

# Data Management Plan - Bibliotheca Eugeniana Digital (BED)

| I General Information  |  |
|--|--|
| <b>I.1 Administrative information</b>                                  | <p>Funding Authority: Austrian Academy of Sciences (ÖAW)</p> <p>Call: <a href="#">Go Digital 3.0</a></p> <p>Project-Number: GD3.0_2021-37_BED</p> <p>Project-Title: Bibliotheca Eugeniana Digital</p> <p>Coordinator: Simon Mayer (ONB)</p> <p>Principal Investigators: Simon Mayer (ONB), Florian Windhager (UWK)</p> <p>Host Institutions: Austrian National Library (ONB), Danube University Krems (UWK)</p> <p>Project Duration: 2022-11-01 until 2024-10-31</p> <p>Version: 1.0; updated DMP for ongoing project</p> <p>Last update: 2024-07-01</p> <p>Project Objectives: Virtual reconstruction of Prince Eugene of Savoy's library using digital humanities and computer science methods. BED will present the results in different access points for exploration and data usage: interactive visualisations, enriched metadata in ONB's open access catalogue, a Linked Open Data set and a digital edition of archival sources.</p>  |
| <b>I.2 Data management responsibilities and resources</b>              | <p>The DMP and its correct implementation is supervised by the principal investigators (PIs) of the project. The DMP will be evaluated and if necessary revised by the PIs every 6 months during project runtime. Each team member can suggest changes to the DMP, which will be discussed and incorporated by the PIs, since each project partner institution will be represented by the respective PI. Versioning of this document will be carried out using GitLab infrastructure of the project (see below).</p>   |
| II Data Characteristics  |  |
| <b>II.1 Data description and collection or re-use of existing data</b> | <p>For the machine learning (ML) tasks, only material that already has been digitised by the Austrian National Library will be processed (Austrian Books Online corpus). The scans are available in the format JPEG2000, which is a widely used standard for image interoperability. Additionally, metadata and OCR data (in hOCR format) are provided. Training data to detect the supralibros of Bibliotheca Eugeniana are created manually. The designated format for storing the training data is either CSV or JSON since these formats are easy to process for machines and understandable to humans. In addition, these formats work very well with versioning and allow for the evolution of training data to be tracked. For training the models, the project will use open-source ML suites (e.g., Tensorflow). Trained models are either stored as binary data or as sets of weights for the neuronal networks. In general, there are two kinds of research data for the ML component of this project which consist of the following components:</p> <ol style="list-style-type: none"> <li>1. ML training <ol style="list-style-type: none"> <li>a. data to train the ML model</li> <li>b. data to test the ML model</li> <li>c. source code for training</li> </ol> </li> </ol> |

|   |  |
|---|--|
|   | <ul style="list-style-type: none"> <li>d. ML model configurations parameters</li> <li>e. the trained model</li> </ul> <p>2. ML results</p> <ul style="list-style-type: none"> <li>a. list of images to test</li> <li>b. results for each image (classification)</li> </ul> <p>Storage needed for the ML task in this project is between 50GB - 100GB. Especially the trained models can take bigger amounts of hardware storage.</p> <p>Data from the transcription and the indices from the Digital Edition of the handwritten Eugeniana catalogue is stored as TEI-XML. TEI is the leading and widely used standard for text encoding and is a requirement for the publication in the <i>Sustainable Infrastructure for Digital Editions at ONB</i> (edition.onb.ac.at). The initial transcription is carried out using the Transkribus platform, which is widely used for handwritten text recognition. Transkribus uses PAGE-XML for storing data, but also provides export options to TEI-XML. The online presentation of the Digital Edition is carried out using XSL transformations that incorporate the use of JavaScript and CSS. Apart from the scans of the handwritten catalogue, the necessary storage for the Digital Edition will be roughly 500MB. Facsimiles of the catalogue will have been produced before project runtime and will be available via an IIIF endpoint. IIIF (International Image Interoperability Framework) is a de-facto standard for image interoperability widely used in the GLAM sector and supported by ONB infrastructure. The designated format of the images is JPEG and roughly 1GB storage is used for them. ONB holds the original of the handwritten catalogue.</p> <p>User-centred design: The protocols and data from the elicitation of user-requirements and the evaluation will be stored as qualitative text and quantitative numerical data in an anonymized form on the project's document sharing system. These raw data will not be shared openly.</p> |
| <b>III Documentation and Data Quality</b> |  |
| <b>III.1 Metadata and documentation</b>   | <p>Bibliographical metadata created by the project with descriptions of identified Eugeniana items will be integrated in ONB's public open access catalogue and accessible for end-users via the user interface Quicksearch. In addition, the project will provide sustainable access to machine-readable enriched bibliographical metadata via SRU, OAI-PMH API and as a LOD set, aligned with the DARIAH collection description model via ONB's Labs infrastructure (labs.onb.ac.at). The LOD set also comprises metadata about its extent, structure, provenance and versions.</p> <p>Metadata on the Digital Edition will be encoded in the TEI-XML header.</p> <p>Documentation for all software code developed will be maintained in ONBs Labs open GitLab repository.</p>   |
| <b>III.2 Data quality control</b>         | <p>Quality of the transcription of the handwritten catalogue is monitored using character error rates as they are calculated by Transkribus. Further processing of the materials for the digital edition is done in TEI-XML where a project specific XML schema guarantees the conformity of the data.</p> <p>Bibliographical metadata stemming from the library catalogue and from the transcription of the handwritten catalogue will partially be checked manually, partially by means of string matching. Subject classifications using ANNIF will be automatically compared to subject headings already available in the library's public open access catalogue. Erroneous elements in the MARC datasets will be detected by automated quality checks, erroneous elements in the TEI transcription by validating the transcription against a TEI schema. The information visualisations generated in BED will also be used for quality assurance (detecting outliers or gaps).</p>  |

|   |  |
|---|--|
|   | <p>All newly developed software components are equipped with adequate unit tests. The tests are stored inside the ONB Lab GitLab repository and are evaluated automatically using continuous integration pipelines for software development.</p> <p>For machine learning tasks, the quality of the trained model is analysed by providing validation data for the model. Only a certain amount of the annotated training data is actually used for training of the ML model. The remaining part is used to test the model and because of the annotations, it will be possible to quantify the error rate of the model. The validation whether the model was trained for the desired image feature needs to be carried out manually by project members.</p> <p>The creation of the questionnaire to elicit user requirements will integrate questions to assess the specifics of the local assembly of data, users, and tasks (Miksch &amp; Aigner, 2014), design options from international state of the art interfaces, and questions for dissemination and exploitation preferences. The analysis of the results - as well as the analysis of evaluation results - will follow a mixed methods approach. To ensure quality of qualitative analyses, two researchers will code the protocols independently from each other. The dealing with - and representation of - historical quality and accuracy of the BE collection data will itself be a central part of the BED project. We aim to make these information layers visible with the means of collection visualisations to support a critical-historical assessment of both the data and its representation - and to strengthen the trust in DH tools of both expert users and public audiences.</p>                                   |
| <b>IV Data Storage, Sharing, and Long-Term Preservation</b>     |  |
| <b>IV.1 Data storage and backup during the research process</b> | <p>Research data is stored and versioned inside the GitLab infrastructure at ONB in multiple projects. Data processing is done either on local computers from project members, on dedicated servers from the institution or on third-party cloud resources provided by the institution. Thus, certain data may be stored temporarily on any of these three kinds of devices during processing for better performance.</p> <p>Scans and OCR data of the ABO Corpus are stored in ONB's Digital Repository. The Digital Repository is maintained and backed up by ONB's Department for IT Services and Solutions. Scans from the books are accessed from the Digital Repository via the IIIF infrastructure at the ONB.</p> <p>GitLab and IIIF infrastructures are maintained by the Department for Research and Development together with the Department for IT Services and Solutions at ONB. Backups for both infrastructures are created on a regular basis.</p> <p>In case of failure, the GitLab infrastructure can be restored using backups. During a downtime, all team members can still work on their local machines and results can be synchronised afterwards. In case of a longer emergency, the individual projects temporarily can be migrated to another git-Service easily.</p> <p>The initial transcription of the handwritten catalogue is stored at the third-party service provider Transkribus and therefore relies on the data recovery plan of Transkribus. After a TEI-XML export, further output of the digital edition is stored in the ONB Labs GitLab repository.</p> <p>All team members have access to the data in the GitLab projects. For each team member authentication via a personal password or authentication token is necessary to access the data.</p> |
| <b>IV.2 Data sharing and long-term preservation</b>             | <p>Core components of this project will be accessible via the GitLab infrastructure provided by ONB Labs. All relevant project data is stored in multiple projects which are associated to the same dedicated namespace. This includes source code developed during the project runtime, training data for ML applications as well as the results. Due to its size, pretrained models are not stored in a Git environment and are published using a web server instead but will be referenced in the GitLab repositories. Further, the configurations to run the developed visualisation as well as the transcription of the handwritten text catalogue are also versioned using the GitLab infrastructure.</p>  |

|                                    |  |
|------------------------------------|--|
|                                    | <p>The conditions for reuse of source code or research data are indicated using a licence file and an according badge in the readme file of each GitLab repository. Resources like source code or other relevant data (ML training data, TEI-XML files for the transcriptions, ...) are published using a CC-BY licence wherever possible.</p> <p>Each repository contains a README file with information about the project and where to find the software documentation of the repository for better reusability. The documentation is either available via the Wiki of the GitLab project or is part of the repository data. Reusability of the developed software is guaranteed by either listing all necessary Python packages to set up an individual Python environment to execute the prepared scripts or by increasing portability using containerised applications with Docker. In the latter case all information on how to build the container to run certain software will be provided but it is not planned to publish prebuilt containers.</p> <p>Image data is accessible via ONB's IIIF infrastructure that guarantees IIIF conformant access to all processed facsimiles.</p> <p>The digital edition will be published in the <i>Sustainable Infrastructure for Digital Editions at ONB</i> that relies on GAMS (Geisteswissenschaftliches Asset Management System), which is a Fedora data repository and is used for all digital edition projects at the ONB. The transcriptions and registers in TEI format are ingested into the repository and can be retrieved by their ID.</p> <p>All objects that can be identified as part of the Bibliotheca Eugeniana are tagged accordingly in ONB's public open access catalogue. Thus, the entire collection is accessible via the SRU and OAI-PMH API of the catalogue and metadata can be retrieved in either MARC-XML or Dublin Core format. The catalogue also guarantees an ID for each object. Extended Linked Open Data support is provided by information in BIBFRAME, JSON-LD and RDA/RDF format.</p> <p>In addition, the whole collection is also registered in the DARIAH-DE collection registry including a description of the dataset and the stakeholder using the predefined vocabulary of the DARIAH-DE collection registry. The resulting entry lists the methods to access the collection data as well.</p> |
| <b>V Legal and Ethical Aspects</b> |  |
| <b>V.1 Legal aspects</b>           | <p>The entire collection of Bibliotheca Eugeniana is public domain and therefore research on this corpus does not underlie further legal conditions.</p> <p>The scans of the books from the ABO corpus have been created in a public private partnership between ONB and Google and can be used for research purposes.</p> <p>The results of the user study by UWK are available for all team members, but personal data is stored only in an anonymised fashion to protect participants' personal rights.</p>   |
| <b>V.2 Ethical aspects</b>         | <p>No ethical issues are expected for this project as there is no processing with regards to the content of the individual books of Bibliotheca Eugeniana planned. Nevertheless, the historical collection will be contextualised and described to give background information about the data for BED.</p> <p>Regarding the user studies for the visualisation interface the ethics committee of the UWK is involved, participants are asked for their informed consent, and results are only stored and published in an anonymous way.</p>  |