

Гайд по структурной разметке

Для того, чтобы начать работу со структурной разметкой (тэгами) необходимо перейти:

Metadata → Structural. Далее для того, чтобы нанести тэг на область/линию/слово нужно щелкнуть на это место на странице или меню Layout (открывается под списком тэгов) и нажать на зеленый плюсик рядом с нужным тэгом.

Для основной модели этого года выбрано 6 тэгов:

Page-number – номер страницы

Marginalia – заметка на полях

Book decoration – украшение

Map – карта

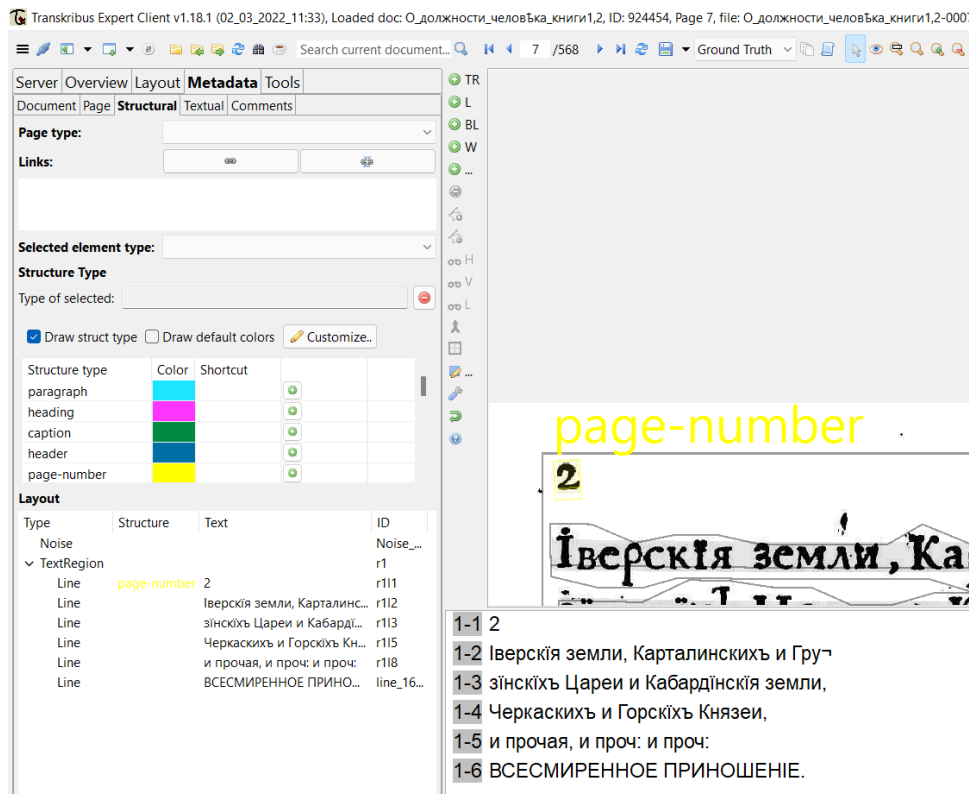
Header – колонтитул

Initial – буква

Как с ними работать?

1. Тэг Page-number наносится на строку, в которой содержится номер страницы. Номер страницы у нас всегда выделяется отдельной маленькой строкой, кажется, случаев, где само число номера страницы придется распознавать как отдельное слово, возникнуть не должно.

Пример выделения тэга (название дока в самом верху картинке):



Transkribus Expert Client v1.18.1 (02_03_2022_11:33), Loaded doc: О_должности_человѣка_книги1,2, ID: 924454, Page 7, file: О_должности_человѣка_книги1,2-0001

Server Overview Layout **Metadata** Tools

Document Page **Structural** Textual Comments

Page type:

Links:

Selected element type:

Structure Type

Type of selected:

☒ Draw struct type ☐ Draw default colors

Structure type	Color	Shortcut
paragraph		
heading		
caption		
header		
page-number		

Layout

Type	Structure	Text	ID
Noise			Noise_...
TextRegion			r1
Line	page-number	2	r111
Line		Іверскія земли, Карталинс...	r112
Line		зінскіхъ Цареи и Кабард...	r113
Line		Черкаскихъ и Горскіхъ Кн...	r115
Line		и прочая, и проч: и проч:	r118
Line		ВСЕСМИРЕННОЕ ПРИНО...	line_16...

page-number

2

Іверскія земли, Ка

1-1 2

1-2 Іверскія земли, Карталинскихъ и Груч

1-3 зінскіхъ Цареи и Кабардінскія земли,

1-4 Черкаскихъ и Горскіхъ Князеи,

1-5 и прочая, и проч: и проч:

1-6 ВСЕСМИРЕННОЕ ПРИНОШЕНИЕ.

2. Тэг Marginalia наносится на область текста (Text region), так как заметка на полях всегда выделяется как отдельная область текста. Если на одной странице сразу несколько отдельных заметок на полях, в отдельную область текста выделяется каждая.

Пример выделения тэга:

Transkribus Expert Client v1.18.1 (02_03_2022_11:33), Loaded doc: Дневная_записки, ID: 885686, Page 19, file: Дневная_записки-0020.jpg [Image Meta Info: (Resolution:1.0, w:h: 2

Server Overview Layout **Metadata** Tools

Document Page **Structural** Textual Comments

Page type:

Links:

Selected element type:

Structure Type

Type of selected:

☒ Draw struct type ☐ Draw default colors [Customize...](#)

Structure type	Color	Shortcut
paragraph		
heading		
caption		
header		
page-number		

Layout

Type	Structure	Text	ID
> Line		18 версть, разстоянiемъ о...	r1130
> Line		отъ дороги, видѣли общи...	r1131
> Line		прозываемое. О семь озе...	r1132
> Line		тсѣ баснь, будто бы тутъ п...	r1134
> Line		Чость I.	r1136
> Line		Б	line_16.
> Line		селе-	r1135
TextRegion	marginalia		region..
> Line		Деревня	r117
> Line		липни.	r119
TextRegion	marginalia		region..
> Line		Село Ун-	line_16.
> Line		далы.	r1114
TextRegion	marginalia		region..
> Line		Поганое	line_16.
> Line		озеро.	r1133

ику: мѣстами просядали
какъ то обыкновенно бы-
. Мы прошли около де-
ли насъ наши подводчи-
ны до деревни Липни,
иржачи въ 28 верстахъ.
ю, и яы, желая награ-
въ Киржачахъ, продол-
езъ двадцать верстъ до
лы весьма пространно,
мѣстѣ. Позади села
аешь рѣка Клязьма, отъ
продолжаются поемные

1-1 1768, ЮЛЯ 11—12. СЕЛО УНДАЛЫ.
1-2 9
1-3 калинику, ивняку, осиннику: мѣстами просядали
1-4 и боры, на которыхъ, какъ то обыкновенно быт
1-5 ваетъ, росъ красной лѣсѣ. Мы прошли около де-
1-6 вяти верстъ, какъ нагнали насъ наши подводчи-
1-7 ки, съ которыми ѣхали мы до деревни Липни,
1-8 отстоящей отъ села Киржачи въ 28 верстахъ.

3. Тэг Book decoration выделяется на орнаментах, картинках, рисунках, печатях (на всех изображениях кроме карт), он наносится на весь регион как и тэг Marginalia. Есть существенное отличие в выделении регионов страницы под этим тэгом. Горизонтально вытянутые орнаменты выделяются регионом Separator. Все картинки, печати и прочее (визуально не разделяющее страницу) выделяется регионом Image.

Примеры нанесения тэга для Separator:

Transkribus Expert Client v1.18.1 (02_03_2022_11:33), Loaded doc: moskovsky_zhurnal_chast_1_1791_.jpg, ID: 866876, Page 1, file: moskovsky_zhurnal_chast

Server Overview Layout **Metadata** Tools

Document Page **Structural** Textual Comments

Page type:

Links:

Selected element type: Separator

Structure Type

Type of selected: book decoration

☒ Draw struct type ☐ Draw default colors [Customize...](#)

Structure type	Color	Shortcut
paragraph		
heading		
caption		
header		
page-number		

Layout

Type	Structure	Text	ID
Image	book decor...		Image...
Separator			Separa...
Separator	book decor...		Separa...
TextRegion			r1

ОСКОВСКОЙ
У РНАЛТ

book decoration

Пример нанесения тэга для Image (здесь же ниже другой пример нанесения того же тэга на t.r. Separator):

Transkribus Expert Client v1.18.1 (02_03_2022_11:33), Loaded doc: Журнал или Дневные записки путешествия капитана Рычкова.pdf, ID: 888950, Page 1, file: p001.jpg [Image Meta Info: (Resolution

Server Overview Layout **Metadata** Tools

Document Page **Structural** Textual Comments

Page type:

Links:

Selected element type: Image

Structure Type

Type of selected: book decoration

☒ Draw struct type ☐ Draw default colors [Customize...](#)

Structure type	Color	Shortcut
paragraph		
heading		
caption		
header		
page-number		

Layout

Type	Structure	Text	ID
Image	book decor...		Image_...
Separator	book decor...		Separa...
TextRegion			r1
Line		ЖУРНАЛЬ	r111
Line		или	r112
Line		ДНЕВНЫЕ ЗАПИСКИ	r113
Line		путешествия	r114
Line		Капитана Рычкова	r115

1-1 ЖУРНАЛЬ

1-2 или

1-3 ДНЕВНЫЕ ЗАПИСКИ

- Тэг Мар наносится исключительно на карты. Выделяется регион Image и на него наносится тэг. Пояснение к карте, которое печатается в углу на белом фоне из региона удаляется.

Пример нанесения тэга Map:

Transkribus Expert Client v1.18.1 (02_03_2022_11:33), Loaded doc: Журнал или Дневные записки путешествия капитана Рычкова.pdf, ID: 888950, Page 6, file: p006.jpg [Image Meta Info: (Resolutio

Server Overview Layout **Metadata** Tools

Document Page **Structural** Textual Comments

Page type:

Links:

Selected element type: Image

Structure Type

Type of selected: map

☒ Draw struct type ☐ Draw default colors [Customize...](#)

Structure type	Color	Shortcut
paragraph		
heading		
caption		
header		
page-number		

Layout

Type	Structure	Text	ID
Image	map		Image_...
TextRegion			region..
Line		Къ путешествію Капит. Ры...	line_16.
TextRegion			region..
Line		Карта	line_16.
Line		Учиненная во время путе...	line_16.
Line		Капитана Рычкова по раз...	line 16.

1-1 Къ путешествію Капит. Рычкова.

2-1 Карта

5. Тэг Header – пока самый опциональный, не уверена в его нужности, наносится на заголовки в самом верху обычных страниц текста (подобие колонтитулов). Они, как правило, содержат либо сведения о дате написания (в дневниках), либо там дублируется название самой книги или ее главы. Наносится на строку текста.

Здесь также важно, что под колонтитулом часто находится горизонтальная жирная черта. Ее мы выделяем регионом Separator (чтобы позже модель не пыталась прочесть в ней буквы), но никакой тэг на нее не наносим (если это простая линия, а не орнамент).

Примеры нанесения тэга Header:

The first screenshot shows the Transkribus Expert Client v1.18.1 interface. The document is 'Журнал или Дневные записки путешествия капитана Рычкова.pdf'. The 'Metadata' tab is active, and the 'Page type' is set to 'Text'. The 'Structure Type' is set to 'header'. The 'Layout' table shows the following structure:

Type	Structure	Text	ID
Separator			Separ.
TextRegion			r2
Line	page-number	4	line_16
Line	header	МАІА МЪСЯЦА 19 ДНІА.	r211
Line		Вышеупомянутое жительство окруже	r214
Line		всѣхъ сторонъ великимъ числомъ земледѣль	r215
Line		хлѣбная цѣна бываетъ обыкновенно дѣше	r216
Line		нежели въ другихъ мѣстахъ: то для сего по	r218
Line		енѣ тутъ казенный винокуранный заводъ ,	r219
Line		рый довольствуется хлѣбомъ...	r2110
Line		онимъ же селѣ по причити...	r2111
Line		ждуя недѣлю тутъ базара...	r2112
Line		щихъ деревень великимъ...	r2113

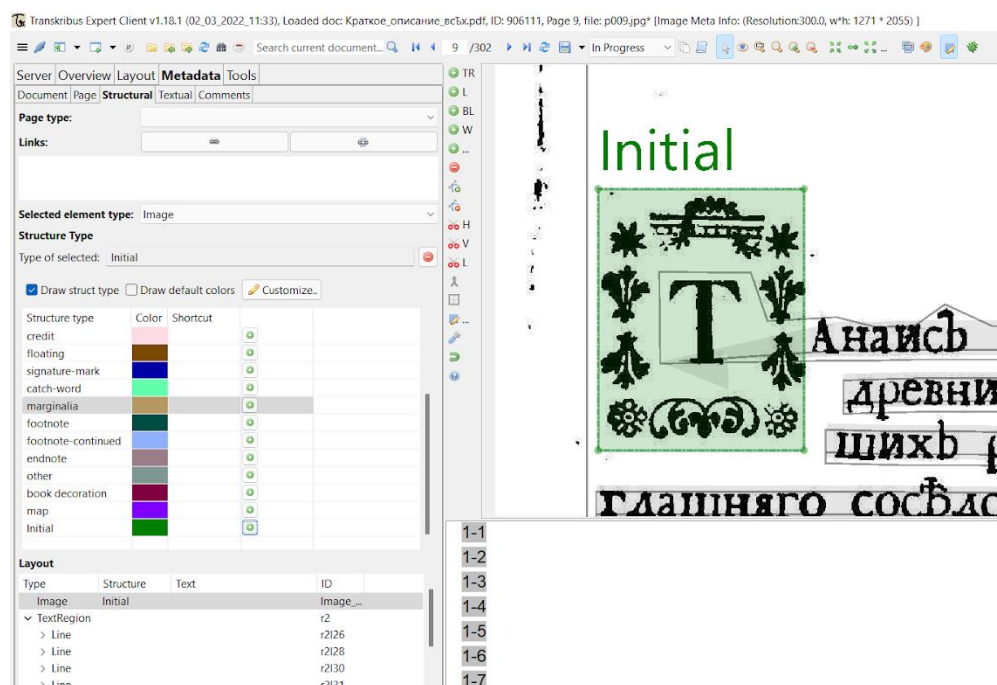
The second screenshot shows the same interface with the document 'Нещастной_француз_или_Жизнь_кавалера_Беликурта_описанная_им_самим.pdf'. The 'Page type' is set to 'Text'. The 'Structure Type' is set to 'header'. The 'Layout' table shows the following structure:

Type	Structure	Text	ID
Separator			Separ.
TextRegion			r1
Line	page-number	4	r12
Line	header	НЕЩАСТНОЙ	line_16
Line		его вшествіе въ сію пространн	r13
Line		человѣческихъ бѣдностей стоило	r14
Line		ши тысячь дивровъ	r15

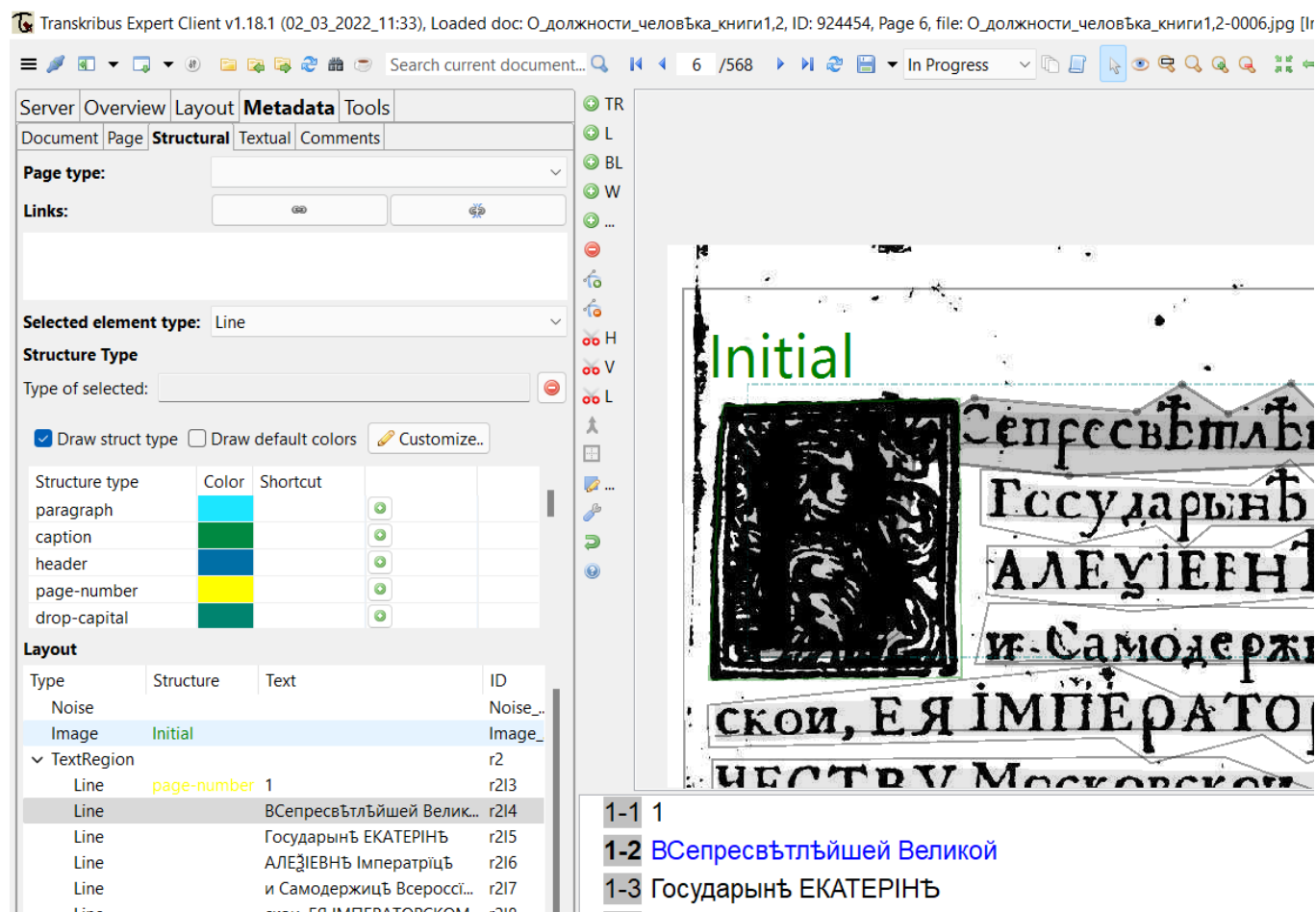
6. Тэг Initial. Совсем не понимаю, как его наносить. Пока есть два варианта.

Первый – выделяем буквицу вместе с первой строкой, в качестве линии. Поверх исключительно на буквицу наносим регион Image и на регион Image наносим тэг Initial. Плюс в том, что так буква не отрывается от последующего слова, минус в том, что буква не всегда хорошо читаема и не всегда удачно вписывается в линию (иногда букву вообще невозможно отличить от картинки).

Пример удачный:



Пример неудачный:



Вариант второй – выделять буквицу отдельной строкой, в нее вписывать одну букву и на строку навешивать тэг Initial. При таком подходе у нас буквица будет отрываться от всего остального слова, зато вероятность распознавания такой буквицы моделью, возможно, будет неплохой.

Transkribus Expert Client v1.18.1 (02_03_2022_11:33), Loaded doc: Краткое_описание_всѣх.pdf, ID: 906111, Page 9, file: p009.jpg* [Image Meta Info: (Resolution:300.0, w*h: 1271

Server Overview Layout **Metadata** Tools

Document Page **Structural** Textual Comments

Page type:

Links:

Selected element type: Line

Structure Type

Type of selected: Initial

☒ Draw struct type ☐ Draw default colors

Structure type	Color	Shortcut
endnote		
other		
book decoration		
map		
Initial		

Layout

Type	Structure	Text	ID
TextRegion			r2
> Line	Initial	T	line_16.
> Line			r2126
> Line			r2128
> Line			r2130
> Line			r2131
> Line			r2132
> Line			r2133
> Line			r2134

Initial

Анаисѣ или
древнихъ
шихъ рѣкъ
гдашняго сосѣдства

1-1 T
1-2
1-3
1-4