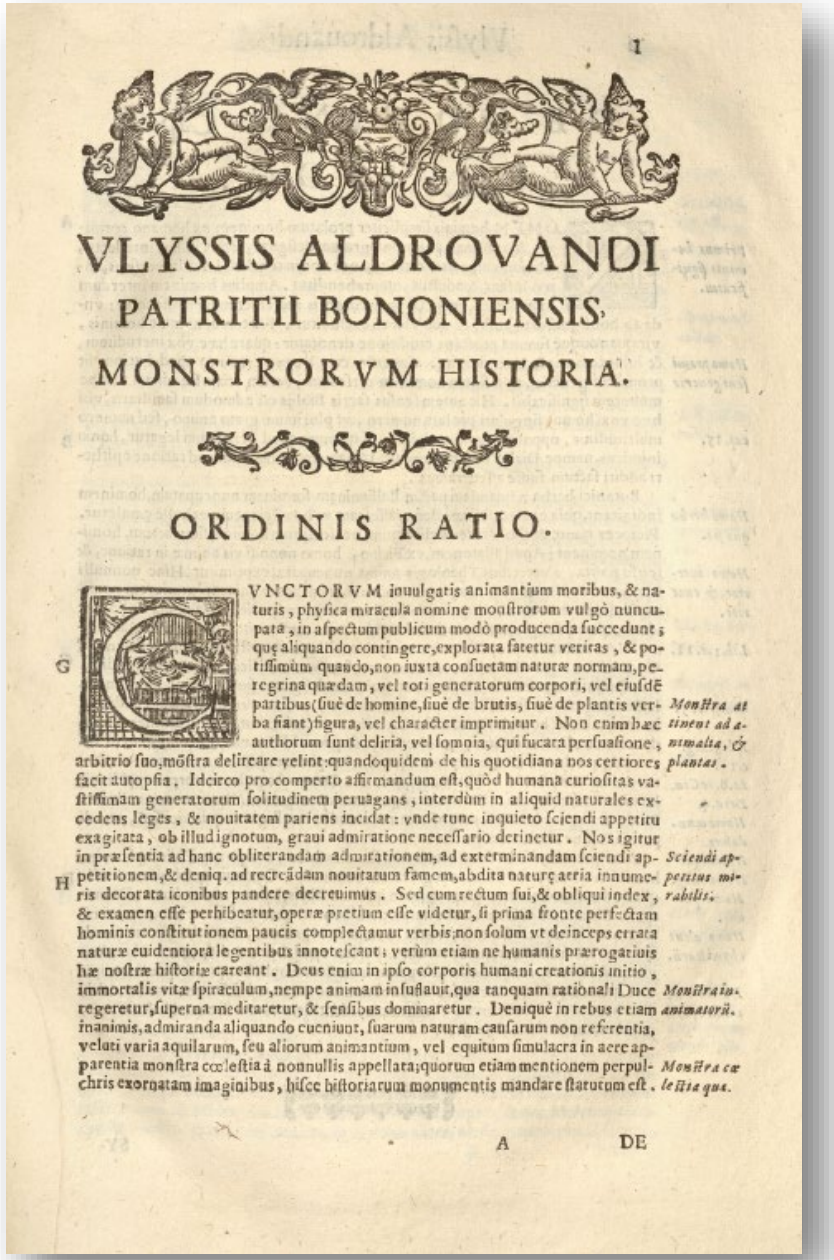
Accessible (Digital) Archives: Why we don't have them and why AI isn't helping

(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



1

VLYSSIS ALDROVANDI
PATRITII BONONIENSIS
MONSTRORVM HISTORIA.

ORDINIS RATIO.

G VNCTORVM inuulgatis animantium moribus, & naturis, phyſica miracula nomine monſtrorum vulgò nuncupata, in aſpectum publicum modò producenda ſuccedunt; quę aliquando contingere, explorata fatetur veritas, & potiſſimùm quando, non iuxta conſuetam naturæ normam, peregrina quædam, vel toti generatorum corpori, vel eiuſdē partibus (ſiuè de homine, ſiuè de brutis, ſiuè de plantis verba fiant) figura, vel character imprimitur. Non enim hæc authorum ſunt deliria, vel ſomnia, qui fucata perſuaſione, arbitrio ſuo, mōſtra delireare velint: quandoquidem de his quotidiana nos certiores facit autopſia. Idcirco pro comperto affirmandum eſt, quòd humana curioſitas vaſtiſſimam generatorum ſolitudinem peruagans, interdùm in aliquid naturales excedens leges, & nouitatem pariens incidat: vnde tunc inquieto ſciendi appetitu exagitata, ob illud ignotum, graui admiratione neceſſario detinetur. Nos igitur in præſentia ad hanc obliterandam admirationem, ad exterminandam ſciendi appetitionem, & deniq. ad recreādam nouitatum famem, abdita naturę aeria innumeris decorata iconibus pandere decreuimus. Sed cum rectum ſui, & obliqui index, & examen eſſe perhibeatur, operæ pretium eſſe videtur, ſi prima fronte perfectam hominis conſtitutionem paucis complectamur verbis; non ſolum vt deinceps errata naturæ euidentiora legentibus innoteſcant; verùm etiam ne humanis prærogatiuis hæ noſtræ hiſtoriæ careant. Deus enim in ipſo corporis humani creationis initio, immortalis vitæ ſpiraculum, nempe animam inſuflauit, qua tanquam rationali Duce regeretur, ſuperna meditaretur, & ſenſibus dominaretur. Deniquè in rebus etiam inanimis, admiranda aliquando cueniunt, ſuarum naturam cauſarum non referentia, veluti varia aquilarum, ſeu aliorum animantium, vel equitum ſimulacra in aere apparentia monſtra cœleſtia à nonnullis appellata; quorum etiam mentionem perpulchris exornatam imaginibus, hiſce hiſtoriarum monumentis mandare ſtatutum eſt.

Monſtra attinent ad animalia, & plantas.

Sciendi appetitus mirabilis.

Monſtra animatorū.

Monſtra cœleſtia quæ.

H

A DE

Sample Starting Image

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

Latin language, hand-written, blackletter font.

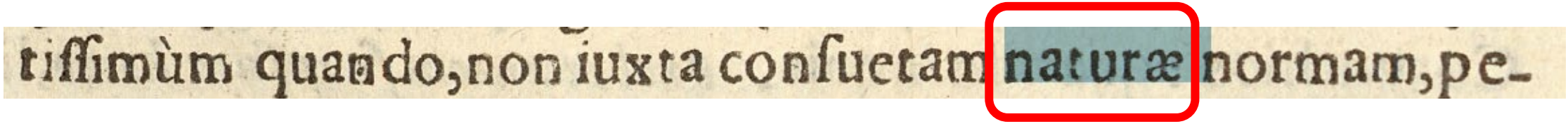
Digitized as 600dpi TIFF image, 62.2 MB.

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.




```
<span class='ocr_line' id='line_1_14' title="bbox 1091 2820 2733 2893; baseline 0.001 -22; x_size 72;
x_descenders 19; x_ascenders 17">
 <span class='ocrx_word' id='word_1_57' title='bbox 1091 2821 1321 2875; x_wconf 20'>tiſſimim</span>
 <span class='ocrx_word' id='word_1_58' title='bbox 1347 2825 1678 2893; x_wconf 66'>quando,</span>
 <span class='ocrx_word' id='word_1_59' title='bbox 1593 2816 1668 2897; x_wconf 89'>non</span>
 <span class='ocrx_word' id='word_1_60' title='bbox 1695 2827 1834 2876; x_wconf 85'>iuxta</span>
 <span class='ocrx_word' id='word_1_61' title='bbox 1851 2820 2144 2875; x_wconf 88'>conſuetam</span>
 <span class='ocrx_word' id='word_1_62' title='bbox 2160 2836 2354 2874; x_wconf 85'>naturæ</span>
 <span class='ocrx_word' id='word_1_63' title='bbox 2383 2837 2620 2885; x_wconf 89'>normam,</span>
 <span class='ocrx_word' id='word_1_64' title='bbox 2639 2839 2733 2892; x_wconf 90'>pe-</span>
</span>
```

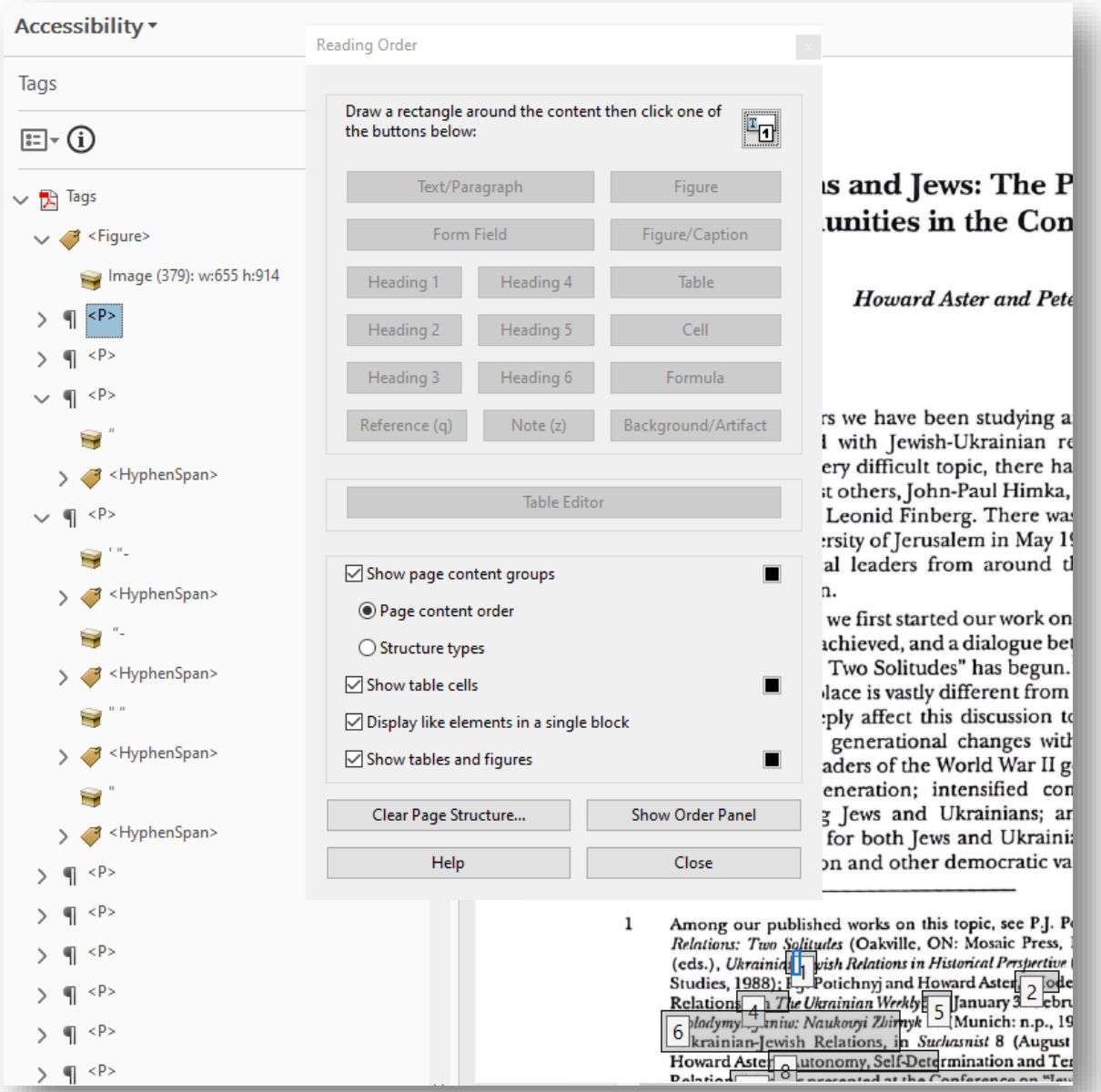
Accessibility Tools: Adobe Acrobat

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

Elements of an accessible PDF:

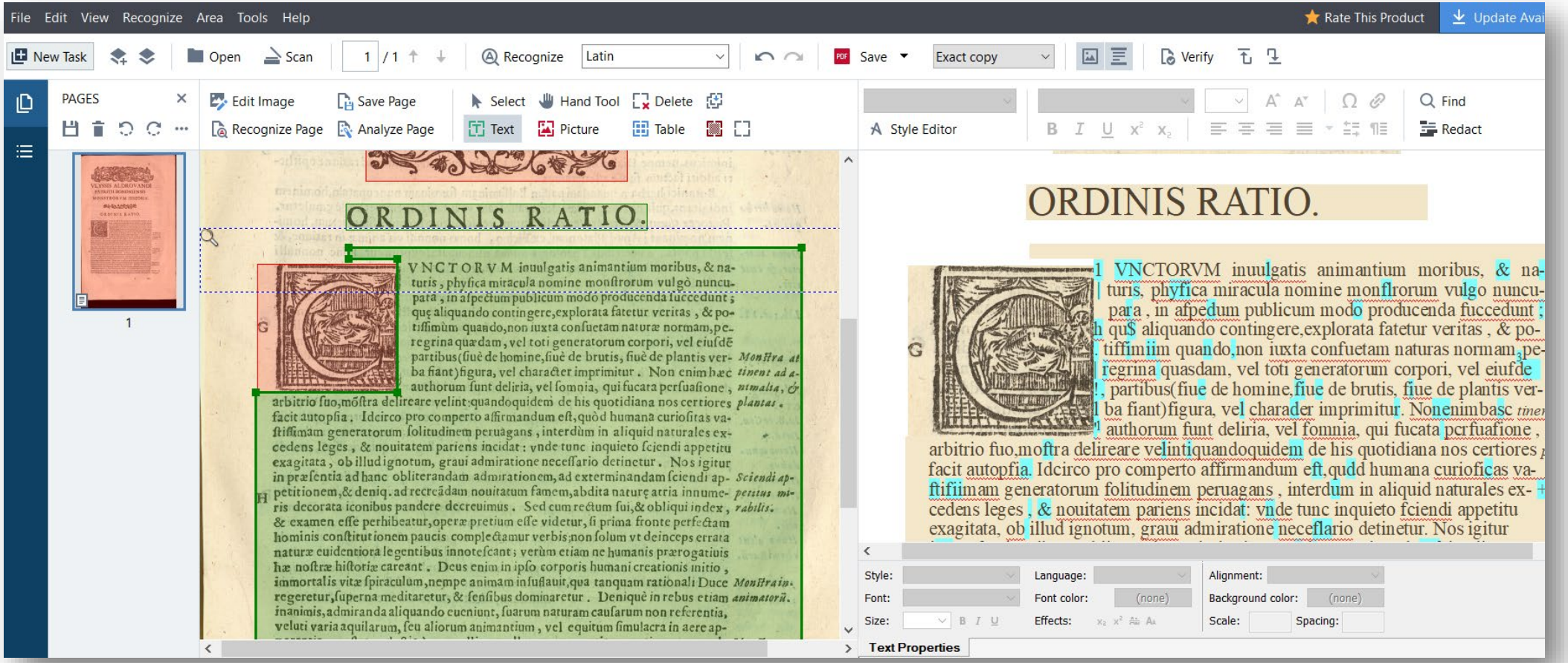
- Tagging – PDF elements employ html tagging just like a website, with opportunities to support modern HTML5 semantic tagging.
- Reading Order – tagged elements must be assigned to numerical order for assistive devices.
- Text overlay – the invisible textual element on top of the source image.
- Alt-text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

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



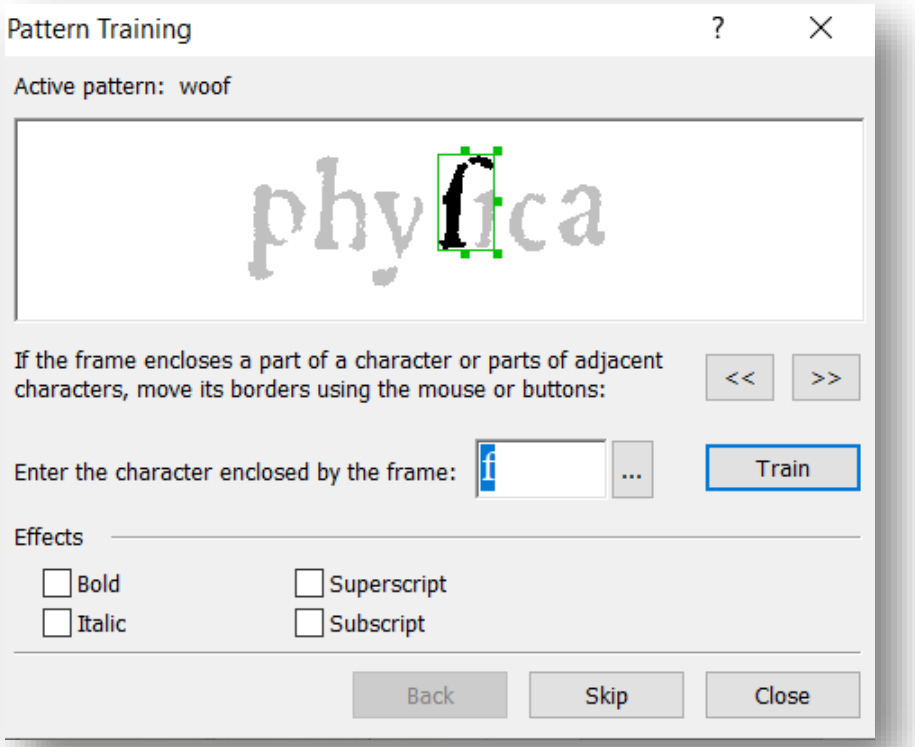
Accessibility Tools: Abbyy OCR Editor

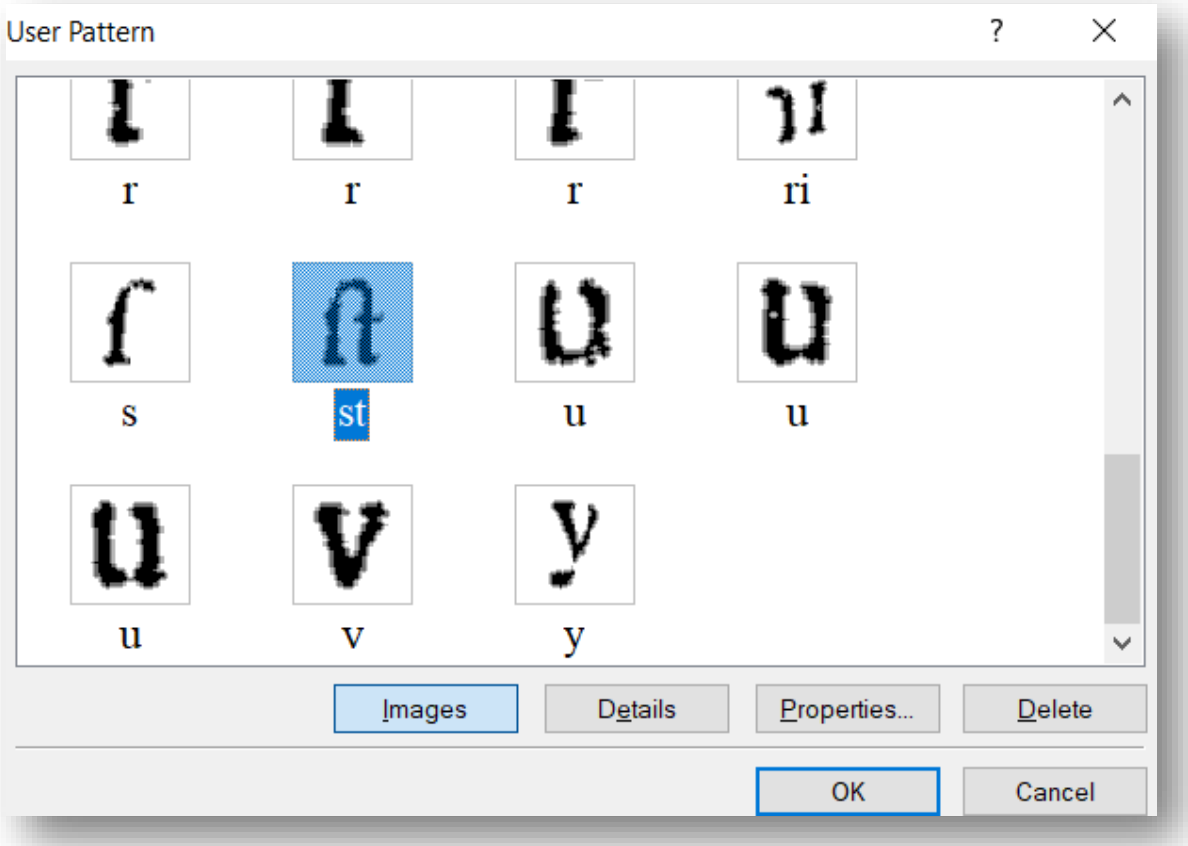
- Text editing, verification, and correction tool for optical character recognition.
- Left side: Image with recognition areas (green for text, red for image)
- Right side: OCR output for manual correction and verification.
- Abbyy does not support accessibility features. You need Acrobat for that.



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)





AI OCR Options:

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

Corporate OCR Options

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

Cloud Vision API

Copilot

ChatGPT

Amazon Textract

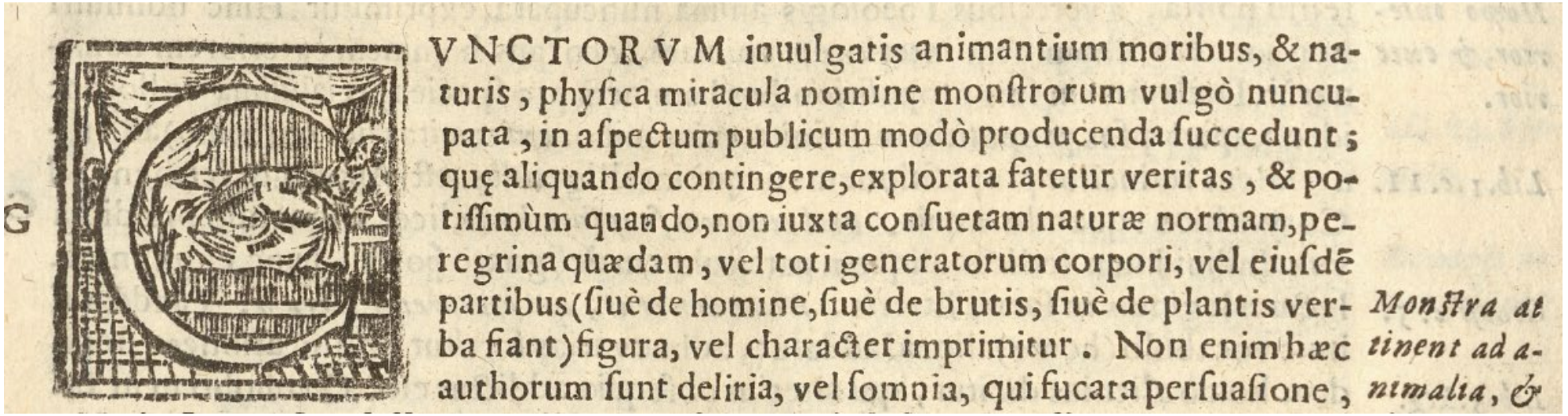
VS

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

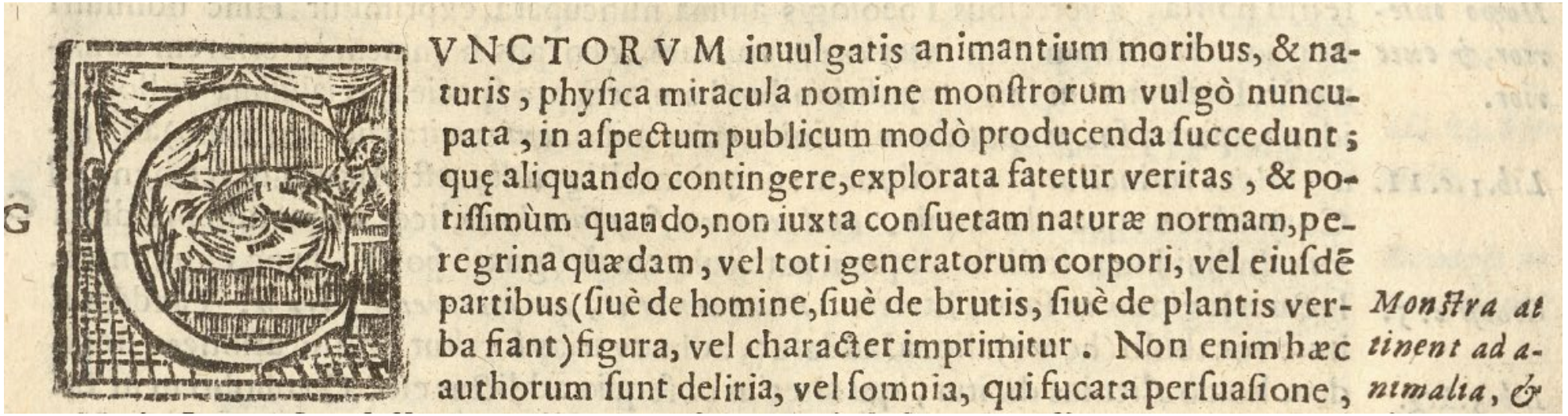
Tesseract OCR Output vs Source Image



VNCTORVM inuulgatis animantium moribus, & naturis, phyſica miracula nomine monſtrorum vulgò nuncupata, in aſpectum publicum modò producenda ſuccedunt; quę aliquando contingere, explorata fatetur veritas, & potiſſimùm quando, non iuxta conſuetam naturæ normam, peregrina quædam, vel toti generatorum corpori, vel eiuſdē partibus (ſiuè de homine, ſiuè de brutis, ſiuè de plantis verba fiant) figura, vel character imprimitur. Non enim hæc authorum ſunt deliria, vel ſomnia, qui fucata perſuaſione, *Monſtra attinent ad animalia, &*

VNCTORV M inuulgatis animantium moribus, & nai
lits0 turis, phyſica miracula nomine monſtrorum vulgoò nuncu-
V 7lll l pata, in aſpectum publicum modò producenda ſuccedunt;
V 7ẽã7 3à5 qus aliquando contingere, explorata fatetur veritas, & potiſſimim
quandon,on iuxta conſuetam naturæ normam, pere
grina quædam, vel toti generatorum corpori, vel eiuſdẽ
partibus(ſiuẽ de homine, ſiuẽ de brutis, ſiuẽ de plantis ver- Moniftra at
ba fiant)figura, vel character imprimitur. Non enimhæc tinent aaà a-
———" authorum ſunt deliria, vel ſomnia, qui fucata perſuaſione, nimalia, &

Chat GPT OCR Output vs Source Image



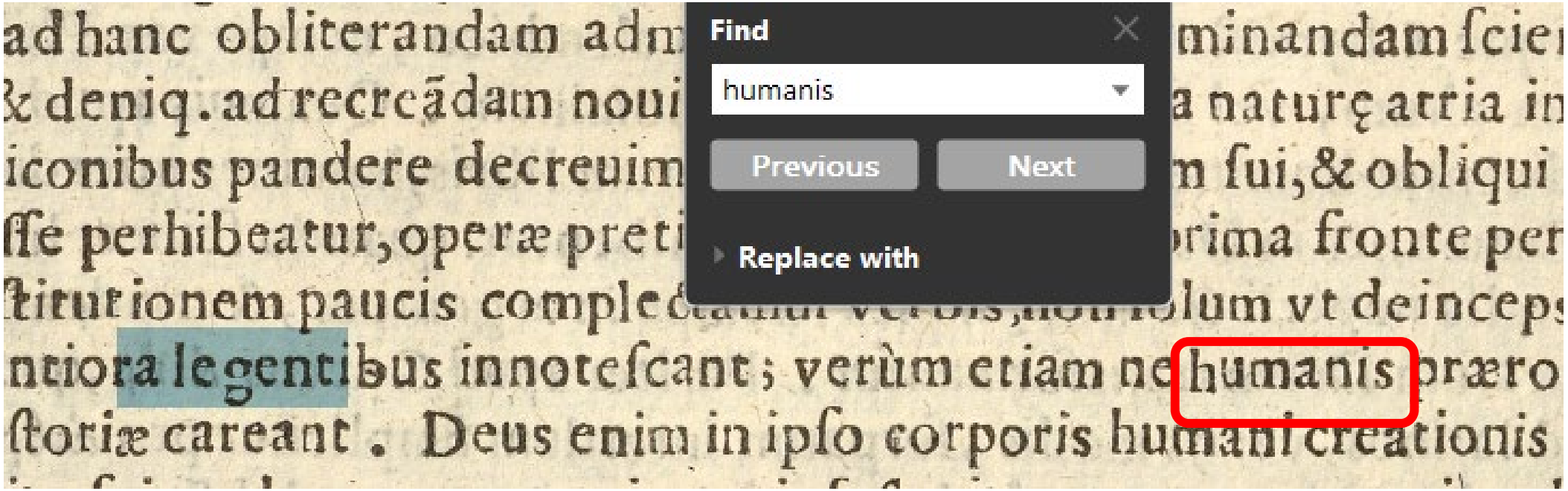
VNCTORVM inuulgatis animantium moribus, & naturis, phyſica miracula nomine monſtrorum vulgò nuncupata, in aſpectum publicum modò producenda ſuccedunt; quę aliquando contingere, explorata fatetur veritas, & potiſſimùm quando, non iuxta conſuetam naturæ normam, peregrina quædam, vel toti generatorum corpori, vel eiuſdē partibus (ſiuè de homine, ſiuè de brutis, ſiuè de plantis verba fiant) figura, vel character imprimitur. Non enim hæc authorum ſunt deliria, vel ſomnia, qui fucata perſuaſione, *Monſtra attinent ad animalia, &*

CVNCTORVM inualgatis animantium moribus, & naturis, physica miracula nomine monſtrorum vulgò nuncupata, in aſpectum publicum modo producenda ſuccedunt: quae aliquando contingere, explorata facituer veritas, & poſtissimum quando, non iuxta conſuetam naturæ normam, petitur regina quædam vel toti generatorum corpori, vel eiusdem partibus (ſiue de homine, ſiue de brutis, ſiue de plantis verſabatur figura), vel character imprimuit. Non enim hæc authorum tantum deliria, vel ſomnia, quæ ſui carent periculo,

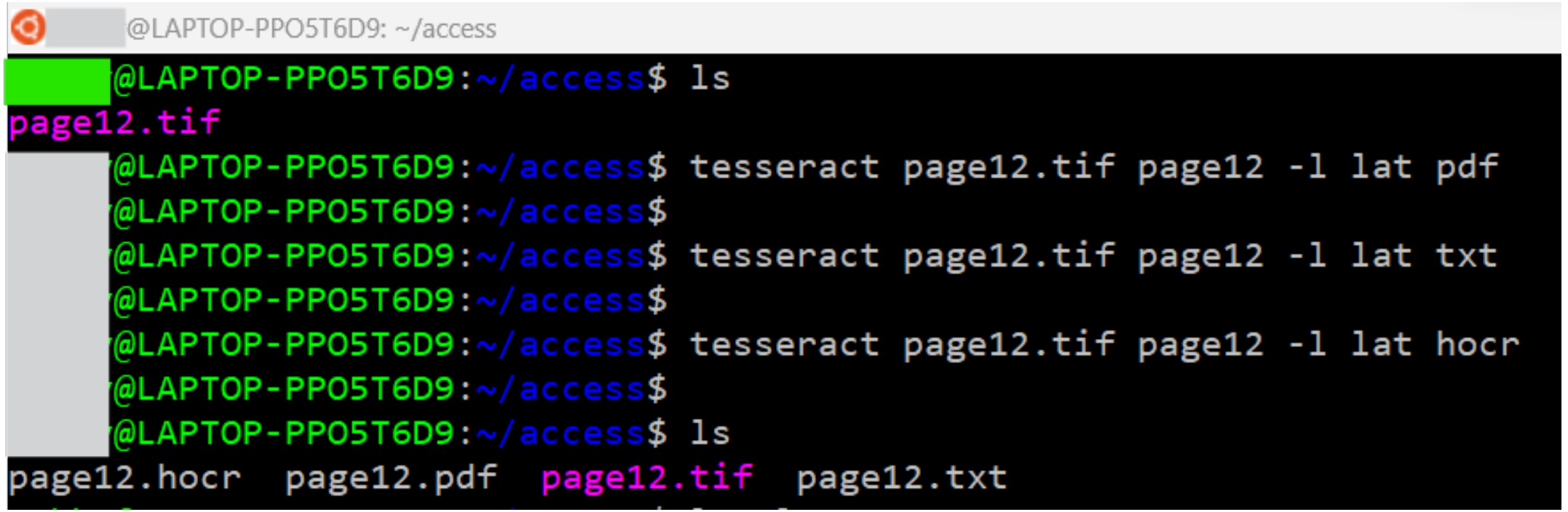
Using AI generated extracted text in a PDF

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 https://latinocr.org/



- **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

VLYSSIS ALDROVANDI
PATRITII BONONIENSIS
MONSTRORVM HISTORIA.

ORDINIS RATIO.

VNCTORVM inuulgatis animantium moribus, & naturis, phyſica miracula nomine monſtrorum vulgò nuncupata, in aſpectum publicum modò producenda ſuccedunt; quę aliquando contingere, explorata fatetur veritas, & potiſſimùm quando, non iuxta conſuetam naturæ normam, peregrina quædam, vel toti generatorum corpori, vel eiuſdē partibus (ſiuè de homine, ſiuè de brutis, ſiuè de plantis verba fiant) figura, vel character imprimitur. Non enim hæc authorum ſunt deliria, vel ſomnia, qui fucata perſuaſione, arbitrio ſuo, mōſtra delireare velint: quandoquidem de his quotidiana nos certiores facit autopſia. Idcirco pro comperto affirmandum eſt, quòd humana curioſitas vaſtiſſimam generatorum ſolitudinem peruagans, interdùm in aliquid naturales excedens leges, & nouitatem pariens incidat: vnde tunc inquieto ſciendi appetitu exagitata, ob illud ignotum, graui admiratione neceſſario detinetur. Nos igitur in præſentia ad hanc obliterandam admirationem, ad exterminandam ſciendi appetitionem, & deniq. ad recreādam nouitatum fāmem, abdita naturę atria innumeris decorata iconibus pandere decreuimus. Sed cum rectum ſui, & obliqui index, & examen eſſe perhibeatur, operæ pretium eſſe videtur, ſi prima fronte perfectam hominis conſtitutionem paucis complectamur verbis; non ſolum vt deinceps errata naturæ euidentiora legentibus innoteſcant; verùm etiam ne humanis prærogatiuis hæ noſtræ hiſtoriæ careant. Deus enim in ipſo corporis humani creationis initio, immortalis vitæ ſpiraculum, nempe animam inſuſtauit, qua tanquam rationali Duce regeretur, ſuperna meditaretur, & ſenſibus dominaretur. Deniquè in rebus etiam inanimis, admiranda aliquando eueniunt, ſuarum naturam cauſarum non referentia, veluti varia aquilarum, ſeu aliorum animantium, vel equitum ſimulacra in aere apparentia monſtra cœleſtia à nonnullis appellata; quorum etiam mentionem perpulchris exornatam imaginibus, hiſce hiſtoriarum monumentis mandare ſtatutum eſt.

Monſtra attinent ad animalia, & plantas.

Sciendi appetitus mirabilis.

Monſtra animatorū.

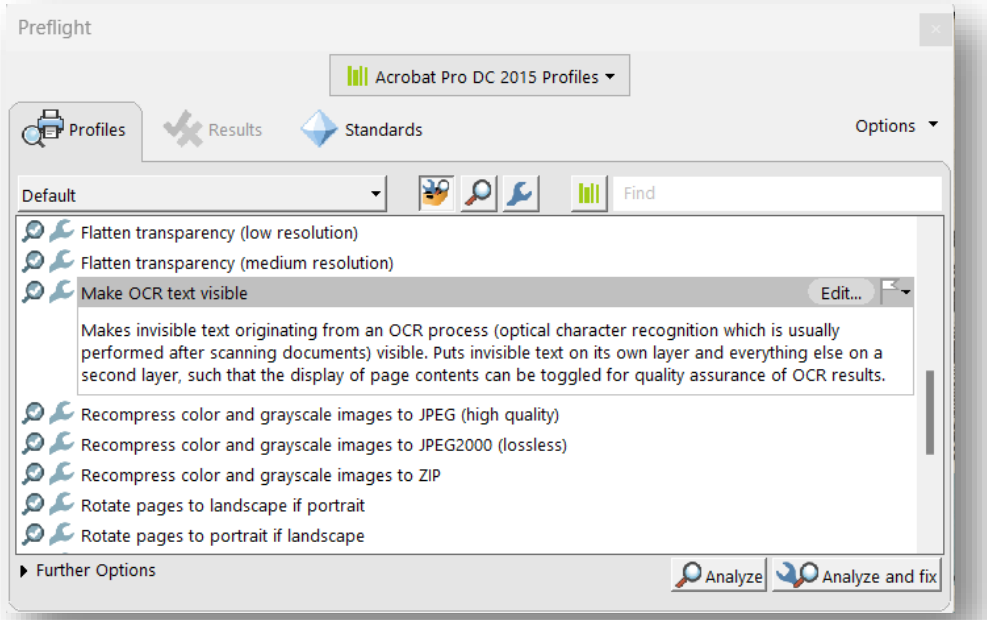
Monſtra cœleſtia quæ.

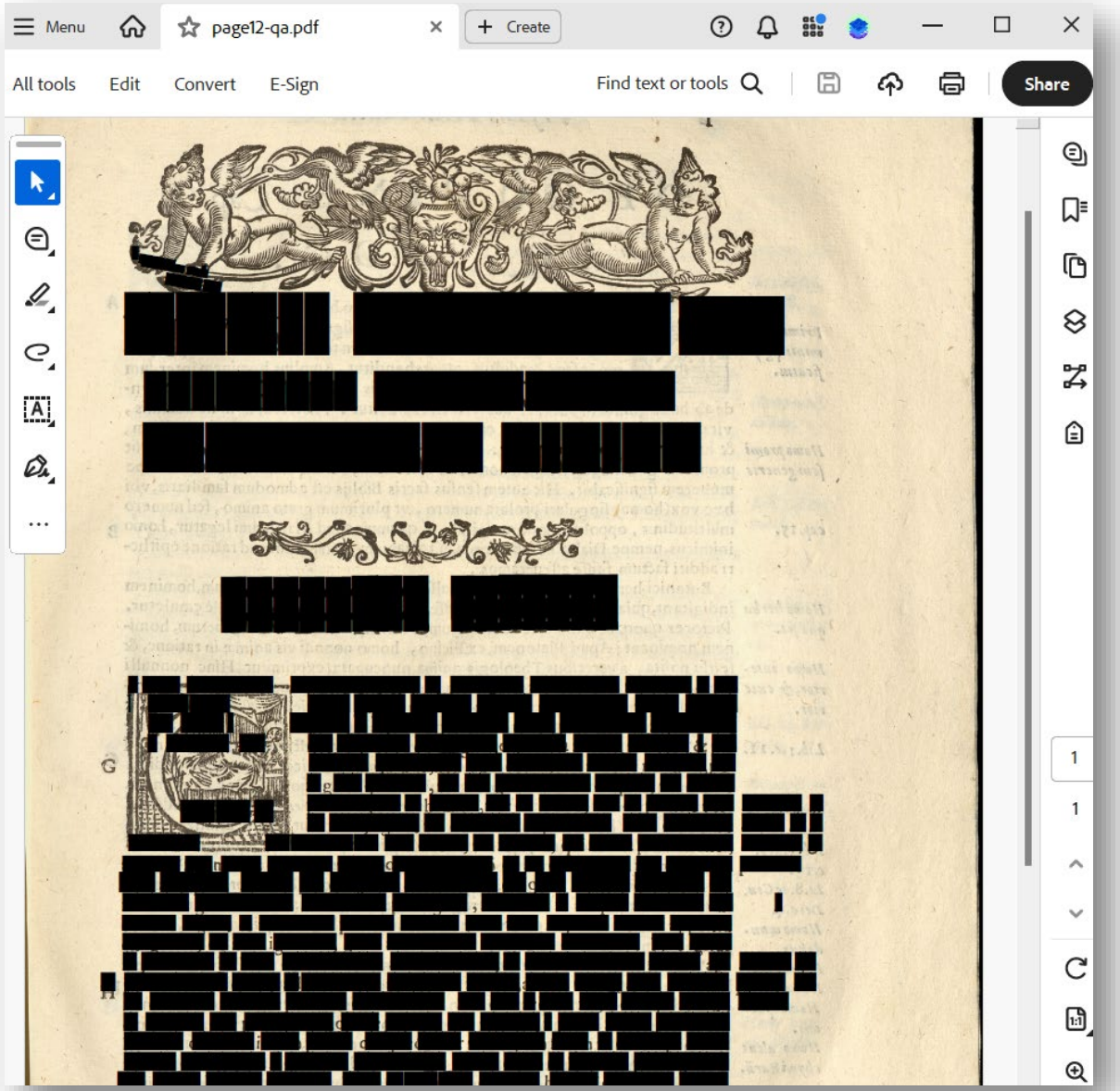
page12.txt

```
S3

D. :
m ſſ m
NT

VLYSSIS ALDROVAN
PATRITII BONONIENSIS
MONSTRORVM HISTORIA.

DI



ORDINIS RATIO.

ſ —— inid  VNCTORV M inuulgatis animantium moribus, & na-
i lits0 turis, phyſica miracula nomine monſtrorum vulgoò nuncu-
V 7lll l pata, in aſpectum publicum modò producenda ſuccedunt;
V 7ēā7 3à5 qus aliquando contingere, explorata fatetur veritas, & po-
tiſſimim quando, non iuxta conſuetam naturæ normam, pe-
re grina quædam, vel toti generatorum corpori, vel eiuſdē
partibus(ſiuē de homine, ſiuē de brutis, ſiuē de plantis ver- Moniſtra at
ba fiant)figura, vel character imprimitur. Non enimhæc tinent aaà a-
-——" authorum ſunt deliria, vel ſomnia, qui fucata perſuaſione, nimalia, &
arbitrio ſuo, mōſtra delireare velint: quandoquidem de his quotidiana nos cer
plantas.
facit autopſia. Idcirco pro comperto aſſirmandum eſt, quòd humanaā curioſitas
ſtiſimam generatorum ſolitudinem peruagans, interdum in aliquid naturales ex-
cedens leges, & nouitatem pariens incidat: vnde tunc inquieto ſciendi appetit
exagitata, ob illud ignotum, graui admiratione neceſſario detinetur. Nos igit
in præſentia ad hanc obliterandam admirationem, ad exterminandam ſciendi ap-
ap-
5 petitionem  & deniq  ad recreādam nouitatum famem  abdita nature atria innu
```

Editing Tesseract PDFs in Acrobat

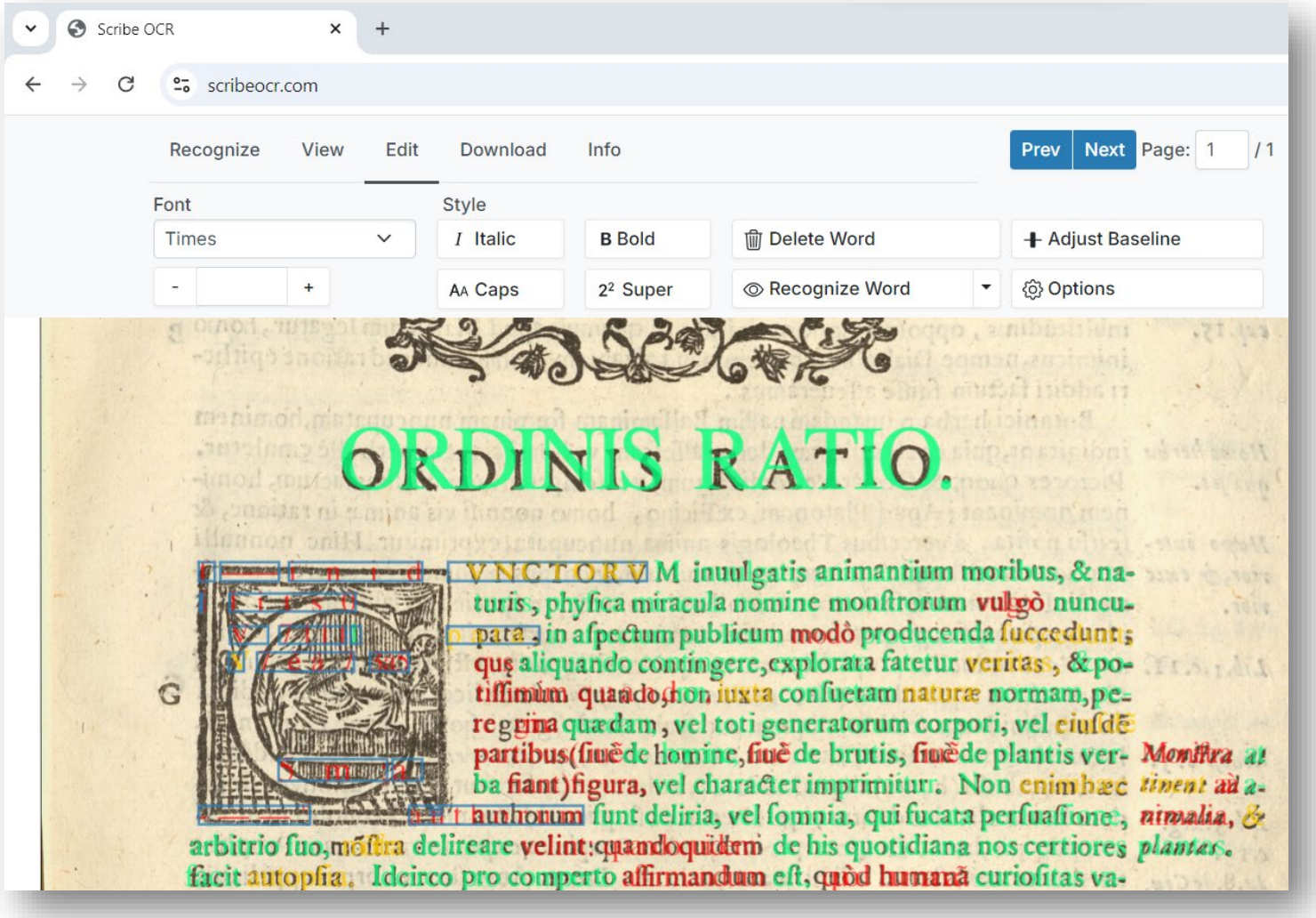
- Tesseract embeds a 'Glyphless' font when creating a PDF.
- PDF editing or verification tools cannot open or understand this font.
- Acrobat displays text as blacked out.
- Abbyy will not open PDFs created with Tesseract.





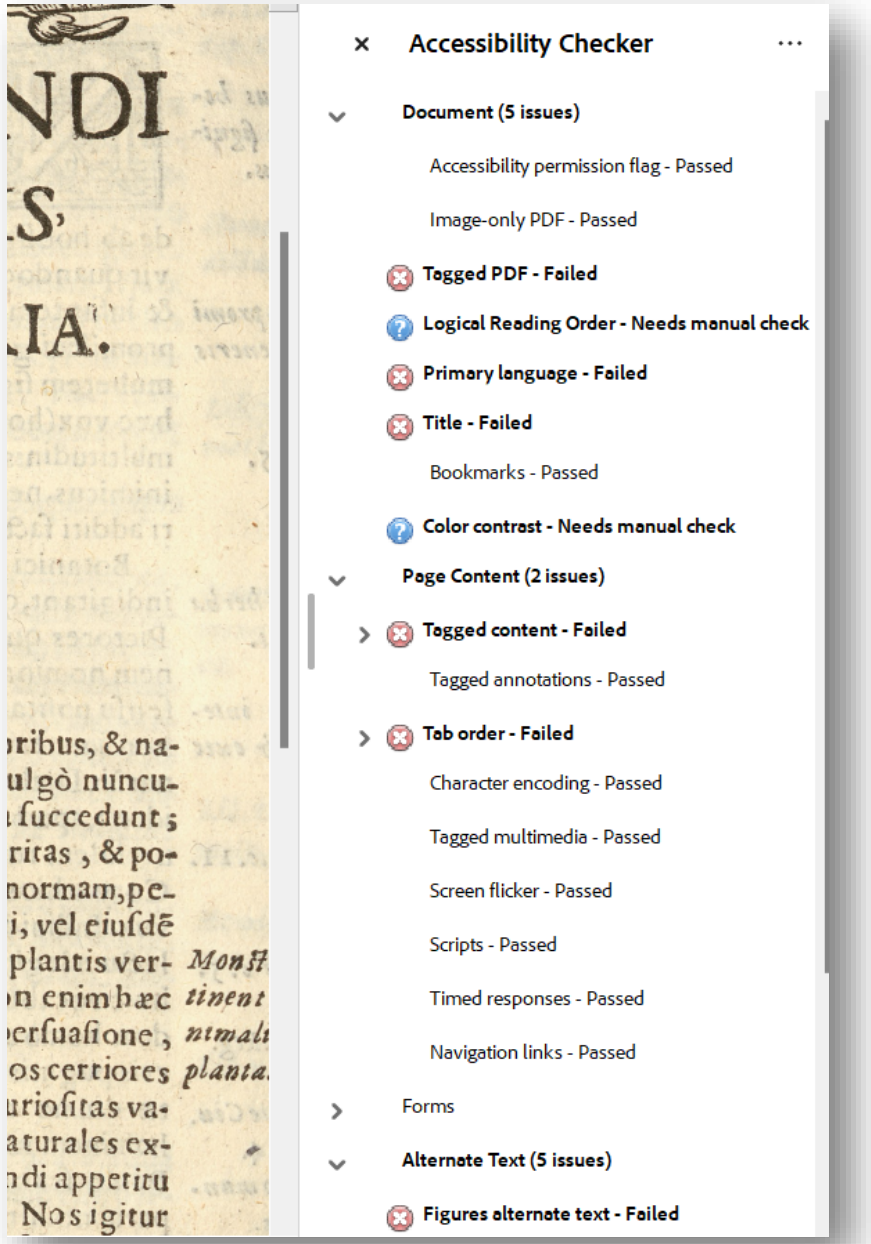
Homegrown Solutions - ScribeOCR

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



So, what do we do?

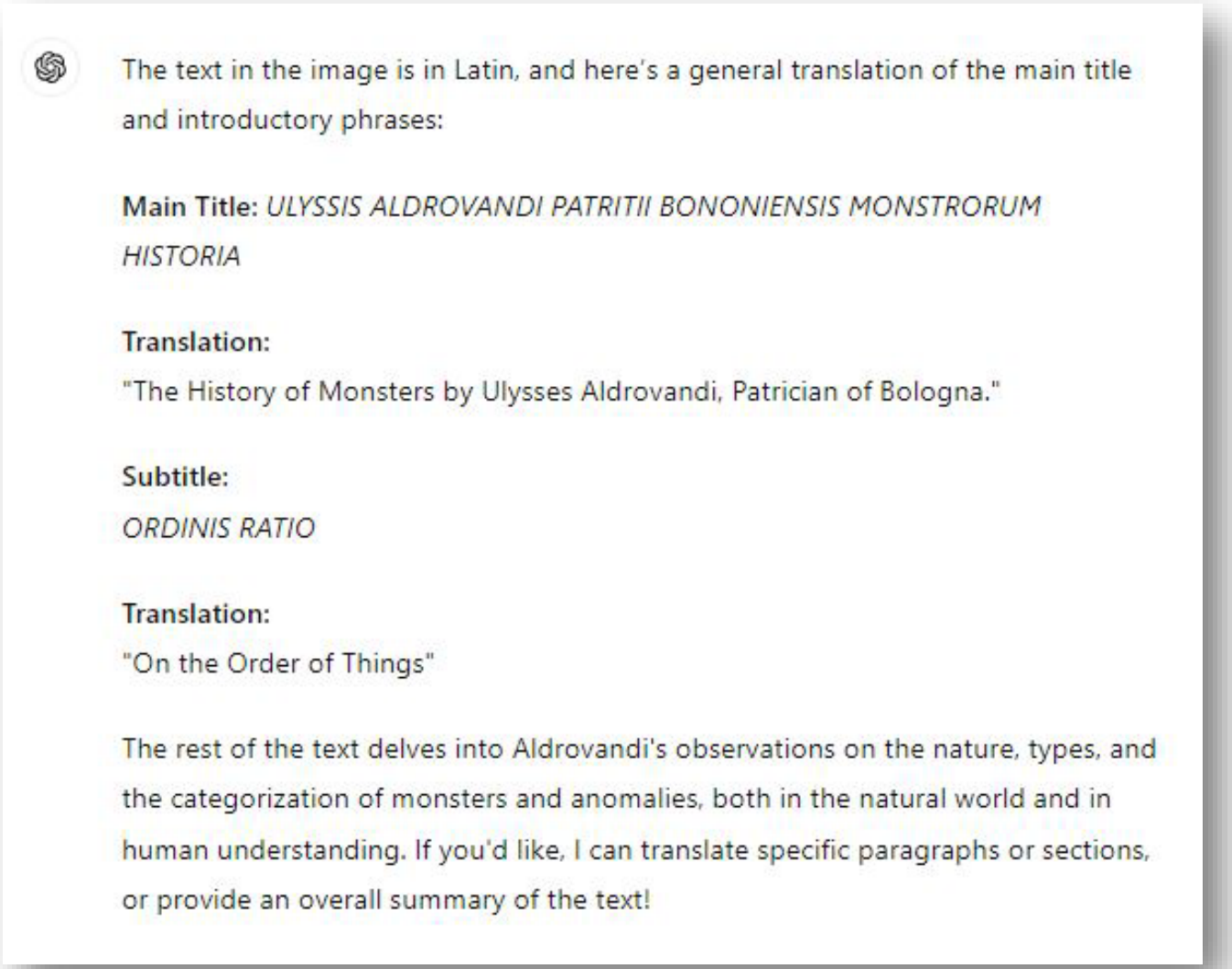
- Meeting PDF accessibility requirements is very laborious.
- Accessibility requests can take too long to fulfill.
- Accessibility software has little development and hasn't advanced in years.
- This is all reliant on Adobe to further develop PDF accessibility features that are sparsely used.
- We don't 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.

ChatGPT



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:

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