This is the second 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; to the [next part](https://kitab-project.org/OpenITI-vs-Fihrist-part-3/) for the analysis and conclusions of this case study.

In this blog, I will describe in detail the methodology I used to analyse the overlap between the OpenITI corpus and the Fihrist.

## A. Tagging of authors and books

In a first step, I took the text of the Fihrist as it is in the OpenITI corpus ([0385IbnNadim.Fihrist.Shia003355-ara1](https://raw.githubusercontent.com/OpenITI/0400AH/master/data/0385IbnNadim/0385IbnNadim.Fihrist/0385IbnNadim.Fihrist.Shia003355-ara1.mARkdown) - this is the Rida Tajaddud edition,),[[1]](#footnote-1) and extended its structural annotation using the [OpenITI mARkdown scheme](https://alraqmiyyat.github.io/mARkdown/). The text had been annotated only two levels deep: the 10 chapters (*maqala*) mentioned [in the previous blog](https://kitab-project.org/OpenITI-vs-Fihrist-part-1/), and their primary subsections (*fann*). In many chapters, however, the text has a much more fine-grained structure, grouping authors together based on commonalities (e.g., generation, regional background, family relationship, student-teacher relationships, etc.) in a generally chronological order (see Shawkat Toorawa’s analysis of the structure of the third *maqala*,[[2]](#footnote-2) and Devin Stewart’s of the sixth *maqala[[3]](#footnote-3)*).

### Person sections

Within this structure, Ibn al-Nadim often lists persons (writers, poets, translators, occasionally people who did not write books but otherwise contributed to the written culture) and books.

I tagged all such listed persons as section headers (even if Ibn al-Nadim lists no further information about them), by inserting at the start of the section the `###` tag followed by a number of pipes `|` that reflects the hierarchical location of the header within the book’s structure, and finally a dollar sign `$` to identify this specific section as a section on a person (see Figure 1). We will refer to such sections as “**person sections**” in what follows.

A picture containing diagram

Description automatically generated

Figure1: Sample of the tagged Fihrist text: the biography of al-Jarmī with the list of his books (from the first subsection (*fann*)of the second chapter (*maqala*) of the Fihrist). The section on al-Jarmī has been tagged as a third-level section header (three pipes after the ### tag), because it is an immediate subsection of the *fann*. The colours are generated by the OpenITI mARkdown highlighting scheme in EditPad Pro.

### Book entries

Many of the sections we called “person sections” contain a list of books written by that person, usually introduced by a statement like *wa-lahu min al-kutub* (“among his books are:”). I tagged all those books with a `### $$ ` tag.[[4]](#footnote-4)

Books are not only listed in person sections in the *Fihrist*: the work also contains lists of books that are not grouped together based on their authorship but based on other criteria. (e.g., *tasmiyat al-kutub al-muṣannafa fī tafsīr al-Qur’ān*, “List of the books written about Qur’anic exegesis”). In such sections, which we will call “**non-person sections**”, I tagged the books with the same `### $$` tag.

In what follows, we will refer to the books tagged in this way as “**book entries**”.

Text, letter

Description automatically generated

Figure2: Sample of the tagged Fihrist text: start of the list of books written on uncommon words in Ḥadīth texts (from the second *fann* of the third *maqala*).

### Persons and books mentioned in passing

Persons and books mentioned by Ibn al-Nadim in passing within another section, without a clear link with the structure of the book, were not tagged – in most cases (but not always), these persons and books have their own sections and entries in the book.

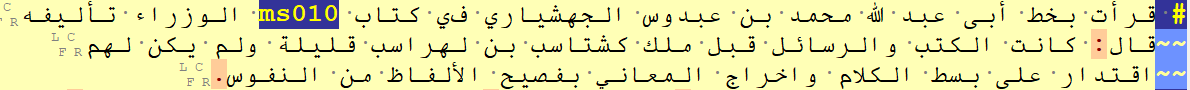


Figure3: Example of a book not tagged: mention of al-Jahshiyārī’s *Kitāb al-Wuzarā*ʾ in the introductory section on Persian literature.

## B. Identifying authors and books

Each book and author in the OpenITI corpus has its own [Unique Resource Identifier (URI)](https://kitab-project.org/docs/openITI#uris--cts-like-folder-structure). An author URI consists of that author’s approximate death date (in the hijri era) followed by a simplified version of his shuhra. For example, the URI for al-Tabari is 0310Tabari. The URI for a book consists of the URI of the author, followed by one or more words from the title. For example, the URI for al-Tabari’s *Tarikh* is 0310Tabari.Tarikh.

In this second step, I loaded the metadata tsv file of the [2021.2.5 release](https://zenodo.org/record/5550338#.YyQqCXbP3cs) of the OpenITI corpus into a [spreadsheet](https://docs.google.com/spreadsheets/d/17VJ1YEnRrYXZnA7A7UwArAsZuvQxEW5vYpVQDiKMDak/edit?usp=sharing) and checked for every author in the corpus who died before 400AH, and for their books, whether they were mentioned in the *Fihrist*.

This check was done mostly manually: for each author in the OpenITI corpus who died before 400AH, multiple searches (*shuhra*, *ism* + father’s name, *kunya*, *nisba*) were carried out using regular expressions in the Fihrist text in our favourite text editor [EditPad Pro](https://kitab-project.org/docs/openITI#editpad-pro). A match was considered secure only if two or more of these name elements matched, and if one or more of the following conditions were found to be true:

* the chapter in which the name was found fit the expectations; and/or
* the death date mentioned in the text fit the death date of the OpenITI author; and/or
* at least one of the author’s books in the OpenITI was found listed under the matched author’s name.

Each secure match was marked in the metadata spreadsheet with a `TRUE` in the author column, and the name of the author in the text was tagged with a tag that includes the relevant author URI: `[@URI@Pxx](https://alraqmiyyat.github.io/mARkdown/#-individuals-persons)` (in which the first x stands for the number of prefix characters that should be ignored, and the second x for the number of tokens (words) the name consists of – the numbers facilitate highlighting the author’s names and extracting them automatically from the text using a script, see Figure 4).

For each book by this author in the corpus, I checked if the book was listed among the author’s books in the Fihrist; if it was, it was marked in the spreadsheet with a `TRUE` mark in the book column and the title was tagged in the text of the Fihrist with a `@URI@SRCxx` tag (see Figure 4). If the book was not listed in the list of books, I ran another search in the text of the Fihrist to check if the book was mentioned anywhere else in the Fihrist; if so, I tagged it there. Each book was tagged only once in the Fihrist, even if it was mentioned more than once.

Table

Description automatically generated with medium confidence

Figure 4: author’s name and book title tagged with OpenITI URIs. The tag `@P03` means “the following **3** tokens are the name of a **P**erson (and that name is preceded by **0** prefix characters)”. Similarly, `@SRC02` means that the following **2** tokens are the title of a book (preceded by **0** prefix characters). Note that the OpenITI highlighting scheme in EditPad Pro highlights the tagged titles in yellow and book URIs in mint green. The highlighting of the author URI and name is not visible here because the section title highlighting is covering them.

After going through all authors who died before 400 AH in the OpenITI metadata, I ran a number of tests to check if I had overlooked any matches. One test consisted of searching for the regular expression `توف[يى]` (to match verbal expressions referring to the death of an author) and checking whether any of the name elements of the author’s name was to be found in the spreadsheet within a 50-year time span around the death date as recorded in the Fihrist.

I also checked if all authors and books that were tagged in the text were also marked in the spreadsheet and vice versa.

Finally, I went through all the books that were titled “Dīwān” or “Shiʿr” followed by the name of the poet in the OpenITI corpus. I had noticed that these collections of poetry were almost never recorded under that title in the Fihrist, but in many cases, the Fihrist does mention that a poet’s collected poems reached a certain number of pages (*waraqa*) or that a specific person had collected and “published” (Ibn al-Nadim uses the term *ʿamala* for this) that poet’s poetry. In this case, only the poet’s name had been tagged in the Fihrist; I added the codes `WARAQA` and/or `CAMALA` in the book column of the spreadsheet for those cases.

This way of tagging authors and books enable us to do different types of computational analysis: we can count the number of authors and books that have their own sections in the Fihrist (or in a specific *maqala* / *fann*); the URIs enable us to [calculate](https://kitab-project.org/OpenITI-vs-Fihrist-part-3#how-many-openiti-books-are-in-the-fihrist-and-vice-versa) how many OpenITI authors/books are in the Fihrist, and vice versa.

## C. Assigning categorical tags

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 based on subject matter (“topic”). 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 that book itself is about topic X, only that Ibn al-Nadim associated its author mostly with this topic.

In order to compare the topic coverage of the Fihrist and the OpenITI corpus, I decided to assign Ibn al-Nadīm’s categories to all pre-400 AH books in the OpenITI corpus.

I first devised a single-word tag for each chapter that (roughly) covers that chapter’s contents:

1. SCRIPTURE: On languages and scripts, revealed scriptures and Qur’an-related writings
2. LANGUAGE: On the Arabic language and grammar
3. HISTORY: On history and genealogy, and on rulers and administrators who wrote books[[5]](#footnote-5)
4. POETRY: On poets and poetry collectors
5. THEOLOGY: On Islamic theology (subdivided into sections on five “schools” of theology)
6. LAW: On Islamic law (again subdivided into sections on eight “schools” of law)
7. SCIENCE: On Greek philosophy and sciences
8. FOLK: On folk stories, magic, dreams, cooking, etc.
9. NONMUSLIM: On the literatures of non-Muslim communities (but not Christians and Jews)
10. ALCHEMY: On alchemy

In addition, I created a couple of additional tags for categories in the OpenITI corpus that did not fit easily into Ibn al-Nadīm’s very wide categories:

* HADITH: for collections of ḥadīths, biographical dictionaries of ḥadīth transmitters etc.
* MISC: for books that do not fit in any of the previous categories.

I then assigned these tags to all pre-400 AH books in the [OpenITI metadata spreadsheet](https://docs.google.com/spreadsheets/d/17VJ1YEnRrYXZnA7A7UwArAsZuvQxEW5vYpVQDiKMDak/edit?usp=sharing). Contrary to Ibn al-Nadim’s practice, I did attach different categorical tags to different books by the same author if that seemed fitting. The decision which tag to apply to a book was based mostly on the genre classifications attached to that book in the source collections (Shamela, eShia, …) from which the book was added to the OpenITI; only in a few cases where this metadata was not conclusive, did I undertake some deeper research into the book in question in order to decide on a tag.

This approach allows us to compare which topics (as defined by Ibn al-Nadim) are covered more heavily in the OpenITI corpus than in the Fihrist and vice versa.

## D. Geographical metadata

In addition to variation in topics, I also wanted to compare the geographical distribution of the authors in the OpenITI and the Fihrist. Thanks to the efforts of Aslisho Qurboniev (see [this blog](https://kitab-project.org/b/)), the OpenITI metadata of every author who died before the year 500 AH now contains a record of his place and region of birth, in the form of URIs from the [Althurayya gazetteer](https://althurayya.github.io/). The OpenITI metadata also contains URIs for places and regions they visited, resided in or died in; but this information has not been used in this case study because it is currently only available for a small number of authors.

Using a Python script I extracted this data into a tsv file in which each author's URI was linked to that author’s region of birth. Note that I used only a single Althurayya URI for each author, even if more were listed in the original metadata (e.g. places of death, residence, etc.).

This allows us to count the number of OpenITI authors that come from each region, both in the pre-400 AH OpenITI corpus and in the Fihrist.

## E. Extracting the data

This type of tagging allows us to make precise counts of OpenITI books and authors in the Fihrist. Because the tags are in the text itself, it makes it easy for myself and others to check the decisions I made assigning a URI to a book or author.

To extract the data from the Fihrist text and OpenITI metadata spreadsheet, and compute the statistics, I wrote a Python script. This script

* Counts the number of books and authors in the OpenITI corpus (pre-400 AH) and the Fihrist
* Counts the number of books by topic in the OpenITI corpus (pre-400 AH) and the Fihrist
* Counts the number of authors by region in the OpenITI corpus (pre-400 AH) and the Fihrist

The outputs generated by the script were then used to create the graphs discussed in the [next blog](https://kitab-project.org/OpenITI-vs-Fihrist-part-3/).

## F. Caveats

As with many computational methods used in the humanities, the exact numbers generated by these scripts should be regarded only as approximate. There is a significant margin of error due to a number of uncertainties introduced at different steps in the process:

* Some books/authors I tagged may not have been in the original text of the Fihrist: the text of the Fihrist as it was established in Rida Tajaddud’s edition is based largely on a single manuscript that was probably copied directly from Ibn al-Nadim’s autograph (see [above](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#tagging-of-authors-and-books)). However, that manuscript is incomplete, and the missing parts have been filled in from other (more distant and later) manuscripts. Moreover, the single manuscript contains many additions by a later user of the manuscript, al-Wazīr al-Maghribī; these additions are not consistently indicated in the Tajaddud edition, which means that some of the books/authors we tagged may not have been in the Fihrist but were in fact added by al-Wazīr al-Maghribī.
* Some OpenITI books/authors that are mentioned in the Fihrist may not have been tagged in the text: authors and books are sometimes known under different names. Moreover, words are frequently undotted in the manuscripts, leading to possible misreadings by the editor.[[6]](#footnote-6) In some cases, this may have led to me missing a reference in the Fihrist to an OpenITI author or book, even after taking measures to mitigate this (described [above](https://kitab-project.org/OpenITI-vs-Fihrist-part-2/#assigning-categorical-tags)).
* Some titles tagged in the Fihrist as books may in fact be names of sections rather than separate works. This is not so much a problem for identification of OpenITI books in the Fihrist text, but may influence the total numbers of books in the Fihrist and its *maqala*s (and hence also any percentages, when we divide the number of OpenITI books in the Fihrist or a section by the total number of books in the book or that section).
* The geographical tags have only a limited validity: for many authors, their place of birth is not known with certainty; some spent an important part of their lives in other places than when they were born. Since we use only one geographical tag per author (usually, their place of birth, as it is the most widely available piece of geographical information), this important information will not be reflected in our graphs and analyses.

1. Rida Tajaddud (1971). Kitab al-Fihrist li-al-Nadim, Tehran. This edition is based on the oldest, and best, surviving manuscript, which was cut in half; its first part is now in the Chester Beatty library in Dublin (MS 3315), the second in Istanbul (MS Şehid Ali Paşa 1934). Unfortunately, the editor does not always consistently indicate what the text of the main author is, and what are additions by a later user of the manuscript, al-Wazir al-Maghribi (d. 418/1027). Ayman Fuad Sayyid’s 2009 edition, which does make this distinction more clearly, is not yet available in digitised form. [↑](#footnote-ref-1)
2. Shawkat Toorawa(2010). “Proximity, Resemblance, Sidebars and Clusters: Ibn al-Nadīm’s Organizational Principles in Fihrist 3.3”, Oriens 38(1-2): 217-247. [↑](#footnote-ref-2)
3. Devin Stewart (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: 370 [↑](#footnote-ref-3)
4. In standard OpenITI mARkdown, the tag `### $ ` (with a single dollar sign) is used to tag entries in dictionary-like books (for example, `### $ ` tags are used to tag persons in biographical dictionaries). Since Ibn al-Nadim has entries on persons and books, I decided to tag persons with one dollar sign and books with two (this diverges from standard. In standard OpenITI mARkdown practice, where the double dollar sign is used for female persons in biographical dictionaries), but here I use it for book titles. In the words of its creator, Maxim Romanov: “despite the existence of a standard scheme, OpenITI mARkdown is meant as an analytical tool that provides an easy way to tag information relevant for current analytical tasks.” [↑](#footnote-ref-4)
5. Many of the books in this chapter are the subject of what is usually called *adab* literature. [↑](#footnote-ref-5)
6. Page upon page of examples are given in Devin Stuart’s review of Ayman Fuad Sayyid’s edition of the text. [↑](#footnote-ref-6)