This is the third blog in a short series of blogs on the overlap between the OpenITI corpus and Ibn al-Nadim’s Fihrist. Please refer to the [first part](https://kitab-project.org/OpenITI-vs-Fihrist-part-1/) for a statement of the problem and a very short introduction to the Fihrist; the methodology I used is described in detail in the [second part](https://kitab-project.org/OpenITI-vs-Fihrist-part-3/).

In this instalment, I will analyse the data generated by tagging all authors and books from the OpenITI corpus in the text of Ibn al-Nadim’s Fihrist (the [methodology](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/) is described in a separate blog), and end with some conclusions.

## How many OpenITI books are in the Fihrist and vice versa?

The first question we are interested in is, how many of the authors and books in the OpenITI corpus are mentioned by Ibn al-Nadim, and vice versa: how many of the authors and books listed by Ibn al-Nadim are present in the OpenITI corpus?

First, the bare numbers - bear with me for analysis.

Counting the [OpenITI authors I tagged in the Fihrist text](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#Tagging-of-authors-and-books), it turns out that 257 OpenITI authors were identified in the Fihrist; 241 of these received their own [person sections](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#Person-sections) in the Fihrist, the other 16 were [named only in passing](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#Persons-and-books-mentioned-in-passing) within other sections. These 257 authors represent about **1 in 3 of the 738 authors in the OpenITI corpus** who died before 400 AH (and whom we could therefore expect to be potentially named in the Fihrist).

If we look at it from the Fihrist side, the 241 authors who have their own [person section](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#Person-sections) in the Fihrist represent only about **1 in 8 of the 1826 author sections I identified in the Fihrist**.

If we look at the books rather than the authors, the overlap between the OpenITI corpus and the Fihrist is even smaller (in relative numbers): 318 OpenITI books were identified in the Fihrist; of these, only 293 were mentioned in formal lists of books in the text. These 318 books represent about **1 in 5 of the 1608 books in the OpenITI corpus** written by authors who died before 400 AH, and the 293 listed books in the Fihrist only less than **4% of the 7662 book entries I tagged in the Fihrist**. Even allowing for a potentially large error rate due to the instability of titles (which makes identification difficult) on the one hand and the fact that some titles may exist in the corpus as chapters within books on the other hand, these numbers flesh out our starting hypothesis that neither the OpenITI corpus nor the Fihrist can be considered very representative for the Arabic written tradition.

A visualisation may drive this point home more clearly than these rather abstract numbers:

Timeline

Description automatically generated

Figure 1: The overlap of the authors and books in the OpenITI corpus (pre-400 AH) and the Fihrist. The red part of each bar represents books/authors found in the OpenITI corpus but not in the Fihrist, the blue part books/authors found only in the Fihrist but not in the OpenITI corpus, and the purple section books/authors found in both.

This visualisation made me gasp when I first saw it… only the tiny purple slivers represent books/authors that are present both in the OpenITI corpus and the Fihrist, making abundantly clear that neither is in any way close to providing exhaustive coverage of the literature of this period. Since the overlap between both is so small, both must clearly contain only a fraction of the written Arabic corpus.

In what follows, we’ll try to get a closer look at these numbers from different angles: chronology, genre, and geography.

## The chronological coverage of the OpenITI corpus vs. the Fihrist.

The OpenITI corpus contains chronological metadata on all of its authors, encoded in each author’s URI: the first four characters of each URI represent the death date (in the hijrī era) of an author (e.g. 0310Tabari: al-Ṭabarī died in 310 AH = 922-923 CE). We can use this information to get an idea of the chronological coverage of the OpenITI corpus in comparison with the Fihrist’s.

Since our dataset does not include death dates for authors who are mentioned in the Fihrist but are not in OpenITI, we can’t break down by century the percentage of OpenITI authors in the Fihrist vs all authors in the Fihrist. What we can do is break down the number of OpenITI authors by century, both in the corpus and in the Fihrist. That yields this table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1st C | 2nd C | 3rd C | 4th C | total |
| All pre-400 AH authors in OpenITI | 57 (7.69%) | 70 (9.45%) | 239 (32.25%) | 375 (50.61%) | 741 |
| OpenITI Authors in Fihrist | 20 (7.84%) | 42 (16.47%) | 102 (40.00%) | 91 (35.69%) | 255 |

Table 1: Number of OpenITI authors (pre-400AH) in the corpus and Fihrist, divided per century based on their death dates. The percentages in the parentheses indicate what fraction of the total number (last column) of (pre-400AH) authors in the OpenITI (first row) and the Fihrist (second row) this number represents. The percentages that are significantly higher are highlighted in green.

Twenty OpenITI authors who died in the first century were found in the Fihrist; these twenty authors represent about 8% of all 255 OpenITI authors mentioned in the Fihrist. In total, the OpenITI contains works by 57 authors who died in the first century; these 57 also represent about 8% of all authors in the OpenITI who died before 400 AH.

In the OpenITI corpus, the number of authors seems to grow steadily every century, and for the 4th century, the corpus contains as many authors as for the three previous centuries combined. The Fihrist, conversely, has the largest percentage of OpenITI authors in the 3rd century, and significantly fewer in the second and fourth centuries compared to the OpenITI corpus. This may be partly due to the fact that Ibn al-Nadim probably stopped writing the Fihrist in the early 380s - In this context, it is important to note that the latest OpenITI author found in the Fihrist died in 398 AH (0398IbnZurca). However, this tailing off of the quantity (and diversity) of information closer to the time of the collector seems to be a more general phenomenon, as remarked by Maxim Romanov for, among others, Ismail Basha (d. 1339/1920-1)’s *Hadiyyat al-ʿArifin* and Brill’s *Index Islamicus*.[[1]](#footnote-1)

The pattern becomes clearer if we break down the death date information by 25-year periods (Figure 2):

Chart, bar chart

Description automatically generated

Figure 2: Percentages of OpenITI authors who died before 400 AH, in the Corpus (in red) vs. in the Fihrist (in blue), grouped by 25-year periods. The label on the X axis reflects the end of each 25-year period (that is, the bar at “100” reflects the percentage of authors who died between 75 and 100 AH). Note that the first bar from the left includes all pre-Islamic authors. Slightly less than 5% of the 741 pre-400 AH OpenITI authors predate the year 25 AH; and about the same percentage of the 255 pre-400 AH OpenITI authors mentioned in the Fihrist.

The percentage of death dates of OpenITI authors in the Fihrist tops in the first quarter of the fourth century (bar labeled “325”), and then drops significantly for the next 75 years, which is presumably the period in which Ibn al-Nadim, who died around 380/990, wrote the Fihrist. This suggests that he included more “dead” authors than living ones. However, since we did not undertake this analysis for all authors in the Fihrist, but only for those who are also in the OpenITI, another interpretation is possible: if a work written before, say, the beginning of the 4th century survived long enough and was spread widely enough to make it into Ibn al-Nadim’s Fihrist, it is more likely that it survives to make it into the OpenITI corpus than a work written during Ibn al-Nadim’s own lifetime.

Chart, bar chart

Description automatically generated

Figure 3: Percentages of OpenITI books written by authors who died before 400 AH, in the Corpus (in red) vs. in the Fihrist (in blue), grouped by 25-year periods. The label on the X axis reflects the end of each 25-year period (that is, the bar at “100” reflects the percentage of authors who died between 75 and 100 AH). Note that the first bar, which includes all books written by pre-Islamic authors, is very small in comparison to the same bar in the authors’ graph; this is because almost all of these authors are poets, and their collections have not been tagged as books in the Fihrist text. About 16% of the 741 pre-400 AH OpenITI authors died in the last quarter of the fourth century; but only about 5% of the 255 pre-400 AH OpenITI authors mentioned in the Fihrist.

A similar pattern can be observed with the books (Figure 3): only here, the graph tops much earlier, in the second half of the third century rather than the first quarter of the fourth. The difference is probably due to the fact that both the OpenITI corpus and the Fihrist contain a lot of (rather short) books by a few authors from this period (e.g., Jahiz (d. 255), Ibn Qutayba (d. 276), Ibn Abi Dunya (d. 281)), compared to the fourth century, where the number of books per author the corpus and the Fihrist have in common is generally smaller.

## The subject matter coverage of the OpenITI vs. the Fihrist

As we have seen [before](https://kitab-project.org/OpenITI-vs-Fihrist-part-1), the Fihrist is divided into 10 chapters that can be seen as roughly defined by subject matter. An author and his books are usually (with few exceptions) listed in only one chapter, even if his books cover multiple topics; so, if a book is listed in chapter X, it does not mean that it belongs itself to topic X, only that Ibn al-Nadim associated its author mostly with this topic.

If we count the authors and books [tagged](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#tagging-of-authors-and-books) in each chapter (*maqala*) in the Fihrist, we can observe that the numbers fluctuate heavily from chapter to chapter:

Chart

Description automatically generated

Figure 4: Number of authors (in blue) and books (in red) tagged in the different chapters (*maqala*s) of the Fihrist text (with tags `### |+ $` and `### $$`, respectively). For the key words assigned to each chapter, see [above](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#assigning-categorical-tags).

Maqala 3, on history and related subjects, contains by far the largest number of tagged books (1687 books), while the smallest number of tagged books is found in chapter 9, on non-Muslim sects (only 86 books). Especially the first and last three chapters contain only small numbers of books and authors. In all chapters, the number of books is much higher than the number of authors (between 4 and 30 times higher), except for chapter 4, where Ibn al-Nadim lists collections of the poems of a specific poet only under the name of the poet, without book titles, and I have refrained from tagging references to such collections as books (see [above](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#identifying-authors-and-books)).

A picture containing chart

Description automatically generated

Figure 5: Sample from the poetry chapter, mentioning poets and how many pages of their poems are extant, but no direct mention of a book.

In order to compare this Fihrist data to the OpenITI, I have added broad topic tags based on the Fihrist chapters to all pre-400AH books in the OpenITI corpus in a [spreadsheet](https://docs.google.com/spreadsheets/d/17VJ1YEnRrYXZnA7A7UwArAsZuvQxEW5vYpVQDiKMDak/edit?usp=sharing) (see the methodology)

Graphical user interface, application

Description automatically generated

Figure 6: pre-400 AH books in the OpenITI corpus, by topic tag. The first 10 tags agree with the Fihrist’s chapters, the other tags were created for books that did not fit into these categories easily.

In order to compare the genre distribution of the Fihrist and OpenITI, it is useful to put both into one graph:

A picture containing chart

Description automatically generated

Figure 7: Number of books for each topic in the Fihrist and the pre-400 AH OpenITI corpus. For the Fihrist, the chapters were taken as topic indicators; for the OpenITI, tags were manually attached to each book. Books that are in the Fihrist but not OpenITI are in red, OpenITI books not found in the Fihrist in blue. OpenITI books found in the Fihrist are in purple; the darker shade of purple indicates that the book was given the tag corresponding to the Fihrist chapter in the OpenITI spreadsheet, the lighter purple that another tag was given to the book in the OpenITI spreadsheet (for example, all Jāḥiẓ’s books appear in the fifth chapter (THEOLOGY), but most were given the HISTORY tag in the OpenITI spreadsheet because that chapter in the Fihrist covers *adab* literature).

A first observation is that the topics from the last three Fihrist chapters are almost entirely absent from the pre-400 AH OpenITI corpus: only one book (Ibn Sīrīn’s book on dream interpretation, which is also in the Fihrist) was given a tag that represents these chapters.

Secondly, HADITH is very much over-represented in the pre-400AH OpenITI corpus: more than 40% of the books in the pre-400 AH OpenITI corpus was given this tag, while the Fihrist does not even devote a chapter to ḥadīth literature. A number of these ḥadīth works are in the Fihrist, though: they are mostly found in the sixth chapter, on law (where the *Aṣḥāb al-Ḥadīth* is one of the 8 law schools discussed). The reason for this state of affairs is obviously that the (individuals and) organisations who took it upon themselves to digitise the Arabic texts that form the basis of the OpenITI corpus (see [my blog on al-Maktaba al-Shamela](https://kitab-project.org/Al-Maktaba-al-Sh%C4%81mila-a-short-history/) for a case study) were much more interested in this type of works than Ibn al-Nadim was.

There may, however, be other factors contributing to the prevalence of pre-400 AH hadith collections in the OpenITI corpus and their scarcity in the Fihrist. Perhaps these hadith collections (often with titles like *hadith fulan* “hadith [transmitted by] X”, *juz’ (hadith) fulan* “fragment of X(’s hadith)”, *amali fulan* “dictations by X”) were not considered real books by Ibn al-Nadīm. It is also possible these hadith collections are later compilations and simply did not exist at Ibn al-Nadim’s time. An example is Ibn al-Nadim’s statement about Sufyan b. ʿUyayna that “no book by him is known, but people used to hear [hadith] from him";[[2]](#footnote-2) The OpenITI corpus contains at least four different selections of hadith transmitted by him.[[3]](#footnote-3) Perhaps the key lies in his use of the word kitab (book);

## The regional background of the OpenITI vs. the Fihrist

We can also break down the data on a regional level, based on the metadata on the primary geographical association of early authors in the OpenITI corpus [collected by Aslisho Qurboniev](https://kitab-project.org/b/) and stored in the OpenITI corpus’ author metadata files. When we count the regions these authors are associated with, the data shows that the vast majority of authors that feature both in the Fihrist and the OpenITI came from Iraq or its neighbouring regions (187 out of the 234 identified locations, ca. 80%). The entire region to the West of Egypt is absent (apart from one author, Ibn Abī Zayd al-Qayrawānī).

Map

Description automatically generated

Figure 8: regional distribution of the authors that feature both in the OpenITI and the Fihrist. In bold: Iraq and neighbouring regions.

At first sight, the geographical metadata suggest that the books known/available to Ibn al-Nadim were mostly limited to those written in the vicinity of Iraq. This was to an extent to be expected. However, it’s important to remember that we aren’t analysing the geographical background of all authors in the Fihrist, only of the OpenITI authors mentioned in the Fihrist, so this could be a property of the OpenITI authors, not the Fihrist authors. To check this, we can compare this data with the geographical metadata of all authors in the OpenITI corpus who died before 400 AH (Table 2). We see a pretty similar distribution, but the percentage of OpenITI authors from Iraq is significantly higher in the Fihrist than in the whole pre-400 AH corpus. The percentage of OpenITI authors from far-flung regions is slightly lower in the Fihrist. This seems to suggest that both the Fihrist and the pre-400 AH OpenITI corpus suffer from a similar (Iraq-centred) geographical bias, in a somewhat higher degree in the Fihrist but a full analysis of all authors mentioned in the Fihrist would be needed to confirm this.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **region** | **Number of OpenITI authors in Fihrist** | **Percentage**  **(Fihrist)** | **Percentage (OpenITI)** | **Number of all OpenITI authors** |
| **Iraq\_RE** | **123** | **46.8** | **36** | **280** |
| **JaziratArab\_RE** | **44** | **16.7** | **16.3** | **127** |
| Khurasan\_RE | 22 | 8.4 | 7.2 | 56 |
| **Sham\_RE** | **14** | **5.3** | **6.2** | **48** |
| **Jibal\_RE** | **13** | **4.9** | **6** | **47** |
| **Aqur\_RE** | **10** | **3.8** | **2.8** | **22** |
| Fars\_RE | 10 | 3.8 | 1.8 | 14 |
| Misr\_RE | 10 | 3.8 | 4.5 | 35 |
| Mawarannahr\_RE | 6 | 2.3 | 3.2 | 25 |
| **Khuzistan\_RE** | **4** | **1.5** | **1.2** | **9** |
| Sijistan\_RE | 2 | 0.8 | 1.3 | 10 |
| Daylam\_RE | 2 | 0.8 | 1.7 | 13 |
| Yaman\_RE | 1 | 0.4 | 1.2 | 9 |
| Maghrib\_RE | 1 | 0.4 | 2.7 | 21 |
| Andalus\_RE | 0 | 0 | 3 | 23 |
| Kirman\_RE | 0 | 0 | 0.3 | 2 |
| Sind\_RE | 0 | 0 | 0.1 | 1 |
| Rihab\_RE | 0 | 0 | 0.1 | 1 |
| Unknown\_RE | 1 | 0.4 | 4.5 | 35 |
| TOTAL | 263 |  |  | 778 |

Table 2: Regional distribution of the authors that feature both in the OpenITI and the Fihrist, compared to the whole pre-400 AH OpenITI corpus. In bold: Iraq and neighbouring regions. Highlighted in green are percentages that are significantly higher than in the other category.

# Conclusions

This case study showed the degree to which the Fihrist and the OpenITI corpus are both very fractional representations of the entire written tradition of the first four Islamic centuries. About 80% of all books in the OpenITI corpus written by authors who died before 400 AH were not found in the Fihrist; and the other way around, more than 95% of the book entries in the Fihrist are not in the OpenITI corpus. Even allowing for the obvious problems with titles in the Fihrist and their identification, these are sobering numbers.

Breaking these numbers down along chronological lines, it became clear that the Fihrist contained much fewer OpenITI books and authors from Ibn al-Nadim’s lifetime than one would expect. The reason for this cannot be established based on the available data (without trying to date all books and authors in the Fihrist, instead of only those that are part of the OpenITI corpus); one possible interpretation is that Ibn al-Nadim has a preference for dead authors; another that authors who died a century or more before Ibn al-Nadim and still made it into the Fihrist, had more chance to survive until today than his contemporaries. Finally, the Fihrist itself may have played a role in deciding which books were preserved: the work was a major source for later bibliographical works, and these bibliographical works have certainly been used by book collectors to decide which books to acquire.

With regards to the subject matter distribution of the Fihrist and the OpenITI corpus, it became abundantly clear that both have their own preferences. Whereas 40% of the pre-400 AH books in the OpenITI corpus are devoted to ḥadīth, ḥadīth literature is not even awarded its own chapter in the Fihrist, and only relatively few ḥadīth works are listed in the Fihrist. Conversely, only one of the books from the last three chapters of the Fihrist is in the OpenITI corpus. For the OpenITI corpus, the reason for this imbalance seems clear: the OpenITI corpus consists to a very large extent of books that were digitised by other projects, and these clearly spent a lot of effort on digitising ḥadīth texts (for religious/ideological reasons), disregarding other genres – especially the more “secular” genres that make up most of the last chapters of the Fihrist. The reason why Ibn al-Nadim does not include many hadith collections ascribed to early authors is less certain. One possible reason is that these collections simply did not exist in his time; another, that these orally/aurally transmitted works, even when written down in the process, did not fit his idea of a book worthy to be recorded in his Fihrist. Finally, it is of course possible that the absence of these hadith collections in the Fihrist reflects Ibn al-Nadim’s personal predilections, and/or his audience’s.

Finally, we saw that the Fihrist (or at least, the subset of books in the Fihrist that are also in the OpenITI corpus) contains significantly more books written by authors from Iraq than the pre-400 AH OpenITI corpus. Moreover, the Islamic West is missing almost entirely in the Fihrist. On the other hand, it is clear that the pre-400 AH OpenITI corpus shares the Fihrist’s focus on books from Iraq and its immediate neighbours to a large extent - the Fihrist itself may even be one of the contributing factors to this geographical bias in the OpenITI corpus.

This kind of research helps us to identify blind spots in the OpenITI corpus. We can then purposely focus on adding works of specific genres, time periods and regions that are under-represented in the corpus in order to fill the gaps. Until now, we have been handicapped by the fact that we were almost entirely dependent on other projects’ digitization efforts (and preferences) for additions to the corpus. With the recent improvements of Optical Character Recognition (OCR) for Arabic script by the [OpenITI Arabic-Script OCR Catalyst Project (AOCP)](https://mith.umd.edu/research/openiti-aocp/), we hope we will be able to add more and more printed books from genres, time periods, and regions that are disregarded by third parties. Books that never made it into print, and are only extant in manuscript, form another hurdle; it is hoped that the [new phase](https://arhu.umd.edu/news/mellon-grant-funds-continuation-persian-and-arabic-digitization-project) of the OpenITI AOCP, which will focus on the development of generalised models for Hand-written Text Recognition (HTR) for Arabic script, will enable us to add more of such missing works to the corpus in the not so distant future .

# Bibliography

Fück, Johann Wilhelm (1969). “Ibn al-Nadīm”, in: *Encyclopaedia of Islam, Second Edition*, edited by: P. Bearman, Th. Bianquis, C.E. Bosworth, E. van Donzel, W.P. Heinrichs, Leiden: Brill. Consulted online at <http://dx.doi.org/10.1163/1573-3912_islam_SIM_3317>

Romanov, Maxim (2017). “Algorithmic Analysis of Medieval Arabic Biographical Collections”, *Speculum* 91/S1: 227-246. <http://dx.doi.org/10.1086/693970>

Soravia, Bruna (2015). “Bibliographies, Arabic”, in: *Encyclopaedia of Islam, THREE*, edited by Kate Fleet, Gudrun Krämer, Denis Matringe, John Nawas, Everett Rowson, Leiden: Brill. Consulted online at <http://dx.doi.org/10.1163/1573-3912_ei3_COM_24336>

Stewart, Devin (2007). “The Structure of the Fihrist: Ibn al-Nadim as Historian of Islamic Legal and Theological Schools”, *International Journal of Middle East Studies* 39(3): 369-387. <https://www.jstor.org/stable/30069526>

Toorawa, Shawkat (2010). “Proximity, Resemblance, Sidebars and Clusters: Ibn al-Nadīm’s Organizational Principles in Fihrist 3.3”, *Oriens* 38(1-2): 217-247. <https://doi.org/10.1163/187783710X536725>

Glossary

*Shuhra*: the name by which a person is best known

*Ism*: the given name of a person

*Kunya*: teknonym, a name given to a person derived from the name of their (first) child; usually takes the form *Abu X* (Father of X) or *Umm X* (Mother of X)

*Nisba*: a name element derived usually from a person’s geographical background

1. Maxim Romanov (2017). “Algorithmic Analysis of Medieval Arabic Biographical Collections”, Speculum 92/S1. https://www.journals.uchicago.edu/doi/10.1086/693970 [↑](#footnote-ref-1)
2. *lā kitāb lahū yuʿraf wa-innamā kāna yusmaʿu minhu* (Ibn al-Nadim, Fihrist ed. Tajaddud p. 282). He makes a similar statement about a certain Fatḥ al-Mawṣilī: “no writing by him is known but his words are being memorized” (*lā kitāb lahu yuʿraf wa-innamā yuḥfaẓ kalāmuhu*; p. 237). Thanks to Lorenz Nigst for providing the first example. [↑](#footnote-ref-2)
3. 0198SufyanIbnCuyayna.JuzFihiHadith.ShamAY0032777-ara1, 0198SufyanIbnCuyayna.JuzHadith.Shamela0026337-ara1, 0265IbnHarbMawsili.HadithSufyanIbnCuyayna.Shamela0029708-ara1, 0492AbuHasanKhulci.HadithSufyanIbnCuyayna.Shamela0029776-ara1. [↑](#footnote-ref-3)