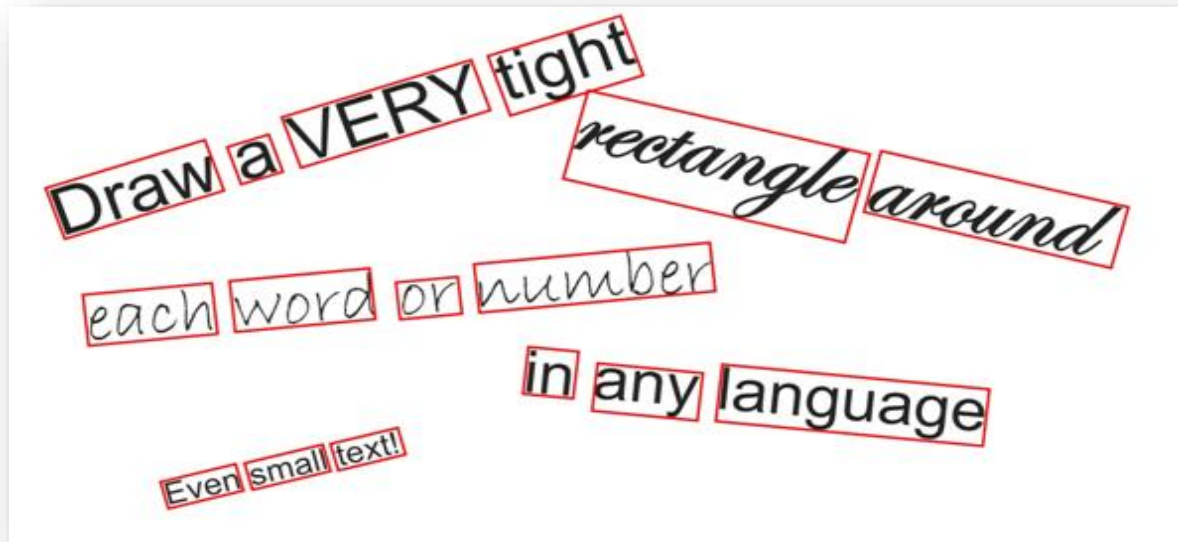


Ocean - Optical Character Recognition (OCR)

Guidelines

What is OCR

Amazon Textract uses Optical Character Recognition (OCR) technology to automatically detect printed text and numbers in a scan or rendering of a document, such as a legal document or a scan of a book.



Task Detail

In this task, you would use the Appen Annotation Tool to draw a VERY tight bounding box around each wordline, word and number set as well as provide transcription.

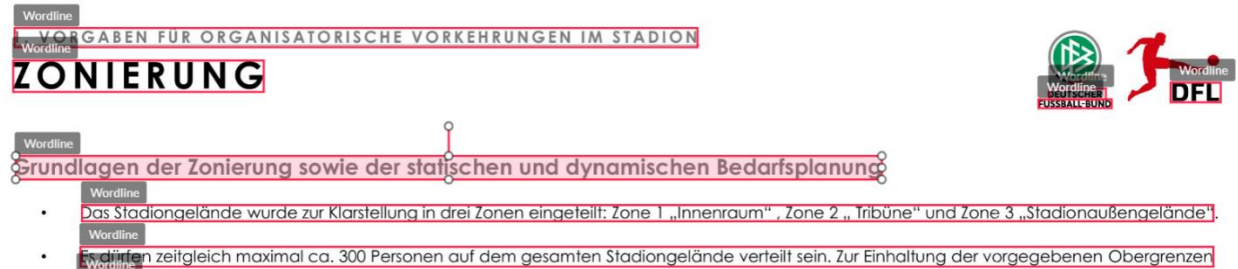
Appen Annotation Tool Tutorial

You can access the tutorial from Academy tab from your Appen Connect Homepage or click https://connect.appen.com/grp/core/vendors/academy_home/view/98

Wordline Guidelines

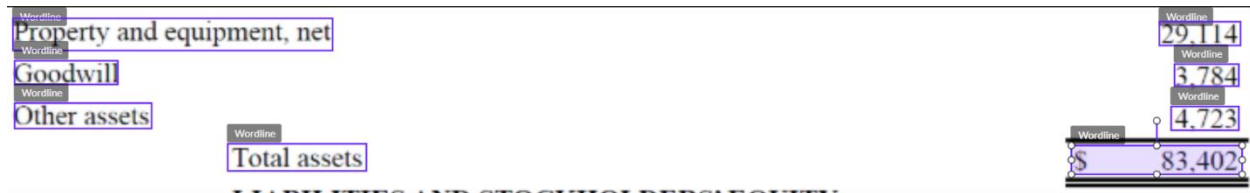
For Wordlines, you will be drawing bounding boxes around each line that there is a horizontal word written on, we place a full box around all words in that line (including if the words are contained within a symbol)

Example



If two sections of the texts are very far apart from each other, please see them as two separate wordlines

Example



Other FAQs

Q1. If page has 2 columns of text, will wordline be across both columns if the text from both columns completely align? Or will they be considered separate wordlines?

Response: They should have separate wordlines (see annotation below)

UNDER CAPITALISM, "the majority of people spoil their lives by an unhealthy and exaggerated attention—are forced, indeed, so to spoil them"

altruistic virtues have really prevented the carrying out of this aim. Just as the worst slave-owners were those who were kind their slaves, so prevented the horror of the system being realised by those who suffered from it, and understood by those who contemplated it, so, in the present state of things in England, the people who do most harm are the people who try to do most good; and at last we have had the spectacle of men who have really studied the problem and know the life—educated men who live in the East End—coming forward and imploring the community to restrain its altru-

are Governments armed with economic power as they are now with political power; if, in a word, we are to have at many people who, having no private property of their own, and being always on the brink of sheer starvation, work of beasts of burden, to do work that is quite uncongenial to tr civilisation, or culture, or refinement in pleasures, or joy of life. From their collective force Humanity gains much in material prosperity. But it is only the material result that it gains, and the man who is poor is in himself absolutely of no importance. He is merely the infinitesimal atom of a force that, so

Annotation and Transcription Guidelines

Be as accurate as possible with the bounding boxes. Do NOT cut text strings off, and do NOT leave too much extra space around any text.

Separate Box Logic

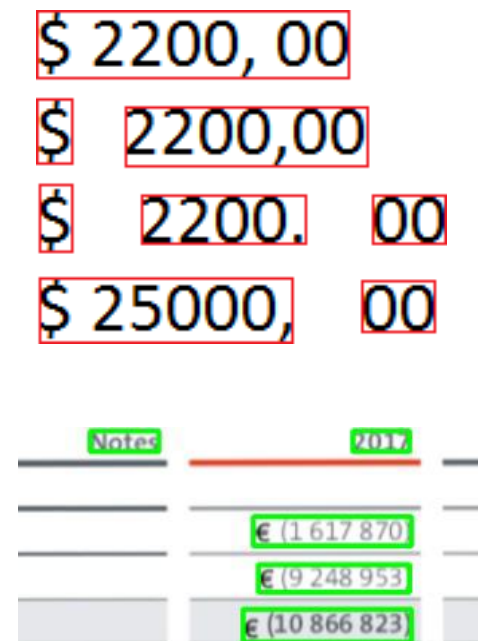
- ❖ For Words : If identify **single** space between words.

Example



- ❖ For numbers/currency: If identify more than 1 **space (> or =2)** between numbers/symbols

Example



- A word is one or more ISO basic Latin script contiguous characters that are on the same line but not separated by spaces or vertical lines.

Good examples	Wrong examples
----------------------	-----------------------

47%	47%
\$25.0	\$25.0

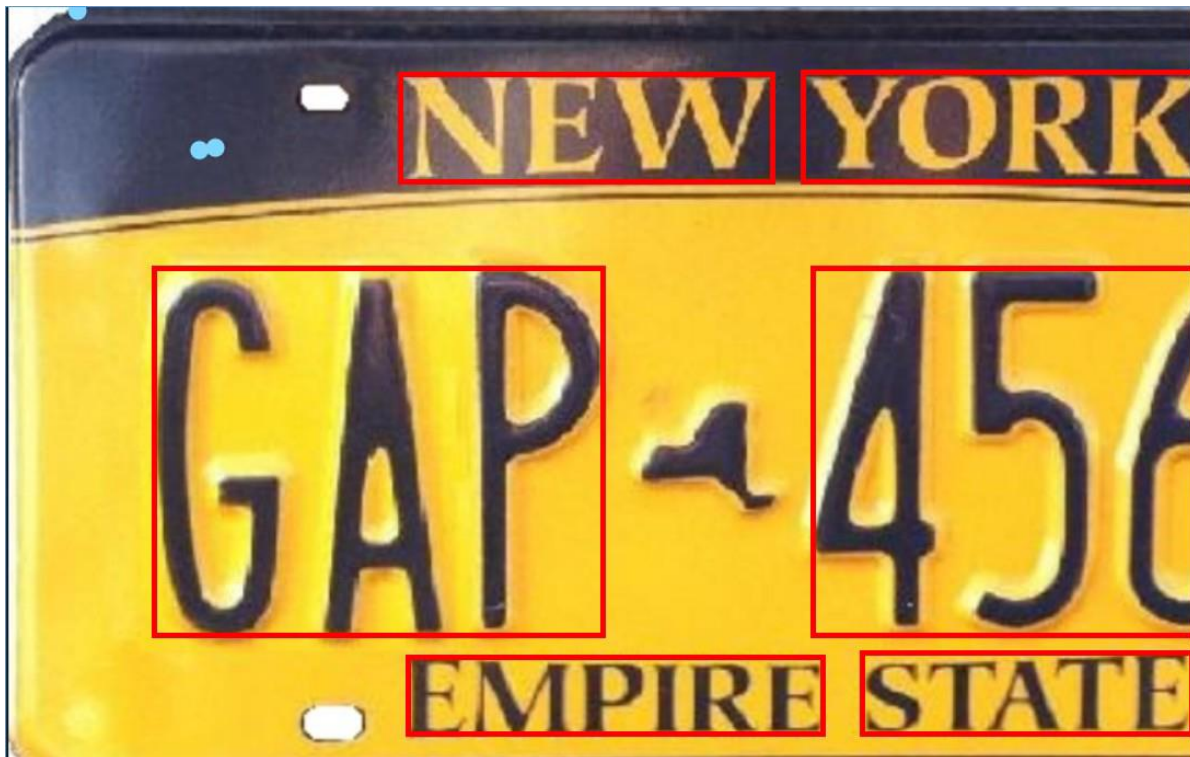
- Draw a VERY tight box around each word and transcribe texts as shown. For example, for the following image, they should be transcribed as “YOUR”, “TEXT”, and “HERE”.



- Draw a bounding box per text string even when there could exist overlapping between boxes.





- Draw tight boxes even with the text strings are cut and include any characters that are legible



- Draw box on pipeline “|”

Examples

	
<p>We can annotate this type of Pipeline (same font)</p>	<p>We do NOT annotate this type of Pipeline (appears to be different font or separator)</p>

➤ Annotate proper transcriptions of UNICODE

- Exception - Check boxes and Radio Button do NOT need to be annotated.

Example: ◆ ▲ → • ●

➤ Draw single bounding box for the power digit if no space available between two numbers (highlighted in yellow)

Penalty on early withdrawal of savings	30	\$ 400	00
Alimony paid to Recipient's SSN ▶ 154 : 32 : 451	31a	\$ 120	00
IRA deduction (see page 31)	32	\$ 150	00
Student loan interest deduction (see page 33)	33	\$ 140	00
Jury duty pay you gave to your employer	34	\$ 200	00
Domestic production activities deduction Attach Form 8903 ²⁵	35	\$ 250	00
Add lines 23 through 31a and 32 through 35			36
Subtract line 36 from line 22. This is your adjusted gross income ▶			37

➤ Do **NOT** draw bounding box on below highlighted dots which looks like a bullet under OCR annotation.

Image Bounding Box Task

✓	Q10	10000	.	5%	10%	.	✓	Q10	15000	.	5%	10%	.
✓	Q11	10000	.	5%	10%	.	✓	Q11	15000	.	5%	10%	.
✓	Q12	10000	.	5%	10%	.	✓	Q12	15000	.	5%	10%	.
✓	Q13	10000	.	5%	10%	.	✓	Q13	15000	.	5%	10%	.
✓	Q14	10000	.	5%	10%	.	✓	Q14	15000	.	5%	10%	.
✓	Q15	10000	.	5%	10%	.	✓	Q15	15000	.	5%	10%	.
✓	Q16	10000	.	5%	10%	.	✓	Q16	15000	.	5%	10%	.
✓	Q17	10000	.	5%	10%	.	✓	Q17	15000	.	5%	10%	.
✓	Q18	10000	.	5%	10%	.	✓	Q18	15000	.	5%	10%	.
✓	Q19	10000	.	5%	10%	.	✓	Q19	15000	.	5%	10%	.
✓	Q20	10000	.	5%	10%	.	✓	Q20	15000	.	5%	10%	.
✓	Q21	10000	.	5%	10%	.	✓	Q21	15000	.	5%	10%	.
✓	Q22	10000	.	5%	10%	.	✓	Q22	15000	.	5%	10%	.
✓	Q23	10000	.	5%	10%	.	✓	Q23	15000	.	5%	10%	.
✓	Q24	10000	.	5%	10%	.	✓	Q24	15000	.	5%	10%	.
✓	Q25	10000	.	5%	10%	.	✓	Q25	15000	.	5%	10%	.
✓	Q26	10000	.	5%	10%	.	✓	Q26	15000	.	5%	10%	.
✓	Q27	10000	.	5%	10%	.	✓	Q27	15000	.	5%	10%	.
✓	Q28	10000	.	5%	10%	.	✓	Q28	15000	.	5%	10%	.
✓	Q29	10000	.	5%	10%	.	✓	Q29	15000	.	5%	10%	.
✓	Q30	10000	.	5%	10%	.	✓	Q30	15000	.	5%	10%	.

- Draw separate bounding box if alphanumeric character visible in Date. Check the below highlighted image for good or bad annotation. Left example is correct - highlight it as the example

DUE DATE OF INSTALLMENTS		
April 15, 20 17	June 15, 20 15	Sept. 15, 20 13

Correct Annotation Incorrect Annotation

- Any mathematical equations should be annotated with a box.

$$e^{i\pi} + 1 = 0$$

- Annotate the Box if word character is available by Curve, Horizontal, or Vertical
 - We will annotate the horizontal, vertical, curved letters, words, digits, and sign in OCR



- In this example, the annotation should be done as a tight quadrilateral (rotated rectangles)



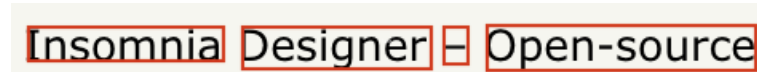
For following vertical text, you draw one **separate** line for each letter – V, E, R, T, I, C, A, L.

V
E
R
T
I
C
A
L

Instruction for special character annotation (ex- *, /, #, - etc)

- For '-' & '—' cases

Will be including the '-' & '—' symbols except for the cases where the above symbols does NOT have any gap from the subsequent contents. Leave vertical space above and below the dash according to the font height of the neighboring crops. (disregard any annotation in this doc that don't follow this requirements, the boxes below are the correct one) For example, in the below image, we have included the '—' as it have proper space from the contents; on the other hand, '-' have been excluded in the second case where it is present alongside the contents without any spaces /gaps.



- For cases where "*" is used besides text

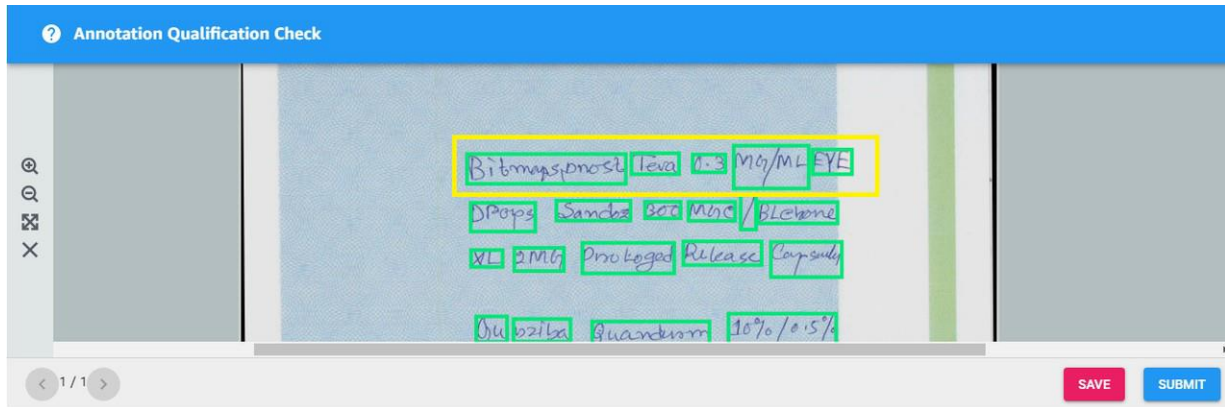



Note: Do NOT annotate if they are used as delimiters like *****

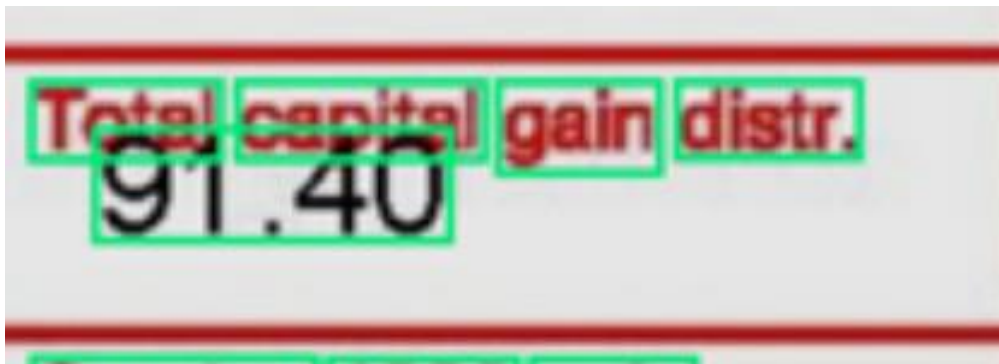
Dataset wise Update

<p>2a Total capital gain distr.</p> <p>\$ 91.40</p>	<p>2b Unrecap. Sec. 1250 gain</p> <p>\$ 0.00</p>
---	--

- In case a word/letter goes to the others portion of the pages, then we should annotate the same.



- Annotate the overlap text if it is readable, skip only those part which is not readable.



- Annotate all the characters which are readable



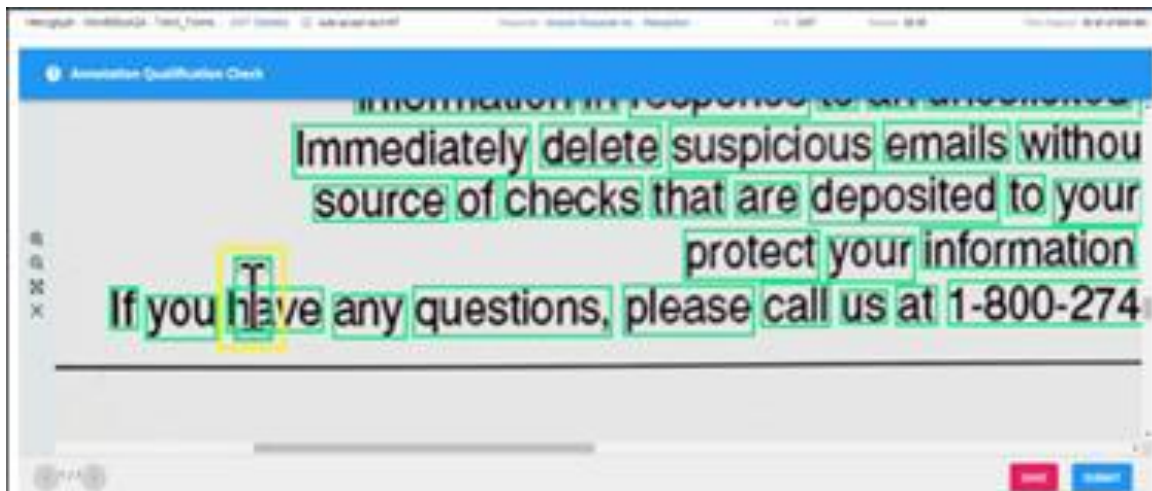
- Annotate the watermarks

DO

NOT

COPY

- Skip the word, which has visibility issue due to cursor or mouse pointer. Do **NOT** draw a bounding box around the cursor (highlighted in Yellow)

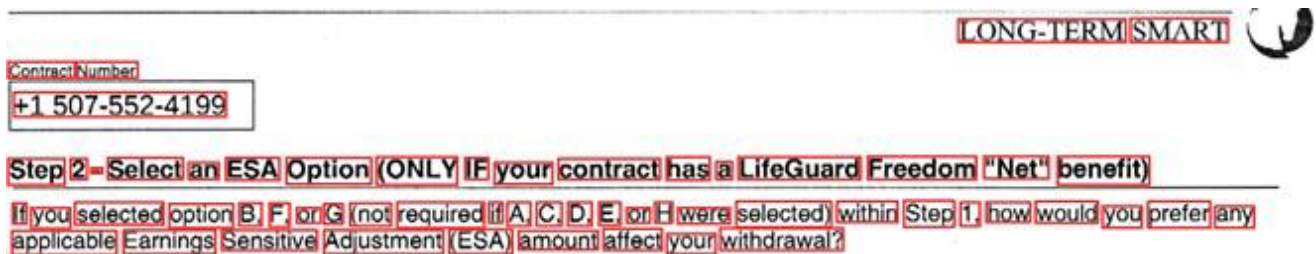


Currency


For all Currency/Symbols the focus will be for Latin Languages.

Samples of Good and Bad Annotations

Sample of Good Annotations



Sample of Bad Annotations Loose Boxes


LONG-TERM SMART 

Contract Number
+1 507-552-4199

Step 2 - Select an ESA Option (ONLY IF your contract has a LifeGuard Freedom "Net" benefit)

If you selected option B, F or G (not required if A, C, D, E or H were selected) within Step 1, how would you prefer any applicable Earnings Sensitive Adjustment (ESA) amount affect your withdrawal?

Missing Boxes


LONG-TERM SMART 

Contract Number
+1 507-552-4199

Step 2 - Select an ESA Option (ONLY IF your contract has a LifeGuard Freedom "Net" benefit)

If you selected option B, F or G (not required if A, C, D, E or H were selected) within Step 1, how would you prefer any applicable Earnings Sensitive Adjustment (ESA) amount affect your withdrawal?

Too small / Tight Boxes

LONG-TERM SMART 

Contract Number
+1 507-552-4199

Step 2 - Select an ESA Option (ONLY IF your contract has a LifeGuard Freedom "Net" benefit)

If you selected option B, F or G (not required if A, C, D, E or H were selected) within Step 1, how would you prefer any applicable Earnings Sensitive Adjustment (ESA) amount affect your withdrawal?

Other FAQs

Q1. What should be done on the highlighted area? Shall we draw the box, or we should ignore it?

Screenshot:



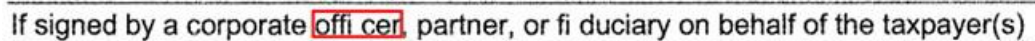
Response

Please ignore.

Justification: it's not a word, it's a logo without easily readable letters.

Q2. How should we draw the box on the red highlighted area? Do we need to consider the word "offi cer" in single box or we should draw a separate box for it as there is space between the word?

Screenshot:



If signed by a corporate **offi cer** partner, or fi duciary on behalf of the taxpayer(s)

Response

Correct annotation: "offi", "cer" because it's a typo and we don't want to correct it. Similar to the word in the same line "fi", "duciary".

Justification: The letters are close to each other (1 space).

Q3. How we should draw the box on the highlighted area? Do we need to draw separate box or we should consider it in a single box as it is a mathematical equation?

Screenshot:

<p>E Percentage (E ÷ D)</p>	<p>G Tangible property credit component previously allowed</p>	<p>H Recaptured tangible property credit component (F × G)</p>
--	---	---

Response

Correct annotations and a flag of “equation” or “difficult”

“(E ÷ D)”

“(F × G)”

Note: Equations are treated as a single annotation, including spaces between numbers and operations.

Q5. How should we draw the boxes over 640.60%?

Screenshot:

<u>6</u> <u>4</u> <u>0</u> . <u>6</u> <u>0</u> %
<u>4</u> <u>1</u> <u>0</u> . <u>4</u> <u>1</u> %
<u>6</u> <u>5</u> <u>1</u> . <u>6</u> <u>5</u> %
<u>4</u> <u>8</u> <u>4</u> . <u>6</u> <u>8</u> %
<u>6</u> <u>7</u> <u>8</u> . <u>6</u> <u>7</u> %
<u>6</u> <u>3</u> <u>3</u> . <u>6</u> <u>3</u> %
<u>4</u> <u>1</u> <u>4</u> . <u>6</u> <u>1</u> %

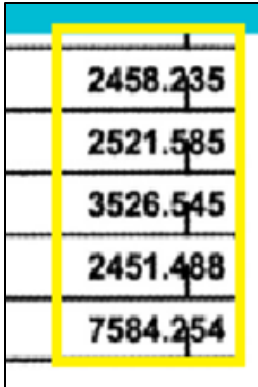
Response

Correct annotation: "640.60%"

Justification: letter proximity and line alignment is very close.

Q6. Shall we consider the highlighted area as in single box or we should draw two different box before and after the partition line?

Screenshot:



2458.235	
2521.585	
3526.545	
2451.488	
7584.254	

Response

Separate word bounding box:

“2458.235”

“2458.585”

Etc

Q7. Should the ‘\$’ sign and the dollar amount be in a single box?

Screenshot:



Response

Correct annotation: “\$3000”

Justification: if the line high of a letter overlaps with the other one 30% or more, then it's considered the same line / word.

Q8. How will we annotate the yellow highlighted area in this below Image?

Screenshot :

Incremental cost (see instructions)	Enter the lesser of column C or 10,000
\$1456465 .00	\$12544 .00
\$2898441 .00	\$9845 .00
\$3456644 .00	\$36245 .00
.....	3 \$6466 .00
.....	4 \$51565 .00
.....	5 \$82265 .00

Response

Correct annotation:

“\$12544”, “.00” because “.00” is far from the rest of the number

Q9. How will we annotate the yellow highlighted area in this below Image?

Screenshot:

7	\$25544 .00
8	\$58745 .00
9	\$54875454 .00

Response

as on 06/06/2019 Separate word bounding box:

“7”, “\$25544”, “.00”

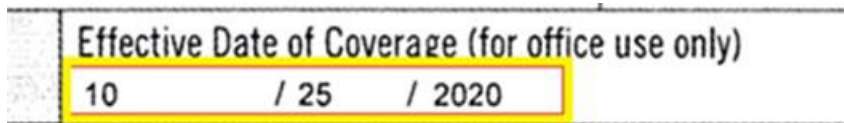
“8”, “\$58745”, “.00”

“9”, “\$54875454”, “.00”

Justification: visually, “.00” seem to be located on a line below.

Q10. How do we annotated the yellow highlighted part? Is it taken under a single box or separate box?

Screenshot:



Response

While the usual guidance to keep such digits together, in this case, we request to keep the separate because they are too far removed:

“10”, “/ 25”, “/2020”

Q11. How can we annotate the below highlighted part? “\$” & “2000” should be taken in a single box or separately?

Screenshot:



Response

Together as the dollar sign line and the number line heights overlap for ~50%: “\$2000”

Q12. How do we annotate the below highlighted part? Is the annotation is correct or not?

Screenshot:



Response

Correct Annotation above

Q13. How do we annotate the yellow highlighted part?

Screenshot:



Response

We will make separate box here as there are maximum space in between. (“ “145” “) “458-5363” “(458” “)” “458-6692”

Q14. How do we annotate the yellow highlighted part?

Screenshot:

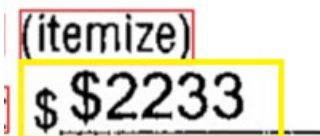


Response

We will make separate box here “10-” “8563252358”

Q15. How do we annotate the yellow highlighted part?

Screenshot:



Response

It will be in separate box “\$” “\$2233”

Q16. What should be done with § and 1102(4).? Bbox should be done separately or all together?

Screenshot:

ect to the best of
 .S. § 1102(4).
 valuation rates and

Response

Separate word BB:

“§”, “1102(4).”

Q17. What should be done in the highlighted area? Altogether in a single box or we need to draw a separate box?

Screenshot:

e taxable year _____ 35%

Response

Separate word BB:

“year”, “35%”

Q18. Do we have to make a single box or we will do a separate box in the highlighted area?

Screenshot:

New Brunswick (NB)		Ontario (ON)	
	10,043.00		10,354.00
+	35.00	+	30.00
+	28.00	+	6.00
+	32.00	+	27.00
	5819		125.00

Response

Separate word BB:

“10,043”, “00”

Justification: visual separation with a delimiter.

Q19. How to annotate the green highlighted area?

Screenshot:

Purchaser Information	Name Nora R. Jenkins		
	Address 2995 Terra Cotta Street Duluth, MN		
	City Portland	State Oregon	Zip Code 30003
	Contact Telephone Number (1 2 4)-5 7 8 -9 9 2 3	E-mail Address NoraRJenkins@dayrep.com	
Seller Information	Name Daniel E. Pahl		Missouri Tax Identification Number 4 5 7 8 9 6 2 1
	Address 3131 Rose Street Westchester, IL		
	City Austin	State Texas	Zip Code 21500
	Contact Telephone Number (4 5 7)-8 9 6 -1 1 2 3	E-mail Address DanielEPahl@rhyta.com	

Response

Correct annotation: “(124)-578-9923”

Q20. How to annotate the yellow highlighted area?

Screenshot:

<p>zero or less, enter -0-. However if Column A report the excess as e 8 are zero you cannot deduct enses. Stop here and attach Form from line 8 in Column B multiply yees subject to Department of of service limits. Multiply mea</p>	8	\$2506	00	
--	---	--------	----	--

Response

Correct Annotation: “ -0- “ Single box

Q21. How to annotate the yellow highlighted area?

Screenshot:

TAXABLE YEAR		California Capital Gain or Loss Adjustment				SSN or ITIN	
2016		Do not complete this schedule if all of your California gains (losses) are the same as your federal gains (losses).				D (540)	
Name(s) as shown on return				SSN or ITIN			
Tracey Maloney				1 0 4-2 1-0 4 7 8			
	(a) Description of property Example: 100 shares of "Z" (\$ stock)	(b) Sales price	(c) Cost or other basis	(d) Loss If (c) is more than (b), subtract (b) from (c)	(e) Gain If (b) is more than (c), subtract (c) from (b)		
1							
a	Land	51,420	42,000	9,420	0		
b	House	148,520	452,100	0	303,580		
c	Flat	31,045	54,000	0	22,955		
d	Shop	785,420	800,000	0	14,580		

Response

Each and every numeric digit will be separate box “1” “0” “4” “-” “2” “1” “-” “0” “4” “7” “8”

Q22. Is the below annotation is correct or not?

Screenshot:



Response

Correct Annotation.

Q23. Can Annotations be made when blank fields are observed



Response

Yes, you can work on pages with blank fields, but please annotate any existing text (even if it's handwritten)

Q24. For transcription, if there seems to be a typo error where a letter is missing a diacritic, should the transcription reflect how the actual text is showing or should transcription show the proper marking?

Response: *Transcribe, as written.*

Q25. Another question about design layout, if the texts are presented intentionally with large space in between, should we consider it as one word or box the letters individually?

Screenshot:



Response: *Single word is preferred*

Q26. How should the art texts be boxed? “GLUTEN” “FREE” or “GLUTEN FREE”?



Response: *Separate bounding boxes for each word - "GLUTEN" "FREE"*

Q27. In a scenario where a printed word seems to have a few letters blurred out for some reason (water damage, for example), should those blurred letters be included in the annotation/transcription (which means transcriber will make an assumption on the word) or should the annotation/transcription only include clearly printed letters?

Response: *If it's legible it should be annotated and if it's not, it shouldn't. When it's in-between, judgement call from the annotators.*