# 2 - Data for HTR

# Images & Annotations

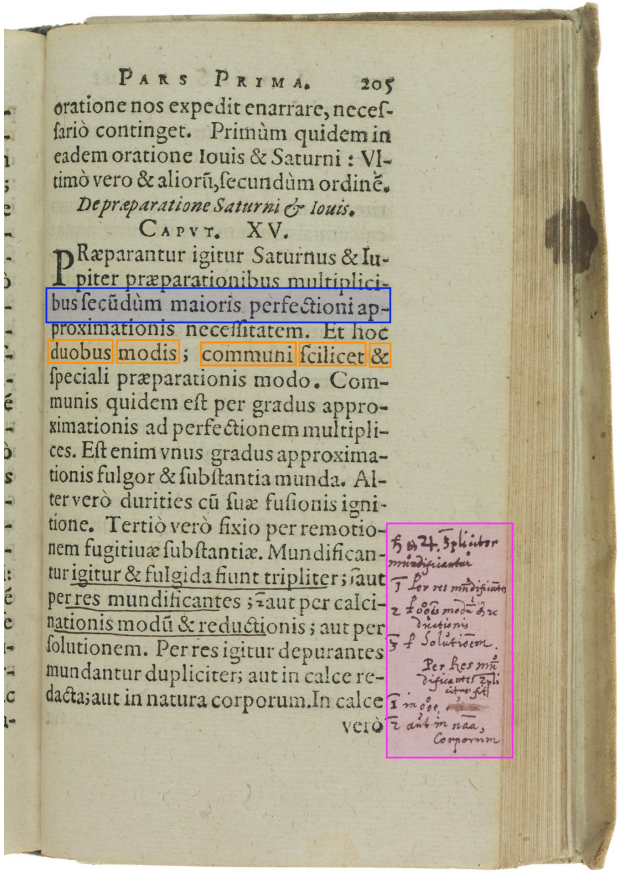
# Images

- typically: scans, photographies
  - very rarely: multi-spectral imaging, volumetric data, ...
- varying colour spaces
  - RGB
  - greyscale
  - binary
  - ...
- image resolution, compression (artefacts)


Image credit: flickr-user **pedrik** (Pedro Mendes), CC-BY 2.0 (2018) https://flickr.com/photos/24388834@N04/39575554842

# Annotations

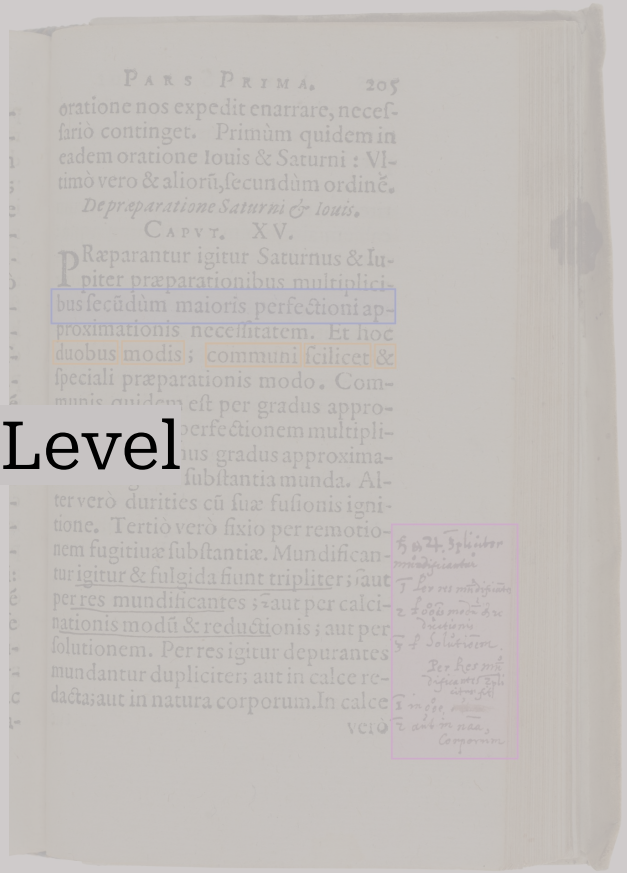
- image coordinates + corresponding transcriptions
- varying level of granularity: page, line, word, character
- common formats:
  - PAGE XML
  - ALTO XML
  - TEI XML
  - homebrew (csv, json, misc. text files)



# Annotations

- image coordinates + corresponding transcriptions
- varying level of granularity: page, line, word, ch
- common formats:
  - PAGE XML
  - ALTO XML
  - TEI XML
  - homebrew (csv, json, misc. text files)

## State of the Art: Line-Level



# Transcription Styles

- level of detail at which annotations record e.g.:
  - expanded abbreviations
  - letter forms
- level of detail for recognition
  - might require text normalisation before training
- discuss as early as possible in the project!
  - may also affect how annotations are collected, stored, etc.