Digital humanists often say they would like to read more work in progress. Our blog posts represent such work. We worked intensively over months to create and interpret this data. There was much back and forth. The data presented here offers a snapshot of our work at a specific point in time.

For detailed information on the data and the files contained therein, please see the [data set](https://zenodo.org/record/8233103) released with these posts. What follows is a summary of the data on which our work on Ibn ʿAsākir is based.

The book files

For the *TMD*, we used the 1995–2001 Dār al-Fikr edition of ʿUmar al-ʿAmrawī and ʿAlī Shīrī. The edition consists of eighty volumes, with volumes 71–74 representing a *mustadrak*, or supplement, added by the editors (including additional biographical entries) and volumes 75–80 containing indexes. This is the first modern scholarly edition of the *TMD*, and it was completed only in 2001, reflecting the challenges the work posed to editors. There was no complete manuscript, and the work’s sheer size and the complexity of its *isnād*s created formidable obstacles. Judging by Ibn Manẓūr’s *Mukhtaṣar*, which substantially reproduces the *TMD* but without *isnād*s, the edition and the manuscript tradition on which it is based contain lacunae.[[1]](#footnote-1)

The machine-readable file of Ibn ʿAsākir’s list of teachers (*Muʿjam al-shuyūkh*) that we used is based on an edition by Shākir Faḥḥām and Wafāʾ Taqī al-Dīn, published in Damascus by Dār al-Bashāʾir in 1421/2000.[[2]](#footnote-2)

To obtain information from Ṭalāl b. Saʿūd al-Daʿjānī’s *Mawārid Ibn ʿAsākir*, we OCR’d volume 1 of the printed text[[3]](#footnote-3) with a Kraken model (version 3.0.0.0b21.dev6) and then extracted a list of author names and book titles to be searched for in the *TMD*. The OCR’d text was lightly checked and corrected against a PDF of the printed text by Hamid Reza Hakimi, who also tagged the author and title names in the text.

*Isnād*s

We relied on paragraph markers in the *TMD* to identify the beginnings of *isnād*s and took the next fifty word tokens following each beginning to constitute the *isnād*. We thus extracted 79,470 *isnād*s. We gave each a unique identification number, recorded the milestone in the text at which it was located and assembled all the *isnād*s in a table.

We then split each *isnād* into names within cells, using transmission terms. In the example in Table 2.1, we used the words in square brackets to split the *isnād* string, which then gave us the list of names shown in the table.

]أنبأنا] أبو الحسن عبد الغافر بن إسماعيل [أنا] محمد بن يحيى بن إبراهيم [أنا] أبو عبد الرحمن السلمي [قال سمعت] الحسين بن أحمد [يقول سمعت] الدقي [يقول سمعت] أبا عبد الله

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| name\_5 | name\_4 | name\_3 | name\_2 | name\_1 | name\_0 |
| أبا عبد الله | الدقي | الحسين بن أحمد | أبو عبد الرحمن السلمي | محمد بن يحيى بن إبراهيم | أبو الحسن عبد الغافر بن إسماعيل |

Table 2.1: An example *isnād* split into names.

We filtered out any *isnād*s that lacked a name in position 0 (that is, the name of a direct informant) as well as *isnād*s that forked, meaning that Ibn ʿAsākir introduced an additional name with another transmission chain after naming his direct informant. Applying these filters left us with 77,231 *isnād*s.

Name List

Ibn ʿAsākir refers to a lot of people. From the *TMD*’s *isnād*s we compiled a list of over 9,238 ‘surface forms’ of names (that is, names as they appear in a text), which mapped to 3,885 distinct identities. As we explain later, we did not map all surface forms – far from it. For comparison, al-Ṭabarī’s three works contained 4,921 surface forms mapped to 3,787 identities. Both Ibn ʿAsākir and al-Ṭabarī sometimes use the same surface form for different people and different surface forms for the same people.

To facilitate handling and filtering of the names, we normalised each to a standard Latin-script form. These forms are based on the transliteration system of the *International Journal of Middle Eastern Studies* (*IJMES*) but without diacritics.

To decide on the normalised form to be used, we looked for patterns, searching, for example, for similar surface forms that differed only in the inclusion or omission of a *kunya*, or for names that consistently appeared in the same position within *isnād*s, with the same names preceding and following it. The normalised name is not always the most commonly cited surface form. Instead, it is often the shortest form or the one that best distinguishes the named individual from other, similarly named people.

Although Ibn ʿAsākir uses thousands of different surface forms, and many for the same person, he refers to his most important direct informants typically in a consistent way. Readers may wish to investigate the data closely on this point.[[4]](#footnote-4)

Transmission terms

Our list of transmission terms in the *TMD* built on an earlier version of such a list initiated by R. Kevin Jaques and expanded for our work on al-Ṭabarī. We augmented the list with additional terms found at the beginnings of paragraphs in the *TMD*. We also created a list of words to remove from the cells once we had split them.

We then grouped semantically related transmission terms together and decided to focus on the six most frequently used terms, as they initiate more than 99% of the *isnād*s.

We also searched within *isnād*s for other transmission terms that Ibn ʿAsākir uses in conjunction with the six to explain his method. These include the adverbial modifiers *idhnan*, *munāwalatan*, *lafẓan, ʿāliyan*, *shifāhan*, *qirāʾatan* and *ijāzatan*, as well as references to a person’s writing (*fī kitābihi*). We chose the terms on the basis of their frequency within the *isnād*s.

Author names and titles of works

We assembled a list of persons whom modern scholars credit with authoring books on which Ibn ʿAsākir relied. We gave each author a unique ID and used these IDs throughout our data set. There are 238 such authors. However, their ID numbers, assigned before normalisation was complete and thus before variant names for the same individual were identified, run to 241. We refrained from renumbering the authors to avoid disrupting links within our data.

To identify the authors, we drew (as noted above) on the works of Scheiner, Mourad and al-Daʿjānī, who gave us other important information, including the titles of the authors’ works (which we used for searches) and the dates of their deaths. We complemented this information with details from the OpenITI and occasionally the third edition of the *Encyclopaedia of Islam*.

Search terms and results

We searched within *isnād*s for titles used in conjunction with author names. This was not very successful.

We also did some general searches – Underwood’s Boolean fishing expeditions. We include the results and discuss these in posts 5 and 6.

Three terms proved illuminating about Ibn ʿAsākir’s working methods and the character of his sources. The term *tasmiya* refers to a type of written source that Ibn ʿAsākir used in preparing the *TMD* (and indeed, the full title of the *TMD* contains the term *tasmiya*). *Qaraʾtu bi-khaṭṭ* denotes his reading of his predecessors’ handwritten text. He introduces written texts, and some books, with the term *dhakara*. We believe that these three terms offer a good representation of the nature of Ibn ʿAsākir’s sources, and we will argue that they merit more attention than they have received.

Text reuse data

We identified thirty-nine texts within the OpenITI corpus that Scheiner cited as sources for Ibn ʿAsākir. We then compared these files with the *TMD* to detect text reuse. We collated the results and graphed them with Power BI software in order to reveal the general pattern of reuse for each work.

All texts in the OpenITI corpus are divided into numbered 300-word chunks, which we call milestones. For each instance of overlap between two books, the reuse comparison files record the milestone location of the alignment in each book.

We also noted the milestone for each *isnād* in our *isnād* listing. This allowed us to investigate whether citations of an author’s name are accompanied by reuse of that author’s text.

The data we present here has not yet been carefully studied. But our general impression is that there is less correlation between citation and reuse than modern scholars might assume.

Challenges

This work was highly iterative. We ran into challenges and then addressed them.

Paragraph markers

The text of the *TMD* in the OpenITI corpus had been annotated previously by team members, but the annotation procedure does not normally involve checking paragraph tags, as this is very time-consuming and is not required for text reuse detection or other corpus-wide studies we are undertaking. We thus felt some light editing would be useful. Hamid Reza Hakimi spent more than ten hours correcting paragraph markers. We know additional work in this area could be done.

A problem that arose from our use of paragraph markers to identify *isnād*s was that we sometimes missed distinctions between separate *isnād*s: Ibn ʿAsākir sometimes follows an *isnād* with a second, distinct *isnād* for the same the *matn* (transmitted text). To give an example, Ibn ʿAsākir always transmitted from al-Khatib through just one intermediary, so al-Khatib’s position in an *isnād* should always be 1, with the intermediary in position 0. However, because of the way our data was extracted, he sometimes appeared in both position 1 and position 2.

To address this issue, we tagged secondary *isnād*s separately, using # for primary *isnād*s and ## for secondary ones. Of the 79,470 *isnād*s we originally extracted, 2,236 (3%) were found to contain a second *isnād*. We are aware that we have not identified all cases of secondary *isnād*s in the text, but their number does not seem to affect our data or interpretation, as we are focused mainly on the big picture of the *isnād*s and not on, say, finding rare and interesting ones. It is always possible to retag the text in the future to account for the additional *isnād*s and then to regenerate the data.

Our final table excludes *isnād*s that contain secondary *isnād*s, as it proved difficult to collect statistics using such data. But we have examined the excluded *isnād*s and provide them in the data set. We do not believe that excluding them alters our conclusions.

Errors

Many problems were rooted in the long history of the text, encompassing both manuscripts and the modern edition. For example, copyists surprisingly often confused Ḥasans and Ḥusayns. The process of transferring the text to print also gave rise to inconsistencies that challenged our methodology. A common issue was spacing: the same pair of words in a name might be written separately or as one word (for example, Abu al-Barakat was sometimes rendered without a space between Abu and al-), which generated two different surface forms of the name.

Our normalisation process addressed such spacing issues, tracing names with divergent spacings to the correct individual. We were generally conservative with normalisation. On rare occasions, we would add an al-, for example, where the text did not have one. And in a handful of cases, we accepted minor variation in author names (e.g. when all major elements of a name were the same) and normalised the variant forms to one and the same name so we could search for author names as expansively as possible. Otherwise, we accepted the text as it was and reckoned its state as a limitation on what we may conclude from it. Ḥasan in Arabic script is normalised as Hasan, even when we suspected that the copyist had made a mistake and that the name should in fact be Ḥusayn. This means that if Ḥusayn was in fact the intended person, we are perpetuating the error and have an invented person on our Name List. In the future, the data set will benefit enormously from the use of natural language processing tools to weed out such errors.[[5]](#footnote-5)

Many surface forms

The normalisation process entailed many difficulties. We often wondered what Ibn ʿAsākir might have been thinking when he referred to the same person in many ways – was he assuming the knowledge of his audience?

Interestingly, when referring to persons further down in the *isnād* chains, he himself occasionally admitted to uncertainty about the identity of individuals so vaguely named. One entry in the *TMD* is dedicated to ‘Aḥmad b. ʿAlī’, about whom Ibn ʿAsākir says, ‘I believe (*aẓunnu*) that he is Abū ʿUmar al-Ṣūfī.’[[6]](#footnote-6) On another occasion, he gives a chain of transmission but suggests a correction to it, arguing that the transmission more likely involved a father rather than a brother, as the chain claims (*ʿan akhīhi aẓunnuhu ʿan abīhi*).[[7]](#footnote-7)

Much of our work on name normalisation involved simple transcription of Arabic surface forms to Latin-script normalisations. However, in many cases some sleuthing was required. We did substantial work on named entity recognition (NER), which is a technique of information extraction that involves identifying and classifying names, and we drew on our knowledge of Arabic onomastics and familiarity with common versus rare names to trace different surface forms to the same person.[[8]](#footnote-8) This was very much a brute-force effort. We have experimented with further ways to disambiguate names automatically, such as by using a clustering algorithm, but have not been able to move beyond where we are now.

Sarah’s initial Name List contained more than 6,000 surface forms and their normalisations. For 89% of *isnād*s, we could not identify a direct informant (judging by the proportion of *isnād*s that still contained only an unnormalised, Arabic-script name in the first position in the *isnād*). Hamid Reza Hakimi joined the effort, and together we developed more computational ways to normalise names. This critical work brought the number of name normalisations to 9,238. We now have at least one normalised name for 91% of Ibn ʿAsākir’s *isnād*s. Further normalisation would take much work and likely yield little benefit, insofar as we seek the big picture.

A particular problem is posed by *isnād*s that contain two distinct names in position 0, as in ‘Abū Ghālib [Abu Ghalib] and Abū ʿAbd Allāh [Abu 'Abd Allah] told us that …’ To partially address this issue, we searched the unnormalised *isnād*s for surface forms associated with the twenty most frequently cited transmitters. We found many partial matches, in which one of these names was embedded in a longer list of names. We counted these as matches but marked them in our data as ‘partial matches’ to signal that the string might contain additional names we have not normalised.

We are satisfied with our Name List for now, as it serves our current purposes. But we anticipate developing better methods for refining it in the future. Ryan Muther has been working on parallel problems, and his work will be helpful. We also expect work on NLP to be useful. For example, names might be extracted from *isnād*s through the use of an automatic parser such as those that exist now for English and other languages. Though various projects are developing such tools for Arabic,[[9]](#footnote-9) we have only begun testing them for historical Arabic.

Uncertainty

In a very few cases, we have normalised different surface forms to a common short name, even when examination of the elements in longer forms of the name seems to point to two distinct persons. For example, Ibn ʿAsākir often cites two transmitters called Umm al-Mujtaba and Umm al-Baha'. He also occasionally names them together, as ‘Umm al-Mujtabāʾ Fāṭima bt. Nāṣir’ and ‘Umm al-Bahāʾ Fāṭima bt. Muḥammad’. Somewhere along the way, however, the genealogies of the two Fāṭimas got mixed up, so the *TMD* sometimes refers to Umm al-Mujtaba as Fāṭima bt. Muḥammad and to Umm al-Baha' as Fāṭima bt. Nāṣir (and also as Fāṭima bt. Muḥammad b. Aḥmad or Fāṭima bt. Muḥammad b. ʿAbd Allāḥ). Given that Ibn ʿAsākir often mentions the two names together, it seems clear that they identify two separate women, so we felt we could safely refer to them by their shortened names and ignore the confusion about their genealogies. This seemed a good route especially in cases in which the short-form names were not themselves very common; the situation would have been very different with, say, an Abū al-Ḥasan or an Abū al-Ḥusayn.

Conventions

We cite our data set throughout the posts that follow, sometimes using the *isnād* number within the data set. We also give milestones for *isnād*s. We use the normalised forms of informants’ names but often refer to authors by their better-known, transliterated names. When translating passages from the text, we give the surface form as it appears in the text, followed by the normalised form in square brackets. Readers can recognise normalised names by their lack of diacritics. Normally, OpenITI blog posts lack diacritics, but for this series, we found them necessary.

1. On the book’s publication history, see Steven Judd and Jens Scheiner, ‘Introduction’, in Judd and Scheiner (eds), *New Perspectives on Ibn ʿAsākir*, 1–3; Nancy Khalek, ‘Prologue: The Publication of the Dār al-Fikr Edition of Ibn ʿAsākir’s *Taʾrīkh Madīnat Dimashq*’, in Judd and Scheiner (eds), *New Perspectives on Ibn ʿAsākir*, 4–8. See also James E. Lindsay, ‘Appendix C: Major Lacunae in *TMD*’, in Lindsay (ed.), *Ibn ʿAsākir and Early Islamic History*, 141–3. See also a further critique of the Dār al-Fikr edition (supporting our exclusion of the *mustadrak*, which Lindsay includes) in Mourad, *Ibn ‘Asakir of Damascus*, 78–80. [↑](#footnote-ref-1)
2. Ibn ʿAsākir, *Muʿjam al-shuyūkh*, ed. Shākir Faḥḥām and Wafāʾ Taqī al-Dīn, 3 vols (Damascus: Dār al-Bashāʾir, 2000), 0571IbnCasakir.MucjamShuyukh.Shamela0012750. [↑](#footnote-ref-2)
3. Al-Daʿjānī, *Mawārid Ibn ʿAsākir*, vol. 1. [↑](#footnote-ref-3)
4. [See Savant and Seydi, ‘Ibn ʿAsākir and His History of Damascus’](https://zenodo.org/record/8233103), ‘Isnads’, table ‘TC\_NameCount\_Arabic’. [↑](#footnote-ref-4)
5. ʿAlī b. al-Ḥasan Ibn ʿAsākir, *Tārīkh madīnat Dimashq: Wa-dhikr faḍlihā wa-tasmiyat man ḥallahā min al-amāthil aw ijtāza bi-nawāḥīhā min wāridīhā wa-ahlihā* [henceforth *TMD*], ed. Muḥibb al-Dīn Abī Saʿīd ʿUmar b. Gharāma al-ʿAmrawī, 80 vols (Beirut: Dār al-Fikr, 1995–2001), 2:412, 0571IbnCasakir.TarikhDimashq.JK000916, ms. 00652. [↑](#footnote-ref-5)
6. Ibn ʿAsākir, *TMD*, 5:81, ms. 01439. [↑](#footnote-ref-6)
7. Ibn ʿAsākir, *TMD*, 2:412, ms. 00652. [↑](#footnote-ref-7)
8. Reading through these names gave us a good chance to think about Arabic onomastics and to see a few rarities. For example, for most of history in the Muslim world, compared to the Christian West, there is generally less tendency for a son to share his father’s given name, but we do have the case of Abu al-Hasan Muhammad b. Muhammad b. Muhammad b. al-Muhtadi. This man went by the *laqab* ‘Father of al-Ḥasan’ but was the son and grandson of Muḥammads. Nothing fundamental is learned from looking at this one name, but looking at thousands gives a sense of the possibilities. [↑](#footnote-ref-8)
9. See, for example, the work of Nizar Habash and his CAMel laboratory’s tools. [↑](#footnote-ref-9)