You have now entered the weeds.

Over the past months, we have been asking ourselves what we would need to tell the story of how al-Tabari (d. 310/923) worked. Our first blog post summarised the procedure that we used to generate our data set. This post describes the data in greater detail.

Our data set is based on al-Tabari’s *Taʾrikh al-rusul wa-l-muluk*, *Jamiʿ al-bayan ʿan taʾwil ay al-Qurʾan* and *Tahdhib al-athar*, included in the [2021.1.4 release](https://zenodo.org/record/4513723) of our data via Zenodo under the identifiers Shamela0009783BK1, Shamela0007798 and JK008250Vols, respectively. The books have been annotated with tags belonging to the OpenITI mARkdown scheme to identify certain structural and content features. The present study made use of these tags. We have used the structural tagging in the above release.

## Data Tables

We produced five tables that display the data gleaned from al-Tabari’s three books. We did not set out to create these five; rather, they emerged in the course of the work. What follows is an account of research in progress. We expect to improve on our method in the future.

### 1. Transmission Chains

We collected all *isnad*s from the *Taʾrikh*, the *Tafsir* and the *Tahdhib* that start a paragraph with either *haddathani* (‘he told me’) or *haddathana* (‘he told us’); we included also related phrases, such as *wa-qad haddathani*.These phrases occur a total of 32,115 times in the three books – 2,065 times in the *Taʾrikh*, 26,234 in the *Tafsir* and 3,816in the *Tahdhib*. The largest number of such citations is thus found in the *Tafsir*, which is also the longest of the three works, containing 2.91 million word tokens to the *Taʾrikh*’s 1.46 million and the *Tahdhib*’s roughly 346,000.

We put each *isnad* into a row in a CSV file.



Image 1: A section of al-Tabari’s *Taʾrikh*, tagged with OpenITI mARkdown encoding (implemented in EditPad Pro), with a transmission chain featuring Muhammad b. Humayd al-Razi (d. 248/862) transmitting to al-Tabari from Jarīr b. ʿAbd al-Hamid (d. 188/804). Muhammad b. Humayd also transmitted material from Muhammad b. Ishaq’s (d. 150/767) *Sira* to al-Tabari.

As noted in our previous post, the style of al-Tabari’s writing and citation in the *Taʾrikh* differs from that in the *Tafsir* and the *Tahdhib*, as al-Tabari invokes *isnad*s less frequently in the *Taʾrikh*. Furthermore, in the latter parts of the *Taʾrikh*, the *haddathani* or *haddathana* formula is extremely rare (as evident from Graph 1, in Post 1). Instead, beginning with his account of the founding of Baghdad, al-Tabari typically uses the passive voice to attribute information (*dhukira ʿan*, ‘It is reported from’ so-and-so that …).[[1]](#footnote-1)

We relied on machine-readable files of al-Tabari’s books in the OpenITI corpus, annotated and vetted by KITAB team members Maxim Romanov, Hamid Reza Hakimi and Masoumeh Seydi. In these files, the beginning of each paragraph is indicated with a #, which enabled us to write our search pattern (‘regular expression’) to detect instances of *haddathani*/*haddathana* specifically at the start of a paragraph (regular expressions are a formalised way of constructing search patterns that is used in many computer languages and text-editing softwares). The reason we excluded non-paragraph-beginning uses of the formula from our search is that when the phrase appears within a transmission chain, it generally introduces not al-Tabari’s direct informant but some earlier authority cited by an intermediate source. In a small number of cases (fewer than twenty), our subsequent review revealed that the machine-readable files required editing, as the # marks had not been placed appropriately (e.g., an *isnad* containing our formula had been erroneously split between two paragraphs).

We also excluded some data because its structure would have complicated our analysis. The biggest such exclusion concerned *isnad*s that feature two or more names in the same node – that is, a piece of information was attributed equally to two different sources. This could occur anywhere within the *isnad*, but it was most problematic when the two sources were named as al-Tabari’s direct informants, as in this case:

حدثنا أبو كريب محمد بن العلاء ومحمد بن يزيد الرفاعي قالا حدثنا حسين بن علي عن فضيل بن عياض عن الأعمش …

Abu Kurayb Muhammad b. al-ʿAlaʾ and Muhammad b. Yazid al-Rifaʿi told us: Husayn b. ʿAli told us from Fudayl, from al-ʿAmash …

Of the 32,115 *isnad*s with which we began, we omitted from our analysis 550 (1.7%) that contained two direct informants. However, this exclusion is not significant and has little or no effect on our general findings.

In the Transmission Chains table, each *isnad* has its own row, and each transmitter is placed in a separate cell. The data set also contains a list of transmissive terms prepared by Kevin Jaques as well as a list of other words that appear in *isnad*s but do not belong to the names. We used these lists to ensure that only the transmitters’ names were included in the table and that separate names were correctly demarcated.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Milestone id** | **Isnad text** | **Name in pos[[2]](#footnote-2) 1** | **Name in pos 2** | **Name in pos 3** | **Name in pos 4** | **Name in pos 5** | **Name in pos 6** | **Source** | **Number of names** |
| ms0185 | حدثني ابن حميد قال حدثنا جرير عن مغيرة عن المسيب بن رافع عن أبي هريرة قال ما كذب ابراهيم ع غير ثلاث كذبات | ابن حميد | جرير | مغيرة | المسيب بن رافع | أبي هريرة | ما كذب ابراهيم ع غير ثلاث كذبات | *Taʾrikh* | 6 |

Table 1: Sample transmission chain from al-Tabari’s *Taʾrikh*, milestone 185. The original passage containing the names is preserved in the table as ‘isnad text’.

At this point, we are only able to deal with relatively simple *isnad*s. We eliminate rows where in position 1 two names occur, as mentioned above.

### 2. Name Normalisation

Next, we created an authorities list that reduced (normalised) variants of a name to a single Latin-script version, without diacritics. The table also provides biographical information, when available, for direct informants. In its present form, the table contains 4,920 individual names in their ‘surface form’, that is, the form in which they appear in the text, found in any position in an *isnad* in any of al-Tabari’s works. Some of these surface forms in fact refer to the same individual, as in the following example, which demonstrates the need for name normalisation:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Surface form (in Arabic)** | **Normalisation (transliteration)** | **Appears in position 1** | **Appears in position 2** | **Appears in position 3** | **Appears only in position 1** |
| إبراهيم بن عبد الله العبسي | Ibrahim b. 'AbdAllah b. Muhammad | True | False | False | True |
| إبراهيم بن عبد الله بن محمد العبسي | Ibrahim b. 'Abd Allah b. Muhammad | True | False | False | True |
| إبراهيم بن عبد الله بن مسلم | Ibrahim b. 'Abd Allah b. Muslim | True | False | False | True |
| إبراهيم بن عبد الله بن مسلم أبو مسلم | Ibrahim b. 'Abd Allah b. Muslim | True | False | False | True |

Table 2: Sample name normalisations, also indicating the position(s) in which the name appears within al-Tabari’s *isnad*s.

The *isnad* positions in which particular names appear help us assess whether two surface forms could refer to a single person.

We then updated the Transmission Chains table with the normalised names.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Milestone id** | **Isnad text** | **Name in position 1** | **Name in position 2** | **Name in position 3** | **Name in position 4** | **Name in position 5** | **Source** | **Number of names** |
| ms0185 | حدثني ابن حميد قال حدثنا جرير عن مغيرة عن المسيب بن رافع عن أبي هريرة قال ما كذب ابراهيم ع غير ثلاث كذبات | Muhammad b. Humayd al-Razi | Jarir | Mughira | al-Musayyib b. Rafi' | Abu Hurayra | *Taʾrikh* | 6 |

Table 1b: Sample normalised transmission chain. The normalised names of transmitters have been inserted and cells without names (in this case ‘Name in position 6’; cf. Table 1) have been omitted. However, the ‘Number of names’ column retains the original values.

### 3. Direct Informants Position Frequencies

For the third table, we calculated the frequency with which each normalised name of a direct informant appears in each position within the *isnad*s (pos 1 = direct informant). Theoretically, the direct informants should only appear at position 1, but for a variety of reasons the names repeat at later positions (see below ‘Data Challenges’).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Count in pos 1** | **Count in pos 2** | **Count in pos 3** | **Count in pos 4** | **Count in pos 5** | **Count in pos 6** | **Count in pos 7** | **Count in pos 8** | **Count in pos 9** | **Count in pos 10** |
| Muhammad b. Humayd al-Razi | 2879 | 4 | 13 | 4 | 1 | 9 | 21 | 9 | 8 | 0 |
| Bishr b. Mu'adh | 2418 | 3 | 1 | 4 | 2 | 10 | 5 | 2 | 0 | 0 |
| al-Muthanna | 2259 | 0 | 8 | 16 | 9 | 39 | 89 | 70 | 19 | 15 |
| Yunus | 2250 | 9 | 162 | 56 | 6 | 7 | 4 | 0 | 2 | 0 |
| al-Qasim | 1699 | 51 | 9 | 20 | 12 | 13 | 8 | 7 | 4 | 0 |

Table 3: Example of direct informants position frequencies, showing the five names al-Tabari cites most frequently in position 1 (i.e. as direct informants).

This table allows us to identify the names of direct informats cited most frequently in each position (as well as the less commonly cited names). It also helps us tackle some of the data challenges involving names discussed below.

### 4. Direct Informants Overview

We extend the Direct Informants Position Frequencies table. It includes the names of al-Tabari’s direct informants and the number of times he cites them (‘count’) in position 1. The table also records the works in which they are cited as well as a cumulative total of the citations and the percentage they represent of all citations (revealing, for example, the portion of all citations accounted for by the top five or ten direct informants).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name in position 1** | **Count** | **Cumulative count** | **Cumulative percentage** | **Count in *Taʾrikh*** | **Count in *Jāmiʿ al-bayān*** | **Count in *Tahdhib*** |
| Muhammad b. Humayd al-Razi | 2879 | 2879 | 9.11855066 | 457 | 2126 | 296 |
| Bishr b. Mu'adh | 2418 | 5297 | 16.776993 | 31 | 2383 | 4 |
| al-Muthanna | 2259 | 7556 | 23.9318405 | 23 | 2233 | 3 |
| Yunus | 2250 | 9806 | 31.05818262 | 26 | 2075 | 149 |
| al-Qasim | 1699 | 11505 | 36.43936275 | 31 | 1667 | 1 |

Table 4: Example of direct informants overview, showing al-Tabari’s top five direct informants. The most frequently cited informant is Muhammad b. Humayd al-Razi, who is cited in all three of the works under study and whose name begins 9 percent of all *isnad*s in our data set. Muhammad b. Humayd al-Razi and Bishr b. Mu'adh together appear as the direct informant in sixteen percent of all *isnad*s in our data set.

It should be clear that we take these numbers as relative, not absolute, indicators of frequency, having been forced to make our textual data conform to a data model that is not always perfectly suited. We lose some information, such as the omitted *isnad*s that contain two or more names in position 1. Furthermore, al-Tabari relies on his top direct informants even more heavily than these numbers indicate. Our data accounts only for cases in which an informant is named at the beginning of an *isnad*, but al-Tabari also refers back to already named sources in ways that are too oblique to be picked up by our search (e.g., ‘he said’).

**5. Texts Name Counts**

Next, we counted how many direct informants al-Tabari cites in each of his works and how many of these informants are cited not just in one but in two or three of the works.

|  |  |
| --- | --- |
| **Source** | **Count of names in position 1** |
| *Taʾrikh* (total) | 347 |
| *Tafsir* (total) | 543 |
| *Tahdhib* (total) | 342 |
| *Taʾrikh*, *Tafsir* | 139 |
| *Tahdhib*, *Taʾrikh* | 110 |
| *Tahdhib*, *Tafsir* | 229 |
| *Tahdhib*, *Taʾrikh*, *Tafsir* | 104 |

Table 5: Count of direct informants who appear in each work, in specific pairs of works and in all three works. We also have comparable data on transmitters cited in positions 2 and 3 (not included in this table).

**Text Reuse and Other Data**

Besides this data, we relied on text reuse alignments generated in February 2021, when we ran the books included in the corresponding OpenITI corpus release (version [2021.1.4](https://zenodo.org/record/4513723)) through passim software. The most important alignments that we discovered were between two works of al-Tabari, rather than between al-Tabari’s writings and those of another author. We also drew on alignments found between the *Taʾrikh* and the *Tafsir* in October 2020; this run used a variation of the passim algorithm that is more useful for uncovering long alignments. We considered all of these alignments within the wider context of the text reuse patterns discovered in the February 2021 data set. Finally, we also relied on data generated by searches across the corpus using regular expressions, especially citations of al-Tabari within *isnad*s in works that post-date his lifetime.

**Data Challenges**

It is important to stress that we focused our energies on the disambiguation of direct informants – not on the identification of names further down in the *isnad*s. Many names that appear in later positions have not been normalised and may thus still refer to the same person. The questions we wished to pose to our data – about al-Tabari’s working methods and source base – did not require us to put the same effort into pinning down the identities of these later persons as they did into identifying reliably al-Tabari’s direct informants. Our scripts, furthermore, are not currently very good at detecting the ends of *isnad*s. Our present work in progress is oriented towards a specific set of questions, but once we publish it, subsequent researchers will be able to build on our data set, pursue their own questions and improve on our methods.

Also, we cannot yet rely on Natural Language Processing tools (which would have speeded up the extraction of names from the *isnad*s). Ryan Muther is working on word-embedding methods for name extraction, and these will also help address the problems identified here.

***Isolating Single Names***

The first major problem we faced was that after we had divided the *isnad*s into single-name cells in the Transmission Chains table, some cells still contained non-Latinised names or non-name words. That was because we built the list of names gradually and because the transmissive terms list was not comprehensive. There were a lot of names; we did not originally plan to Latinise as many as we did. Also, many *isnad*s contained further details that needed to be excluded, such as the location where a piece of information was transmitted. In response, whenever we came across a non-Latinised name, we added it to the Name Normalisation table. And when we found non-name words, we added these to a growing list of words to be purged from the table (we kept this list for further study).

In terms of omissions, as already noted we deleted from our data set *isnad* rows that featured two or more names in the first position, though we saved these rows in a spreadsheet for future study. At subsequent positions within the *isnad*s, we removed cells that contained two or more names, as well as cells that marked the beginning of a second (branching) *isnad* with the word *wa* (‘and’) prefacing the parallel chain. Our data extraction method erroneously treated these branches as a single *isnad* (with the result that in the Names Position Frequency table, a direct informant would appear in two positions). We did not remove all such rows and cells since identifying them was time consuming. We also removed within the cells some words (including transmissive terms) to better isolate the names. In general, we focused on removals (of either rows or cells) where a direct informant was cited at least five times.

Whenever we removed a name cell for either of the above reasons, we also removed all further cells to the right of it, as the *isnad* was effectively broken from that point onwards. This procedure preserved the information on al-Tabari’s direct informant but lost the information on that informant’s sources. This occurred with 1,787 rows in this way (marking each omission with ##### in the data set).

***Matching Names to People***

The Direct Informants Position Frequencies table allowed us to address two different but related problems. We can call them the one name/multiple personsproblem and the multiple names/one person problem.

In terms of the first problem, we have not yet come up with a way to encode a single name to point to two different persons depending on, for example, the name’s position within the *isnad*. The name Yunus in the Transmission Chains table illustrates the dilemma. When al-Tabari mentions a Yunus in position 1, he means Yunus b. ʿAbd al-Aʿla al-Sadafi (d. 264/877–8), an Egyptian scholar of Quran and Hadith, but when Yunus appears in position 2, it refers to Yunus b. Bukayr (d. 199/815), who transmitted material from Ibn Ishaq’s biography of the Prophet. We could not automatically normalise instances of Yunus in position 2 to Yunus b. Bukayr without the normalisation’s applying also to Yunus b. ʿAbd al-Aʿla al-Sadafi.

The case of Muhammad b. Saʿd, who appears in both position 1 and position 2 in al-Tabari’s *isnad*s, is similar. In the first position, al-Tabari uses the name for Muhammad b. Saʿd b. Muhammad b. al-Hasan b. ʿAtiyya b. Saʿd b. Junada al-ʿAwfi (d. 276/889), who transmitted information through a family *isnad* that goes back to Ibn ʿAbbas (d. c. 68/687–8). Al-Tabari cites him and this particular family *isnad* many times. In the second position, however, ‘Muhammad b. Saʿd’ refers to the author of *Kitab al-Tabaqat al-kabir* (d. 235/830).

This one name/multiple persons problem results in undercounting of people, particularly when we look across the *isnad*s. Counts of individuals in a single position within the *isnad*s (e.g. direct informants) rather than across positions are generally more accurate. The problem might be addressed with a word-embedding method: the computer ‘learns’ the sequences of names that occur in *isnad*s across al-Tabari’s books and can then predict which Yunus or Muhammad b. Saʿd is meant on the basis of the position in which the name appears. However, the programme would need to be able to handle transmitters who genuinely occupy different positions in different *isnad*s, reflecting the varying routes through which al-Tabari acquired information conveyed by such transmitters. Ryan Muther is now working on developing such word-embedding methods with the project team.

While we cannot yet resolve this disambiguation problem, we were able to deal with the much more frequent problem of one person’s being referred to in multiple ways. The problem is less severe in al-Tabari’s texts than it is in many later works (especially Ibn ʿAsakir’s *Taʾrikh madinat Dimashq*). But it is a problem nonetheless and had to be addressed to avoid the artificial inflation of our direct informant count through counting the same informant multiple times under different names.

Accordingly, we selected standard Latin-script names for al-Tabari’s informants and then assigned all the relevant surface-form names to them. We identified the individuals behind apparently different names by observing patterns in the sequences of names preceding and following them in *isnad*s: slightly varying names that typically appeared between the same pair of names could be assumed to refer to the same person. Again, we prioritised reconciliation of al-Tabari’s direct informants and made only passing efforts to unify surface forms of names appearing in other positions in the *isnad*s.

We normalised the names of al-Tabari’s direct informants even when the normalisation resulted in some ambiguity. For example, ‘al-Qasim’ and ‘al-Qasim b. al-Hasan’ are both cited as direct informants (i.e. in position 1). The *isnad*s that follow these names contain the same transmitters, indicating that the two names refer to the same person. We thus normalised both as ‘al-Qasim’ in the Name Normalisation table to facilitate accurate counting of al-Tabari’s direct informants. However, a (possibly different) al-Qasim further down in an *isnad* (say, in position 2) now has the same normalised name as al-Qasim b. Muhammad in position 1, and the two individuals cannot be distinguished from one another. We have accepted this ambiguity as preferable to the error that would follow from differentiating the two names.

Somewhat similarly, the word *abi* (‘my father’) appears frequently in *isnad*s, when transmitters report having received information from their fathers without naming the latter. We let the word stand in each case, rather than identifying the father in question. Although this means that the normalised ‘name’ Abi now represents multiple people, retaining it has the benefit of preserving information on family relationships among the transmitters.

We found many cases of names that occurred only once or twice but closely resembled other names that appeared frequently in the same positions. For example, al-Tabari cites al-Husayn b. ʿAmr b. Muhammad al-ʿAnqazi several times in the *Tafsir* and the *Tahdhib*, but once he cites al-Husayn b. Muhammad al-ʿAnqazi, indicating that this al-Husayn was a son rather than a grandson of Muhammad al-ʿAnqazī. The transmission chains that the two al-Husayns start contain the same names further down the *isnad* (*abi* – Asbat – al-Suddi). Although it is true that the omission of an element in a genealogy is not uncommon, for now we chose not to unify these names. When analysing our data, we looked at the frequency of names, especially in the first position, and generally disregarded names that occur infrequently.

Our approach was to avoid normalisation when in doubt. For example, while filtering the *isnad*s we discovered that two of our most commonly cited direct informants (al-Muthanna and ʿAlī b. Dāwūd, the latter normalised to ‘Ali) had many common nodes in their citation chains. They may well be the same person, but since none of our secondary sources, either medieval or modern, seem to have equated them, we did not do so either.

This data is our starting point and subject to further consideration. The methods we are developing should be discussed and debated. The iterative processes can be improved. But the following posts go on to investigate what the data can show now.

1. Discussed in Sarah Bowen Savant, ‘Al-Ṭabarī’s Unacknowledged Debt to Ibn Abī Ṭāhir Ṭayfūr’ (forthcoming). [↑](#footnote-ref-1)
2. Pos = position. [↑](#footnote-ref-2)