@Title: The challenges of visualising OpenITI mARkdown: comparing the structures of Ibn al-Dawādārī’s *Kanz al-Durar* with Ibn Taghrībirdī’s *Nujūm al-Ẓāhira* @

As historians we often engage with parts of a text. I research Fatimid history and to do so I often engage with short sections of very long chronicles. However, when reading these texts it is important to have an understanding of how the parts relate to the whole. In the case of Fatimid history, chronicles that focus on Egyptian history can provide very short entries for under years corresponding to Fatimid history. This might correspond to a source problem. Very little Fatimid-era historiography survives and this might have impacted how later chroniclers were able to document the period. It might conversely correspond to an authors’ ideological bias. The Fatimids were Shīʿī Ismāʿīlīs and the authors of many chronicles were Sunnīs. It is possible that they preferred not to document Fatimid history so thoroughly. Both of these claims assume that sections of chronicles corresponding to Fatimid history are comparatively shorter than those corresponding to other periods. In other words, before making such claims we need to understand how these part fit into the broader chronicle.

OpenITI structural mARkdown is the tool that allows us to pursue such questions. The KITAB team has made extensive efforts to add structural annotation to the OpenITI texts, through collating the digital texts with their original printed editions. This structural annotation is valuable in a large number of ways. Crucially, it makes our works more manageable to read. Annotated headings can be collapsed in EditPad Pro, allowing us to quickly find relevant sections of text. I have been thinking about working with these these structural markers statistically and graphing them. The next few paragraphs will make some preliminary remarks on these matters, as they relate to two Mamluk-era chronicles: Ibn al-Dawādārī’s (fl. eighth century) *Kanz al-Durar* and Ibn Taghrībirdī’s (d. 874) *Nujūm al-Ẓāhira*.

***Finding aids and highly-structured texts***

The structural annotation used in OpenITI mARKdown is particularly valuable for understanding how authors structured their works through the use of contents pages and finding aids. From around the eighth-century increasingly structured their texts to facilitate the easy location of specific topics, sections or parts. Most prominently this featured the use of red ink to indicate headings or important words and sometimes the use of contents pages. This was accompanied by complex organisational structures that sometimes used nesting to clearly identified topics and their subtopics. In al-Qalqashandī’s (d. 821) *Subḥ al-ʾAʿshā* this nesting is indicated in his use of language.[[1]](#footnote-1) See, for example, his categorisation of dynasties in Yemen and their history (I have used indentation to make the nesting more explicit):

The Second Object [*al-maqṣid*]: On the kingdoms of the Arabian peninsular outside of the places of the Egyptian region.

The first region:

The first part [*qism*]: The Tihāmas.

The first group [*jumla*]: it mentions what [Yemen] comprises in [terms of] centres and towns.

The first centre: Taiz

The second centre: Zabid

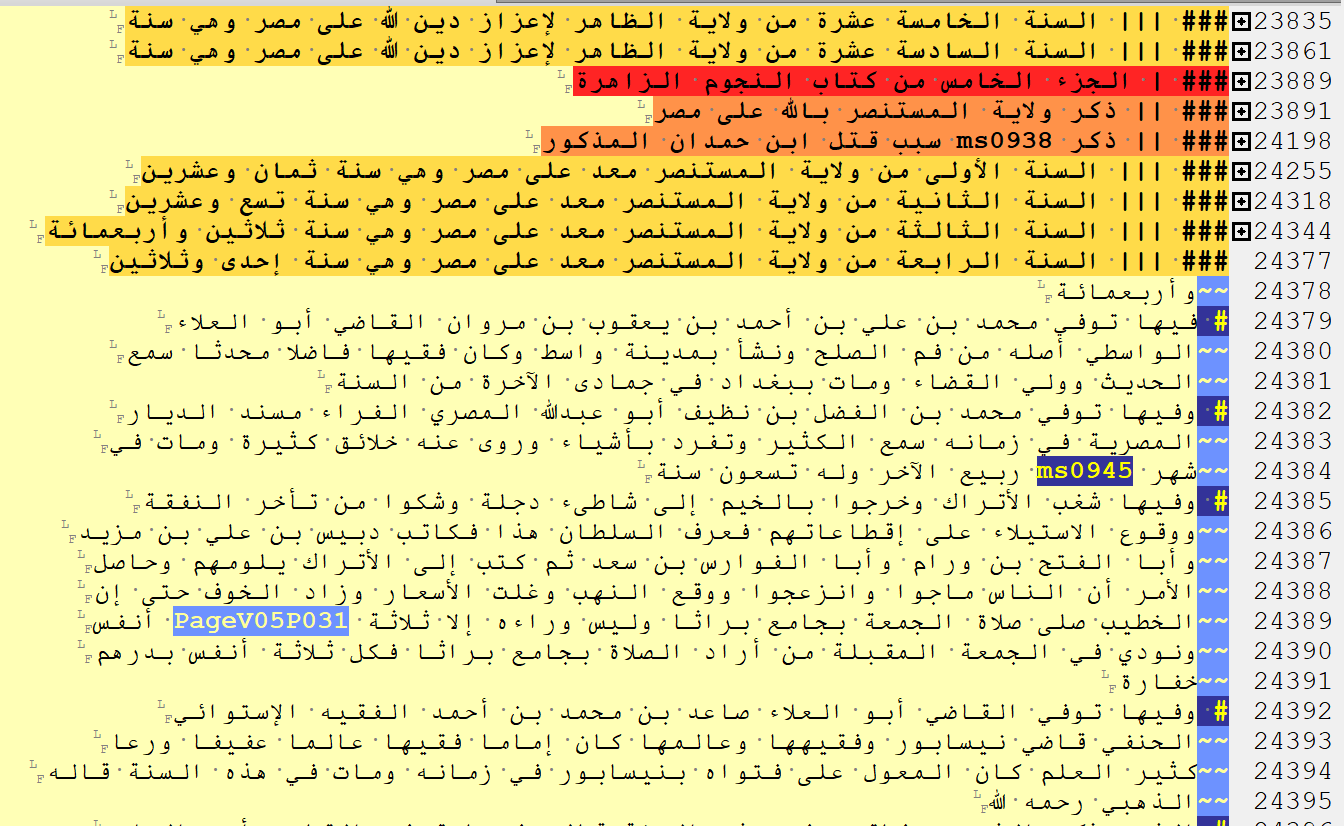
The second group [*jumla*]: it mentions its animals, grains, fruits, herbs, its commerce and its prices.[[2]](#footnote-2)

Al-Qalqashandī, therefore, describes the geography and politics of the Yemen by category, providing a clearly-signposted nested structure that in this case goes five levels deep. Few authors are so explicit about the nested structure of their works, but there are a large number of works that feature similar levels of complexity.

Ibn Taghrībirdī’s chronicle of Egyptian history, the *Nujūm al-Ẓāhira*, uses two levels of nesting. At the first level he notes the Egyptian rulers, providing a summary of that ruler’s life (typically referencing and comparing earlier historical accounts). Then he provides headings for each of the years of the reign. The summary of the reign often describes events more particular to the reign and Egypt. By comparison, under each year Ibn Taghrībirdī describes events specific to the year (but not necessarily the ruler), gives short obituaries for the year and concludes with the record of the Nile flood.

Ibn al-Dawādārī’s universal chronicle, the *Kanz al-Durar* adopts a similar format, but with slightly deeper nesting. Ibn al-Dawādārī, like Ibn Taghrībirdī, gives summaries of rulers at the first level. In this case, Ibn al-Dawādārī does not focus on just Egyptian rulers. Thus for the period where there are both Abbasid and Fatimid Caliphs, he gives summaries of both rulers at the relevant points. At the next level, Ibn al-Dawādārī gives the years of the reign (occasionally combining two years into one heading). Immediately under the year Ibn al-Dawādārī gives a record of the Nile flood. Then under this, he provides another heading under which he summarises the events of the year and occasionally adds other key events under specific headings. A description of events may then occasionally lead Ibn al-Dawādārī to make a digression (which might be considered to be nested underneath the event). For example, after he describes the Seljuk seizure of Baghdad (and their deposition of the Buyids), Ibn al-Dawādārī gives a list of the Buyid rulers of Baghdad.

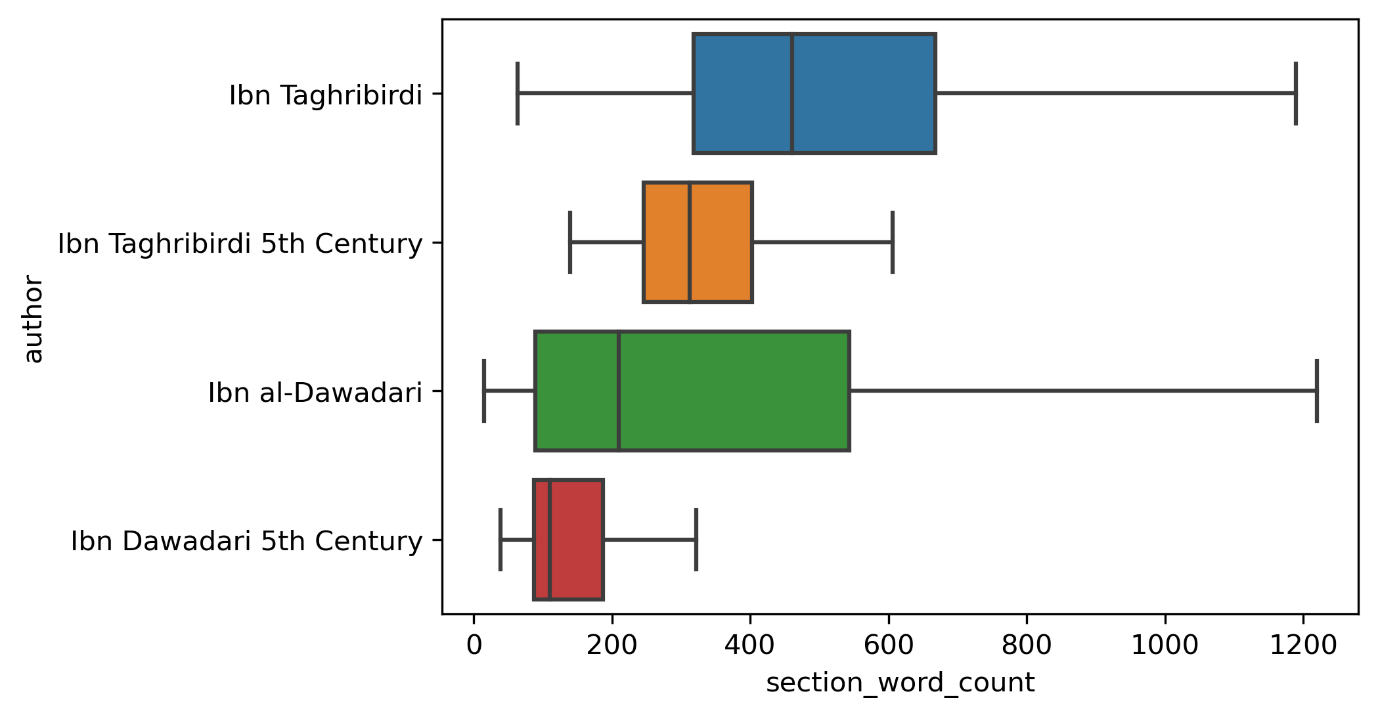
This nesting is recorded in OpenITI mARkdown using the ‘pipe’ notation (### | means level 1, ### || means level 2, and so on). Figure 1 gives an example of how this notation looks for Ibn Taghrībirdī’s *Nujūm* (note that the volume numbers of the work are given at level 1).

****

***Figure 1.*** *A clip of OpenITI mARkdown for Ibn Taghrībirdī’s Nujūm as it is seen in EditPad Pro. All of the headings have been collapsed apart from the final one.*

This annotation gives us insight into how highly structured texts function. In texts like those written by Ibn al-Dawādārī, al-Qalqashandī and Ibn Taghrībirdī, the author had a clear intention for organising their work. This might have been guided by their working methods or by their intended audience. Analysis of the mARkdown can, therefore, tell us something about how these texts functioned. From my reading of the parts of these chronicles related to the fifth century (that is the period of Fatimid rule), it appears that Ibn al-Dawādārī’s annual sections are shorter than those provided by Ibn Taghrībirdī. Analysis of the mARkdown allows me to see how representative this is of the whole work, which can in turn inform my understanding of these authors and their interests.

To perform this analysis, I extracted the headings from each work related to the fifth century and counted the number of words that appeared under each of these sections. As noted above, both authors uniformly use one level of heading for sections related to a specific year. I, therefore, compared the counts for the sections corresponding to the fifth century to the number of words given under every section of their works at the same mARkdown level. It should be noted that authors do not use this level of heading for only dated sections. In the parts of their work concerning events prior to the Hijri calendar (and in a few other cases), they describe key events under this heading level. Figure 2 provides a boxplot that shows the distribution of word counts for these sections in each of the works.

****

***Figure 2.*** *A boxplot comparing the spread of section lengths for level 3 sections in Ibn Dawādārī’s Kanz and Ibn Taghrībirdī’s Nujūm (that is the sections that predominantly deal with dates) with those sections concerning only the fifth century.*

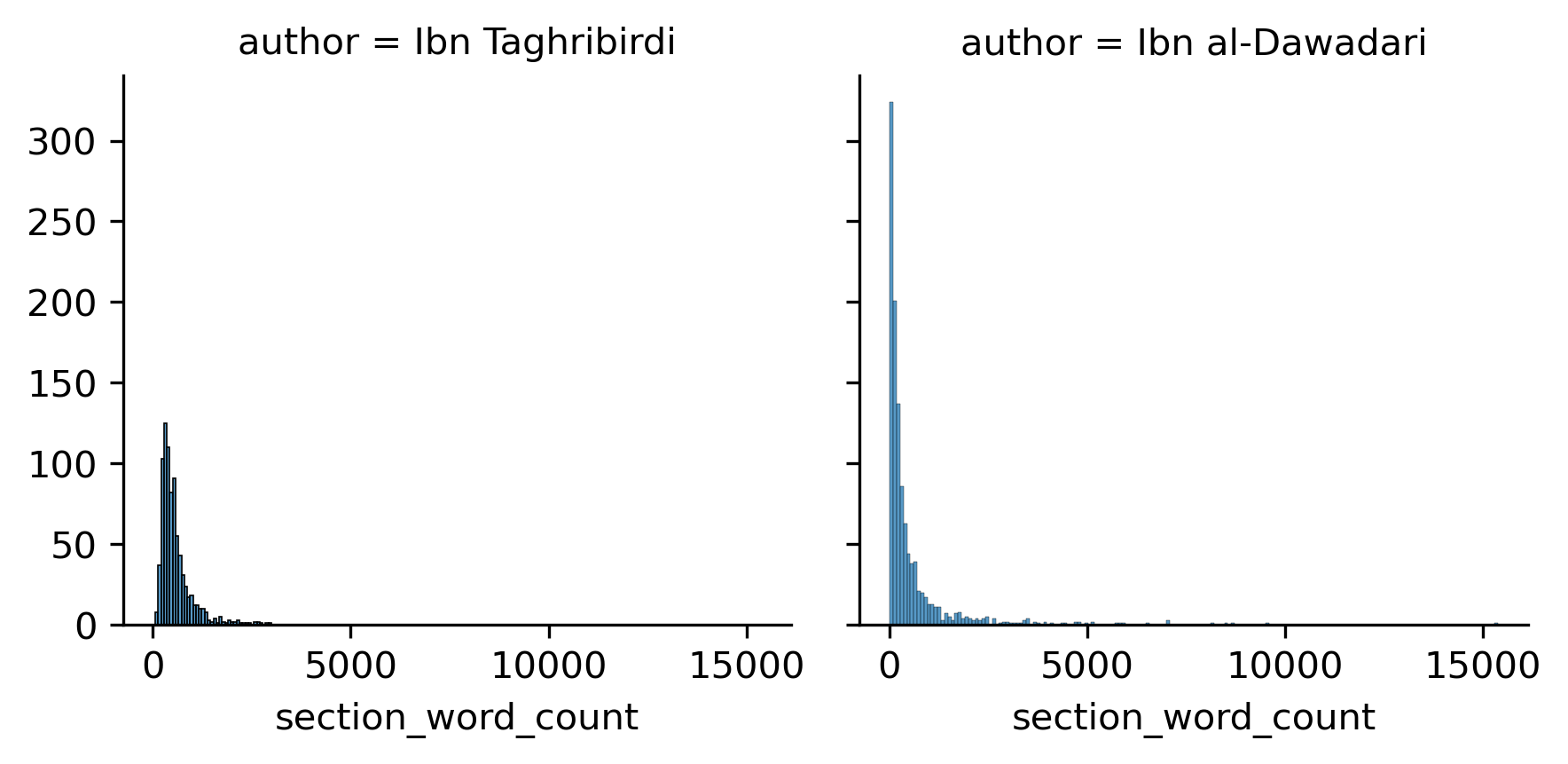
Boxplots are a valuable tool for understanding uniformity or variation in a dataset and comparing between datasets. The left of the box gives the lower quartile and the right of the box gives the upper quartile. In other words, the box corresponds to 50% of the datapoints in the dataset. The central line gives the median (the middle datapoint) and the data between outer whiskers constitutes 95 % of all datapoints in the dataset.

Boxplots are often given with outliers (that is any datapoints in the remaining 5%). To ease our reading of the data, the outliers have been excluded from the plot, but there were a number of them. For example, one of Ibn al-Dawadārī’s headings was 16,000 words long (this was outside the fifth century). The large number of outliers speaks in part to the nature of our dataset. History is not uniform. As was noted above, although Ibn al-Dawādārī uses this level of heading for dates, earlier in the work he uses this level for important events in pre-Islamic and early-Islamic history. Some of these subjects (such as ones relating to key events in the Prophet Muhammad’s life) are likely to have spanned a large number of words.

It is nonetheless helpful to focus the 95% of datapoints shown in figure 2. They appear to confirm my assumption that Ibn al-Dawādārī generally provides shorter sections under headings at this level than Ibn Taghrībirdī. However, the subset of data that focusses on the fifth century shows that this difference is more pronounced for headings relating to this century. Both authors appear to provide shorter sections on average when describing this period, but Ibn al-Dawādārī’s sections are much shorter. This helps to contextualise the work I am doing on the fifth century and it can inform my further study of these works.

***Understanding the structure of the work: the nesting problem***

There are, however, three main problems with focussing on the distribution of datapoints as we have in figure 2. Firstly, as noted above, there is the issue of outliers. To make the boxplots readable we needed to exclude the outliers. However, this is not scientific data and the outliers are likely to be of enormous interest to us (they are not simply anomalies that we can ignore). They might indicate an error in the annotation, but they might also indicate a heading of enormous importance. One solution is to use a histogram, which would show the outliers without drowning out the important datapoints. However, comparing across histograms is difficult. Figure 3 gives an example of comparative histograms for level 3 headings in Ibn Dawādārī’s and Ibn Taghrībirdī’s works. For the purpose of comparison, the two charts share an axis. However, because Ibn al-Dawādārī has a much higher frequency of short headings, while Ibn Taghrībirdī has a more even distribution, the graph for Ibn Taghrībirdī is difficult to read.

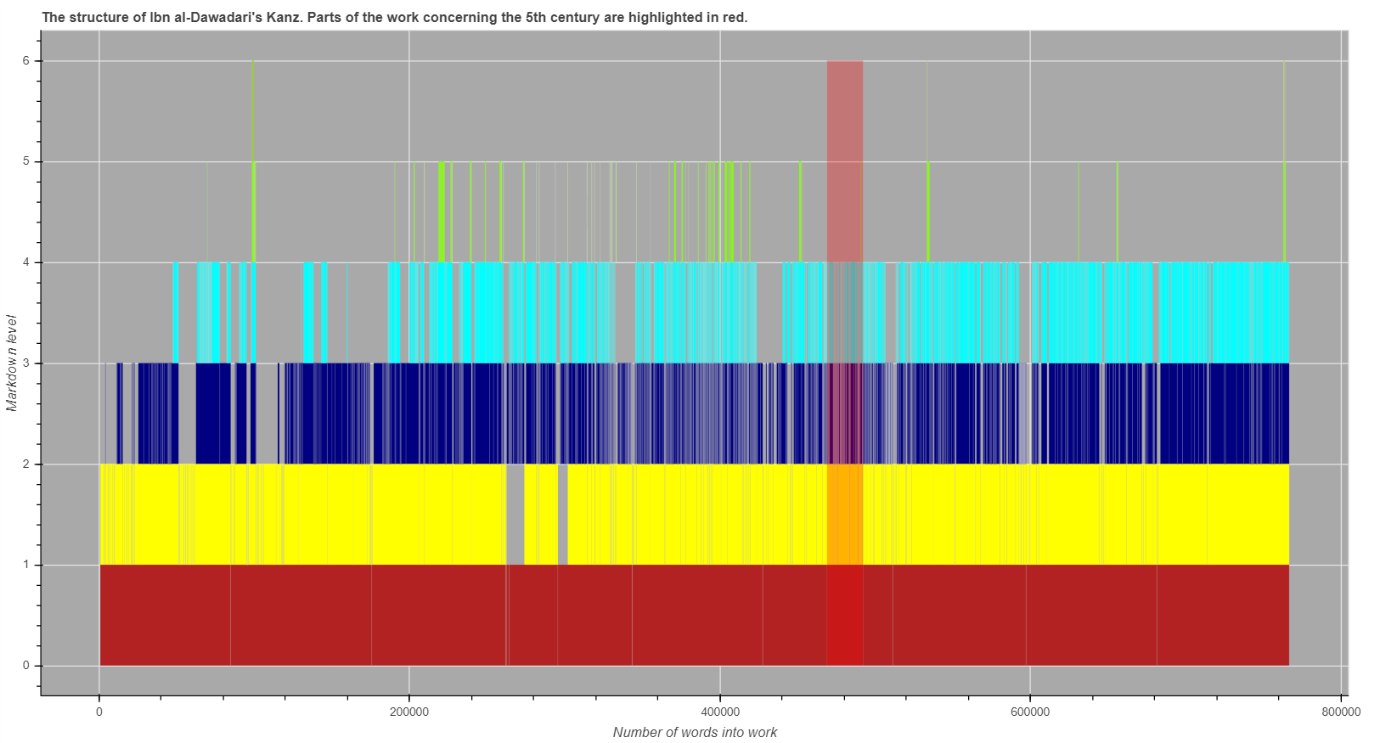
****

***Figure 3.*** *Histograms comparing the distribution of word counts for all level 3 sections in Ibn al-Dawādārī’s Kanz with Ibn Taghrībirdī’s Nujūm.*

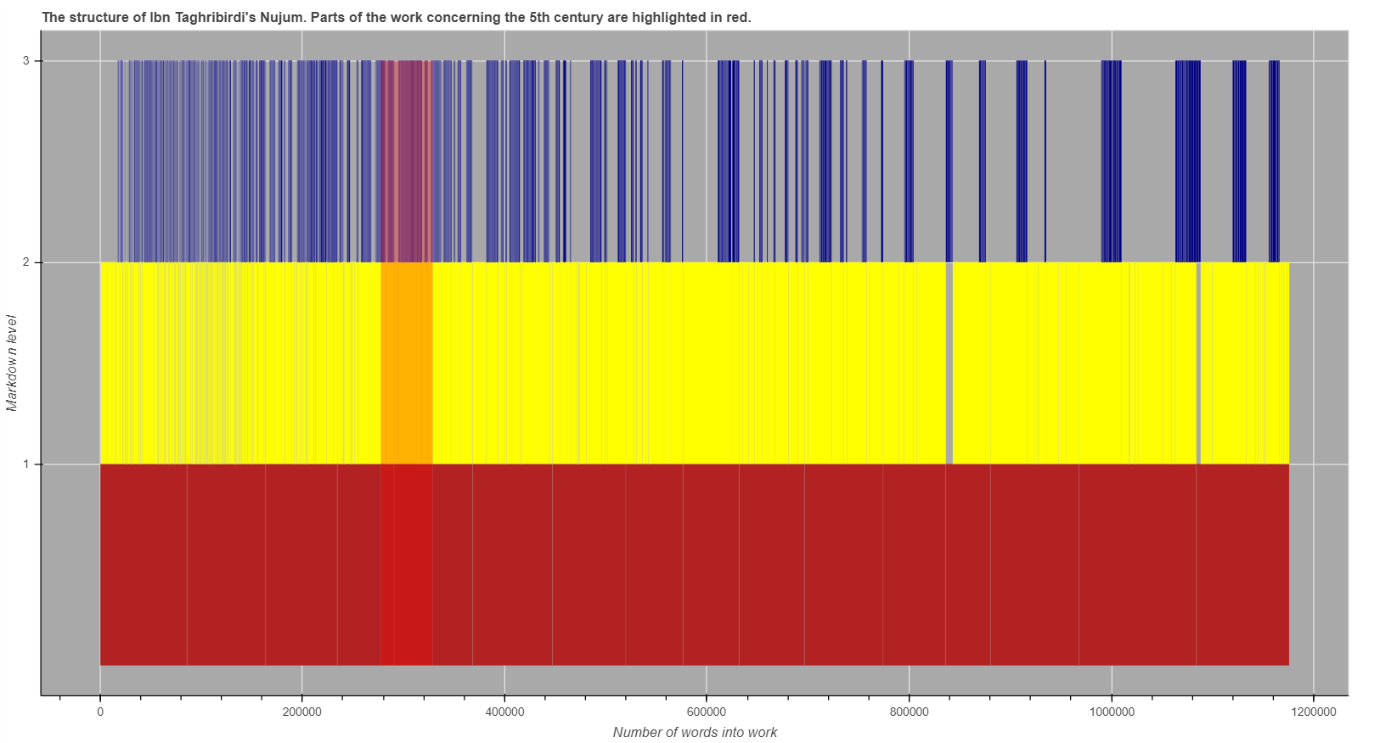
The second problem cannot be solved through histograms. All techniques for visualising data distributions abstract the data from their original contexts, making it impossible to know the distribution of the data throughout the original work. One solution is to slice the data into segments based on the arrangement of the work (as I did for the sections related to the fifth century in figure 2, and one could take this further to create separate boxplots for slices of 100 headings). However, in a dataset with around 1000 datapoints, this approach would still divorce the data from its original structural context.

The third problem relates to the second and it concerns nesting. At the start of this blog, I stressed that nesting matters enormously in these works. Focussing on one level of that nesting collapses all of the headings underneath into one word count and ignores any headings at the level above. In the case of Ibn al-Dawādārī and Ibn Taghrībirdī this could create misleading views of their works. Both provide historical summaries of the lives of rulers, into which the annual sections are nested. By focussing on the annual sections, we ignore the potential significance of that higher tier. Moreover, Ibn al-Dawādārī nests topics underneath each year. Focussing on the dated-sections does not, therefore, take account of how Ibn al-Dawādārī divides up those sections into sub-topics.

For this reason, I have been experimenting with ways of mapping the whole structure of OpenITI works. These visualisations are at the moment only experimental, but I show them here because they highlight the complexity of the nesting problem and its importance for these two works. Figure 4 gives a map of the headings in Ibn al-Dawādārī’s *Kanz* and figure 5 those for Ibn Taghrībirdī’s *Nujūm*. The y axis gives the mARkdown level, and the colour coding also corresponds to the mARkdown level. The x axis plots the word count of the work from start to finish. These figures, therefore, give a better understanding of how these two works are constructed and how the headings are nested into one-another.

****

***Figure 4.*** *A map of the structure of Ibn al-Dawādārī’s Kanz. The red highlighted section shows the part of the work corresponding to the fifth century.*



***Figure 5.*** *A map of the structure of Ibn Taghrībirdī’s Nujūm. The red highlighted section shows the part of the work corresponding to the fifth century.*

In both of these graphs, the dark blue bar represents the level 3 sections that are typically used by both authors for sections corresponding to years. Moreover, in both the yellow represents the level 2 sections that are used for recording summaries of the lives of rulers. For reference, the red shaded area of each graph corresponds to sections of the work dealing with the fifth century. These graphs show how Ibn al-Dawādārī generally provides shorter dated sections than Ibn Taghrībirdī (note that the latter’s work is nearly 200,000 words longer than the former, which makes the former’s sections appear a little longer). There are, however, a few sections that are much longer than Ibn Taghrībirdī’s. These constitute the outliers that were excluded from the boxplot and they often appear earlier in the work (although there are few corresponding to the later period, which would need further investigation). This partly confirms our suspicions that they correspond not to dated sections, but level 3 sections relating to pre and early Islamic history. These graphs also allow us to compare nesting. It shows, for example, how Ibn Taghrībirdī provides much longer summaries of the rulers before proceeding to describe the years of those reigns.

These graphs are only experimental, but they show the power of reading texts through mARkdown. The brief study presented here has used a very naïve reading of the mARkdown relying on texts that have quite formalised structures. I made no attempt here to look in detail at the content of the sections and have instead assumed their content based on the relatively uniform structure of the texts. The next step would, however, be to use the language of section headings to understand the structure of works more finely. One could, for example, identify headings that mention dates and even use a script to extract the dates (on identifying dates, see Maxim Romanov’s blog on the subject). One could, moreover, identify formulae associated with particular types of heading – those mentioning rulers, for example - and use that to identify headings mentioning rulers. Alternatively one could tag the content of the sections (for example, by identifying the parts of each year related to events, pilgrimages, obituaries and the Nile flood) and see if an author follows a uniform organisation even where they do not use headings. In short, this is only the beginning. However, at the very least, I hope for the moment to have shown how mARkdown can allows us to work with large chronicles and understand how their parts relate to the whole.

1. See, for example, the discussion in Hirschler, *The Written Word,* 17-20 (who notes al-Qalqashandī’s approach to structure). [↑](#footnote-ref-1)
2. Al-Qalqashandī, Subḥ, V:5-16. [↑](#footnote-ref-2)