Katharina Ost

Transkriptionsrichtlinien „Digitalisierung und Volltexterkennung der ehemals Reichenauer Inkunabeln“

2024

Empfohlene Zitierung:

Ost, Katharina: Transkriptionsrichtlinien „Digitalisierung und Volltexterkennung der ehemals Reichenauer Inkunabeln“, RegionaliaOpen 2024, https://doi.org/10.57962/regionalia-22875.

Transkriptionsrichtlinien „Digitalisierung und Volltexterkennung der ehemals Reichenauer Inkunabeln“

Stand: 31.03.2024.

Im Rahmen des Projektes „Digitalisierung und Volltexterkennung der ehemals Reichenauer Inkunabeln“ digitalisierte die Badische Landesbibliothek die 243 Titel umfassende Inkunabelsammlung aus der ehemaligen Bibliothek des Klosters Reichenau und erschloss diese im mit Hilfe des Texterkennungssystems Transkribus. Die Digitalisate und Volltexte sind über die Digitalen Sammlungen der Badischen Landesbibliothek verfügbar (https://digital.blb-karlsruhe.de/topic/view/7530707).

Nachfolgende Transkriptionsrichtlinien wurden innerhalb des Projektes für die computergestützte Transkription von Inkunabeln und Frühdrucken definiert. Insbesondere liegen sie dem Trainingsmaterial der auf der Transkribus-Plattform veröffentlichten Texterkennungsmodelle

- „Latin Incunabula (Reichenau)“ (Modell-ID 61337),
- „Latin/German Bilingual Incunabula (Reichenau)“ (Modell-ID 61316) und
- „German Incunabula (Reichenau)“ (Modell-ID 61285)

zu Grunde.

Das Projekt wurde von der Stiftung Kulturgut Baden-Württemberg gefördert.

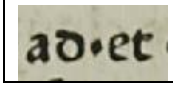
Text

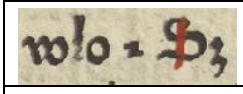
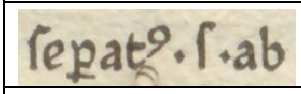
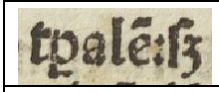
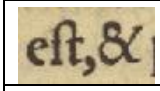
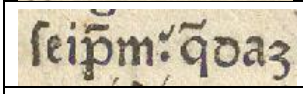
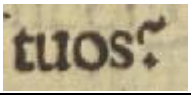
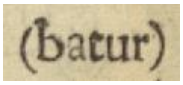
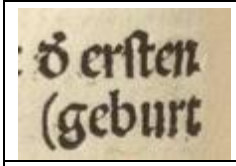
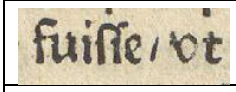
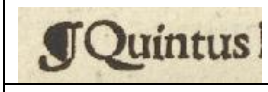
Allgemeine Prinzipien:

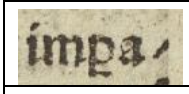
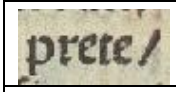
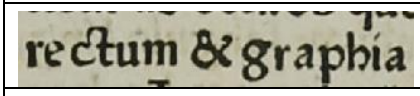
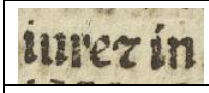
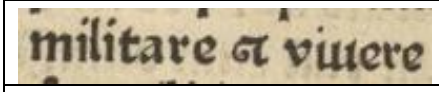
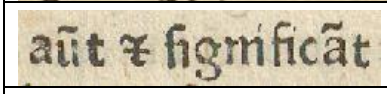
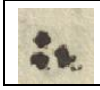
- Glyphen werden sprach- und kontextabhängig transkribiert.
- Abbreviaturen werden nicht ausgeschrieben. Zu ihrer Wiedergabe werden ausschließlich solche Unicode Zeichen verwendet, die in gängigen Schriftarten darstellbar sind.
- Konsonantische Ligaturen werden in der Transkription aufgelöst.
- Satzfehler bleiben i. A. unkorrigiert. Eine Ausnahme bilden kopfüber gesetzte Buchstaben.
- Es werden s/ſ, U/V, u/v, i/j unterschieden. Andere Buchstabenvarianten (z.B. ꝺ/d, ꝛ/r, I/J, verschiedene M-Formen, ...) werden im Sinne einer höheren Les- und Durchsuchbarkeit vereinheitlicht.
- Handschriftliche Eintragungen werden nur dann transkribiert, wenn sie für das Texterkennungsmodell (nach Binarisierung) voraussichtlich nicht als solche erkennbar sein werden.

Interpunktion

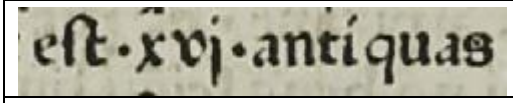
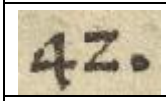
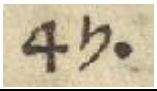
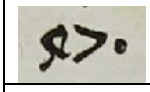
Grundsätzlich werden Satzzeichen an das vorangehende Wort gezogen. Auf sie folgt ein Leerzeichen.

Beispiel	Transkription	Unicode	
	vniuerſalis. Valet	U+002E	FULL STOP
	ad. et	U+002E	FULL STOP

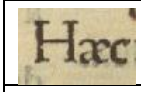
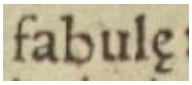
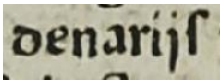
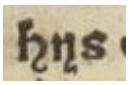
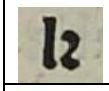
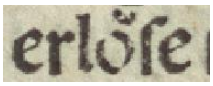
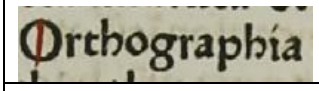
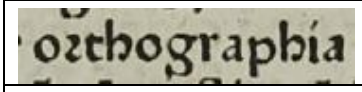
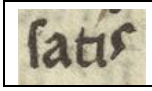
	volo. Sꝫ	U+002E	FULL STOP
Auch Punkte, die leicht oberhalb der Grundlinie stehen, werden mit U+002E transkribiert.			
	ſepat⁹ .ſ. ab		FULL STOP
Bei Abbreviaturen, die konventionell durch beidseitige Punkte markiert sind, wie z.B. für *scilicet* und *enim*, werden beide Punkte – von der Grundregel abweichend – an die Abbreviatur gezogen.			
	tꝑalē: ſꝫ	U+003A	COLON
	eſt, &	U+002C	COMMA
	ſuccendi; Et	U+003B	SEMICOLON
	ſeiꝑm; q̄daꝫ	U+003B	SEMICOLON
Der punctus elevatus wäre nur über die Private Use Area abbildbar (MUFI: F161, PUA-8), daher wird mit Semikolon transkribiert.			
	tuos?	U+003F	QUESTION MARK
	(batur)	U+0028 U+0029	LEFT PARENTHESIS RIGHT PARENTHESIS
	d erſten (geburt	U+0028	LEFT PARENTHESIS
Auch Klammern, die den Text-„Übertrag“ einer angrenzenden Zeile markieren, werden transkribiert.			
	detur/ et	U+002F	SOLIDUS
	fuiſſe/ vt	U+002F	SOLIDUS
	¶ Incipit	U+00B6	PILCROW SIGN
	¶ A quo	U+00B6	PILCROW SIGN
	¶ Quintus	U+00B6	PILCROW SIGN
	¶ De	U+00B6	PILCROW SIGN
	¶ Quero	U+00B6	PILCROW SIGN
⁋ (U+2E3F) wird nicht verwendet. Handschriftliche Paragraphenzeichen werden mittranskribiert, da die Texterkennungsmodelle sie von gedruckten nur schlecht unterscheiden können.			

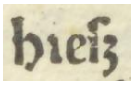
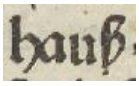
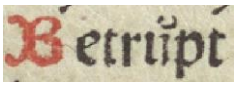
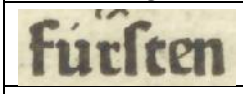
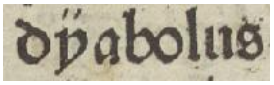
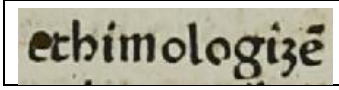
	☞ Sigmundus	U+261E	WHITE RIGHT POINTING INDEX
	ha¬	U+00AC	NOT SIGN
	Au¬	U+00AC	NOT SIGN
	impera¬	U+00AC	NOT SIGN
	prete¬	U+00AC	NOT SIGN
U+00AC wird nur verwendet, wo auch ein grafisches Trennzeichen vorhanden ist.			
	rectum & graphia	U+0026	AMPERSAND
	iure ⁊ in	U+204A	TIRONIAN SIGN ET
	militare ⁊ viuere	U+204A	TIRONIAN SIGN ET
	aūt ⁊ ſignificāt	U+204A	TIRONIAN SIGN ET
			
Q; vnus deus eſt. ij.			
Auch größere Abstände werden als einzelnes Leerzeichen abgebildet.			
	⁂	U+2042	ASTERISM

Zahlen

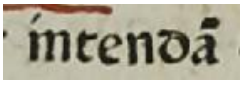
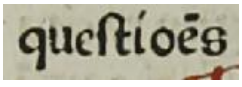
Beispiel	Transkription
	eſt. xvj. antiquas
	.iiii.
Nur bei isoliert stehenden Zahlen werden flankierende Punkte beidseitig an das Zahlzeichen gezogen.	
	42.
	45.
	47.
Arabische Zahlen werden unabhängig von der jeweiligen Form als moderne arabische Zahlzeichen wiedergegeben.	

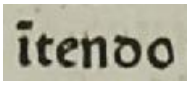
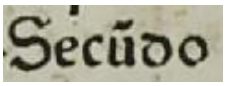
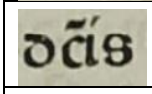
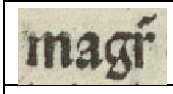
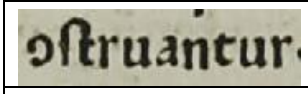
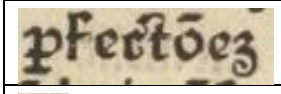
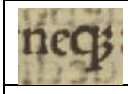
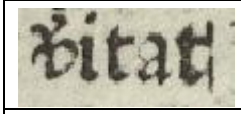
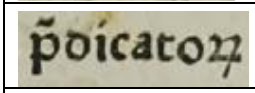
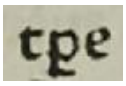
Buchstaben

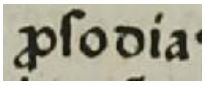
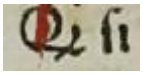
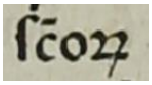
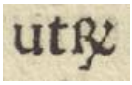
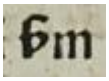
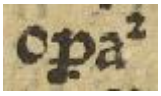
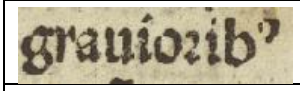
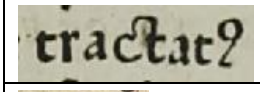
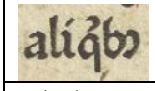
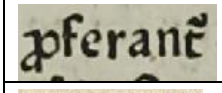
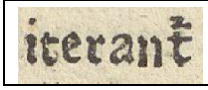
Beispiel	Transkription	Unicode	
	Hæc	U+00E6	LATIN SMALL LETTER AE
	Ęſchynes.	U+0118	LATIN CAPITAL LETTER E WITH OGONEK
	fabulę	U+0119	LATIN SMALL LETTER E WITH OGONEK
	Cępi	U+0119	LATIN SMALL LETTER E WITH OGONEK
	Iuris	U+0049	LATIN CAPITAL LETTER I
U+004A (LATIN CAPITAL LETTER I) wird nicht verwendet.			
	denarijſ	U+0069 U+006A	LATIN SMALL LETTER I LATIN SMALL LETTER J
Zwischen i und j wird aber differenziert.			
	hijs	U+0069 U+006A	LATIN SMALL LETTER I LATIN SMALL LETTER J
Die Ligatur wird hier aufgelöst. Steht allerdings die Glyphe y für ii/ij, wird mit U+0079 (LATIN SMALL LETTER Y) transkribiert.			
	k	U+006B	LATIN SMALL LETTER K
Das typografisch aus l und 2 zusammengesetzte k wird als einzelner Buchstabe transkribiert.			
	dō̊nent	U+0364	COMBINING LATIN SMALL LETTER E
	erlō̊ſe	U+0364	COMBINING LATIN SMALL LETTER E
	Fœcerat	U+0153	LATIN SMALL LIGATURE OE
	Orthographia	U+0072	LATIN SMALL LETTER R
	orthographia	U+0072	LATIN SMALL LETTER R
Zwischen geradem r und r rotunda wird nicht unterschieden, einzige Ausnahme ist die Abbreviatur q2.			
	fatiˢ		

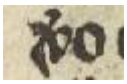
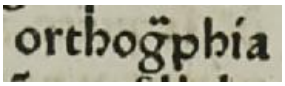
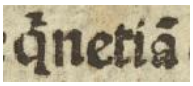
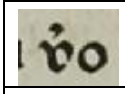
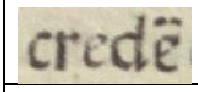
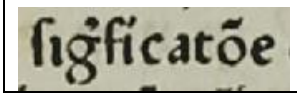
	hieß	U+00DF	LATIN SMALL LETTER SHARP S
	hauß	U+00DF	LATIN SMALL LETTER SHARP S
Hier greift die sprachabhängige Transkription.			
	Ultimo	U+0055	LATIN CAPITAL LETTER U
	Vel	U+0056	LATIN CAPITAL LETTER V
Bei runder Buchstabenform wird U, bei spitzer Buchstabenform wird V transkribiert.			
	růffen	U+0075 & U+0364	COMBINING LATIN SMALL LETTER E
	Betrůpt	U+0075 & U+0364	COMBINING LATIN SMALL LETTER E
	fündē	U+00FC	LATIN SMALL LETTER U WITH DIAERESIS
	üppigkait	U+00FC	LATIN SMALL LETTER U WITH DIAERESIS
Sind zwei getrennte Kreissegmente erkennbar, wird ü transkribiert.			
	fürſten	U+00FC	
	nüt	U+00FC	
	zů	U+0075 & U+0366	COMBINING LATIN SMALL LETTER O
	dÿabolus	U+00FF	LATIN SMALL LETTER Y WITH DIAERESIS
	ethimologizē	U+007A	

Abbreviaturen

Beispiel	Transkription	Unicode	
	intendā	U+0101	LATIN SMALL LETTER A WITH MACRON
	queſtioēs	U+0113	LATIN SMALL LETTER E WITH MACRON

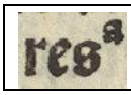
	ītendo	U+012B	LATIN SMALL LETTER I WITH MACRON
	cōſtructō	U+014D	LATIN SMALL LETTER O WITH MACRON
	Secūdo	U+016B	LATIN SMALL LETTER U WITH MACRON
Die Transkription unterscheidet nicht, ob die Linie gerade oder leicht geschwungen ist.			
	beniamȳ	U+0233	LATIN SMALL LETTER Y WITH MACRON
	om̄e	U+0304	COMBINING MACRON
	dc̄is	U+0304	COMBINING MACRON
	mgr̄	U+0304	COMBINING MACRON
Abkürzungszeichen, die nicht mit Vokalbuchstaben kombiniert sind, werden durch U+0304 wiedergegeben.			
	9ſtruantur	U+A76F	LATIN SMALL LETTER CON
	placꝫ	U+A76B	LATIN SMALL LETTER ET
	ꝑfectōeꝫ	U+A76B	LATIN SMALL LETTER ET
	ſꝫ	U+A76B	LATIN SMALL LETTER ET
	neqꝫ	U+A76B	LATIN SMALL LETTER ET
Auch wenn der visuelle Eindruck eher ʒ (U+0292, LATIN SMALL LETTER EZH) entspricht, wird dem Sinn entsprechend ꝫ gesetzt.			
	deitatſ	U+A76D	LATIN SMALL LETTER IS
	ỿitatſ	U+A76D	LATIN SMALL LETTER IS
	p̄dicatoꝛ	U+0070 & U+0304	COMBINING MACRON
	p̨diolum	U+0070 & U+0304 & U+0328	COMBINING MACRON & COMBINING OGONEK
	tꝑe	U+A751	LATIN SMALL LETTER P WITH STROKE THROUGH DESCENDER

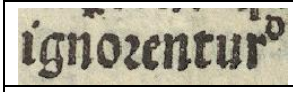
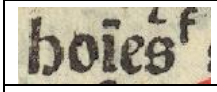
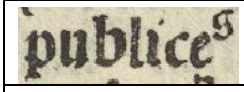
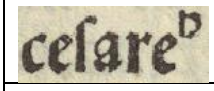
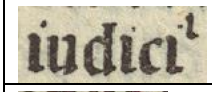
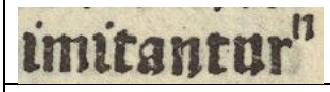
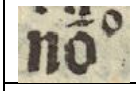
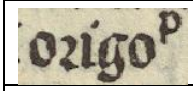
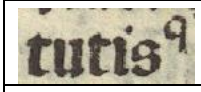
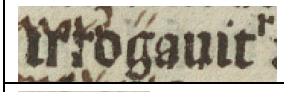
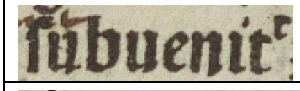
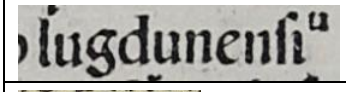
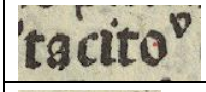
	ꝓſodia	U+A753	LATIN SMALL LETTER P WITH FLOURISH
	Ꝙ ſi	U+A758	LATIN CAPITAL LETTER Q WITH DIAGONAL STROKE
	ꝙ	U+A759	LATIN SMALL LETTER Q WITH DIAGONAL STROKE
	ꝗd	U+A757	LATIN SMALL LETTER Q WITH STROKE THROUGH DESCENDER
	qꝛ	U+A75B	LATIN SMALL LETTER R ROTUNDA
	Ꝛº	U+211F	RESPONSE
	ſc̄oꝛꝝ	U+A75D	LATIN SMALL LETTER RUM ROTUNDA
	utꝛꝝ	U+A75D	LATIN SMALL LETTER RUM ROTUNDA
	ẜm	U+1E9C	LATIN SMALL LETTER LONG S WITH DIAGONAL STROKE
	opaꝛ	U+02B3	MODIFIER LETTER SMALL R
Für die hochgestellte r rotunda als Abkürzung von *tur* wird U+02B3 (und nicht wie bei *ur* U+0309) verwendet.			
	grauiorib⁹	U+A770	MODIFIER LETTER US
	tractat⁹	U+A770	MODIFIER LETTER US
	aliꝗb⁹		
Ob die -us Abbreviatur auf der Grundlinie oder hochgestellt steht, wird nicht unterschieden.			
	ꝓferan̓t	U+0309	COMBINING HOOK ABOVE
	iteran̓t	U+0309	COMBINING HOOK ABOVE

Da U+1DD1 (COMBINING UR ABOVE) in Standardschriftarten nicht umgesetzt ist, wird stattdessen U+0309 verwendet.			
	ỷo	U+A75F	LATIN SMALL LETTER V WITH DIAGONAL STROKE
Kreuzt das Abkürzungszeichen das v, wird U+A75F verwendet, ansonsten wird mit U+0313 kombiniert (s.u.).			
	orthog̓phia	U+0363	COMBINING LATIN SMALL LETTER A
Da U+1DD3 (COMBINING LATIN SMALL LETTER FLATTENED OPEN A ABOVE) in Standardschriftarten nicht umgesetzt ist, wird stattdessen U+0363 verwendet.			
	ↄc͛	U+0363	COMBINING LATIN SMALL LETTER A
In der Transkription der Abkürzung für *et cetera* wird das hochgestellte a stets dem c zugeordnet.			
	q̇netiā	U+0365	COMBINING LATIN SMALL LETTER I
	q̇buſcūqȝ	U+0365	COMBINING LATIN SMALL LETTER I
	antiq̊ꝙ	U+0366	COMBINING LATIN SMALL LETTER O
	v̇o	U+0313	COMBINING COMMA ABOVE
	int̓	U+0313	COMBINING COMMA ABOVE
Die Verwendung von U+0313 hängt nicht davon ab, was der Haken abkürzt (*er*, *re*, *i*, ...).			
	credė	U+0313	COMBINING COMMA ABOVE
Auch hier wird dem Sinn folgend U+0313 (fü *re*) gesetzt.			
	vocal	U+0313	COMBINING COMMA ABOVE
	dr	U+0313	COMBINING COMMA ABOVE
	ſig̓ficatōe	U+0313	COMBINING COMMA ABOVE

Hochgestellte Buchstaben

Für hoch- und nachgestellte Buchstaben, wie sie z.B. für Verweise auf Kommentarstellen verwendet werden, wird das nachfolgende Alphabet verwendet:

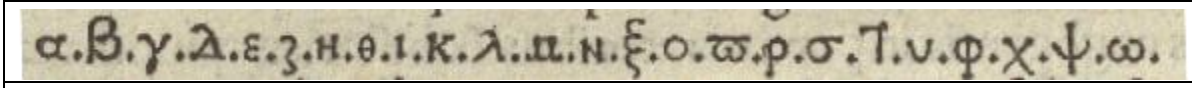
	res^{a}	U+00AA	FEMININE ORDINAL INDICATOR

	actionesᵇ	U+1D47	MODIFIER LETTER SMALL B
	ꝑſoneᶜ	U+1D9C	MODIFIER LETTER SMALL C
	ignorenturᵈ	U+1D48	MODIFIER LETTER SMALL D
	Ummaᵉ	U+1D49	MODIFIER LETTER SMALL E
	hoĩesᶠ	U+1DA0	MODIFIER LETTER SMALL F
	publiceᵍ	U+1D4D	MODIFIER LETTER SMALL G
	ceſareʰ	U+02B0	MODIFIER LETTER SMALL H
	iudiciⁱ	U+2071	SUPERSCRIPT LATIN SMALL LETTER I
	¶ Exᵏ	U+1D4F	MODIFIER LETTER SMALL K
	diuturniˡ	U+02E1	MODIFIER LETTER SMALL L
	ꝙſenſuᵐ	U+1D50	MODIFIER LETTER SMALL M
	imitanturⁿ	U+207F	SUPERSCRIPT LATIN SMALL LETTER N
	nōº	U+00BA	MASCULINE ORDINAL INDICATOR
	origoᵖ	U+1D56	MODIFIER LETTER SMALL P
	tutisᑫ	U+0071	LATIN SMALL LETTER Q
Für das hoch- und nachgestellte q gibt es leider kein passendes Unicode-Zeichen.			
	irrogauitʳ	U+02B3	MODIFIER LETTER SMALL R
	ſineˢ	U+02E2	MODIFIER LETTER SMALL S
	ſubuenitᵗ	U+1D57	MODIFIER LETTER SMALL T
	lugdunenſiᵘ	U+1D58	MODIFIER LETTER SMALL U
	tacitoᵛ	U+1D5B	MODIFIER LETTER SMALL V
	aliaˣ	U+02E3	MODIFIER LETTER SMALL X

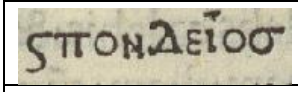
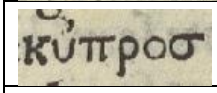
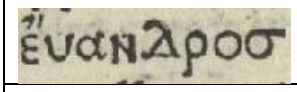
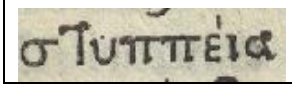
	¶ Om̃e^{y}	U+02B8	MODIFIER LETTER SMALL Y
	ꝑſonas^{z}	U+1DBB	MODIFIER LETTER SMALL Z

Griechische Buchstaben

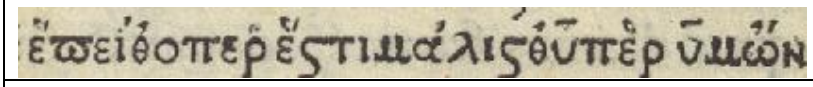
Bei griechischen Buchstaben wird zwischen π/ϖ und σ/ς differenziert, die übrigen Buchstabenformen werden vereinheitlicht wiedergegeben.


α. β. γ. δ. ε. ζ. η. θ. ι. κ. λ. μ. ν. ξ. ο. ϖ. ρ. σ. τ. υ. φ. χ. ψ. ω.

Akzente und Spiritus werden so abgebildet, wie sie gesetzt sind – unabhängig davon, ob dies korrekt ist oder modernen Gepflogenheiten entspricht.

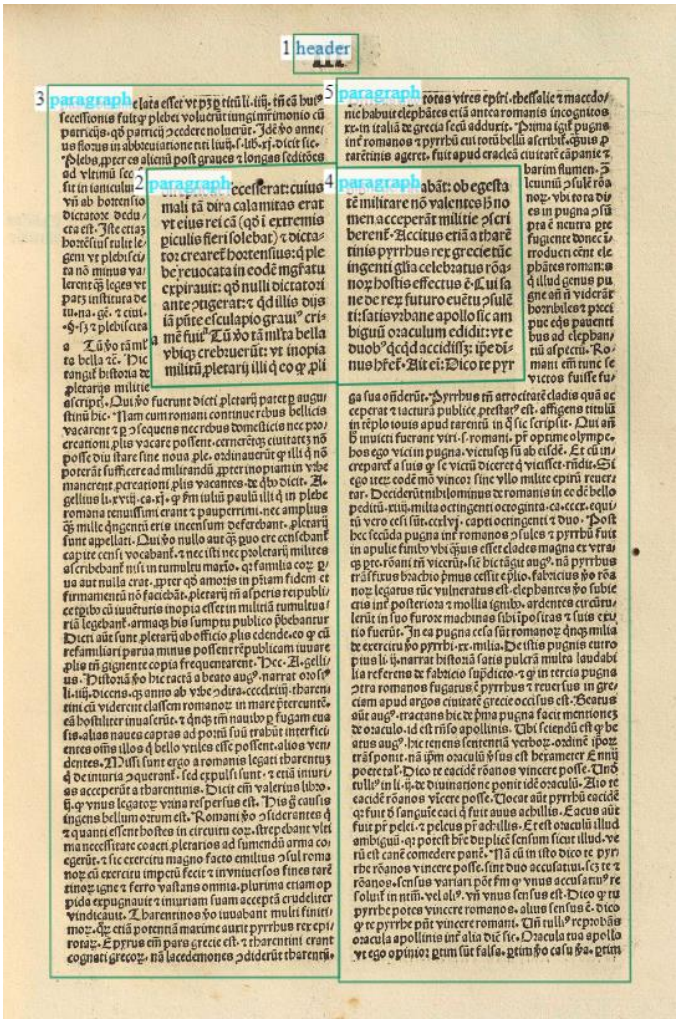
	ςπονδεῖοσ
	κύπροσ
	ἔυανδροσ
	στυππέια

Die Worttrennung ist so zu setzen, dass ein sinnvoller Text approximiert wird.


ἔωειθ᾽ οπερ ἔςτι μάλιςθ᾽ ῡπὲρ ῡμῶν

Layout

Das Seitenlayout wird nur insoweit annotiert, wie dies für die eigentliche Texterkennung unabdingbar ist. Abbildungen und Schmuckelemente werden nicht kenntlich gemacht.



Textregionen

- Text wird möglichst großflächig in Textregionen zusammengefasst, eine Unterscheidung nach typografischen Paragraphen erfolgt nicht. Bei mehrspaltigen Layouts erhält jede Textspalte korrekte eine eigene Textregion, um die Lesereihenfolge der zugehörigen Zeilen sicherzustellen.
- Für Textregionen wird nur zwischen den Strukturtypen (type Attribut in Page XML) *paragraph* und *header* unterschieden. Diese Unterscheidung wurde vorgenommen, um eine verbesserte Anzeige der Transkripte zweispaltiger Seitenlayouts vorzubereiten.
- Die Lesereihenfolge wird im Allgemeinen so gesetzt, dass Elemente auf derselben Höhe von links nach rechts geordnet sind (das entspricht dem „smartseder sorting“ in Transkribus). Eine Ausnahme bilden Klammerglossen und vergleichbare Seitenlayouts: Hier steht erst der Haupttext, dann der zugehörige Kommentar.
- Gedruckte Marginalien können je nach Umfang als eigene Textregionen abgebildet werden (eignet sich v.a. für mehrzeilige Marginalien) oder als eigene Zeile in das Transkript des Textblockes eingeschoben werden.
- Bei der Wiedergabe von Diagramminhalten über Textregionen wird auf eine möglichst sinnvolle Lesereihenfolge geachtet.
- Tabellen werden je nach Inhalt und Aufwand werden Tabellen entweder bei der Transkription ignoriert oder als einzelne Textregion mit zeilenweise gebündelten Inhalten wiedergegeben.

Zeilen

- Schmuckinitialen und Platzhalter für handschriftliche Initialen werden ignoriert.
- Die Erfassung von Seitenzahlen und Lagensignaturen ist optional.
- Erstrecken sich ganze Wörter über zwei Zeilen (z.B. Lemmata in Wörterbüchern), sind sie in eine separate, in der Lesereihenfolge vorangeordnete Zeile auszugliedern.