REGION ANNOTATION GUIDELINES
DH25 TEAM

Introduction

This document provides guidelines for annotation of regions i.e. segmentation of Persian manuscripts in the 20th century originating in Central Asia and South Asia. Regions are classified as zones. The manuscripts currently being digitized by the DH25 team are important publications of the Shia Imami Ismaili community in Gilgit-Baltistan. They are being digitized in order to utilize them in broader studies of the history of Ismailism and study of Islam in South and Central Asia. For this purpose, a segmentation ML model has been trained using selected pages from the two manuscripts, which can then apply segmentation on all pages of the manuscripts and facilitate transcription.

Segmentation model:

urseg_dh25_59p.mlmodel

See below (page#2 to page#10) for definitions and examples of each category of annotation of regions.

NumberingZone: is a zone containing the page number, with no regard for the mark's origin (scribe, curator, etc). The zone usually is at the top of the page. Note that this numbering region is strictly for pagination.
See Figure 1 below for an example of this region.

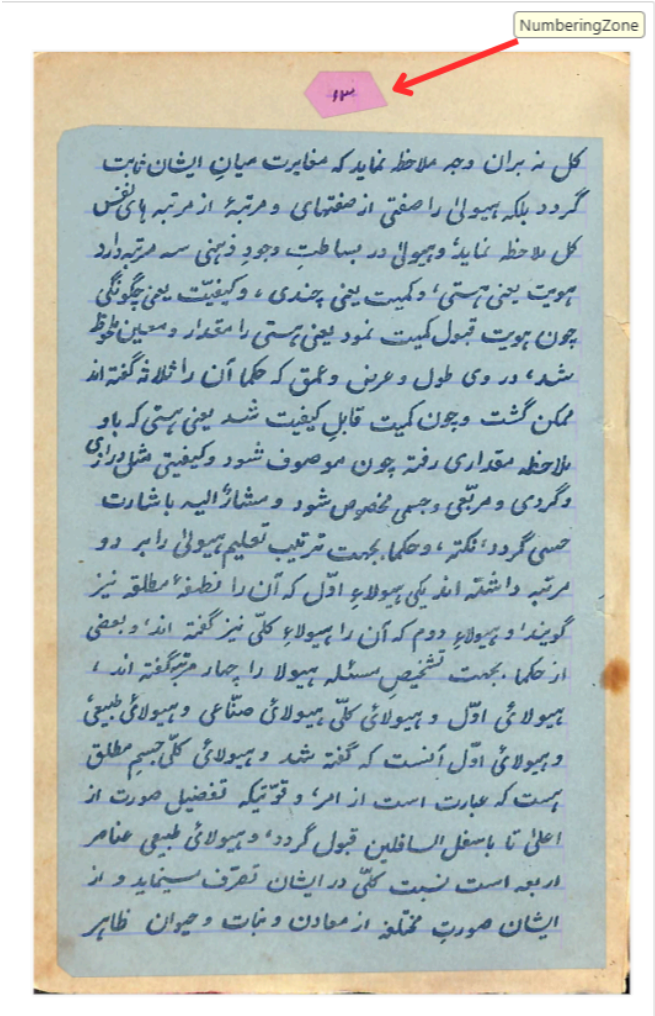


Figure 1: NumberingZone region shown in pink. The red arrow indicates the position of the region for the label.

MainZone: is the main area containing the text, excluding any paratext, and it is either a single block or multiple columns. For our purposes, elements such as transcribed poems, folionations, and in text-titles are included in this zone.

See figures 2,3 and 4 below for examples.

Figure 2: MainZone region shown in blue. MainZone is the text body on a page. The image shows the MainZone region in blue.

Any titles or headings within the text body are also a part of the MainZone. See Figure 3 below.

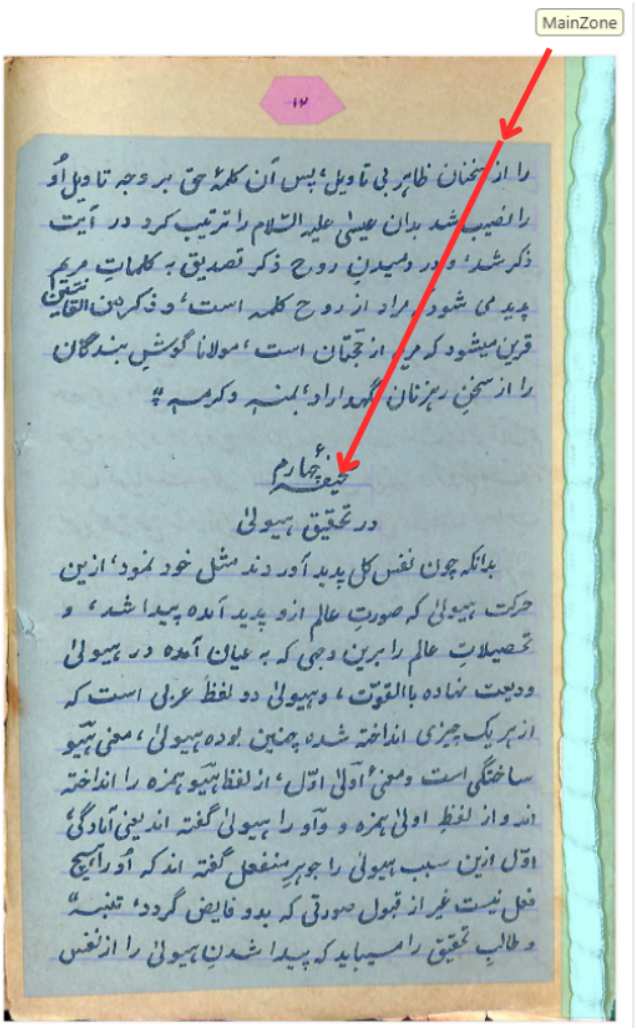


Figure 3: See extended red arrow indicating titles/headings included within the MainZone region. MainZone is the text body on a page. The image shows the MainZone region in blue.

Any text in columns that are part of the main text on a page, such as verses of a poem, fall within the MainZone. See Figure 4 below.

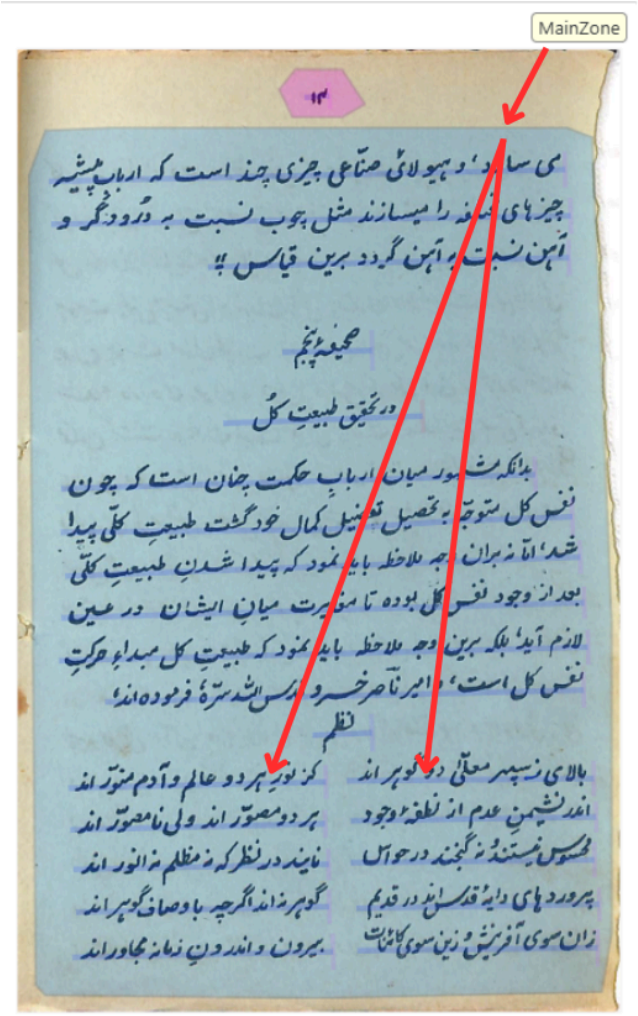


Figure 4: See extended red arrows indicating columns for verses of a poem included within the MainZone region. MainZone is the text body on a page. The image shows the MainZone region in blue.

MarginTextZone: characterises any text zone contained in the margins no matter its position on the page (upper, lower, inner or outer), including the space between two columns.

MarginTextZone:note: refers to the designated area at the bottom of a page where explanatory notes, references, or citations are placed, clearly separated from the main text body.

See figure 6 below for an example of this region.

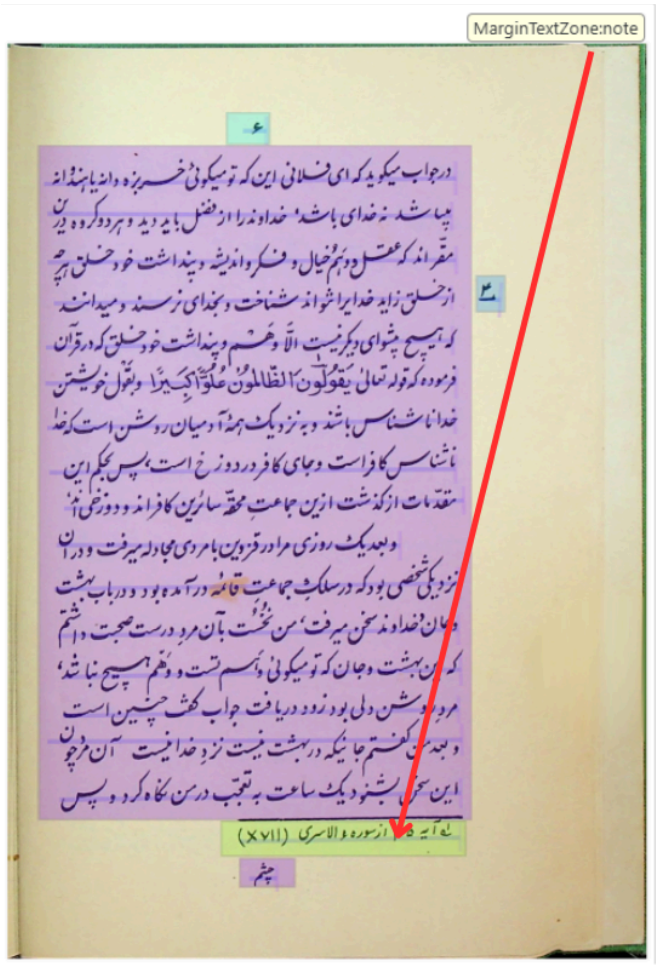


Figure 6: MarginTextZone:note shown in yellow. See red arrow indicating footnote on the page.

TitlePageZone: characterises an entire page in the front of our documents containing headings (chapter title, act or scene number, etc.) and bibliographic or identifying information, such as the title of the work, the production date, the names of the printer(s), publisher(s) and author(s), etc.

See Figure 7 below for an example of this region.

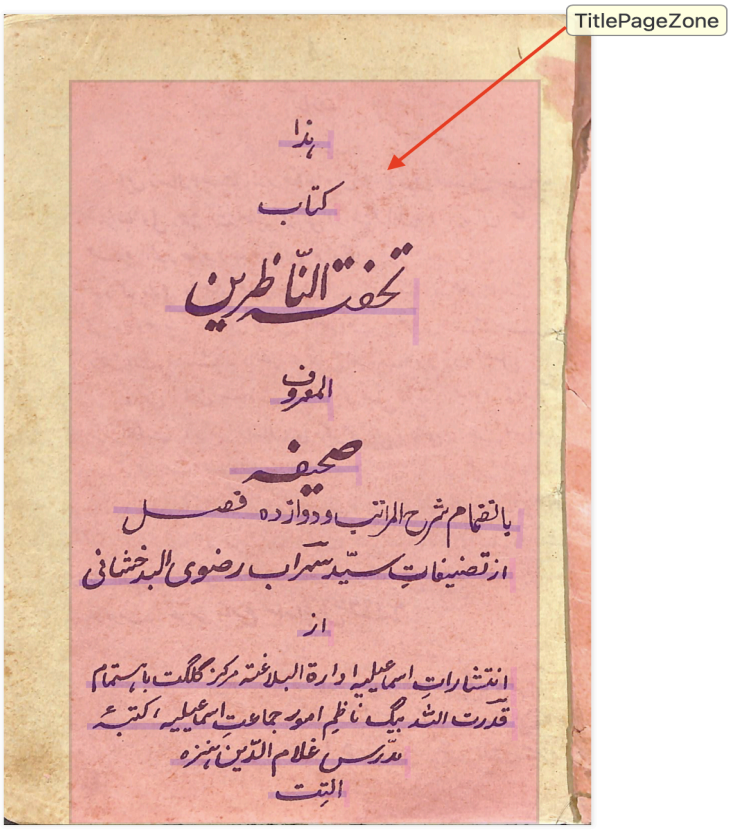


Figure 7: TitlePageZone region shown in pink. The red arrow indicates the position of the region for the label.

NumberingZone:other: refers to the designated area of a page or section used for displaying non-body-text numbering systems, such as annex numbering, appendix references, table/figure numbering, or special section identifiers.

See Figure 8 below for an example of this region.

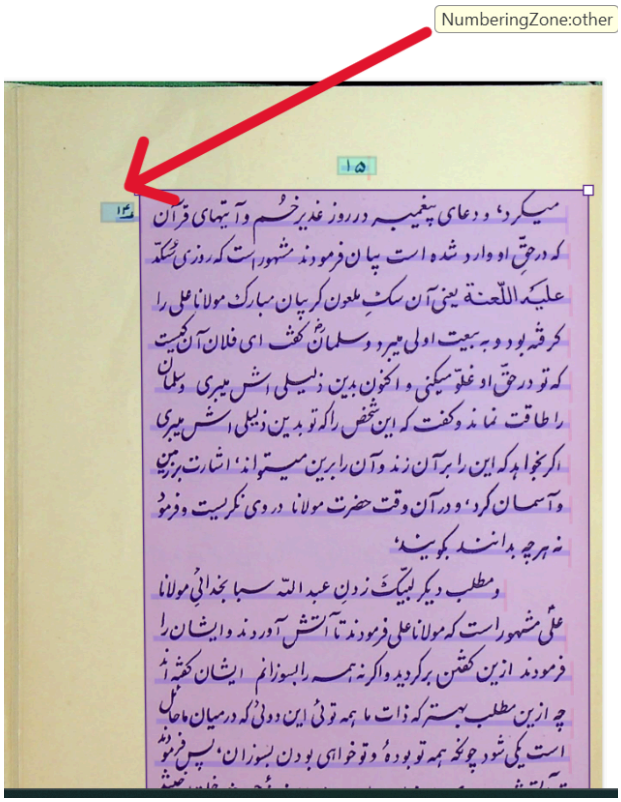


Figure 8: NumberingZone:other region shown in green. The red arrow indicates the position of the region for the label.

QuireMarksZone:catchwords: is a zone which contains a word, usually placed on the margins of the text, which is the same as the first word on the next page.

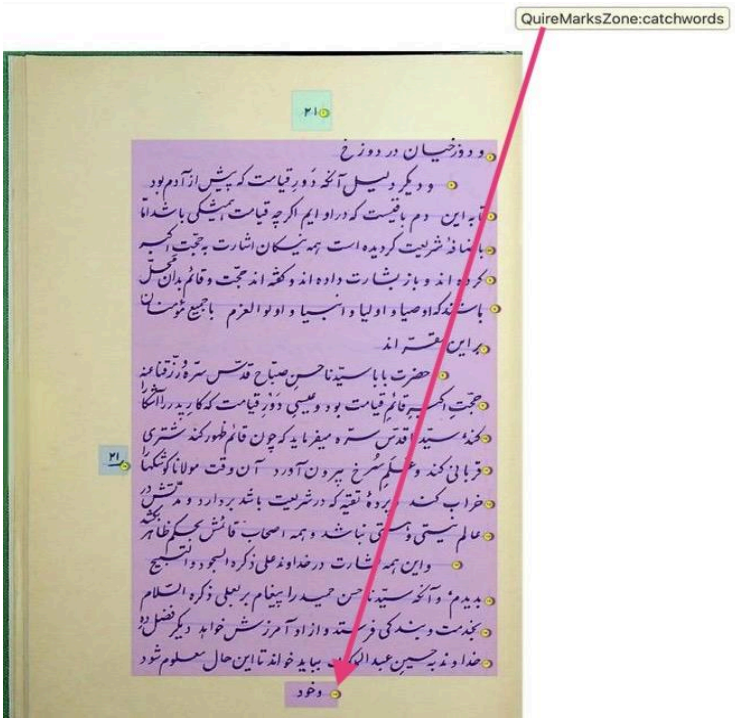
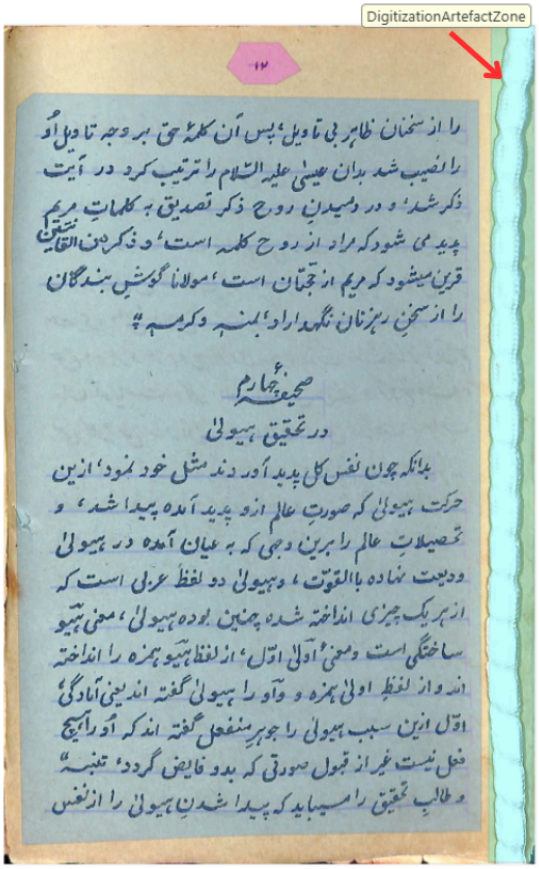


Figure 9: QuireMarksZone:catchwords region shown in purple. The red arrow indicates the position of the region for the label.

DigitizationArtefactZone: contains any type of item external to the document itself present on the image because of the digitisation process. It can be a ruler to measure the document or a test card (or color table/target) to calibrate colours for the camera.

See Figure 10 below for an example of this region.



را از سخنان ظاهر بی تاویل، پس آن کلمهٔ حق بر وجه تاویل او
را نصیب شد بدان عیسی علیه السلام را ترتیب کرد در آیت
ذکر شد، و در دمیدن روح ذکر تصدیق به کلمات مریم
پدید می شود که مراد از روح کلمه است، و ذکر بین التقیین
قرین میشود که مریم از جملهٔ ان است، مولانا گوشش بندگان
را از سخن رهزنان نگهدارد، بمنه وکرمه"

صحیفهٔ چهارم
در تحقیق هیولی

بدانکه چون نفس کل پدید آورد و مثل خود نمود، ازین
حرکت هیولی که صورت عالم از و پدید آمده پیدا شد، و
تحصیلات عالم را بدین وجهی که به عیان آمده در هیولی
ودیعت نهاده بالقوّت، وهیولی دو لفظ عربی است که
از هر یک چیزی انداخته شده چنین بوده هیولی، معنی هیو
ساختگی است ومعنی اولی اوّل، از لفظ هیو همزه را انداخته
اند و از لفظ اولی همزه و واو را هیولی گفته اند یعنی آمادگی
اوّل ازین سبب هیولی را جوهر منفعل گفته اند که او را هیچ
فعل نیست غیر از قبول صورتی که بدو فایض گردد، تنبیه"
و طالب تحقیق را میباید که پیدا شدن هیولی را از نفس

Figure 10: DigitizationArtefactZone region shown in green. In this case it is a white snake that was used to hold the book open and the page to be scanned flat. The red arrow indicates the position of the region for the label.