

Introduction

The present dataset stems from research pursued within the SNSF funded Starting Grant *EGRAPSA: Retracing the evolutions of handwritings in Graeco-Roman Egypt thanks to digital palaeography* (SNSF grant n° 211682). The aim is to provide solid ground truth for **palaeographic dating**, and **detection and recognition of characters of cursive handwritten ancient Greek**.

It is composed of **194 images of 157 papyri that are precisely dated** (within two years) from the Hellenistic period (3rd to 1st c. BCE, more precisely from -310 to -3). The chronological coverage is balanced around 50 papyri per century over the considered period (3rd to 1st c. BCE); only the earliest decades are not covered, and the decade 250s is overrepresented. Most documents come from Egypt, but there are a few outsiders from Near East. The dataset also includes the **annotation to the character level of these images**, every time the preservation of the original writing and the quality of the image allowed for an annotation. It is divided into two subsets, a training set comprising 176 images and a test set comprising 18 images, following the same division used in De Gregorio et al. 2024.

For each papyrus, **the following identifiers are used**:

- TM numbers: unique identifiers of texts according to the Trismegistos database (https://www.trismegistos.org/about_how_to_cite.php)
- Checklist identifiers: short abbreviations indicating the publication of the edition of the papyrus (<https://papyri.info/docs/checklist>)

Dataset structure

The Hell-Date.zip archive contains the following files:

1. **data.csv** gives access to the 194 images with, for each image, a standard name, the location, collection name, inventory number, link to access online the file, and license attached to the image.
 - Names are standardised across the csv file as TMnumber_checklistAbbreviation. Some papyri are in more than one image, in that case the name contains additional information to distinguish the various images (e.g., two fragments of the same papyrus preserved in different collections, or the recto and verso of the same papyrus);
 - A python script is joint with the csv file to automatize the download process.
2. **downloader.py** allows to download automatically all the images of the dataset taking each of them from the original archive.
3. **How_to_download_the_dataset.pdf** briefly describes the simple procedure to download the images using the downloader.py script.
4. **requirements.txt** lists the requirements for the python environment to run the script correctly; it is needed to run downloader.py.
5. **metadata.csv** contains metadata for each image. Each column of the file represents the following metadata:
 - image_name: name of the file for the image of the papyrus;
 - checklist: checklist identifier of the papyrus (usual way to refer to the papyrus in papyrology);
 - TM: TM number as unique identifier of the text; as one text can have multiple images, images can share their TM number;
 - Year post quem: i.e. the year before which the papyrus cannot have been written;
 - Year ante quem: i.e. the year after which the papyrus cannot have been written;
 - Production Nome (supposed): the geographical region where the papyrus was written;

- Function: the type of document (e.g. a contract, or a letter. This item could be a comma separated list);
 - Subset: the subset to which the image belongs, i.e. the training set or the test set;
 - Annotated: indicates whether the image is at least partially annotated at the character level.
6. **annotations_training.json** and **annotations_test.json** contain the annotation of the images in the coco-json standard for the training and the test sets. Each annotation identifies one character on an image. The file is structured as such:
- Categories, identifying the type of character through their UNICODE value; the list includes Greek letters (24 alphabetical signs + 3 letters with numeric value: ϛ “stigma” = 6, Ϡ “qoppa” = 90 and ϡ “sampi” = 900), one category “symbol”, and one category “unknown” for unidentified characters;
 - Licences, repeating the licences already mentioned in data.csv;
 - Images, identifying the images obtained through the downloader;
 - Annotations, identifying the area on the image in which one specific character appears; following information is given:
 - Bbox identifying the surface of the image where the character appears;
 - Category_id identifying the category to which the character belongs;
 - Unique id of the annotation;
 - Image_id identifying the image to which the character belongs;
 - BaseType tag, identifying the preservation status of the character:
 - Bt1: letter perfectly preserved;
 - Bt2: letter partially preserved but unequivocally identifiable;
 - Bt3: letter poorly preserved, traces ambiguous;
 - Bt4: letter partially preserved that allows for maximum two-three identifications; not used in this dataset, but see Marthot-Santaniello et al. 2024;
 - Bt5: letter deformed in its original appearance (e.g. squeezed, enlarged, modified to alter its meaning);
 - Zone, identifying the type of writing area in which the letter appears:
 - 0: Body: the main part of the text;
 - 1: Paratext: additional textual elements around the body;
 - 2: Addition: additional textual elements added by another scribe in another moment;
 - Rotation, indicating the rotation of the cliplet compared to the original image.
 - Zones, giving the name for the zone_id of the annotations.
7. **Description.pdf**, repeating the present description.

Specificities in the data collection

- The images are downloaded from the WWW or directly provided by the following collections:
 - [Archäologische Sammlung der Universität Zürich](#).
 - [Berliner Papyrusdatenbank](#).
 - [British Library Explore Archives and Manuscripts](#).
 - [Hamburger Kulturgut Digital](#).
 - [Heidelberg historic literature – digitized](#).
 - [Institut de Papyrologie de la Sorbonne](#).
 - [Kölner Papyri](#).
 - [Louvre Collections](#).
 - [Österreichische Nationalbibliothek Digital](#).
 - [Papiri dell'Università di Genova](#).
 - [Papyri in the Department of Papyrology, University of Warsaw](#).
 - [Photographic Archive of Papyri in the Cairo Museum](#).
 - [PSIonline](#).
 - [The Duke Papyrus Archive](#).

- [The Morgan Library and Museum Corsair Online Catalog](#).
- [The Oxyrhynchus Papyri](#).
- [The University of Manchester Library Digital Collections](#).
- [Turin Papyri Online Platform](#).
- [University of California Berkeley Library Digital Collections](#).
- [University of Columbia APIS](#).
- [University of Michigan Library Digital Collections](#).
- [Washington University Digital Gateway](#).
- The images are neither pre-processed nor harmonised concerning their resolution, colour scale, scale. However, we have converted the images to jpeg format for ease of downloading.
- Five images, labelled _GreekOnly, were cropped to remove their large Egyptian Demotic text and focus on the Greek text.
- The image of TM3563 was divided into nine individual images, one for each column of text.
- For some papyri, one of the two images has very little text on it. If pertinent, one can exclude it from the dataset.
- We have annotated as many characters as possible; however, not all characters are annotated. We have avoided annotating characters that were barely readable, and small characters on low-resolution images. Moreover, for the very long TM3563, we only annotated the first column.

Licenses

Users of this dataset must comply with the licenses provided by the various websites that give access to the images. Please take note that some of them do not allow reuse, or commercial reuse, of the images, and that credits are mostly required. By using this dataset, you confirm that you have read and understood the following licenses:

- Ann Arbor, Michigan University: CC Public Domain 1.0, <https://creativecommons.org/publicdomain/mark/1.0/>
- Berkeley, University of California: Permissions policies at the UC Berkeley Library, <https://www.lib.berkeley.edu/about/permissions-policies>
- Berlin, Staatliche Museen zu Berlin: Berlpap Nutzungshinweise, <https://berlpap.smb.museum/nutzungshinweise/>
- Cairo, Egyptian Museum Cairo: unknown license
- Cologne, Universität zu Köln: CC BY 4.0, <https://creativecommons.org/licenses/by/4.0/>
- Durham (NC), Duke University: CC BY-NC 3.0, <https://creativecommons.org/licenses/by-nc/3.0/>
- Florence, Biblioteca Medicea Laurenziana: PSOnline Autorizzazioni e diritti, <https://psi-online.it/rightpermission>
- Genova, Università di Genova: PUG Contatti, <http://www.pug.unige.net/Home/Contatti>
- Hamburg, Staats- und Universitätsbibliothek Hamburg: CC Public Domain 1.0, <https://creativecommons.org/publicdomain/mark/1.0/>
- Heidelberg, Universität Heidelberg: Heidelberg Institute for Papyrology Terms of Use, <https://www.uni-heidelberg.de/fakultaeten/philosophie/zaw/papy/images.html>
- London, British Library: Public Domain, <https://creativecommons.org/public-domain/pdm/>
- Manchester, John Rylands Library: unknown license
- New York, Columbia University: CC BY-NC 3.0, <https://creativecommons.org/licenses/by-nc/3.0/>
- New York, Pierpoint Morgan Library: unknown license
- Oxford, Art Archaeology and Ancient World Library: Rightsstatements in Copyright, <https://rightsstatements.org/page/InC/1.0/?language=en>
- Paris, Musée du Louvre - Antiquités égyptiennes: Louvre Terms of Use, <https://collections.louvre.fr/en/page/cgu>
- Paris, Sorbonne Université - Institut de Papyrologie: CC BY-NC 4.0 Sorbonne, <https://papyrologie.sorbonne-universite.fr/la-collection/conditions-dutilisation-des-photos/>

- St. Louis, Washington University: CC Public Domain 1.0, <https://creativecommons.org/publicdomain/mark/1.0/>
- Turin, Museo Egizio: CC0, <https://creativecommons.org/public-domain/cc0/>
- Vienna, Österreichische Nationalbibliothek: OeNB Nutzung, <https://www.onb.ac.at/nutzung>
- Warsaw, University of Warsaw: University of Warsaw Papyri Copyright, <http://www.papyrology.uw.edu.pl/copyright.htm>
- Zurich, Archäologische Sammlung der Universität Zürich: © Archaeological Collection, University of Zurich, inv. 1904. Photo: Frank Tomio. CC BY-NC 4.0, <https://creativecommons.org/licenses/by/4.0/>

References

The research behind Hell-Date would not have been possible without the data provided by Papyri.info (<https://papyri.info/>, CC BY 3.0), Trismegistos (<https://www.trismegistos.org/>, CC BY-SA 4.0) and PapPal (<https://pappal.info/> - many thanks to Rodney Ast for sharing the data).

For more details about Hell-Date and to credit it, please quote, in addition to this repository, the following publication: G. De Gregorio, L. Ferretti, R.C.G. Pena, I. Marthot-Santaniello, M. Konstantinidou, J. Pavlopoulos, “A New Framework for Error Analysis in Computational Paleographic Dating of Greek Papyri”, in Mouchère, H., Zhu, A. (eds) *Document Analysis and Recognition – ICDAR 2024 Workshops. ICDAR 2024. Lecture Notes in Computer Science*, vol 14936. Springer, Cham, 2024. https://doi.org/10.1007/978-3-031-70642-4_7.

For a related dataset, see I. Marthot-Santaniello, O. Serbaeva, S. White, S. Agolli, M. Seuret, G. Carrière, D. Rodriguez-Salas, V. Christlein, “ICDAR2023 Competition on Detection and Recognition of Greek Letters on Papyri”, Zenodo, 2024. <https://doi.org/10.5281/zenodo.13825619>.