

OpenITI mARkdownMSS: Basics of the Format

A *Technical Note*: the highlighting scheme for mARkdownMSS is a part of the new mARkdownSimple scheme, which will replace the initial mARkdown. The scheme is activated by the *magic value* #OpenITI#

The latest highlighting scheme for EditPad Pro is available here: https://github.com/OpenITI/mARkdown_scheme

The implementation of this scheme in development can be found here: <https://github.com/OpenITI/mARkdownMSS>; for examples of automatically generated editions, see <>

Main Text

```
#=# for a diplomatic transcription
#~# for an edited version

### for a comment
#+# for an insertion into the text
```

For each line of MSS text there must be two lines of transcription, which can be complemented with additional lines---one for comments, footnotes, etc.; another---for insertions of any kind.

1. #=# is the diplomatic transcription aimed at representing the text as close as possible to the witness; ideally, a transcriber should maintain a separate record to keep track of how specific cases are handled (this comment can be stored in the metadata head of the text file)
2. #~# is for an edited version with all the corrections that an editor deems relevant; this version should include structural and analytical tags
 - *important*: the edited transcription of the text should still closely follow the diplomatic version; if words must be added or removed, this should be done through comments or insertions. If this symmetry is broken, auto-collation will not be done correctly.
3. ### is for any kind of annotation or comment can be added through this line, using A[nchor tag]
 1. can be used to annotate a range of text or a specific location
 2. Range example:
 - #~# A33 this is the range of text to annotate A33
 - ### A33 :: This is a text of a comment to the range marked with the anchor tag A33
 3. Specific location example:
 - #~# a comment will be for this word A33
 - ### A33 :: there is really nothing special about this word
4. +#+ is for any kind of insertions that the editor deems necessary
 1. Example of an insertion:
 - #=# The butcher, the baker, the A55 maker
 - +#+ A55 :: candlestick

Multiple lines would look like shown below. Note an empty line between two-line blocks.

```
#=# for a diplomatic transcription
#~# for an edited version

#=# for a diplomatic transcription
#~# for an edited version

#=# for a diplomatic transcription
#~# for an edited version
```

Comment and insertion lines can be added either at the end of the document, or, perhaps more conveniently, right after the lines that they are connected to. Each one of these lines must be separated from any other line by an empty line. For example:

```
#=# for a diplomatic transcription
```

```
#~# for an edited version

## a comment

## an insertion

## for a diplomatic transcription
#~# for an edited version

## for a diplomatic transcription
#~# for an edited version

## a comment

## an insertion
```

or like this:

```
## for a diplomatic transcription
#~# for an edited version

## for a diplomatic transcription
#~# for an edited version

## for a diplomatic transcription
#~# for an edited version

## a comment

## an insertion

## a comment

## an insertion
```

A note: a file with `mARkdown` text will be automatically generate from the `mARkdownMSS` --- this file is then to be used for any kind of computational analysis. The files will maintain different extensions: `.mARkdown` and `.mARkdownMSS` respectively.

Marginal Notes

For transcribing marginal notes, use the same approach, but with the following beginning-of-the-line tags: `===#` & `#~#` (i.e., the middle element is repeated twice).

- Placing a note within the transcription document: let's say the comment begins around line 10 of the main text block; in this case, the commentary is to be added *after* the transcription block of line 10;
- Each comments must be transcribed as a single unit, no matter how many "lines" it takes; you can insert "|" to demarcate lines, if you consider that necessary.

Using Anchor Tags

- Anchor tags have a simple structure: `A + digits`
- how to pick numbers for anchor tags?
 1. technically, any number can be used, but the number must be unique, i.e. the same number can only be used once in the entire document;
 2. *practical suggestion*: the easiest way to ensure that numbers do not repeat is to use the line number shown in the EditPad Pro; if you have two or more ranges on the same line, you can use the `LineNumber-1`, `LineNumber-2` etc. The key point is that these numbers are not repeated and each anchor tag is unique.
- After you assigned and placed an anchor tag somewhere in the edited text, you need to add a comment or insert lines whose format is as follows:
 - for a comment: `## A33 :: your comment`
 - for an insertion: `## A55 :: your insertion`

- In other words:
 - it must be a new line, separated from other content with an extra empty line (before and after)
 - the line starts with `##` or `#+`
 - followed by space, two colons, and another space
 - after that you add your comment of insertion element.

Structural Tags

Structural tags include headers, beginnings of serial units (like biographies in biographical dictionaries; individual *ḥadīths* in *ḥadīth* collections, etc.), paragraph breaks.

Headers

Since headers are likely to span multiple lines, the following tag should be used:

- `HL_XX`, where:
 - `H` is the indication of header
 - `L` is the level of the header
 - `XX` is the unique number, for which, like with `A[nchor]` tags, you can use the number of the line in the `mArKdown` document.
 - for example: `H1_22 Text of the header H1_22`, is the header of level 1, which starts on line 22.

Serial units

Examples of serial units: *the same as in mArKdown*, but without `$`, just English letters: `BIO`, `BIO_MAN`, `BIO_WOM`, etc. These tags are to be used when it is necessary to mark just the beginning of a serial unit; its end will be automatically determined by the beginning of the next serial unit or a header.

§-breaks

- Insert `BR` (just these two capital letters) where you would want to have a paragraph break.

Poetry

- `BR` can also be used for poetry lines, when they are not written in the verse-per-line format; in such cases `BR` is to be inserted wherever a new line should start, and `%` inserted between hemistiches or at the beginning of a hemistich, if there is only one.
- if there is a `SEP` used in the manuscript to separate hemistiches, please add it in the edited line of text after `%`

Additional elements

Punctuation

Punctuation can be added to the edited version of the text. Simply add it wherever you want to have it. Keep in mind, "punctuation" cannot contain alpha-numeric characters. For these cases, use `A[nchor]` tag for insertions.

Folio Tags

- `FolioV00F000A` and `FolioV00F000B` are tags for folio numbers; the final `A` (*recto/wajh*) and `B` (*verso/zahr*) stand for the front and back pages, respectively, where:
 - `V00` is the volume number; use `V1` if your manuscripts only has 1 volume; use `V01`, `V02`, etc., if there are more than 10 volumes in the actual manuscript.
 - `F0000` is the folio number, where the length of the number should correspond to the length of the largest number of folios in MSS. That is: if there are 500 folios, this part of the tag should look like: `F001`, `F002`, ... `F020`, etc. If there are 80 folios, the tag can look like: `F01`, `F02`, etc.
 - Both `V00` and `F00` elements must be of the same length throughout the transcription.
- insert these tags in the `mArKdownMSS` document at the beginning of a manuscript folio — on the separate line (with an empty line before and an empty line after) right before the first line of text on that folio.

Lacunae

`...` is to be used to indicate an unclear, illegible, or missing section in the original scan. Use sets of three periods up to the approximate amount of illegible text. For example, if three words are illegible, you should insert: `... ..`

Named Entities (persons, toponyms, etc.)

You can tag the same types of entities which are included in `mArKdown`, however, since the structure of `mArKdownMSS` documents is different, the tags are also slightly different: 1) small letters instead of capital (e.g., `@p02` instead of `@P02`, as in `mArKdown`); 2) the entity will not be highlighted like in `mArKdown` (because of the `mArKdownMSS` structure, an entity may be split between multiple lines, which makes highlighting impossible);

- To tag a person, use `@pYX` tags (note the small letter `p`), where `Y` is the length of a prefix in characters, while `X` is the length of the entity in words.
 - For example, if you need to tag *Muḥammad bn Aḥmad*, you can place `@p03` right in front of the name: `@p03 Muḥammad bn Aḥmad`; if you are tagging the phrase *wa-Muḥammad bn Aḥmad*, the `0` in the tag should be changed to `1` to indicate *wa-*: `@p13 wa-Muḥammad bn Aḥmad`
- For toponyms: `@tYX`
- etc.









Note: upon conversion to `mArKdown`, these tags (`@p01` and `@t01`) will be updated into standard ones (`@P01` and `@T01`) and color highlighting will work in the `mArKdown` document.

Graphic Elements from MSS

In-Text Elements: Separator Tags

You can use separator tags to indicate in-text graphical elements (non-alphanumeric separator) in the text of your transcription. The structure of the tag is: `SEPX`, where `X` is a number. Each separator tag must indicate one *type* of separators. For example, if you have circle separators and three-dot separators, you can use `SEP1` to indicate one and `SEP2` to indicate the other. This tag is to be used in the diplomatic transcription only.

For each separator tag there should be a corresponding image of a separator cropped from the original image. Ideally, the original resolution should be preserved. Each image must be named with the code used in the text (for example, if there are `SEP1`, `SEP2` in the text, there must be corresponding images: `SEP1.jpg`, `SEP2.jpg`); this approach will allow us to use original separators in the final version (images can be converted into black and white and “schematized” in order to better fit into the text).

Original Image	Schematized Image
	
	
	
	

Figures, Miniatures, Images

You can use a figure tag in order to indicate a graphical element like a figure, miniature, image. The tag had the following structure: `FIG_LLRR_HI_NNN` where:

- `FIG` is a *figure* indicator (does not change).
- `_LLRR` is the horizontal dimension of the image (on the dozen scale), where `LL` is the left margin of an image and `RR` is the right one; the scale is from 0 to 12 (as it allows to have: a half, a third, a quarter):
 - `_0006` means that the image takes left half of the page; `_0012` - the image takes the full width; `_0612` - the right half; `_0004` - left third; `_0812` - right third. (see “*Line scale examples (LLRR)*” below for more details).
- `_HI` is a number indicating the height of the image in lines.
- `_NNN` is the unique number of the image (you can use the folio number + a letter, if there are more than one image on that folio, i.e. 1, 2, 3a, 3b, etc.)

Place this tag *after* the line of text where non-alphanumeric signs or figures appear---not in the text line itself. For each figure there should be a corresponding image (of a figure cropped from the original image). Ideally, the original resolution should be preserved. Each image must be named with the code used in the text (for example, if there are `FIG1`, `FIG2` in the text, there must be corresponding images: `FIG1.jpg`, `FIG2.jpg`); this approach will allow us to use original figures in the final version (images can be converted into black and white and “schematized” in order to blend better with text)

Original Image	Schematized Image
----------------	-------------------

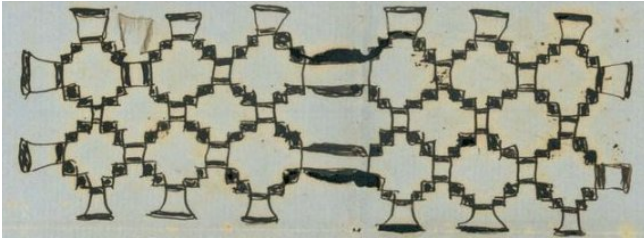
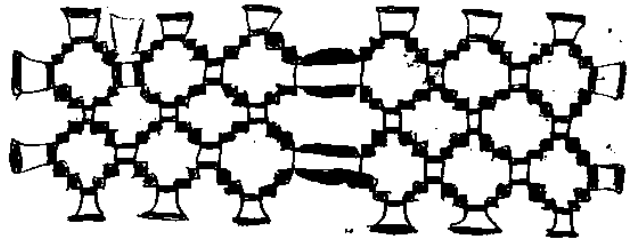
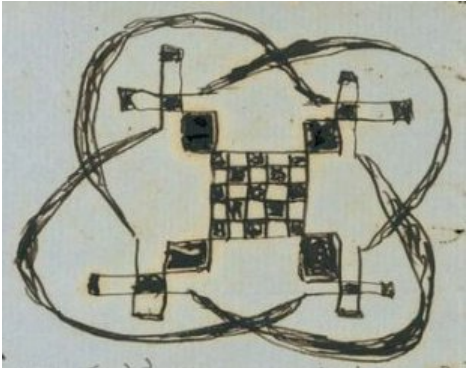
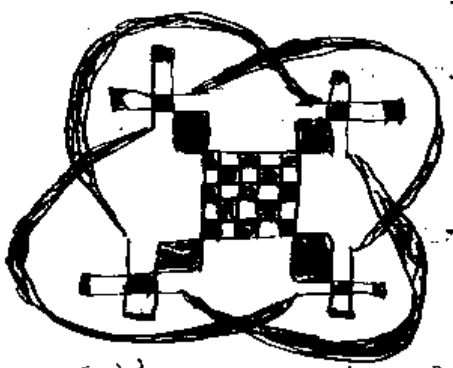
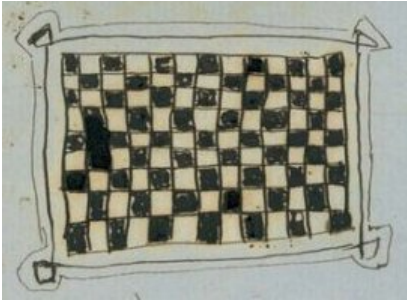
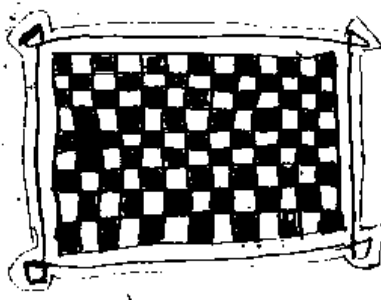
Original Image	Schematized Image
	
	
	

Figure Tag Examples::

- [FIG_0012_05_10](#) is an image that takes the full width of the text block, 5 lines in height and appears on folio 10.
- [FIG_0006_10_100b](#) is an image that takes left half of the text block, is 10 lines in height, and is the second (b) image that appears on folio100
- [FIG_0612_2_5](#) is the image that takes the right half of the text block, is 2 lines in height and appears on folio 5.

Line scale examples (LLRR):

1. [0004](#): left third

```

texttexttexttexttexttex
texttexttexttexttexttex
X X X X texttexttexttex
X X X X texttexttexttex
X X X X texttexttexttex
X X X X texttexttexttex
texttexttexttexttexttex
texttexttexttexttexttex

```

2. [0912](#): right quarter

```

texttexttexttexttexttex
texttexttexttexttexttex
texttexttexttextt X X X
texttexttexttextt X X X
texttexttexttextt X X X
texttexttexttextt X X X
texttexttexttexttexttex
texttexttexttexttexttex

```

3. [0012](#): full width

```

texttexttexttexttexttex
texttexttexttexttexttex
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
X X X X X X X X X X
texttexttexttexttexttex
texttexttexttexttexttex

```

4. 0408: a third in the middle

```

texttexttexttexttexttex
texttexttexttexttexttex
texttex X X X X texttex
texttex X X X X texttex
texttex X X X X texttex
texttex X X X X texttex
texttex X X X X texttex
texttexttexttexttexttex
texttexttexttexttexttex

```

Issues

- how to deal with **split words**, which do occur in MSS?
 - keep as is in the diplomatic transcription;
 - keep as is in the edited, but add pluses: one directly after the first part (no space); another directly before the second part (no space);
 - the word can be merged in mARkdown with the help of those pluses as well as in the final visual representation of the edited text

Notes

- [Getting Started with OpenITI mARkdown \(Google Doc\)](#).
- [OpenITI mARkdownMSS \(Google Doc\)](#).

Up to date

There is a whole second set of issues that Maxim is going to go through and create a script to automatically fix. Most of these are issues or modifications that we thought needed to be made to make the OpenITI mARkdown schema clearer or more precise. Actually, working with you on this project really helped us think through several of these issues. Below are the issues that Maxim is going to write a script to automatically fix:

(1) We realized that mARkdown highlighters only work in non-vocalized text so Maxim is going to automatically generate a non-vocalized version of your edited/corrected lines. This line will be marked with: **#m#**

- **UPDATE:** keep the original two-line transcription: diplomatic (**#=#**) and edited (**#~#**); *Reason:* mARkdown can be automatically generated from **edited**; using English letters messes things up the visual directionality (it works ok on EditPadPro, but will not work in Kate, which we want to adopt as it works natively on Mac and Linux, and is open source)

(2) On a related note, we are going to change the line tags to **#d#** (diplomatic transcription), **#e#** (edited/corrected transcription with vocalization), and **#m#** (edited/corrected transcription with mARkdown and without vocalization) for the sake of clarity. - **UPDATE:** there are issues with using Latin characters (they may mess up the visual directionality in text editors); keeping non-letter characters seems a better option.

(3) The file names should be the same as the URI names in the metadata. Maxim is going to fix a few issues in the URIs you provided and then re-name the file names with them. - IDs should be shorter

(4) For some reason there are a lot of extra spaces in the document so Maxim is going to strip those out. - can be fixed automatically

(5) The additional two transcriptions below the mARkdown documents are not necessary. We only included them in the sample file for demonstration sake. Those two renders of the text can be easily reproduced from the mARkdown text itself. - we can add it as an *appendix* at the end of the file, after: **#mARkdownAUTOGENERATED#####**, which will be regenerated every time the script is run; it can also be placed into a separate file, where only mARkdown is used.

(6) Finally, Maxim is going to change the text anchor name from Q to A for clarity.

Checklist

- ☒ Correct URIDs
- ☒ Correct TAGs, spacing, etc.
- ☐ Converter
- ☐ Need to add processing of margin comments...

Issues

- some way to deal with merged words in the diplomatic transcription:
 - DT "wordword" (will be automatically removed)
 - ED "word word"
- missing words in the diplomatic, i.e. they must be there, but dropped (not the case of illegible words):
 - this is what #++ is for, but insertion is only in the edited text: TAG, TAG+insertion

Inconsistencies

- no catalog ID for 1280CumarIbnSayyid.AyatTaylor.SCHAD2-ara1
- *hamzats* are often omitted in the edited version;
- names are tagged with *waw* attached and with *waw* detached, while they are attached in the diplomatic transcription;
- most commonly, spaces are missing where they must be --- the highlighting scheme is designed to *highlight* those issues: if tags are not changing colors, it means you typed it wrong;
- the comment and addition lines are often typed incorrectly, which is partially because of the issues with directionality, but again, color highlights must be used as a test that format is correctly;