Internet Technologies and Systems
Prof. Dr. Christoph Meinel
Multimedia and Machine Learning Research Group
Master Project Summer Term 2022

# Developing Language Identification for Art-Historical Documents

## Background

Art-historical archives often contain large amounts of documents, such as auction catalogues, letters, or newspapers. These documents enhance our understanding of both the artists themselves and their work. Providing them in a digital way can make them more accessible to art enthusiasts and researchers worldwide. In addition, the digital format allows for new ways of searching and filtering, provided the necessary metadata is available.

## Problem

Machine learning methods can help acquire this metadata automatically, without human work. One such method is language identification. It is useful both on its own, e.g., by enabling the filtering of documents by language, and as a preprocessing step for other tasks, such as optical character recognition (OCR).

Many state-of-the-art machine learning methods rely on large amounts of labeled training data. However, while we do have a large dataset of historical documents given to us by our project partner, the Wildenstein Plattner Institute (WPI), we do not have labels available. Thus, this project aims to solve the language identification task without labels.

## Goal

The goal of this project is to develop a method for language identification that does not require a large, labeled dataset. For this we are going to:

- Familiarize ourselves with the state of the art in the field
- Compare different approaches with varying degrees of supervision and decide on the most feasible
- Implement the selected approach and evaluate its performance on the data given to us by the WPI
- Develop a prototype to showcase the results

## Studiengänge

- IT-Systems Engineering
- Data Engineering
- Digital Health
- Cybersecurity

## Advisors

Prof. Dr. Christoph Meinel
Jona Otholt, Hendrik Rätz, PD Dr. Haojin Yang