# Accessible (Digital) Archives: Why we don't have them and why AI isn't helping

(sensationalist title explained: AI is helping, but only with certain aspects of digital archives workflows, leaving critical areas of accessibility unchanged and greatly in need of improvement)

Mark Pellegrino Digitization Services Librarian McMaster University October, 2024



### States and the second

#### ORDINIS RATIO.



V NC T O R V M inuulgatis animantium moribus, & naturis, phyfica miracula nomine moulfrortun vulgö nuncupata sin afpectum publicum modò producenda fuccedunt s que aliquando contingere, explorata fatetur veriras , & potifimùm quasdo, non iuxta confuetam natura normam, peregrina quaedam, vel toti generatorum corpori, vel eiuldé partibus (fuè de homine, fuè de brutis, fuè de plantis ver- MonHra at ba fiant) figura, vel character imprimitur. Non enimbae timent ad aauthorum funt deliria, vel fonnoia, qui fucata perfuatione, nimalia, cr

arbitrio fuo, moftra delireare velint quandoquidem de his quotidiana nos certiores plantas . facit autopfia. Ideirco pro comperto affirmandum eff, quòd humana curiofitas vaftiffimam generatorum folitudinem peruagans , interdum in aliquid naturales excedens leges, & nouitatem pariens incidat : ynde tunc inquieto fciendi appetitu exagitata, ob illudignotum, graui admiratione necessario detinetur. Nos igitur in præfentia ad hanc obliterandam admirationem, ad exterminandam feiendr ap- Sciendi ap-H petitionem,& deniq. ad recreadam nouitatum famem, abdita nature atria innume- petitur miris decorata iconibus pandere decreuimus . Sed cum rectum fui,& obliqui index , rabilis; & examen effe perhibeatur, operæ pretium effe videtur, fi prima fronte perfectam hominis conftitutionem paucis complectamut verbis;non folum vt deinceps errara natura cuidentiora legentibus innotescant; verum etiam ne humanis prarogatiuis hæ noftræ hiftoriæ careant . Deus enim in ipfo corporis humani creationis initio . immortalis vitæ fpiraculum, nempe animam infußauit, qua tanquam rationali Duce Monitrainregeretur, fuperna meditaretur, & fenfibus dominaretur. Denique in rebus etiam animatorit. inanimis, admiranda aliquando cucniunt, fuarum naturam caufarum non referentia, veluti varia aquilarum, feu aliorum animantium, vel equitum fimulacra in acre apparentia monftra cocleftia à nonnullis appellata; quorum etiammentionem perpul- Monfira car chris exornatam imaginibus, hifee hiftoriarum monumentis mandare flaturum eft . le itia que.

#### DE

## **Sample Starting Image**

Twelfth page of the *Monstrorum historia* By Ulyssis Aldrovandi, 1642. 1 of 956 pages.

Latin language, hand-written, blackletter font.

Digitized as 600dpi TIFF image, 62.2 MB.

#### Very tough situation for OCR

- Drawn adornments around image.
- Decorative drop cap (starting letter).
- Supporting text in margins.
- Single letters around paragraph (G,H, A, DE).

## **OCR and hOCR (HTML Optical Character Recognition)**

- 1. Recognize the characters in the image.
- 2. Overlay those digital character over the images of those character using a coordinate system called hOCR and embed it in your PDF.
- 3. Predict the font in the image as a TTF and embed it in your PDF.

# tissimum quando, non iuxta consuetam naturæ normam, pe-

<pre><span class="ocr_line" id="line_1_14" title="bbox 1091 2820 2733 2893; baseline 0.001 -22; x_size 72;&lt;/pre&gt;&lt;/th&gt;&lt;/tr&gt;&lt;tr&gt;&lt;td&gt;&lt;pre&gt;x_descenders 19; x_ascenders 17"></span></pre>
<pre><span class="ocrx_word" id="word_1_57" title="bbox 1091 2821 1321 2875; x_wconf 20">tiffimim</span></pre>
<pre><span class="ocrx_word" id="word_1_58" title="bbox 1347 2825 1678 2893; x_wconf 66">quando,</span></pre>
<pre><span class="ocrx_word" id="word_1_59" title="bbox 1593 2816 1668 2897; x_wconf 89">non</span></pre>
<pre><span class="ocrx_word" id="word_1_60" title="bbox 1695 2827 1834 2876; x_wconf 85">iuxta</span></pre>
<pre>span class='ocrx word' id='word 1 61' title='bbox 1851 2820 2144 2875; x wconf 88'&gt;confuetam</pre>
<span class="ocrx_word" id="word_1_62" title="bbox 2160 2836 2354 2874; x_wconf 85">naturæ</span>
<pre><span 2383="" 2628="" 2837="" 2885;="" 89="" class="ocrx_word" ld="word_l_b3" title="bbox" x_wcont="">normam,</span></pre>
<pre><span class="ocrx_word" id="word_1_64" title="bbox 2639 2839 2733 2892; x_wconf 90">pe-</span></pre>

## Accessibility Tools: Adobe Acrobat

Why PDF? It's a versatile, universally supported way to encapsulate and disseminate information.

#### Elements of an accessible PDF:

- Tagging PDF elements employ html tagging just like a website, with opportunities to support modern HTML5 semantic tagging.
- Reading Order tagged elements must be assigned to numerical order for assistive devices.
- Text overlay the invisible textual element on top of the source image.
- Alt-text for describing images.
- Embedded Metadata (title, author, subject, etc).

If you're interested, check out The Accessibility Guy on YouTube

Accessibility •	Reading Order			
Tags				
≞- (ì	Draw a rectangle a the buttons below	round the conten :	t then click one of	
🗸 🄁 Tags	Text/Par	ragraph	Figure	s and Jews: The
✓ 💜 <figure></figure>	Form	Field	Figure/Caption	unities in the Co
🥞 lmage (379): w:655 h:914	Heading 1	Heading 4	Table	Howard Actor and Po
> ¶ <p></p>	Heading 2	Heading 5	Cell	110wuru Aster und Fe
> ¶ <p></p>	Heading 3	Heading 6	Formula	
✓ ¶ <p> ₩</p>	Reference (q)	Note (z)	Background/Artifact	rs we have been studying I with Jewish-Ukrainian
> 🗳 <hyphenspan></hyphenspan>				ery difficult topic, there h
✓ ¶ <p></p>		Table Edito	pr	Leonid Finberg. There w
🦉 <sup>+</sup> "-	Show page con	tent groups		al leaders from around
> 🗳 <hyphenspan></hyphenspan>	Page content	order		n. we first started our work o
🦉 <sup>11</sup> -	⊖ Structure typ	es		ichieved, and a dialogue b
> 💜 <hyphenspan></hyphenspan>	Show table cell	s		lace is vastly different from
ii "	Display like eler	ments in a single l	block	ply affect this discussion
> 🇳 <hyphenspan></hyphenspan>	Show tables an	d figures		aders of the World War II
🗑 "	Clear Page Str	ucture	Show Order Panel	g Jews and Ukrainians;
> ¶ <p></p>	Help		Close	on and other democratic
> ¶ <p></p>			1 Among and	Nichard works on this topic see P.I.
> ¶ <p></p>			Relations: Two S	alitudes (Oakville, ON: Mosaic Press
> ¶ <p></p>			Studies, 1988); I Relation	Potichnyj and Howard Aster
> ¶ < <sup>p</sup> >			6 krainian-Jewis	sh Relations, in Suchasnist 8 (Augu
> ¶ <p></p>			Howard Aster	utonomy, Self-Determination and T

## Accessibility Tools: Abbyy OCR Editor

- Text editing, verification, and correction tool for optical character recognition.
- Left side: Image with recognition areas (green for text, red for image)
- Right side: OCR output for manual correction and verification.
- Abbyy does not support accessibility features. You need Acrobat for that.



## Using Abbyy's built-in Pattern Training

- Create a unique character map, specific to this text.
- Increases accuracy of recognition.
- Struggles with ligatures and unconventional characters (like the long-s)

Pattern Training	?	$\times$	User Pattern				?	×
Active pattern: woof			r r	r	r	<b>)]</b> ri		^
If the frame encloses a part of a character or parts of adjacent characters, move its borders using the mouse or buttons: Enter the character enclosed by the frame:	<<	>> Frain	f s	st	u u	u		ļ
Effects Bold Superscript Italic Subscript				v	y y			~
Back Skip	0	Close		<u>I</u> mages	D <u>e</u> tails	<u>P</u> roperties	<u>D</u> elet	te
						ОК	Cance	el

## **AI OCR Options:**

None of these will interface with our Abbyy and Acrobat system.

VS

#### **Corporate OCR Options**

- Highly accurate.
- Only produces extracted text.
- Can not produce PDFs or add accessibility features.
- Privacy concerns.



#### Free, Open-Source Option

- Can create PDF with OCR overlay and hOCR coordinates.
- Runs locally, with private training data
   = no privacy issues.
- Can create and share you own training data.

Tesseract OCR

## **Tesseract OCR Output vs Source Image**



VNCTORVM inuulgatis animantium moribus, & naturis, phyfica miracula nomine monftrorum vulgò nuncupata, in afpectum publicum modò producenda fuccedunt; que aliquando contingere, explorata fatetur veritas, & potiffimùm quando, non iuxta confuetam naturæ normam, peregrina quædam, vel toti generatorum corpori, vel eiufde partibus (fiuè de homine, fiuè de brutis, fiuè de plantis ver-Monstra at ba fiant) figura, vel character imprimitur. Non enim hæc timent ad aauthorum funt deliria, vel fomnia, qui fucata perfuafione, nimalia, &

VNCTORV M inuulgatis animantium moribus, & nai lits0 turis, phyfica miracula nomine monftrorum vulgo o nuncu-V 7lll pata, in afpectum publicum modo producenda fuccedunt; V 7ẽã7 3à5 que aliquando contingere, explorata fatetur veritas, & potiffimim quandon, on iuxta confuetam naturæ normam, pere grina quædam, vel toti generatorum corpori, vel eiufdẽ partibus(fiuẽ de homine, fiuẽ de brutis, fiuẽ de plantis ver- Moniftra at ba fiant)figura, vel character imprimitur. Non enimhæc tinent aaà a--------" authorum funt deliria, vel fomnia, qui fucata perfuafione, nimalia, &

## **Chat GPT OCR Output vs Source Image**



VNCTORVM inuulgatis animantium moribus, & naturis, phyfica miracula nomine monftrorum vulgò nuncupata, in afpectum publicum modò producenda fuccedunt; que aliquando contingere, explorata fatetur veritas, & potiffimùm quando, non iuxta confuetam naturæ normam, peregrina quædam, vel toti generatorum corpori, vel eiufde partibus (fiuè de homine, fiuè de brutis, fiuè de plantis ver-Monstra at ba fiant) figura, vel character imprimitur. Non enim hæc tinent ad aauthorum funt deliria, vel fomnia, qui fucata perfuafione, nimalia, &

CVNCTORVM inualgatis animantium moribus, & naturis, physica miracula nomine monftrorum vulgò nuncupata, in afpectum publicum modo producenda fuccedunt: quae aliquando contingere, explorata facituer veritas, & politissimum quando, non iuxta confuetam naturæ normam, petitur regina quædam vel toti generatorum corpori, vel eiusdem partibus (fiue de homine, fiue de brutis, fiue de plantis verfabatur figura), vel character imprimuit. Non enim hæc authorum tantum deliria, vel fomnia, quæ fui carent periculo,

## Using AI generated extracted text in a PDF

Why not just copy and paste it into your OCR Editor?

- Sadly, it doesn't work like that. It makes for really sloppy PDFs.
- There's no way to generate the hOCR coordinates with extracted text.
- Text won't match the source image.
- Searches may not work.
- Won't be able to highlight and copy + paste the correct words.



## Using Tesseract with Training Data from <a href="https://latinocr.org/">https://latinocr.org/</a>

@LAPTOP-PPO5T6D9: ~/access\$ ls
page12.tif
@LAPTOP-PPO5T6D9: ~/access\$ tesseract page12.tif page12 -1 lat pdf
@LAPTOP-PPO5T6D9: ~/access\$
@LAPTOP-PPO5T6D9: ~/access\$ tesseract page12.tif page12 -1 lat txt
@LAPTOP-PPO5T6D9: ~/access\$
@LAPTOP-PPO5T6D9: ~/access\$ tesseract page12.tif page12 -1 lat hocr
@LAPTOP-PPO5T6D9: ~/access\$
@LAPTOP-PPO5T6D9: ~/access\$
page12.hocr page12.pdf page12.tif page12.txt

- **pdf** Create PDF containing compressed version of original image with hidden, searchable, and selectable textual overlay.
- **txt** extracted text in plain text .txt file.
- **hocr** extracted text as xhtml character position information in .html file

#### PDF textual overlay compared to extracted text



## **Editing Tesseract PDFs in Acrobat**

- Tesseract embeds a 'Glyphless' font when creating a PDF.
- PDF editing or verification tools cannot open or understand this font.
- Acrobat displays text as blacked out.
- Abbyy will not open PDFs created with Tesseract.

reingite		
	Acrobat Pro DC 2015 Profiles 🔻	
Profiles Results	↔ Standards	Options 🔻
Default	🗸 😼 🔎 🖌 🔢 Find	
🔎 🔎 Flatten transparency (low	resolution)	
🔎 🔎 Flatten transparency (med	lium resolution)	
🖸 🔎 Make OCR text visible		Edit 🖂 🚽
		Edition
Make OCK text visible Makes invisible text origin performed after scanning second layer, such that th	nating from an OCR process (optical character recognition documents) visible. Puts invisible text on its own layer an e display of page contents can be toggled for quality ass	n which is usually id everything else on a surance of OCR results.
Make OCK text visible Makes invisible text origin performed after scanning second layer, such that th	nating from an OCR process (optical character recognition documents) visible. Puts invisible text on its own layer an e display of page contents can be toggled for quality ass ayscale images to JPEG (high quality)	n which is usually d everything else on a surance of OCR results.
Make OCK text visible Makes invisible text origin performed after scanning second layer, such that th Recompress color and gr	nating from an OCR process (optical character recognition documents) visible. Puts invisible text on its own layer an e display of page contents can be toggled for quality ass ayscale images to JPEG (high quality) ayscale images to JPEG2000 (lossless)	n which is usually id everything else on a surance of OCR results.
Make OCK text visible Makes invisible text origin performed after scanning second layer, such that th Recompress color and gr Recompress color and gr Recompress color and gr	nating from an OCR process (optical character recognition documents) visible. Puts invisible text on its own layer an e display of page contents can be toggled for quality ass ayscale images to JPEG (high quality) ayscale images to JPEG2000 (lossless) ayscale images to ZIP	n which is usually id everything else on a surance of OCR results.
Make OCK text visible Makes invisible text origin performed after scanning second layer, such that th Recompress color and gr Recompress color and gr Recompress color and gr Recompress color and gr Recompress color and gr	nating from an OCR process (optical character recognition documents) visible. Puts invisible text on its own layer an e display of page contents can be toggled for quality ass rayscale images to JPEG (high quality) rayscale images to JPEG2000 (lossless) rayscale images to ZIP we if portrait	n which is usually id everything else on a surance of OCR results.
Make OCK text visible Makes invisible text origin performed after scanning second layer, such that th Compress color and gr Compress color and gr	nating from an OCR process (optical character recognition documents) visible. Puts invisible text on its own layer an e display of page contents can be toggled for quality ass ayscale images to JPEG (high quality) ayscale images to JPEG2000 (lossless) ayscale images to ZIP re if portrait f landscape	n which is usually id everything else on a surance of OCR results.



## Homegrown Solutions - ScribeOCR

- One of the only hOCR editing tools.
- Completely web-based.
- Accepts image + corresponding hOCR files from Tesseract.
- Has limitations:
  - $\circ$  Single developer.
  - Does not support special characters like the long-s.
  - Exports can be slow.



## So, what do we do?

- Meeting PDF accessibility requirements is very laborious.
- Accessibility requests can take too long to fulfill.
- Accessibility software has little development and hasn't advanced in years.
- This is all reliant on Adobe to further develop PDF accessibility features that are sparsely used.
- We don't have to use PDF. Web-based systems like the IA Reader are surpassing PDF for accessibility, with support for multiple download formats, various views, and built-in screen readers.
- HathiTrust avoids this by providing a text version of documents, but not as a PDF text overlay.
- Students will likely be able to use tools themselves to improve their own accessibility experience.



# Using ChatGPT for Accessibility Support

- Students and researchers are probably just going to do it themselves with ChatGPT.
- ChatGTP can translate languages, create summaries, or generate transcriptions in just a few seconds.
- It's an easy to use, user-facing experience (with obvious environmental and ethical concerns).
- Doesn't mean that we don't still have an obligation to distribute content without meeting accessibility guidelines.



The text in the image is in Latin, and here's a general translation of the main title and introductory phrases:

Main Title: ULYSSIS ALDROVANDI PATRITII BONONIENSIS MONSTRORUM HISTORIA

#### Translation:

"The History of Monsters by Ulysses Aldrovandi, Patrician of Bologna."

Subtitle:

ORDINIS RATIO

#### Translation: "On the Order of Things"

The rest of the text delves into Aldrovandi's observations on the nature, types, and the categorization of monsters and anomalies, both in the natural world and in human understanding. If you'd like, I can translate specific paragraphs or sections, or provide an overall summary of the text!

# Thanks, Let's talk.

Mark Pellegrino Reach me at: pellem11@mcmaster.ca

Some great sources if you're interested:

- <u>The Accessibility Guy</u> on Youtube
- Internet Archve PDF workflows
- Investigating OCR and Text PDFs from Digital Collections. Jonathan Rockhind, 2023
- <u>ScribeOCR</u>
- Latin OCR
- <u>Tesseract GitHub</u> thread about Glyphless fonts.
- <u>Tesseract Google Group</u>
- hOCR Standard