconds of Islamic History** (Version 2.x) — can be found online at http://maximromanov.github.io/2013/02-07.html. Like static maps, each frame shows three decades AH with each following map overlapping with the previous two, i.e. frame 1 shows 41–70 AH; frame 2: 51–80 AH; frame 3: 61–90 AH, etc. It is worth watching this visualization more than once, each time concentrating on a specific region.

The interpretation of these toponymic data is tricky and this explanation is just a preliminary attempt to weave them into a bigger picture. It is worth noting, however, that the toponymic data from *Taʾrīkh al-islām* is quite similar to the representation of the Islamic world in *Shadharāt al-dhahab* of Ibn ʻImad (d. 1089/1679)[1].

[1] See, Graph I in: *Bulliet Richard W.* Conversion to Islam in the Medieval Period: An Essay in Quantitative History. Cambridge, 1979. P. 8. Ibn ʻImad's biographical collection covers the entire first millennium of Islamic history, but includes only ~8,500 biographies, of which only ~6,100 provide information on the places of origin of their subjects. Ibn ʻImad was also a Damascene, but he spent a lot of time in Cairo; unlike al-Dhahabī, he belonged to the Ḥanbalī school of law.

Exploratory analysis: "Descriptive Names"

In combination with already available variables, heuristically extracted "descriptive names" will help to get insights into different social, religious, tribal and other groups that flourished in the Islamic world. As geographical maps were helpful for the visualization of top 100 toponyms over time, word clouds will help to deal with large numbers of "descriptive names". The principle of the word cloud is quite simple: the more frequent the word, the larger it is in the cloud. One of the most efficient tools for generating word clouds, Wordle is an effective exploratory tool that allows to get insights into large quantities of text[1]. With temporal, toponymic and onomastic data already extracted one can zoom in to any group of individuals that can be described with these kinds of data: specific legal school, specific "secular" occupation, specific region, specific period or some combination of the above. Looking into the co-occurrences of "descriptive names" one can get a good idea about the composition of the entire group.

For now, however, let's take a broad look at the identities of notable Muslims and how they changed over time. Figure 9 shows that the majority of Muslims during the period of 41–140 AH/661–757 CE were strongly affiliated with the cities of Kufa (al-Kūfī), Medina (al-Madanī) and Basra (al-Baṣrī); to a lesser extent with Mecca (al-Makkī) and the Syrian cities of Damascus (al-Dimashqī) and Homs (al-Ḥimṣī), and also with Egypt in general (al-Miṣrī). Clearly visible are the names that emphasize affiliation with the Prophet (most prominently, al-Anṣārī, "the Helper of the Prophet"; the words *rasūl*, "messenger", and *nabī*, "prophet", convey the same meaning).

Most striking, however, is the abundance of tribal affiliations with the tribe of the Prophet understandably prevailing: *banū* Quraysh (al-Qurashī), *banū* Asad (al-Asadī), *banū* Umayya (al-Umawī), *banū* Azd, or Asd (al-Azdī), *banū* Makhzūm (al-Makhzūmī), *banū* Thaqīf (al-Thaqafī),

[1] A free online tool, Wordle.net has already become the subject of over a dozen academic articles that explore how it can benefit both teaching and research. It creates aesthetically pleasing visualizations of quantitative data without the use of numbers, the selling point for the humanists who usually do not have "a head for numbers".

banū Taym (al-Taymī), *banū* Hāshim (al-Hāshimī), *banū* Sulaym (al-Sulamī), *banū* Hamdān (al-Hamdānī), plus at least two more dozen other tribes and clans. This is by no means surprising, since this was the tribal period of the Islamic society and non-Arab converts had to become clients of Arab tribes: the word *mawlá*, "client", is the third most prominent "descriptive" name of this period. Religious knowledge is marginally reflected in Muslim names (al-Qāḍī, al-Faqīh); such word as *amīr*, "commander", seems to emphasize the military nature of the early Islamic community.

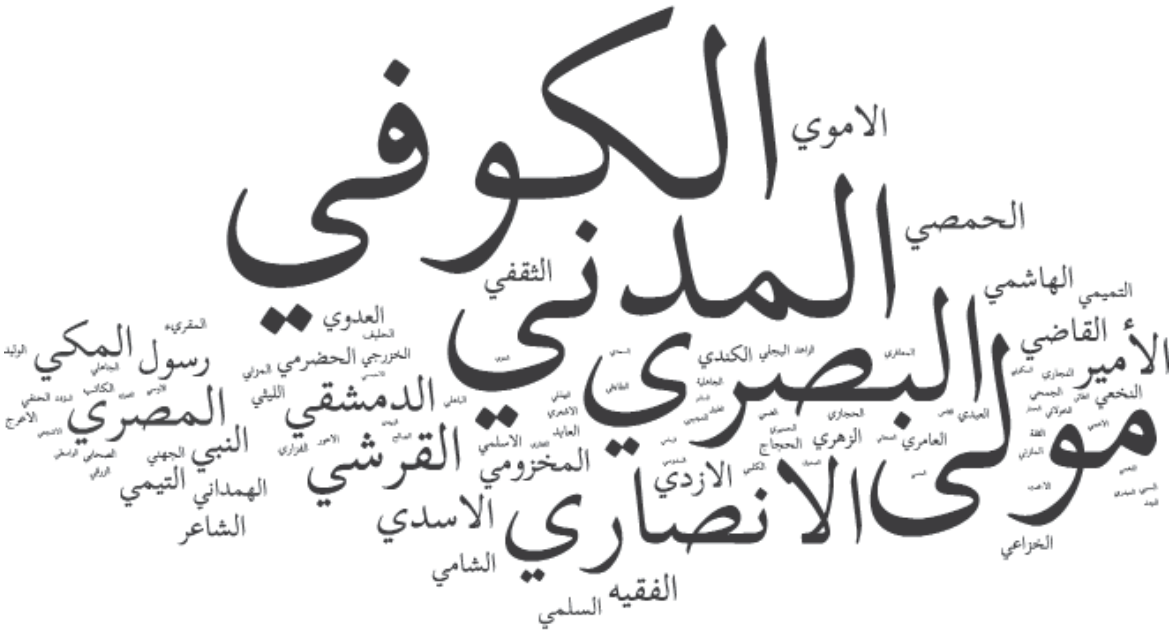


Figure 9: *The Wordle of Islamic Names (41–140 AH/661–757 CE). Top 10 : 382 : al-Kūfī; 316 : al-Madanī; 295: mawlá; 251 : al-Baṣrī; 178 : al-Anṣārī; 104 : al-Qurashī; 089 : al-Miṣrī; 084: amīr; 079 : al-Dimashqī; 072 : al-Makkī*

The overall situation in the middle of our period, 321–420 AH/933–1029 CE (Figure 10), is quite different. First of all, the geography of names changed: while a significant number of Muslims is now strongly affiliated with Baghdad (al-Baghdādī), one can also see a significant presence of other regions: Iran with its cities of Nishapur (al-Naysābūrī), Isfahan

(al-Iṣbāhanī), Herat (al-Harawī), Rayy (al-Rāzī), Merw (al-Marwazī) and many others; Spain (al-Andalusī) with Cordoba (al-Qurṭubī); Syria with Damascus (al-Dimashqī); and Egypt — still without any specific urban centers (al-Miṣrī).

Figure 10: *The Wordle of Islamic Names (321–420 AH/933–1029 CE). Top 10: 642 : al-Baghdādī; 370 : al-Naysābūrī; 349 : faqīh; 260 : qāḍī; 254 : al-Qurṭubī; 252 : al-Miṣrī; 246 : muqriʾ; 222 : ḥāfiẓ; 194 : al-Iṣbāhanī; 159 : al-Dimashqī*

While tribal identities fade away, religious knowledge becomes an essential part of Islamic identity. In addition to jurists (al-Faqīh) and judges (al-Qāḍī), who became much more prominent in this period, one can clearly see reciters of the *al-Qurʾān* (al-Muqrīʾ), transmitters of the Prophetic tradition (al-Ḥāfiẓ, al-Muḥaddith, al-Thiqa), pietists of all kinds (al-Zāhid, al-Ṣāliḥ, al-ʻĀbid, al-Ṣūfī), "public preachers" (al-Wāʻiẓ), muezzins (al-Muʾadhdhin), Friday-preachers (al-Khaṭīb). The development of Islamic administration is attested by the appearance of "ministers" (al-Wazīr) and scribes (al-Kātib).

Secular occupations also become a part of the Islamic onomasticon — al-Bazzāz ("textile dealer"), al-Warrāq ("paper seller", "bookseller", "copyist"), al-Tājir ("merchant"), al-Qaṭṭān ("cotton dealer"), al-ʻAṭṭār ("perfume mer-

chant", "druggist"), al-Ṣayrafī ("money-changer"), al-Ṣaffār ("coppersmith"), although a different Wordle would be more efficient to visualize these data. Affiliations with legal schools — al-Shāfi'ī, al-Mālikī, al-Ḥanafī — begin to gain prominence.

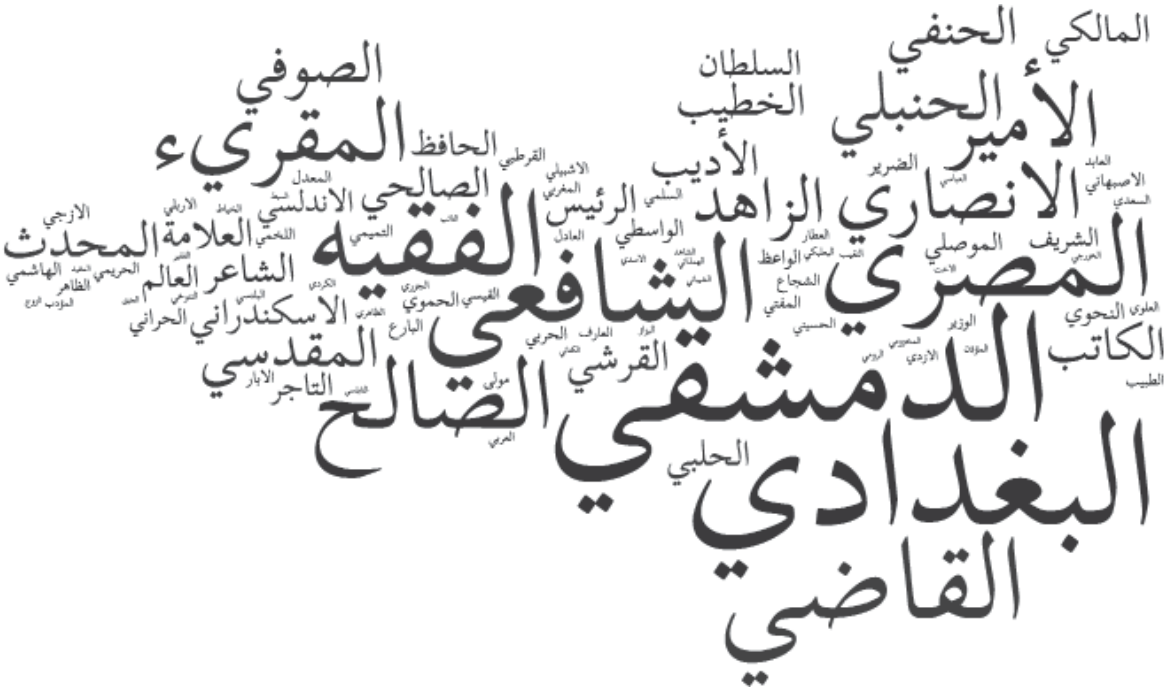


Figure 11: *The Wordle of Islamic Names (601–700 AH/1204–1300 CE). Top 10:; 709 : al-Dimashqī; 677 : al-Baghdādī; 554 : faqīh; 538 : al-Shāfi'ī; 522 : qāḍī; 498 : al-Miṣrī; 435 : ṣāliḥ; 388 : muqri'; 363 : amīr; 325 : al-Anṣārī; 298 : al-Ḥanbalī*

The names of the last century of our period, 601–700 AH/1204–1300 CE (Figure 11), tell yet another story. The geography of names shifted again: Damascus and Baghdad are the main cities of this period. Greater Syria with its numerous cities is now the most prominent region. The importance of Egypt (al-Miṣrī) is growing together with its city of Alexandria (al-Iskandarānī). Muslim Spain continues to fade away, while eastern provinces practically disappear.

Religious identities are now expressed primarily through affiliation with legal schools — al-Shāfi'ī, al-Ḥanbalī, al-Ḥanafī, al-Mālikī. Piety be-

comes even more important part of Islamic identity (al-Ṣāliḥ, al-Zāhid, al-ʻĀbid); one can also see that mysticism (al-Ṣūfī) gains momentum. The importance of religious knowledge is still expressed through jurisprudence (al-Qāḍī, al-Faqīh), recitation of the *al-Qurʾān* (al-Muqrīʾ) and transmission of the Prophetic tradition (al-Muḥaddith, al-Ḥāfiẓ), but the priorities seem to have shifted. The lineage to the Prophet and his close companions — al-Anṣārī, al-Qurashī and al-Sharīf — becomes a significant part of social capital. In the previous period, "public preachers" (al-Wāʻiẓ) were quite prominent, now they disappear passing the baton to Friday-preachers (al-Khaṭīb). Such names as al-Amīr, "commander", and al-Sulṭān, "sultan", seem to reflect the contribution of the military élites — the Zangids, Ayyūbids and Mamlūks — to the religious and social life of the Islamic community.

Conclusion

The main goal of this paper was to demonstrate the potential of the text-mining method and emphasize its importance. Decades ago, the appearance of printed editions of pre-modern Arabic sources made a significant impact on the development of the field of Islamic studies — a great number of sources that previously were trapped in their manuscript form became available to the scholars of Islam over a relatively short period of time. This development must have been one of the main reasons that lead to the significant increase in publications on the Islamic world that began in the 1980s (see, Figure 1 above). These printed editions allowed scholars to work with a much wider range of primary sources and ask questions that were difficult if possible to ask when manuscripts of these sources were scattered all over the world.

The appearance of historical sources in digital format has already made a strong impact on historical studies in general. Digital humanities is a rapidly growing field now — the 126th Annual Meeting of the American Historical Association (Chicago, 2012) featured a series of nearly two dozen sessions on digital history. Titled "The Future is Here", this series particularly emphasized the value of text-mining and geographical in-

formation systems (GIS) for working with "big data". It would be logical to expect that the digitalization of Islamic sources will make a similar impact on our field. This will affect how we study these sources, what research questions we ask of them, and how we train new generations of scholars. The ability to work with "big data", so amply supplied by generations of Muslim authors, will be crucial for the future of Islamic studies as a discipline.