www.snf.ch
Wildhainweg 3, Postfach 8232, CH-3001 Bern

Data management plan (DMP)

1 Data collection and documentation

1.1 What data will you collect, observe, generate or reuse?

1. Digitised primary sources:
1.1. <*Source 1*>: 20'000 pages, png, 50 GB
1.2. <*Source 2*>: 600+ pages, png, 1-2 GB

2. Transcriptions of primary sources:
2.1. Ground truth of <*Source 1*>: 300-500 pages, xml, 100-150MB
2.2. OCR-output: 25'000+ pages, xml, 5GB

3. Named Entity Recognition (NER)/Relation Extraction:
3.1. Ground truth for NER: 300-500 pages, txt, volume negligible
3.2. NER-output: amount depending on output, txt, volume negligible
3.3. Ground truth for relation extraction: 300-500 pages, txt, volume negligible
3.4. Relation extraction output: amount depending on output, txt, volume negligible

4. AI-models:
4.1. OCR-model: Volume depends on architecture (several GB)
4.2. language model finetuned for NER: Volume depends on architecture (several GB)
4.3. language model finetuned for relation extraction: Volume depends on architecture (several GB)

5. Re-used datasets:
5.1. <*Source 3*>: 5'000+ pages, png , 12+ GB
5.2. Pre-trained large language model for finetuning: exact model to be evaluated, several GB
5.3. Text corpora in middle high German / early modern German for model finetuning: corpora TBD, several GB.

1.2 How will the data be collected, observed or generated?

1. Digitized primary sources:
Primary sources will be photographed, preferably using a ScanTent. <*Source 1*> will be grouped by <*Properties*>. For <*Source 2*> further organizing will not be necessary at this stage.

2. Transcriptions of primary sources:
Ground truth (2.1) will be manually produced using the synoptic editor in Transkribus, which allows for export to xml standards relevant for further usage. OCR-output will be generated using the trained OCR-model (4.1). Transcriptions are also organised by type of book and year.

3. NER/Relation Extraction:
Ground truth production (3.1 & 3.3) is carried out by manual annotation. In this process I will follow annotation guidelines to be agreed on with <*Collaborator*> and other scholars working on sources posing similar problems for annotation. 3.2 & 3.4 are generated using the respective models (4.2 & 4.3).

4. AI-models:
All models will be trained based on the respective training data, constantly evaluated, and perpetually improved. For version control, git (GitHub) will be used.

5. Re-used datasets:
<*Source 3*> was already digitised by <*Archive*> and can be reused. The parts to be used will be grouped by <*Properties*>.

Pre-trained language models (5.2) are published on huggingface (https://huggingface.co/) and can be reused from there. The text corpora to be used for finetuning (5.3) can be accessed through the respective project portals.

1.3 What documentation and metadata will you provide with the data?

All high level datasets will be provided with their own documentation.
In addition to the usual metadata such as contributor(s), version number and date of training, each model (4.1-4.3) comes with their distinct readme files informing the user about the purpose of the model, the pre-trained model it is based on, the data used for finetuning, the amount of tokens it contains, as well as its performance measures.
The documentation of ground truth subsets (2.1, 3.1 and 3.3) will include their provenance, structure and an indication of the applied transcription/annotation guidelines, in addition to the standard metadata (contributor, persistent identifier, date etc).
The documentation will be shared alongside the respective datasets as outlined in section 4.1 of this DMP. In this step, additional metadata will be generated in concordance to the respective repositories.

2 Ethics, legal and security issues

2.1 How will ethical issues be addressed and handled?

As no data and models collected, generated and trained in the course of the proposed project includes sensitive information about people that are still alive, there are no ethical concerns that need adressing.

2.2 How will data access and security be managed?

As outlined in Section 2.1 of this DMP, no sensitive data will be collected or generated.

2.3 How will you handle copyright and Intellectual Property Rights issues?

All data collected for the proposed project is based on primary sources, which are so old they are not covered by copyrights or archival retention terms anymore.
As an employee of the University of Bern, it is likely that the University will be the owner of the intellectual property rights of the data collected and generated in my project. For information about the licensing scheme of the data to be shared, please refer to section 4.1 of this DMP.
The pre-trained language models (5.2) are usually open source publications and can be re-used with very little limitations that should pose no problems to comply to.
Additional text corpora (5.3) will only be re-used in accordance to the licence they are published under.

3 Data storage and preservation

3.1 How will your data be stored and backed-up during the research?

During the research project, all data will be stored and backed-up threefold: on my laptop, an external SSD and on a Campus Storage or Research Storage share provided by the University of Bern. Automatic backups to the Server will occur daily. Backups to the external SSD will be carried out at least twice a week.

3.2 What is your data preservation plan?

After the project has been completed, all datasets and models (except the third-party datasets and models) will be archived by the Campus Archive infrastructure of the University of Bern, which guarantees safe archiving of the data for 10 years. The contact persons responsible for data preservation, as well as the data preservation beyond those 10 years are yet to be determined.
They will be stored in the file format they have been generated in (png, txt, xml), as these are suitable for long-term storage.

4 Data sharing and reuse

4.1 How and where will the data be shared?

Datasets and subsets 1, 2 and 3 will be shared on an open repository like BORIS Portal or Zenodo. The exact repository is still to be determined.

The trained models will be published under a MIT licence on huggingface (https://huggingface.co/), the most

popular model sharing platform for AI-tasks.

All manually and automatically produced transcriptions (2.1 & 2.2) will additionally be shared on transcriptiones (https://transcriptiones.ch/). While transcriptiones is not a repository, but rather a platform for sharing and editing of transcriptions, publishing transcriptions there too will increase the visibility and reusability of these datasets. This also means that the transcriptions will be shared under a CC0 license.

4.2 Are there any necessary limitations to protect sensitive data?

There are no limitations to data sharing.

All data required for understanding and reproduce the findings published over the course of the project will be made available at the time of publication.

4.3 All digital repositories I will choose are conform to the FAIR Data Principles.

Yes

4.4 I will choose digital repositories maintained by a non-profit organisation.

Yes