## Sloane Lab HTR Model

**Description**

This repository contains Handwritten Text Recognition training data (layout segmentation and transcriptions) for the Sloane Lab HTR model. The HTR model is trained on the handwriting of Hans Sloane (1660-1753).

The Sloane Lab HTR model was developed in Transkribus (desktop client) and is based on the images and transcriptions of the Sloane’s Catalogue of Miscallanies (folio 3-152, recto and verso). The Catalogue of Miscallanies was chosen for the training data creation because it is known to be predominantly written by Sloane. The transcriptions for the training data were created by AEL Data Service as part of the Enlightenment Architectures project for the [digital edition of the Catalogue of 'Miscellanies, Antiquities, Seals, Pictures, Mathematical Instruments, Agate handles, Agate Cups, Bottles, Spoons](https://enlightenmentarchitectures.reconstructingsloane.org/cataloguemiscellanies/index.html)'

Sloane’s catalogues and their transcriptions contain historic racist language. We have written about the additional problems raised by digital technologies in Ortolja-Baird, A., & Nyhan, J. (2022) <https://doi.org/10.1093/llc/fqab065>. For more information, please see our website [sloanelab.org](https://sloanelab.org).

**Sources:**

Sloan, K., Ortolja-Baird, A., Nyhan, J., Pickering, V., & Fleming, M. (Eds.). (2019). Sir Hans Sloane’s Miscellanea which comprises his catalogues of Miscellanies, Antiquities, Seals, Pictures, Mathematical Instruments, Agate Handles and Agate Cups, Bottles, Spoons (Digital Edition). <https://enlightenmentarchitectures.reconstructingsloane.org/cataloguemiscellanies/index.html>

**Copyright:**

British Museum digital images of Miscellanea: © The Trustees of the British Museum. Shared under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) licence

**Training data creation**

For the HTR model creation transcriptions were mapped to the baselines identified by Transkribus. Whenever possible all transcribed information from the digital edition were included and the baselines corrected. However, the complexity of the historic catalogue’ layout made some compromises necessary.

The HTR training data differs from the digital editions’ transcriptions in the following ways:

|  |  |
| --- | --- |
| Exclusion of overwritten text from training data | A close-up of a document with numbers  Description automatically generated with low confidence |
| Exclusion of text not identified by the automated layout recognition (indicated in red) | A close-up of a document  Description automatically generated with low confidence |
| Exclusion of faded text (CCXXXIII) | A close-up of a piece of paper  Description automatically generated with low confidence |
| Inserted words are treated as separate text lines | A close up of a letter  Description automatically generated |
| Exclusion of textual features such as dotted lines | A close up of a document  Description automatically generated with low confidence |
| Base line separation for text written apart | A close up of a note  Description automatically generated |

Images: © The Trustees of the British Museum

**Models**

All models were trained with Transkribus’ PyLaia HTR engine (May 2023). Settings were left to default.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Number of words in training set** | **English Handwriting M3 base model (BM)** | **Epochs set up** | **Epochs trained** | **Character Error Rate** |
| Sloane\_Lab\_HTR\_Model | 32,103 | No | 100 | 89 | 11.10% |
| Sloane\_Lab\_HTR\_Model\_BM | 32,103 | Yes | 100 | 100 | 7.30% |

**Related work**

Humbel, M., & Nyhan, J. (2019). The Application of HTR to Early-modern Museum Collections: A Case Study of Sir Hans Sloane’s Miscellanies Catalogue. Digital Humanities 2019. Digital Humanities 2019, Utrecht. <https://discovery.ucl.ac.uk/id/eprint/10072160/1/HTR_Sloane_MH_JN.pdf>