SHIFT

MetamorphoSis of cultural Heritage
Into augmented hypermedia assets
For enhanced accessibiliTy
and inclusion

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Document info

Document ID:	D4.3 - Tools for Cultural Asset Curation and features extraction - final version
Version date:	31.03.2025
Total number of pages:	56
Abstract:	This deliverable presents the development of advanced tools for pre-processing, feature extraction, and curation of cultural heritage assets across textual, visual, acoustic, and haptic modalities, contributing to the creation of an enriched SHIFT information corpus. It details the exploration of diverse machine learning models tailored to extract key features from cultural assets. Furthermore, the deliverable emphasizes the integration of external knowledge repositories and the interlinking of cultural assets, enabling the creation of meaningful connections and enhancing the storytelling potential of cultural content. Through these efforts, this deliverable advances the SHIFT project's goal for better accessibility and inclusion of cultural heritage through innovative computational approaches.
Keywords:	Cultural Heritage, Memory Twin, Pre-Processing, Feature Extraction, Interlinking, Multimedia Curation, Smell Extraction, Emotion Classification, Sensory Experience Curation, API Integration

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Authors

Name	Organisation	Role
Björn Schuller	TUM-MED	Researcher
Anika Spiesberger	TUM-MED	Research Assistant
Iosif Tsangko	TUM-MED	Research Assistant
Beatriz Aretz	TUM-MED	Research Assistant
Katerina Valakou	FORTH	Researcher
Allister Carter	MDS	Researcher
Dionyssos Kounadis-Bastian	AUD	Researcher
Felix Burkhardt	AUD	Director of Research

Reviewers

Name	Organisation	Role
Dionyssos Kounadis-Bastian	AUD	Researcher
Katerina Valakou	FORTH	Researcher
Andreas Bienert	SMB	Researcher
Moritz Maier	DBSV	Researcher

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Version History

Version	Description	Date
0.1	Table of Content	02.12.2024
1.0	First Draft	10.03.2025
1.1	Revision First Draft	17.03.2025
2.0	Second Draft	24.03.2025
2.1	Revision Second Draft	28.03.2025
3.0	Final Version	31.03.2025

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

EXECUTIVE SUMMARY

In the convergence of cultural heritage (CH) and cutting-edge technology, SHIFT seeks to realize a transformative initiative that unravels the profound narratives embedded within CH assets. Through the meticulous development of tools for pre-processing, feature extraction, and information curation, our endeavour contributes to the SHIFT information corpus, reshaping how cultural entities are explored.

Deliverable 4.3 (D4.3) builds upon the work presented in Deliverable 4.1 (D4.1) by expanding the scope of cultural asset analysis to encompass a more advanced approach to curation and interlinking of multimedia cultural assets. While D4.1 focused on pre-processing techniques and feature extraction, D4.3 shifts its focus toward the curation of CH assets, interlinking them across different dimensions, and enhancing their storytelling capabilities.

A key component of D4.3 is the development of curation tools that allow for automatic and enhanced storytelling based on visual, textual, auditory, and sensory attributes. These tools enable the creation of richer and more immersive experiences by allowing cultural artefacts to be linked not only to one another but also to external CH repositories. The ability to associate cultural assets thematically, stylistically, or by sensory experiences provides an enriched and multifaceted understanding of the content, transforming isolated artefacts into interconnected narratives that span across time and context.

Within the scope of feature extraction, we further refined our methodologies, building advanced models to detect and extract key characteristics from the cultural assets in our dataset. These include the development of models for textual analysis, image recognition, and audio processing.

In addition, a key contribution of D4.3 is the integration of external knowledge repositories (such as Europeana) into our curation tools. By linking cultural artefacts to these external databases, we create a rich context for users, enabling them to engage with artefacts not only as standalone objects but as part of a much larger narrative that spans across cultures, regions, and periods. This connection not only enhances the depth of information available to users but also facilitates the discovery of new insights through cross-referencing and thematic exploration.

Through the combination of these pre-processing, feature extraction, and curation tools, D4.3 lays the foundation for a more scalable and interactive CH experience. By creating workflows that link cultural assets across different domains, we enhance the accessibility and discoverability of cultural content, offering new opportunities for museums, educators, researchers, and the broader public to engage with our shared cultural history in innovative ways.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Ultimately, this deliverable significantly contributes to the overarching goal of the SHIFT project: to create a robust, interoperable, and future-proof framework for the curation and exploration of CH that is as rich, diverse, and interconnected as the cultural assets themselves.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

TABLE OF CONTENTS

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Abbreviations and Acronyms

A/D/V	Arousal, Dominance, and Valence
API	Application Programming Interface
BS-Net	Background Subtraction using DL
CH	Cultural Heritage
CRM	Conceptual Reference Model
CRMsci	Scientific Observation Model
CRMdig	Digital Provenance Model
DL	Deep Learning
EDM	Europeana Data Model
GMM	Gaussian Mixture Model
KNN	K-Nearest Neighbors
LIDO	Lightweight Information Describing Objects
MDS	Museum Documentation System
ML	Machine Learning
MOG2	Mixture of Gaussians Version 2
OCR	Optical Character Recognition
SAM	Segment Anything Model
SER	Speech Emotion Recognition
TTS	Text-to-Speech
UAR	Unweighted Average Recall
WP	Work Package
YOLO	You Only Look Once

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

LIST OF FIGURES

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

1. INTRODUCTION

Cultural Heritage (CH) represents the collective memory of societies, encompassing tangible and intangible artefacts that define human history, art, and identity. In the digital age, cultural asset preservation, analysis, and dissemination have become increasingly reliant on advanced computational techniques. The SHIFT project is at the forefront of this transformation, developing innovative methods for processing and curating cultural content that enhance accessibility and foster inclusion, particularly by supporting curators in creating enriched, multi-perspective narratives.

This deliverable, focusing on Work Package (WP) 4 and specifically Tasks 4.1 and 4.2, presents the final version of tools developed for the pre-processing, feature extraction, and curation of multimedia cultural assets. These tools enable the automated extraction of meaningful information, such as text recognition, metadata generation, and content association, while integrating state-of-the-art Deep Learning (DL) methods and external knowledge repositories. By leveraging these technologies, the SHIFT project contributes to the creation of *memory twins* – digital counterparts of cultural assets that preserve not only their visual and historical attributes but also their contextual significance – which can support different levels of content interpretation and engagement.

By leveraging these methodologies, SHIFT enables the curation of cultural assets that are interlinked across various dimensions – such as theme, style, and sensory experience – while ensuring interoperability with external CH databases. This approach facilitates more inclusive curation practices, enabling curators to organize and present cultural assets in ways that accommodate diverse audiences, including those with disabilities. This deliverable outlines the tools and approaches used to enhance the accessibility, contextual relevance, and overall usability of CH materials.

The tools presented in this deliverable support the SHIFT goal of enhancing the accessibility, interpretability, and value of CH assets for diverse audiences. While these tools primarily serve curators and CH professionals, they contribute to broader accessibility goals by enabling more nuanced and multimodal representations of cultural content, which are leveraged in the tools developed in WP3 to improve engagement for different user groups.

1.1. SCOPE AND OBJECTIVES

The scope of this deliverable is to present the final version of the tools developed for the pre-processing, feature extraction, and curation of cultural assets within the SHIFT project. These tools are designed to support the systematic analysis and enrichment of CH materials, enabling their integration into the SHIFT information

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

corpus. The focus lies on interlinking cultural assets based on visual, textual, auditory, olfactorily, and haptic characteristics.

The key objectives of this deliverable are:

- To develop and refine pre-processing techniques for extracting meaningful information from CH artefacts.
- To implement feature extraction methodologies that capture a broad range of characteristics, facilitating the enrichment of cultural assets.
- To enable association by design by developing tools that interconnect cultural artefacts based on thematic, stylistic, and sensory attributes, thereby enhancing storytelling and contextual understanding.
- To integrate external knowledge sources as ontologies and controlled vocabularies to enrich metadata and create meaningful links between digital heritage objects.

By achieving these objectives, this deliverable establishes a robust and scalable framework for curating CH assets in a way that maximizes their usability and interconnectivity.

1.2. STRUCTURE OF THE REPORT

This report is structured to provide a comprehensive overview of the tools and methodologies developed for cultural asset pre-processing, feature extraction, and curation within the SHIFT project. The document is organized as follows:

1. Introduction: Provides an overview of the deliverable, outlining its scope, objectives, and structure.
2. Pre-Processing: Describes the techniques used to prepare CH materials for further use.
3. Feature Extraction: Explores the methodologies developed to analyse and extract characteristics from multimedia content, enabling its enrichment.
4. SHIFT Curation Tools: Presents the tools designed for the structured curation of cultural assets, covering different approaches such as thematic grouping, content-based association, sensory-based classification, and interlinking using external databases.
5. Conclusion: Summarizes the key findings and contributions of this deliverable and highlights the development of cultural asset curation within the SHIFT project.

This structure ensures a logical progression from foundational processing techniques to advanced curation methodologies, providing a clear and detailed account of the developments achieved in WP4.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2. PRE-PROCESSING

Pre-processing is a crucial step in the pipeline of cultural asset analysis, serving as the foundation for subsequent feature extraction, content enhancement, and curation. Raw data in the form of images, text, audio, and video is often not suitable for direct use in planned tasks. By applying targeted pre-processing techniques, we ensure that cultural assets are in an optimal format for further processing, improving both the efficiency and reliability of downstream tasks.

In the context of the SHIFT project, pre-processing involves a variety of tasks tailored to the specific nature of each data type. For image data, we apply resizing methods to standardize input sizes for Machine Learning (ML) models, ensuring consistent analysis across the dataset. Text data from scanned books is processed using Optical Character Recognition (OCR) to extract machine-readable text, enabling further linguistic and semantic analysis. Audio data is carefully cleaned and prepared for Speech Emotion Recognition (SER) while the SER model on its own has been re-developed during SHIFT to be robust in real-world actual-life non-acted emotion recognition. These steps lay the groundwork and provide textual input for the development of an affective Text-to-Speech (TTS) tool for audio-narration of CH assets in deliverable D3.6.

These pre-processing steps are essential for handling the diverse and often unstructured nature of CH data. They allow us to extract meaningful information, harmonize datasets, and establish links between artefacts. By implementing robust pre-processing workflows, we facilitate high-quality feature extraction in Task 4.1 and effective asset curation in Task 4.2.

The following sections provide a detailed overview of the specific methods and tools employed for image resizing, text extraction, and audio data cleaning, illustrating their impact on enhancing data usability within the SHIFT platform.

2.1. RESIZING IMAGES

Image data in CH archives comes in diverse resolutions, aspect ratios, and formats. To ensure consistency and facilitate efficient processing, we apply a standardized resizing procedure while preserving the original aspect ratio where possible. This is essential for maintaining the integrity of visual features while ensuring compatibility with our feature extraction and curation workflows.

The primary objective of resizing is to create a uniform input size for ML models and analytical tools. Variations in image dimensions can lead to inconsistencies in feature representation, increased computational costs, and potential distortions in learned embeddings. However, the resizing approach varies depending on the requirements of subsequent algorithms:

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

- Aspect Ratio-Preserving Resizing: For some pretrained classification models, such as Artemis, it is crucial to maintain the original aspect ratio to prevent distortions. In these cases, images are resized such that the longer side does not exceed a predefined maximum (e.g., 600px), ensuring uniform scaling while retaining visual integrity.
- Fixed-Size Resizing: Some ML models, particularly object detection frameworks like YOLO, require input images of a fixed size due to the constraints of their training process. For these tasks, all images are resized to a standard dimension (e.g., 620×620 pixels) regardless of their original aspect ratio. This ensures compatibility with pre-trained models while maintaining consistency across the dataset.

Beyond computational benefits, resizing also plays a role in harmonizing datasets. Many DL architectures and visualization tools expect standardized input dimensions, and a consistent image size facilitates easier integration with text, audio, and metadata. Additionally, for datasets containing high-resolution historical artworks or scanned documents, resizing prevents unnecessary redundancy while retaining the most salient visual details.

This pre-processing step ensures that all images used in subsequent analysis are well-structured and appropriately formatted for each task, maximizing the effectiveness of our DL workflows.

2.2. EXTRACTING TEXT FROM SCANNED BOOKS

Many historical texts and literary works exist in scanned formats, often as PNG or JPG images embedded in PDFs. These must be converted to Unicode's as if they would have been typed from a keyboard, because Language Model's input modality is universally Unicode's characters rather than jpg images. While these scanned documents preserve the visual integrity of the original sources, they lack machine-readable text, making automated analysis and retrieval challenging. To enable text processing methods as implemented in WP3, it is essential to extract textual information from these scanned books using OCR techniques.

To achieve high accuracy in text recognition, OCR models must be fine-tuned on domain-specific datasets. In the SHIFT project, OCR is applied to historical Romanian texts, which include both Latin and Cyrillic scripts and are partially handwritten. Traditional OCR models like Tesseract [smith07] struggle with archaic fonts and degraded print quality, requiring solutions through DL-based approaches [lombardi20], such as TrOCR [li23b], and EasyOCR[1]. Fine-tuning these models involves training on synthetic datasets generated to simulate realistic historical document conditions (e.g., faded ink or font styles). Additionally, the OCR pipeline

[1] https://github.com/JaidedAI/EasyOCR

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

integrates post-processing techniques such as lexicon-based corrections and language modelling to further improve accuracy. Once OCR extracts text, NLP methods refine and process the content.

OCR allows us to convert scanned images into structured text, facilitating tasks such as linguistic analysis, metadata generation, semantic linking, and text processing (see Deliverable 3.5). This process is particularly relevant for historical documents, where text may appear in varying fonts, layouts, and even handwritten scripts.

Extracting text from scanned books not only enhances searchability and accessibility but also enables advanced applications like automated translations, sentiment analysis, and content-based curation. This step plays a crucial role in integrating textual content into the SHIFT platform.

2.3. AUDIO DATA PREPARATION AND CLEANING VIA DISTILLATION TRAINING

Pre-processing of audio datasets is internalised in the sense of having a large teacher model annotating raw audio for student training for developing the SHIFT TTS SER component of the Audio tool of Deliverable 3.6.

This section outlines the rationale of employing knowledge distillation as an automated audio pre-processing procedure instead of manual curation/filtering of audio dataset.

Today, SER is shifting towards dimensional annotations of Arousal, Dominance, and Valence (A/D/V) [fontaine07], as instance-level measures as the L2 distance prove unsuitable for evaluating A/D/V accuracy due to non-converging consensus of annotator opinions [schlossberg54]. People focus increasingly on correlation of SER model with cross-lingual annotator agreement [morgan19]. Even further SER requires high computational resources to overcome the scarcity of well annotated audio datasets [goncalves24].

However, dimensional annotations of arousal, dominance, and valence (A/D/V) where arousal indicates voice excitement, dominance reveals surprise, and valence shows pleasantness or perceived positivity/negativity [fontaine07]. Dimensional annotations (A/D/V) reveal more stable correlations compared to annotations of discrete emotional categories of “Anger”, “Happiness”, etc. [morgan19].

As instance-level measures as the L2 distances between predicted/annotated A/D/V values are still unstable for evaluating A/D/V accuracy due to subjective/non-converging consensus of annotator opinions [wagner23]. Concordance Correlation Coefficient (CCC) has risen as a measure where A/D/V predictions are evaluated to match a whole dataset’s correlation for

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

annotations/model-predictions rather than L2 distances or predicted/annotated A/D/V values for individual audio samples.

This enables cross-lingual/cross-cultural SER neural networks to be trained. Recent studies have shown that Wav2Vec2/WavLM architectures achieve today's highest CCC for real world non-acted emotional speech [wagner23, goncalves24]. The Wav2Vec2/WavLM family has a high computational footprint.

Training SER models on ground-truth emotion annotations is difficult because augmentation invalidates the label: If we are provided by a speech-audio saying, "yeah sure" and we infuse a silence, so it sounds as "yeah,__,sure" the perceived emotion has switched from happiness to contempt, while re-annotation is non-trivial. Thus, augmentations of audio do not preserve the human annotation of perceived emotion, whereas in computer-vision by resizing an image it still preserves the label "car" if there was a "car" in the original image.

Without augmentation of SER datasets, it is difficult to train deep learning SER models as there are not enough annotated audio samples. However, there is a way to allow augmentations of audio for SER, by using an additional "Teacher" model that runs in-parallel during training and provides A/D/V predictions for augmented audio. Those predictions of Teacher are used as annotations of emotion for the augmented audio, during training of a student A/D/V model. During SHIFT, as we present in D3.6 we build a novel Teacher model that reaches state-of-the-art A/D/V performance [kounadis-bastian24] as well as 5 low-computational resources student models.

For SHIFT we use our A/D/V Teacher for benchmarking/assessing the affectivity of the TTS voices produced for SHIFT TTS tool, presented in detail in deliverable D3.6. In parallel we also researched the sources of uncertainty of SER models in [schrufer24]. As well as filtering annotation bias in audio data for our publication [burkhardt24]. For a unified presentation of the A/D/V Teacher for the SHIFT TTS tool and its application for quantitative evaluation of affective TTS voices via SER, we refer the reader to D3.6.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

3. FEATURE EXTRACTION

The extraction of meaningful features is a central objective of Task 4.1. The goal is to use these features to enrich cultural content, enhance storytelling by providing deeper insights into artefacts, and facilitate curation by design. Features, in this context, refer to specific characteristics, properties, and attributes that can be derived from cultural assets through computational methods.

A wide range of features were identified, categorized, and selected for further investigation (see D4.1 for a more extensive account). Since CH assets span different modalities – e.g., text, images, video, and audio – the extraction methods must be adapted accordingly.

In our previous work, we categorized features into factual, contextual, and subjective information. Factual details, such as names, dates, materials, and styles, serve as the foundation for organizing and linking artefacts across repositories. Contextual information adds layers of meaning by capturing relationships between depicted elements, historical references, and associations within a cultural framework. Subjective features can introduce an emotional dimension, capturing for example displayed and evoked emotions in paintings, music, and other media. This aspect is particularly relevant for the creation of enhanced memory twins, as well as for the development of curation applications such as accessibility enhancements (e.g., descriptions for visually impaired users) and adaptive storytelling.

To better define categorized features for use in text, audio, and visual extraction, systems for automated feature recognition were implemented. Notably this included a system to define foreground and background features, thus enabling accurate action sequence recognition, and distinctions of specific 2D objects.

A further aspect of our approach has been the selection of high-quality datasets to train feature extraction models effectively. To ensure comprehensive cultural content analysis, we explored diverse resources across multiple modalities.

In the music domain, we found the Jamendo dataset [bogdanov19] for genre, instrument, and mood classification, while the emoMusic dataset [soleymani13] provided valuable insights into musical emotions based on the arousal/valence scheme [fontaine07].

For text-based feature extraction, we investigated the Standardized Project Gutenberg Corpus [gerlach20], the English Web Treebank [bies12], Penn Treebank corpus [marcus93], Tycho Brahe dataset [galves17], Penn Parsed Corpora of Historical English [kroch20], as well as the LitBank dataset [bamman19, sims19, bamman20].

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

For visual and artistic content, a broad range of datasets were explored including WikiArt[2], OmniArt [strezoski17], MAMe [pares22], and StyleBabel [ruta22] which are labelled to analyse artist details, styles, techniques, and creation periods. The SemArt dataset [garcia18] provides additional interpretative insights with textual artistic comments, while ArtDL [milani21], IconArt [gonthier18], printArt [carneiro12], and the Rijksmuseum Challenge [mensink14] include iconography classification and bounding boxes.

Beyond traditional features, we also explored multi-sensory and emotional aspects of cultural assets. The SniffyArt dataset [zinnen23] offers novel perspective by associating olfactory information with paintings, while the Artemis dataset [achlioptas21] provides an extensive collection of emotions and textual explanations linked to artworks. Finally, the ArtBench dataset [liao22] facilitates benchmarking and generation of artistic imagery.

Building upon these foundational resources, the following sections will detail the implementation of specific feature extraction techniques, focusing on the following sub-chapters:

- Enriching Metadata Using Online Databases: Here we discuss the integration of external databases to enhance metadata, making cultural assets more contextually rich.
- Object Detection in Paintings: This section details the application of object detection models to identify and classify objects within artworks, contributing to a deeper understanding of visual content.
- Extracting Smells from Paintings: Here, we explore the integration of olfactory data with visual representations, offering a novel approach to multi-sensory content extraction.
- Extracting Emotion from Speech: This section discusses how SER models can capture perceived emotional content in audio, and how these can enhance cultural content analysis.
- Evoked Emotions: Here we outline how we can predict which emotions will be evoked in a person when looking at a specific painting, which can enhance the understanding for a blind user.

These approaches serve as the foundation for memory twins, allowing digital assets to be represented and retrieved based on both their objective propertied and their emotional impact thus contributing to a more immersive and contextually enriched representation of CH assets within the SHIFT project.

[2] https://www.wikiart.org/de

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

3.1. ENRICHING METADATA USING ONLINE DATABASES

Documenting cultural assets presents ongoing challenges for museums and cultural institutions. Metadata plays a fundamental role in digitally preserving and analysing cultural assets. To enhance metadata quality, we developed an enrichment that combines multiple specialized databases. This serves several key purposes: making artworks easier to find through better search parameters, providing richer context through extended descriptions, supporting international research with multilingual information, and enabling complex connections between artworks for curatorial purposes.

Data Sources

Our metadata enrichment framework is built upon two primary data sources, Europeana[3] and MuseumPlus RIA[4], both of which provide unique metadata fields.

Europeana hosts one of Europe's largest digital cultural archives. The Application Programming Interface (API) pulls metadata from many European cultural institutions and organizes it using the Europeana Data Model (EDM). The Search API enables structured queries based on metadata fields such as title, creator, material, description, iconclass-notations, and collection.

The Museum Documentation System (MDS) of the Staatliche Museen zu Berlin (SMB), based on the widely used software application MuseumPlus RIA, serves as the main integrated documentation and collection management system for the entire 15 Berlin state museums. It covers collections from the early mankind to contemporary art including the Nationalgalerie, Gemäldegalerie, Kupferstich-kabinett, and the Plakatsammlung der Kunstbibliothek. The system provides detailed metadata including the artist, when and where it was created, what materials were used, its physical characteristics, how it was acquired, and its iconographic classification. There are two ways to access this system: through an API (requiring a verified account) or through XML-LIDO (Lightweight Information Describing Objects) data export, which allows customized data extraction for local processing. All records are also available via the Europeana search engine. They have been populated by the German aggregator DDB (German Digital Library).

[3] https://www.europeana.eu
[4] https://recherche.smb.museum/

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

```
{
  "identifier": {
    "lidoRecID": "",
    "objectPublishedID": ""
  },
  "title": "",
  "title_en": "",
  "provider_url": "",
  "measurements": {
    "dimensions": "",
    "frame_dimensions": null
  },
  "production": {
    "creator": "",
    "creator_dates": "",
    "creator_nationality": null,
    "date": "",
    "place": "",
    "materials": ""
  },
  "acquisition": {
    "method": "",
    "date": null
  },
  "iconography": [
    {
      "type": "",
      "subject": "",
      "notation": ""
    }
  ],
  "image": {
    "url": "",
    "rights": "",
    "credit": ""
  },
  "description": "",
  "additional_works": [
    {
      "title": "",
      "year": "",
      "reference_number": ""
    }
  ],
  "similar_works": [
    {
      "title": "",
      "creator": "",
      "reference_number": ""
    }
  ]
}
```


Figure 1. JSON output structure.

Implementation

We developed a Python-based solution to automate the metadata enrichment pipeline. This solution retrieves metadata from the Europeana Search API, extracts structured information from MDS exports, and integrates the data into a standardized JSON format (see Figure 1).

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

```
{
  "apikey": [redacted],
  "success": true,
  "requestNumber": 999,
  "itemsCount": 1,
  "totalResults": 3,
  "items": [
    {
      "completeness": 0,
      "country": [
        "Germany"
      ],
      "dataProvider": [
        "Picture Gallery, Berlin State Museums"
      ],
      "dcCreator": [
        "Rembrandt Harmensz van Rijn (1606 - 1669, Maler)"
      ],
      "dcCreatorLangAware": {
        "de": [
          "Rembrandt Harmensz van Rijn (1606 - 1669, Maler)"
        ]
      },
      "dcTitleLangAware": {
        "de": [
          "Selbstbildnis mit einem Samtbarett und einem Mantel mit Pelzkragen"
        ]
      },
      "edmConcept": [
        "http://data.europeana.eu/concept/48"
```


Figure 2. API results from Europeana.

Our workflow follows the following steps:

1. **Search for an artwork in Europeana using its title, artist, and/or other search query fields**

We begin by constructing a precise search query to locate the desired artwork in the Europeana database. The query can include the artwork's title, artist name, and specific collection details to ensure accurate results. In our example, the query is: "Selbstbildnis mit Samtbarett und einem Mantel mit Pelzkragen AND Gemäldegalerie Berlin". To query in the browser for testing, the following link can be used: https://api.europeana.eu/record/v2/search.json?wskey=YOUR_KEY&query=Selbstbildnis+mit+Samtbarett+und+einem+Mantel+mit+Pelzkragen&rows=1 replacing YOUR_KEY with the API key requested from the Europeana website. The result is shown in Figure 2.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

The system executes this query using the Europeana Search API, which returns a list of matching records. With every request, the system returns a message stating input values such as API-key and search query, system messages like potential errors and the number of results. In the case that no result is found, the system returns "0 results". When multiple results are found, the system presents them to the user for selection, displaying basic information like title, creator, and the amount of available metadata to help make an informed choice.

2. **Extract basic metadata from Europeana**

After receiving the API response, our system parses the JSON data to extract key metadata fields. These include the artwork's title, creator, date, dimensions, provider information, and rights status. The Europeana API provides a standardized set of fields based on the EDM model, making it possible to extract consistent information across different collections.

For example, from Rembrandt's "Self-Portrait with a Velvet Beret and a Fur Collar", we extract fields such as:

- Title: "Selbstbildnis mit Samtbarett und einem Mantel mit Pelzkragen"
- Creator: "Harmensz van Rijn, Rembrandt"
- Date: "1634"
- Provider: "Gemäldegalerie, Staatliche Museen zu Berlin"
- Rights: "http://creativecommons.org/publicdomain/mark/1.0/"

3. **Find the matching MDS entry**

With the basic information obtained from Europeana, we identify the corresponding entry in the MuseumPlus RIA system. This is done by matching the provider number extracted from the Europeana response with the system number in the MDS database.

The system number is typically embedded in the provider URL field. For example, from a URL like "https://www.smb.museum/en/museums-institutions/gemaeldegalerie/collection-research/collection-highlights/865646/", we can extract "865646" as the system number.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

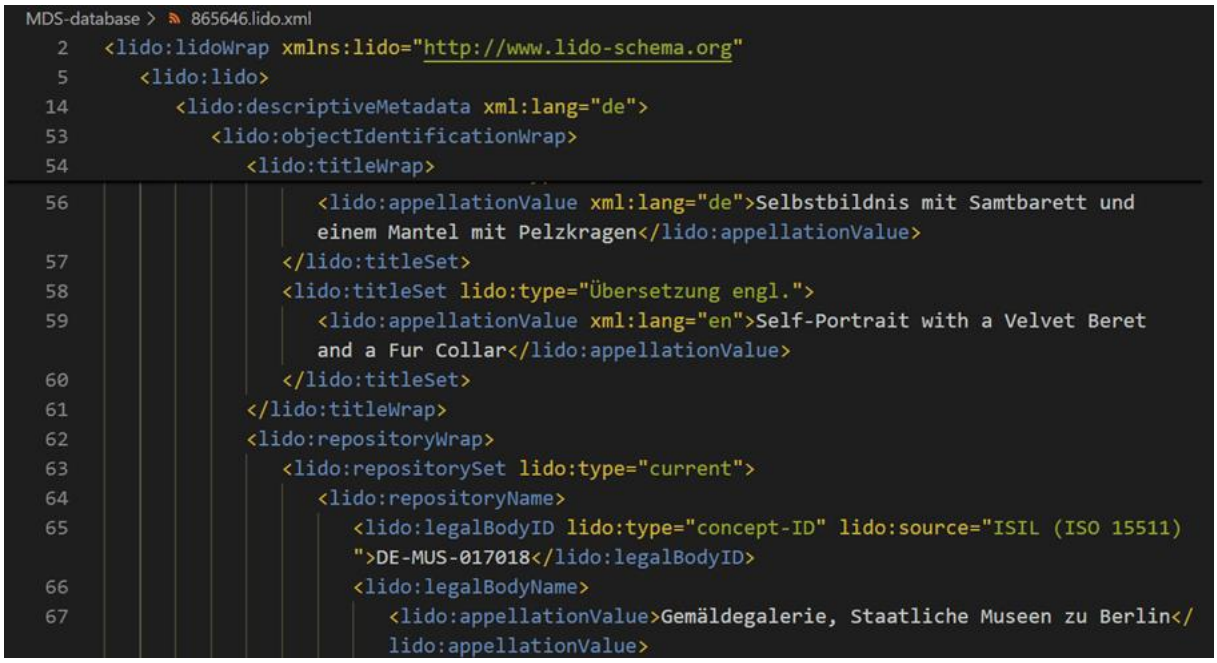

Figure 3. Example of XML file in LIDO database.

4. Get and process the XML-LIDO export

Once the matching entry is found, we extract the detailed metadata from the XML-LIDO export. The LIDO format is an XML schema such as in Figure 3 designed specifically for museum objects, providing a rich set of fields for CH items.

The export contains detailed information organized in sections such as:

- Object identification (titles, repository, measurements)
- Event information (production, acquisition)
- Subject classification (iconography)
- Rights information
- Resource links (images)

5. Merge detailed metadata fields from both sources

We combine the metadata from both Europeana and MuseumPlus RIA to create a complete metadata record. This includes iconographic classification, detailed measurements, materials, multilingual descriptions, and historical context that might not be available in the Europeana record alone.

6. Search additional artworks from the same artist

To provide contextual information, our system performs an additional search in the Europeana database to find other works by the same artist. This helps establish connections between artworks and provides a broader understanding of the artist's

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

oeuvre. For Rembrandt, this might include other self-portraits or notable works like "The Anatomy Lesson of Dr. Nicolaes Tulp" or "The Night Watch". The system collects basic information about these works, such as title, year, and reference number, and includes them in the final metadata record (Figure 4).

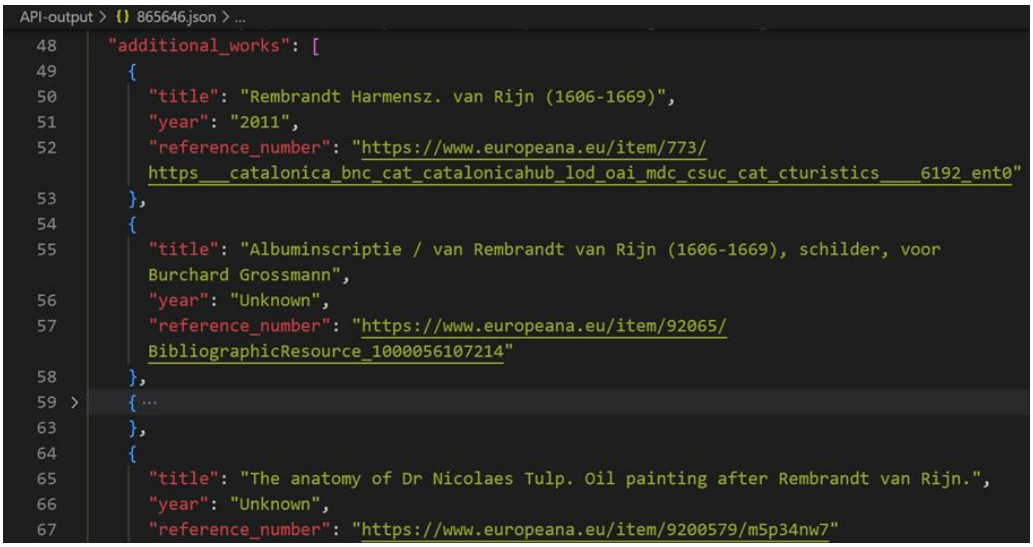

Figure 4. Additional work added to JSON for the search query: "Self-Portrait with a Velvet Beret and a Fur Collar".

7. Extract image as jpg

The system extracts the image URL from the MDS data and downloads the high-quality image of the artwork, saving it in JPG format for further use. The image filename follows the system number convention, such as "865646.jpg" for Rembrandt's self-portrait (Figure 5).

This provides a visual reference that can be used for display purposes, analysis, or integration with other applications. The downloaded image maintains the copyright and usage rights specified in the metadata.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 5. Image "Self-Portrait with a Velvet Beret and a Fur Collar" (1634) by Rembrandt Harmensz van Rijn.

8. Combine all data into a final JSON output

Finally, all the gathered metadata is structured into a standardized JSON format, creating a rich, comprehensive record of the artwork that combines information from multiple sources in a file named after the system number, such as "865646.json" (see Figure 6). The JSON structure groups related information together, making it easily parsable by other applications or services. This enhanced metadata can now be used for advanced search, analysis, or display purposes.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

API-output > {} 865646.json > ...


```
{
  "identifier": {
    "lidoRecID": "DE-MUS-017018/865646",
    "objectPublishedID": "https://id.smb.museum/object/865646"
  },
  "title": "Selbstbildnis mit einem Samtbarett und einem Mantel mit Pelzkragen",
  "title_en": "Self-Portrait with a Velvet Beret and a Fur Collar",
  "provider_url": "https://www.europeana.eu/item/2064108/
  Museu_ProvidedCHO_Gem_ldegalerie__Staatliche_Museen_zu_Berlin_DE_MUS_017018_865646",
  "measurements": {
    "dimensions": "Bildmaß: 58,4 x 47,7 cm",
    "frame_dimensions": null
  },
  "production": {
    "creator": "Rembrandt Harmensz van Rijn (1606 - 1669), Maler*in",
    "creator_dates": "1606 - 1669",
    "creator_nationality": null,
    "date": "1634",
    "place": "Holland, None",
    "materials": "Eichenholz"
  },
```


Figure 6. Final JSON output for search query Object Detection in Paintings.

3.2. OBJECT DETECTION IN PAINTINGS

Object detection plays a crucial role in analysing cultural assets by identifying and localizing key elements within images. In our approach, we utilize YOLO (You Only Look Once) [redmon16] family of models, a state-of-the-art project for object detection known for its speed and accuracy. YOLO processes entire images in a single pass through a convolutional neural network, making it highly efficient for large-scale datasets.

To integrate YOLO into our workflow, we followed a structured approach:

Pre-processing: All images were resized to a fixed 620×620 pixels, as YOLO models are trained on a standardized input size. Aspect ratio preservation was not maintained, leading to potential stretching or distortion. However, this did not affect our workflow significantly, as our primary goal was object detection rather than fine-grained visual details. Moreover, YOLO normalizes bounding box coordinates relative to the input dimensions, making it straightforward to re-project the detected objects back onto the original aspect ratio if needed.

Model Selection and Training: We initially used YOLOv8 [jocher23], but after unsatisfactory results, we experimented with the newer YOLOv11 [jocher24] when it was released. Different model sizes were tested, ranging from small (Nano, Small) to large (Large, XLarge) configurations, aiming to balance detection accuracy with computational efficiency.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Object Detection and Feature Extraction: YOLO was applied to detect objects in each image, producing bounding boxes and confidence scores. The detected objects were stored as structured metadata, enabling further analysis and potential integration into curation workflows. We also saved the pictures overlayed with the bounding boxes for further inspection.

Evaluation and Challenges: Despite testing multiple YOLO versions and model sizes, the detection results were not sufficiently reliable for direct application in CH curation. A key limitation arises from YOLO's training on real-world imagery, making it struggle with unfamiliar artistic elements. For example, as shown in Figure 7, a Pegasus in a classical painting was misclassified as an elephant, and the wings of a Cupid were incorrectly labelled as an umbrella.

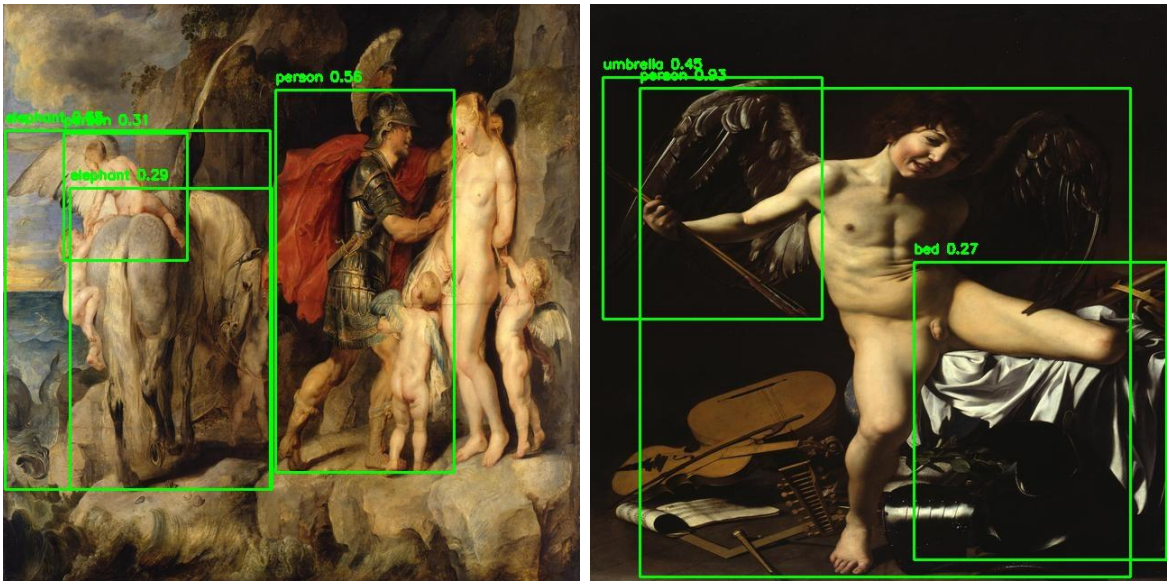

Figure 7. Mislabelling of artistic elements in" Perseus frees Andromeda" (1620/1622) by Peter Paul Rubens and "Cupid as Victor" (1601/1602) by Michelangelo Merisi gen. Caravaggio.

Another major issue is that YOLO performs poorly on landscape paintings, often detecting no objects at all due to the lack of distinct, real-world object boundaries (see Figure 8).

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 8. No detection of objects in landscape paintings like "Der Watzmann" (1824/1825) by Caspar David Friedrich, "Unwetter in der römischen Campagna" (1829) by Carl Blechen, and "Die Welle" (1869/1870) by Gustave Courbet.

Even when YOLO detects objects, it often fails to identify all relevant elements in complex scenes. As demonstrated in Figure 9, where it misses a significant portion of the depicted figures and objects.

Figure 9. Many people and objects are not recognized in crowded paintings like "The Fountain of Youth" (1546) by Lucas Cranach and "The Dutch proverbs" (1559) by Pieter Bruegel.

These challenges suggest that additional fine-tuning, dataset adaptation, or alternative object detection approaches may be necessary to improve performance for CH applications. Given the resource constraints within the SHIFT project, it was not feasible to label a sufficiently large dataset for fine-tuning YOLO on paintings. The manual annotation of a diverse set of artistic styles and subjects requires significant expertise and time, making it beyond the scope of the current project. However, we mitigate these limitations by combining other metadata sources, such

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

as descriptions and iconclass-notations to enhance the available information. This approach allows for meaningful curation despite the current challenges in object recognition. Additionally, we anticipate that advances in domain-specific model training and alternative architectures will further improve detection capabilities. Since the core integration of object-based similarity measures has already been developed within the SHIFT platform, future improvements in object detection can be seamlessly incorporated even beyond the project's completion.

3.3. EXTRACT ASSOCIATED SMELLS FROM PAINTINGS

The intersection of smell and visual art introduces a unique and historically significant dimension to cultural assets, expanding their sensory interpretation beyond the visual realm. Olfactory experiences play a crucial role in shaping human perception and memory, and some museums have already begun incorporating scent into exhibitions to enhance visitor engagement. Our initial efforts in this area focused on fine-tuning object detection models to identify smell-eliciting objects in paintings. The goal was to enrich storytelling by incorporating olfactory elements into the analysis of artworks.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 10. Example image from Musti. Image credit: Laid Table with Cheese and Fruit. 1610. Floris van Dyck. Public Domain, via Wikimedia Commons.

To achieve this, we experimented with the "MUSTI – Multimodal Understanding of Smells in Texts and Images" dataset[5] (see Figure 10), which provides annotated data linking objects to olfactory perceptions. However, extracting smell-related information from paintings presented significant challenges. The complexity of associating visual elements with specific scents, combined with inconsistencies in the available training data, led to unreliable detection results. Despite adjustments and fine-tuning, we were unable to reach a level of accuracy sufficient for meaningful integration into our feature extraction pipeline.

As smell is a fundamental component of human experience and cultural memory, it was essential for us to ensure that any extracted information was reliable. Given the current limitations of available models, we determined that the risk of misinterpretation was too high to justify further integration within the project. While we will not be able to pursue this direction further, the importance of

[5] https://multimediaeval.github.io/editions/2022/tasks/musti/

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

olfactory elements in cultural heritage remains undisputed. Future advancements in multimodal AI may eventually provide the necessary reliability to incorporate scent-based interpretations into digital curation in a meaningful way.

3.4. EXTRACT EMOTION FROM SPEECH

Today's SER models focus on recognizing perceived emotions to reach a high concordance correlation coefficient with a consensus of human annotators/listeners that produced the datasets for SER training [Schuller18]. Even today two predominant paradigms for SER exist, the one involves recognition of categorical emotion-labels such as "anger", "happiness", "sadness", and "neutrality", while the second paradigm focus on continuous emotion-levels recognition and is known as dimensional SER [fontaine07]. During SHIFT and following the recent advancements in SER we delved into dimensional SER as it reveals further details of emotion, alleviating human limited-options opinion errors arising from limiting the possible annotating options to few categories. Surprisingly humans perform well on labelling emotional levels, e.g., of Arousal (i.e., excitement), Dominance (i.e., surprise), and Valence (i.e., pleasantness) in speech better than labelling discrete emotions [morgan19].

For the SER model of SHIFT we introduced various mixing background noises that aided our SER model disentangle emotion from Speech adulterated by various background classes. During the course of SHIFT, we developed a novel SER model that reaches state-of-the-art performance in dimensional SER [kounadis-bastian24] which we describe in detail in Deliverable 3.6 and use for benchmark of SHIFTs Affective text and video to speech production tool, in D3.6. One of today's ongoing benchmarking procedures for synthetic speech relies on human listener evaluations, such as Naturalness Mean Opinion Scores [li23a]. Our aim was a reproducible and quantitative evaluation of TTS speech, beyond subjective evaluations of few human raters. For this reason, we turned our attention to SER. By developing our SER model, we obtained an automatic "rater" of synthetic speech's perceived emotion. Hence a quantitative rater of affectiveness/pleasantness of SHIFT TTS tool's speech output.

For SER benchmarks of TTS speech produced by the SHIFT TTS tool we apply both categorical and dimensional SER as together they reveal the perceived emotion evoked for the listener of speech, indnt of the speech that is difficult to assess objectively, because perceived emotion may be diluted by the language understanding, and is highly subjective, whereas language-agnostic SER models can be more holistically reveal TTS emotionality.

As D3.6 is dedicated to the affective TTS tool and the associated SER model, we refer the reader there.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

3.5. EXTRACT EVOKED EMOTION FROM PAINTINGS

One of the fundamental goals of art is to evoke emotions in the viewer. This is more difficult in the case of blind users. We therefore worked on a method to automatically extract evoked emotions from paintings to integrate this into descriptions tailored towards this user group.

The ArtEmis dataset [achlioptas21] extends the WikiArt dataset, which comprises 80,031 artworks from 1,119 artists across 27 art styles and 45 genres, by providing annotations of the emotions evoked by each painting. Each artwork was assessed by at least five annotators, who recorded their dominant emotional response and provided an explanation. A total of 454,684 responses were collected, covering the emotions 'amusement', 'awe', 'contentment', 'excitement', 'anger', 'disgust', 'sadness', and 'fear', with an additional category, 'something else', allowing for emotions outside this predefined set. Some examples can be found in Figure 11.

Figure 11. Examples from the ArtEmis dataset [achlioptas21].

The original ArtEmis paper introduced models for emotion classification and affective neural speakers to generate emotional captions for artworks. The classification task aimed to predict the dominant emotion felt when viewing a painting, with a baseline model built using ResNet34 [kaiming16] model as the backbone. This model applies an adaptive average pooling layer followed by a fully connected neural network with a hidden layer (100 neurons, batch normalization, ReLU activation, and dropout at 0.3) and an output layer of nine neurons (one per possible emotion). The model was trained for 25 epochs with a batch size of 128, using Adam optimization (learning rate 0.0005) and early stopping.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Evaluation considered both accuracy and Unweighted Average Recall (UAR) due to the dataset's highly imbalanced label distribution. The original paper reported an accuracy of 60.2%, but our training resulted in 57-58% accuracy and a UAR of 29.1% (chance level: 11%).

Our objective was to enhance the baseline model for emotion classification and examine challenges within this dataset. To achieve this, we conducted four independent experiments, each addressing a different aspect of the model, while keeping all other training parameters constant. Given the dataset's class imbalance, the focus was on improving UAR, particularly for underrepresented emotions like 'anger' and 'disgust'.

Experiment 1: Dropout Rates

We tested dropout rates ∈ [0, 0.1, 0.2, 0.3, 0.4, 0.5] in the model's fully connected layers. Results (see Table 1) indicate that a dropout rate of 0.1 yielded the highest UAR (34.3%), while accuracy remained largely unchanged

Table 1: Results Dropout Rates

Dropout value	0.0	0.1	0.2	0.3	0.4	0.5
Accuracy	**58.8%**	57.9%	57.7%	57.6%	56.5%	57.2%
UAR	31.9%	**34.3%**	32.4%	31.1%	29.9%	31.6%

Experiment 2: Optimizers

We compared various optimizers, including Adam (baseline), AdamW, and Adamax. AdamW improved UAR by 2.2% (to 34.3%), while Adamax increased accuracy by 0.6% (to 58.1%), as shown in Table 2.

Table 2: Results Optimizers

Optimizer	Adam	AdamW	Adamax
Accuracy	57.5%	57.0%	**58.1%**
UAR	32.1%	**34.3%**	32.2%

Experiment 3: Loss Methods

We tested different loss functions, including Mean Squared Error (with Softmax activation), Cross Entropy (weighted and unweighted, without activation), and Kullback-Leibler divergence (used in the baseline). Kullback-Leibler divergence achieved the highest accuracy (58.2%) but weighted Cross Entropy loss yielded the best UAR (37.1%), despite a lower accuracy (55.8%) (see Table 3).

Table 3: Results Loss Methods

Loss Method	Kullback-Leibler divergence loss	Mean Squared Error	Cross Entropy Loss (unweighted)	Cross Entropy Loss (weighted)

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Accuracy	**58.2%**	57.9%	57.7%	55.8%
UAR	32.2%	31.1%	31.4%	**37.1%**

Experiment 4: Backbones

We tested alternative backbone architectures, namely ResNet50 [kaiming16] (marginally improving both accuracy and UAR), High-Resolution Net [wang20], and Vision Transformer [dosovitskiy21] (performing significantly worse, likely due to overfitting). Since Vision Transformer requires fixed image sizes, all paintings were resized to 224 × 224 pixels as part of pre-processing for these models. For the full results see Table 4.

Table 4: Results Backbones

Backbone	ResNet34	ResNet50	High-Resolution Net	Vision Transformer
Accuracy	58.1%	**58.2%**	50.2%	50.0%
UAR	31.3%	**31.9%**	19.8%	18.1%

The best-performing model for UAR incorporated:

- Dropout rate: 0.2
- AdamW optimizer
- Weighted Cross Entropy loss
- ResNet50 backbone

Although overall accuracy did not improve, UAR increased to 34.4%, indicating better handling of underrepresented classes.

Figure 12 compares confusion matrices for the baseline and best model. Improvements were observed for 'amusement', 'awe', 'fear', and 'sadness', but 'excitement', 'anger', and 'disgust' remained largely unclassified. This could be due to 1) emotional overlap: 'excitement' might blend with other positive emotions like 'amusement' or 'awe' and 2) class imbalance: 'anger' and 'disgust' were significantly underrepresented, likely leading to poor generalization.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

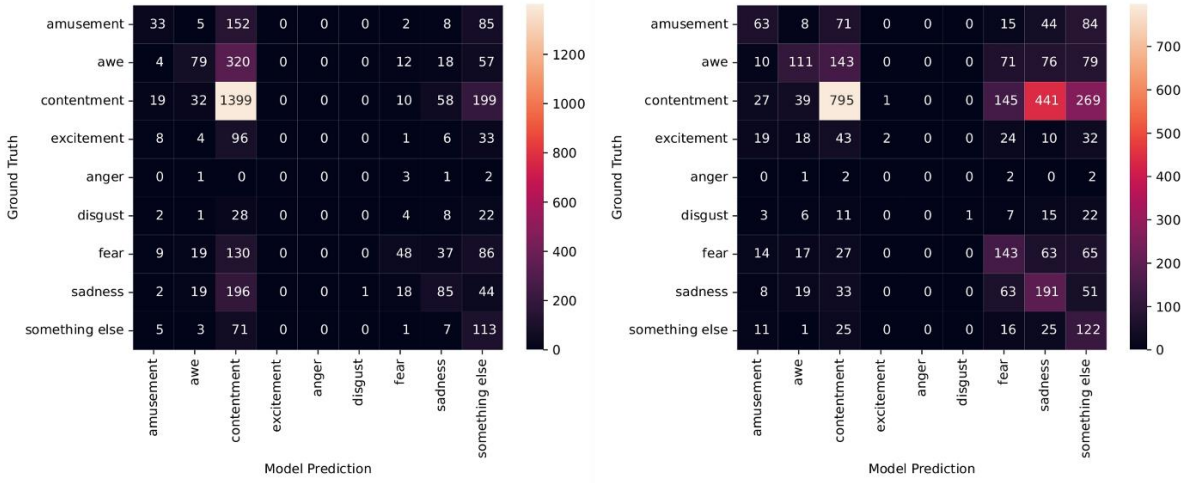

Figure 12. Confusion matrices for the baseline model and the best model. The left confusion matrix belongs to the baseline model. The right confusion matrix belongs to the best model.

While we improved UAR, trade-offs emerged, such as increased false negatives for the dominant class, 'contentment'. Emotion classification from paintings remains challenging, given the dataset's class imbalance and the subjectivity of emotions. Data augmentation was not explored here, as it risks altering the emotional intent of artworks, but it remains a potential avenue for future research.

Although we did not use extracted emotions for enriching textual descriptions, the results contributed to curation tasks, which will be discussed in the next chapter.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

4. SHIFT CURATION TOOLS

In the previous chapter, we explored several key approaches to enriching and analysing cultural assets, from object detection in paintings to the extraction of evoked emotions and the enrichment of metadata using online databases. Building on these efforts as well as on insights from WP2 and WP3, Task 4.2 focuses on establishing meaningful correlations between cultural assets. The goal is to interlink artefacts across multiple dimensions, enriching their contextual relevance and enhancing storytelling capabilities within the SHIFT platform.

In the initial phase of the project (completed over a year ago), we concentrated on developing ontological structures to facilitate better interaction with CH data. Ontologies provide a formal framework for defining concepts, relationships, and properties within a domain, allowing machines to organize, interpret, and share knowledge. During this stage, our primary focus was on tangible CH, particularly artworks. We developed an ontology based on data provided by project partners, enabling users to describe and interlink paintings through attributes such as title, artist, type, style, material, and keywords.

Since then, our approach has evolved to leverage DL-driven methods for curation. Rather than expanding ontologies, we have shifted towards using embedding-based similarity measures. These methods allow us to compute distances between assets based on learned representations from feature extraction tasks, moving beyond predefined relationships to capture more nuanced, dynamic connections between artworks. This transition offers greater scalability and adaptability, as embeddings provide a more flexible means of identifying relationships that may not be explicitly defined in traditional ontologies. Additionally, we have integrated external APIs such as Europeana to enrich metadata and establish links between artefacts.

The following sections detail the implementation of these innovative approaches, including our work on embedding-based similarity calculations, emotion-based curation, curation based on common themes, and integration with external databases. Through these efforts, we aim to foster a more interconnected and inclusive representation of CH, enhancing accessibility and engagement across diverse user groups.

4.1. CURATION USING EXTERNAL DATABASES

The digital representation of CH assets requires not only comprehensive metadata but also contextual relationships between artefacts. By identifying and presenting similar artworks, we create meaningful connections that reveal artistic influences, stylistic developments, and thematic continuities. This approach transforms isolated digital objects into nodes within a rich network of cultural significance,

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

enabling both scholarly research and serendipitous discovery across institutional boundaries.

Europeana Recommendation API

For implementing this functionality, we use the Europeana Recommendation API, which is based on advanced ML technologies. The core of the recommendation engine uses vector embeddings, a technique from artificial intelligence, to compute similarities between information resources based on a selective number of parameters/dimensions.

Implementation

The recommendation process follows a structured workflow:

1. **Identification of Source Artwork**: We formulate a targeted query using the unique record ID of the primary artwork.
2. **Querying the Recommendation API**: We send an HTTP request to the endpoint[6] with parameters for page size and random seed.
3. **Processing Recommendation Results**: The API returns a prioritized list of similar objects generated based on multidimensional similarity calculations.
4. **Extraction of Relevant Metadata**: From each recommended object, we extract key information fields:
    a. Title of the work
    b. Name of the artist/creator
    c. Unique reference number for access
5. **Integration into JSON Structure**: We integrate the processed recommendations into the similar_works array of the final JSON output.

Similarity Algorithm

The similarity calculation in the Europeana Recommendation API is based on a vector approach. Each cultural object is represented by a multi-dimensional vector that encodes various aspects of the work. The similarity between objects is determined by calculating distance measures like cosine similarity in the vector space.

The Europeana API documentation[7] explains that the recommendation system uses "vector embeddings to compute similarity between information resources based on a selective number of parameters/dimensions." This technology allows

[6] https://api.europeana.eu/recommend/record/[RECORD_ID]
[7] https://api.europeana.eu/recommend

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

for a nuanced understanding of "similarity" that goes beyond simple metadata matches and can capture latent relationships between artworks.

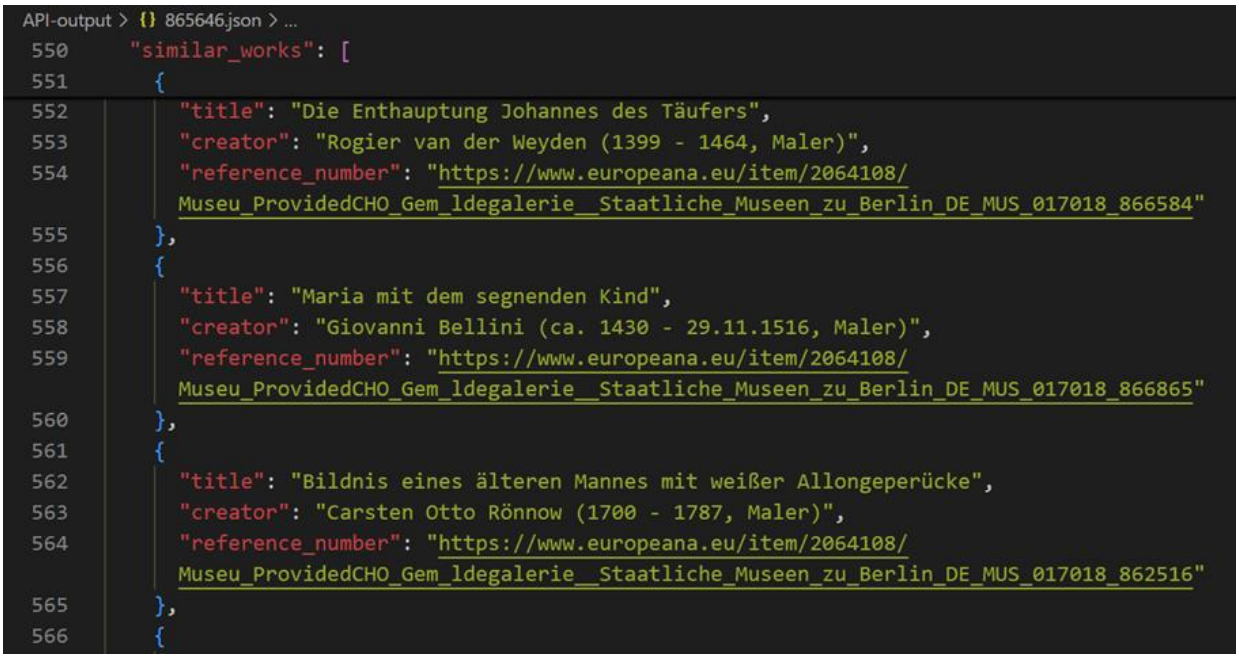

Figure 13. Results for searching similar works for Rembrandt's self-portrait using the API Recommendation API.

Application Example

For Rembrandt's self-portrait, the Recommendation API identifies a collection of similar works that extend beyond the artist's own portfolio. Looking at the actual JSON data, we find diverse connections (see Figure 13) including:

- "Die Enthauptung Johannes des Täufers" by Rogier van der Weyden
- "Maria mit dem segnenden Kind" by Giovanni Bellini
- "Bildnis eines älteren Mannes mit weißer Allongeperücke" by Carsten Otto Rönnow

The algorithm considers both portrait-specific characteristics (compositional elements, subject matter) and broader artistic connections across different time periods and schools as these examples show. This creates a rich contextual framework for understanding the position of a work within the broader art historical landscape.

4.2. CURATION BASED ON COMMON THEME

Curation of visual objects based on common theme involved the utilization of digital tools to enhance the curation potential of artefacts beyond their existing

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

potential while retaining the importance and integrity of the objects in their original function. The use of a common theme allows curators to tell a story involving several objects from a collection that tell a holistic story and can be more readily engaged by viewers. To enhance the capabilities of objects within a collection, digital asset creation from the segmentation of objects allows clusters of similar artefacts to expand their original potential.

The segmentation of 2D objects and detection of defined foreground and background objects allows for the rendering of virtual 3D objects to provide new perspectives on original objects, without altering the original function of the artefact. As a result, pieces can be formed from a single object, and visually focused curations can centre around the enhancement of clusters of similar artefacts or even single objects enhanced in different manners.

Model Evaluation

To determine the distinctions between foreground and background objects, five models utilizing machine learning algorithms were tested in order to determine effectiveness. The results of each are as follows:

1. Gaussian Mixture Model (GMM)

GMM models each pixel's intensity as a mixture of Gaussian distributions, adapting dynamically to changes in the scene.

- Strengths: Robust to gradual illumination changes and repetitive movements like tree branches.
- Limitations: Struggles with sudden illumination changes and requires parameter fine-tuning.

Outcome: GMM provided satisfactory results in controlled environments but needed enhancements to handle complex scenes.

2. Mixture of Gaussians Version 2 (MOG2)

MOG2 is an improved version of GMM that introduces shadow detection and background updating mechanisms.

- Strengths: Better handling of dynamic scenes and shadows.
- Limitations: Computationally expensive for high-resolution videos.

Outcome: MOG2 showed significant improvement over GMM in terms of accuracy, particularly in dynamic environments.

3. K-Nearest Neighbors (KNN)

Using KNN algorithm to classify each pixel as background or foreground based on its neighbourhood.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

- Strengths: Simple and effective for relatively static scenes.
- Limitations: Computationally intensive and sensitive to noise.

Outcome: KNN was effective for static environments but less so for dynamic ones.

4. Background Subtraction using DL (BS-Net)

A convolutional neural network-based approach for foreground detection.

- Strengths: High accuracy and adaptability to complex scenes.
- Limitations: Requires significant computational resources and a large dataset for training.

Outcome: BS-Net performed exceptionally well in high-complexity scenarios, making it a strong candidate for integration into the SHIFT pipeline.

5. Segment Anything Model (SAM)

A segmentation model designed for general-purpose object and region identification.

- Strengths: Highly adaptable to various segmentation tasks, including foreground extraction.
- Limitations: Over-segmentation in cluttered scenes.

Outcome: SAM demonstrated strong potential for integration, particularly when combined with post-processing techniques.

Hybrid Pipeline

Based on the outcomes of the algorithm evaluations, we tested a hybrid pipeline combining the strengths of MOG2 and BS-Net. This pipeline introduces:

MOG2: For real-time background updating and initial segmentation.

BS-Net: For refining the segmentation output, ensuring high accuracy and robustness.

The models were tested on datasets from the SHIFT project, including samples provided by Somkl Magyar Nemzeti Muzeum, SMB Stiftung Preussischer Kulturbesitz, and the Balkan Museum Network.

The results of tested models allow for curators to select and utilize those which best fit their needs and depending on object complexity, detail and capabilities in post-processing enhancement. With these tools available, curators can expand the potential of objects in tandem with other post-processing tools and enhancement technologies such as 3D rendering, and action detection. These tools can create additional output and add complexity to visually focused, themed exhibitions and group with other visual outputs such as video enhancements.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

4.3. CURATION BASED ON EVOKED EMOTIONS IN PAINTINGS

In our curation approach, we focus on utilizing the emotional responses evoked by paintings as a means to create meaningful connections between cultural assets. This process leverages the baseline emotion classification model from the ArtEmis dataset (see chapter 3.5), which enables us to quantify the emotional impact of artworks.

As part of the pre-processing workflow, we resize all images while maintaining their aspect ratio, with a maximum of 600px on the longer side (as described in chapter 2.1), ensuring consistent input dimensions. We then extract embeddings for the paintings using the encoder from the best-performing model, as detailed in the emotion classification experiments.

Using embeddings has the advantage that we do not rely on the final probability evaluations for the predefined emotions, which are subject to classification errors, but instead obtain a continuous, learned representation of the artwork's emotional characteristics. This approach allows us to capture subtle emotional nuances beyond rigid emotion categories, making the curation process more flexible and robust. Given that our classification experiments demonstrated limitations in accuracy, working directly with embeddings ensures that emotional similarities are preserved without depending on potentially misclassified emotion labels. These embeddings, which encapsulate the emotional essence of each artwork, are stored in a CSV file for easy access and continued expansion.

The curation system is designed to allow users to curate based on evoked emotions by specifying a particular painting. Once a painting is selected, the system calculates the Cosine distance between the embeddings of the chosen painting and those of all others in the dataset. This method of comparison is particularly well-suited for analysing the emotional characteristics of artworks because cosine distance measures the cosine of the angle between two vectors, which allows the system to focus on the directionality (or emotional "feel") of the embeddings rather than their absolute magnitude. This is ideal when comparing the emotional impact of paintings, as it better reflects the relative emotional similarities of artworks, even if they differ in scale or intensity. In contrast, Euclidean distance, which calculates the straight-line distance between points in space, is less effective for high-dimensional data like image embeddings. By using Cosine distance, we ensure that the curation system focuses more accurately on identifying artworks that evoke similar emotional responses, which helps preserve the nuances of the emotional experience of each painting.

This allows curators to identify and retrieve the painting that is most similar in terms of the emotional response it evokes. For an example, see Figure 14.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

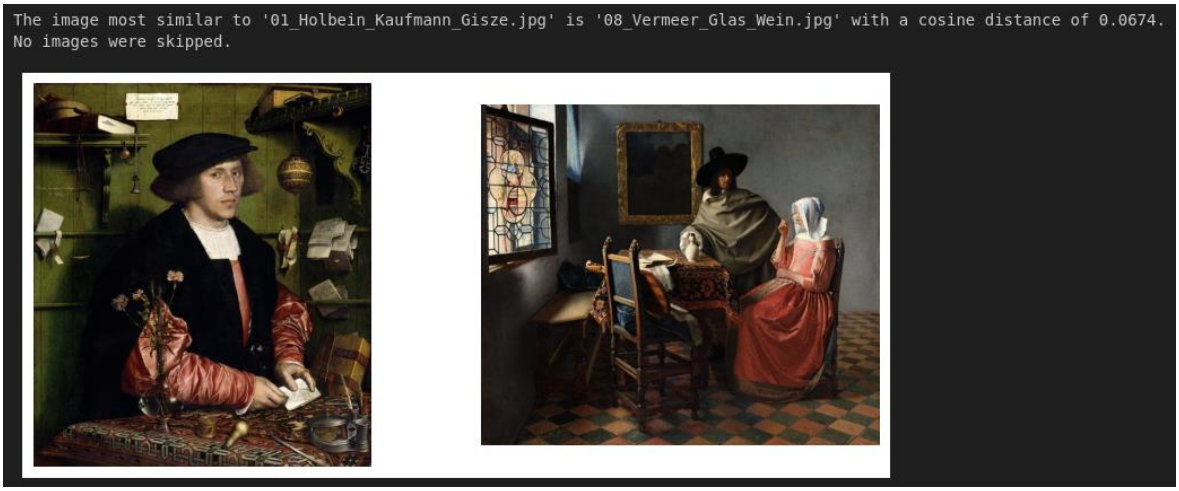

Figure 14. The image in the dataset most similar to Hans Holbein's "The merchant George Gisze" (1532) with regard to its evoked emotion is "The Glass of Wine" by Johannes Vermeer.

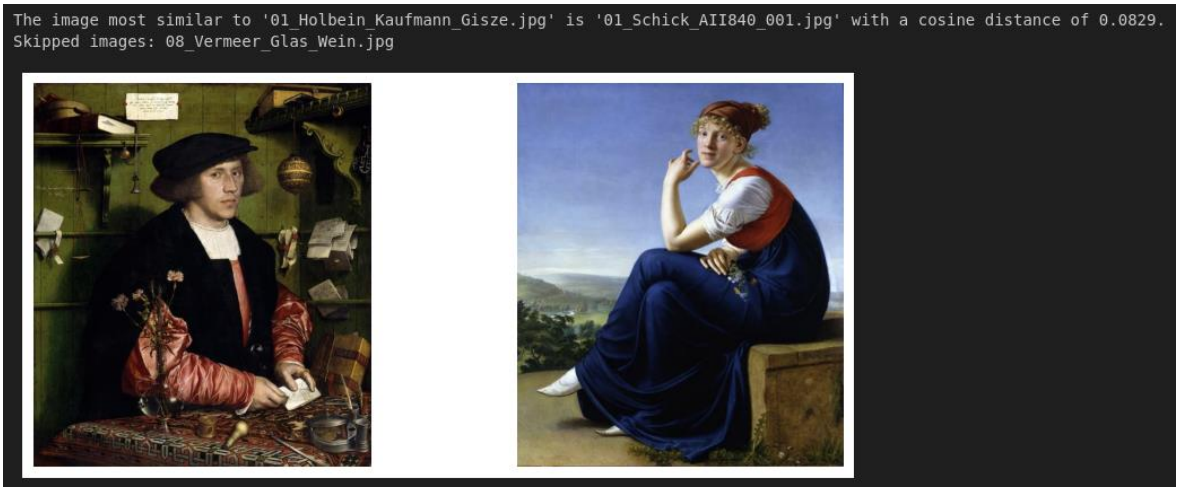

Figure 15. If we ignore the closest image "The Glass of Wine" by Johannes Vermeer, the image in the dataset most similar to Hans Holbein's "The merchant George Gisze" with regard to its evoked emotion is the image of Heinrike Dannecker by Gottlieb Schick.

To provide additional flexibility, the system includes an option for curators to skip certain paintings during the curation process, offering them more control over the results. For an example, see Figure 15.

By combining the power of emotional classification with a distance-based curation method, we create a dynamic framework for selecting artworks with similar emotional impacts, enriching the storytelling and contextual relevance of cultural assets.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

4.4. CURATION BASED ON LANGUAGE AND AUDIO STYLES

In the SHIFT project, the curation of CH assets through language and audio styles plays a pivotal role in enhancing accessibility and engagement for diverse user groups. Specifically, the SHIFT tools enable the generation of user group-targeted textual descriptions, which can be transformed into audio narration using our TTS tool. This integrated process ensures that cultural assets are presented in a way that is accessible to a wide range of audiences, including visually impaired individuals, children, people on the autism spectrum, and professionals.

The curation process begins with the provision of cultural assets, such as paintings, by project partners. Using the feature extraction tools presented in chapter 3 we enrich these assets with detailed information that enhances the understanding of the object. These are then used to create descriptions that are tailored to meet the specific needs of different user groups. For example, descriptions for blind and visually impaired people provide detailed descriptions of the paintings while descriptions for children focus on simple, engaging language. For a detailed overview of the SHIFT text generation system please refer to D3.5.

Once the textual descriptions are generated, they are seamlessly integrated into the SHIFT Text & Video to Affective Speech system, which converts the text into speech and overlays the speech to videos. The TTS tool supports over 200 curated voice styles. These voices are designed to convey different emotional tones and cater to diverse preferences, enhancing the auditory experience for the end-user. The selection of an appropriate voice style ensures that the narration aligns with the specific needs of the user group, creating a more personalized and immersive experience.

For a deeper understanding of the text generation workflows, Deliverable 3.5 outlines the description generation tools used to create text that is both accurate and engaging for various audiences. Similarly, for more information on the generation of TTS voices and their integration into the system, we refer to Deliverable 3.6, which provides comprehensive details about the TTS voice styles and their curation process.

By combining the generation of tailored text descriptions with the conversion of these texts into audio narration, SHIFT's approach to curating CH assets facilitates multi-sensory engagement. This process not only ensures that CH is accessible to individuals with diverse needs but also contributes to a more inclusive and enriched experience of the world's cultural treasures.

4.5. CURATION BASED ON SENSORY EXPERIENCES

Integrating sensory experiences in CH is crucial for improving accessibility, engagement, and emotional connections with artefacts. Haptic experiences, in particular, provide unique opportunities to make cultural assets more immersive

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

and inclusive, especially for individuals with visual impairments. This section focuses on the curation of cultural assets based on their haptic properties – such as material, texture, and tactile sensations – to support exhibition design and exploration. By identifying and cataloguing assets with tactile descriptions or characteristics, curators will be able to search, filter, and present collections based on the haptic experience they offer.

To facilitate this, curated assets are interlinked using descriptors such as smooth, rough, soft, cold, or metallic, enabling a deeper understanding of their sensory qualities. These associations are derived from established entities like CIDOC-CRM[8] alongside external resources such as Europeana[9], DBpedia[10], and Wikipedia[11]. Local databases also contribute with textual descriptions, ensuring that the SHIFT information corpus can support a variety of exhibition needs. This approach not only enhances accessibility but also redefines how visitors engage with CH, highlighting the importance of multi-sensory experiences.

Methodology for Sensory Experience Curation

The methodology for curating cultural assets based on haptic experiences follows a structured, top-down approach that begins with established CH documentation frameworks, followed by the exploration of external databases and concludes with the refinement of local datasets. This results in sensory attributes such as material, texture, and tactile properties being systematically identified, categorized, and interlinked. A similar approach was used by Popovinci et al. [popovici11], though it did not include interaction methods.

Our approach ensures that curators can effectively search for and present assets based on their haptic properties, enabling the design of exhibitions that emphasize multi-sensory engagement. The first step involves establishing a conceptual framework using CIDOC-CRM, which provides a standardized model for documenting material and tactile properties. The Scientific Observation Model (CRMsci) extension is considered for cases where scientific observations of tactile properties are available, ensuring that the representation of haptic data aligns with semantic heritage standards. Additionally, external datasets were analysed, particularly Europeana, Wikipedia, and DBpedia, which provide curated

[8] https://cidoc-crm.org/
[9] https://www.europeana.eu/el
[10] https://www.dbpedia.org/
[11] https://www.wikipedia.org/

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

descriptions of cultural artefacts and their material characteristics[12,13]. These repositories serve as the foundation, offering comparative insights that enhance the categorization of haptic descriptors. In this step, patterns were identified, in how materials such as marble, wood, fabric, or metal are described across multiple sources and how different repositories document sensory attributes such as smooth, rough, soft, cold, warm, or metallic.

After validating the external dataset and mapping it to CIDOC-CRM, we conduct a thorough review of local databases, including those from BMN, SMB PK, and SOM. This manual review involves extracting references to haptic properties from existing text descriptions of cultural assets, focusing on both explicit material mentions and indirect references to sensory experiences, thus validating the precision of our curation datasets. By structuring this information within the CIDOC-CRM framework and cross-referencing it with external findings, the SHIFT dataset is refined and enriched with expanded haptic metadata, ensuring that previously overlooked tactile qualities are systematically documented. Finally, the curated assets are organized and interlinked based on their haptic properties, creating semantic associations that allow curators to explore and group assets by shared tactile experiences.

Haptic descriptions from external databases

CIDOC-CRM

CIDOC-CRM (Conceptual Reference Model)[14] serves as a widely recognized ontology for structuring and linking CH information. Developed by the International Council of Museums, it provides a standardized framework for documenting relationships between events, objects, and historical contexts. Furthermore, it delivers definitions and a formal structure for describing the implicit and explicit concepts and relationships used in CH documentation that are of general interest for the querying and exploring such data. While CIDOC-CRM is primarily focused on the semantic representation of CH, emphasizing conceptual, spatial, and material relationships, it does not explicitly define haptic properties as a distinct sensory modality. However, several of its entities and properties can be interpreted as relevant to describing tactile attributes, surface characteristics, and material properties. Additionally, CIDOC-CRM extensions such as CRMsci and CRMdig (Digital Provenance Model) offer documentation and classification of measured tactile properties and digital haptic reproductions.

---

[12] https://www.europeana.eu/en/item/9200579/nvv8uyzb?q=proxy_dc_creator:%22Science+Museum,+London%22and+proxy_dc_subject:%22statue%22

[13] https://www.europeana.eu/en/item/9200579/n85tafb3

[14] https://cidoc-crm.org/

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Although CIDOC-CRM does not have a dedicated category for haptics, its core structure includes concepts that can be applied to document the sensory dimensions of CH objects. The entity E25 Man-Made Feature[15] represents intentional modifications made to objects, which include engravings, scratches, or carved details that affect how an artefact feels when touched. This is particularly relevant in sculptures, architectural decorations, or coins where tactile features are integral to their interpretation. For example, a coin with raised reliefs can be described using E25, capturing both the artistic intent and the tactile characteristics that result from the minting process. Similarly, E26 Physical Feature[16] encompasses natural and artificial surface characteristics that define an object's haptic properties, such as the roughness of an unfinished stone sculpture or the smooth polish of a marble bust. Since physical features are inherent to an object, they provide an essential category for describing the material qualities that influence the way an artefact is perceived through touch.

The entity E24 Physical Human-Made Thing[17] is highly relevant to our work because it encompasses all human-made physical objects, including those that possess distinct material, tactile, and sensory properties. This entity serves as a higher-level category under which various CH artefacts, such as sculptures, paintings, textiles, and tools, can be classified. Since our study is focused on curation based on sensory experiences, E24 provides a structured way to document both the material composition and the physical attributes that influence an object's tactile and sensory perception. E24 Physical Human-Made Thing is a superclass that includes both E22 Human-Made Object (discrete artefacts like paintings, sculptures, and tools) and E25 Human-Made Feature (modifications or physical characteristics such as engravings, textures, or reliefs). E24 describes human-made objects that involve modifications to pre-existing materials, which directly ties into how an artefact's materiality affects its tactile experience. For example, a marble bust may be polished to a highly smooth and reflective surface, while a stone relief may retain roughness. The ability to track transformations and modifications through CIDOC-CRM also aligns with our analysis of how production techniques influence haptic properties.

To establish relationships between artefacts and their tangible features, P56 bears feature (is found on)[18] can be used to specify the presence of physical elements that contribute to an object's haptic identity. This property allows for the documentation of texture, relief, or any noticeable tactile surface characteristics. In the case of a manuscript contains embossed text or decorative elements that

[15] https://cidoc-crm.org/html/cidoc_crm_v7.1.3.html#E25
[16] http://www.cidoc-crm.org/cidoc-crm/E26_Physical_Feature
[17] http://www.cidoc-crm.org/cidoc-crm/E24_Physical_Human-Made_Thing
[18] http://www.cidoc-crm.org/cidoc-crm/P56i_is_found_on

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

create a raised effect on the page, P56 can be used to link the book (E22 Man-Made Object) to the specific physical feature (E25) that defines its sensory interaction. Similarly, ”P45 consists of” (inverse: is incorporated in) is a CIDOC CRM property used to specify the materials that constitute a given object. While not directly representing haptic qualities, this property can be used to record material composition, which is often closely tied to tactile experiences. An oil painting, for example, consists of canvas and paint, both of which have distinct haptic attributes. Beyond describing the materials themselves, E29 Design or Procedure allows for the inclusion of artistic and manufacturing techniques that influence the sensory qualities of an object. This entity is particularly relevant when considering production methods such as embossing, engraving, weaving, or polishing, all of which contribute to how an artefact feels when handled. The property P32 used general technique (was technique of) links an object to the processes applied during its creation. A wooden sculpture that has been hand-carved, sanded, and lacquered would have multiple associated techniques recorded using P32, showing how the transformation of raw material resulted in its final haptic qualities.

For a more structured and scientific approach to documenting haptic attributes, S4 Observation within the CRMsci extension provides a way to record empirical data about an object’s physical characteristics. If a conservation scientist conducts an analysis of a textile’s surface roughness, this observation event can be documented as an instance of S4, which is then linked to the artefact being studied. The property O8 observed (was observed by) establishes the relationship between the observation event and the tangible aspects of the object being examined, while S15 Observable Entity represents the specific property under analysis. If a scientist measures the porosity of a limestone statue, S15 captures the feature being observed, allowing for more precise documentation of sensory data. Additionally, O13 triggers (is triggered by) can be relevant in cases where haptic interactions elicit a response, such as pressure-sensitive textiles that change colour when touched or sound-producing surfaces in interactive exhibits. The role of digital reproductions in CH is increasingly important, and D7 Digital Machine Event within CRMdig provides a means to document the digitization of haptic properties. If a 3D scan records the fine details of a stone relief, D7 represents the scanning event, while L22 recorded (was recorded by) establishes the connection between the process and the resulting digital model. Digital simulations of tactile experiences, such as force-feedback models in virtual reality, can be described using D10 Software Execution, which allows for the documentation of interactive digital experiences based on haptic feedback. If a museum creates a VR experience where users can “feel” the texture of ancient artefacts through haptic gloves, this process can be recorded using D10, ensuring that the sensory dimensions of heritage objects remain an integral part of digital preservation efforts.

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

Europeana database

To enhance the local dataset and address gaps identified during the analysis of records, Europeana, a comprehensive digital repository for CH, was utilized. The curated datasets available on Europeana served as a resource for identifying additional material descriptors and sensory properties that complemented and enriched the existing collection. The platform's extensive array of cultural assets, encompassing artworks, sculptures, textiles, and other artefacts, provided an opportunity to explore a diverse range of materials and haptic attributes systematically. The process began with the selection of relevant datasets from Europeana. We prioritized datasets that were well-documented and aligned with our focus on material and haptic properties. Specifically, 6 datasets, associated with artefacts such as paintings and sculptures were selected, with a total of 126 artefacts. These datasets were chosen based on their detailed metadata and relevance to sensory analysis. Our selection criteria included material descriptions, references to production techniques, and additional contextual information that could enhance the interpretive value of the artefacts. The same procedure of manually reviewing each dataset was applied, noting information related to haptic sensory attributes and material properties for each record. While the materials documented in these datasets were consistent with those recorded in the local databases, the curated datasets from Europeana provided more detailed and nuanced descriptions, particularly regarding the sensory properties of the artefacts.

One key observation was that, although artefacts were often made from the same materials, their tactile properties varied significantly depending on the curation and context of the text descriptions. For instance, marble sculptures in both the local database and the Europeana datasets were consistently described as smooth and cold, characteristics inherent to marble as a material. However, the Europeana descriptions provided additional references to polished surfaces with a reflective sheen, variations in smoothness and subtle irregularities from hand-carving. Such descriptors highlighted the individuality of each artefact, and the artisanal techniques involved in their creation. For example, a marble bust from the local database was described simply as "smooth, focusing on the refined finish typical of classical sculpture. In contrast, a Europeana record for a marble relief panel included descriptors like delicately etched lines or finely chiseled details, which added textural variety and suggested a more tactile interaction with the object. Furthermore, the Europeana text noted the presence of tool marks left intentionally by the sculptor to enhance the artwork's character, creating areas that feel slightly rougher or less uniform. Similarly, smoothness as a tactile property was not uniform across all marble artefacts. A Europeana entry for a highly polished statue might describe its surface as gleaming and mirror-like, achieved through extensive labor and repeated polishing. This contrasts with a rough-hewn statue, where only

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

certain areas were polished for emphasis, leaving other parts with a deliberately grainy or unfinished texture. Similarly, wooden sculptures and objects were described in both datasets as grainy and carved. However, the Europeana datasets frequently distinguished between types of wood, describing some as dense and fine-grained (e.g., mahogany) and others as coarse and porous (e.g., oak).

For paintings, the Europeana datasets contributed valuable insights into the interplay between the medium and surface texture. For example, oil paintings on canvas were universally described as smooth, glossy, and slightly raised due to the layers of paint and varnish. However, the curated text from Europeana occasionally highlighted unique features, such as areas where thicker brushstrokes created a more textured surface or where fine cracks in the varnish altered the tactile and visual qualities of the painting over time. Another significant enhancement came from the description of context-specific variations in sensory properties. While both datasets included materials like fabric, Europeana's descriptions often emphasized how the intended function or context of the material influenced its tactile qualities. For instance, fabric drapery in sculptures was often described as delicate and flowing, while work clothes in painted scenes were characterized as coarse and heavily textured.

Haptic descriptions from local databases

The final phase of the followed methodology focused on datasets available to SHIFT, where a systematic review of local metadata and descriptive records was conducted to extract and refine haptic features. This involved a manual search for key descriptors related to material, texture, and tactile sensations, prioritizing terms such as smooth, rough, soft, cold, warm, metallic, fabric-like, and wooden. Descriptions mentioning specific materials—such as marble, metal, stone, fabric, wood, or ceramic—were particularly emphasized, as material composition often plays a fundamental role in shaping tactile perception. Where possible, sensory descriptions were aligned with CIDOC-CRM structures to ensure that data from local repositories was consistent with international standards. To ensure consistency and usability, all identified sensory features were systematically documented and categorized within a standardized metadata format that facilitates search and retrieval. The final dataset allows curators to filter, group, and interpret cultural artefacts based on their sensory characteristics, supporting multi-sensory exhibition design and accessibility-driven initiatives. This structured approach ensures that haptic metadata is not only extracted but also contextualized and interlinked, forming a comprehensive and validated foundation for sensory-based curation within the SHIFT database.

The manual search conducted within the SHIFT local databases resulted in the identification of 42 records, including paintings, statues, and other cultural artefacts, that featured detailed descriptions related to their material and haptic

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

properties. These records provided insights into the tactile and sensory dimensions of the assets, forming the foundation for this phase of the project. The systematic review of these records revealed recurring patterns in the documented materials and their associated descriptors regarding their haptic sensations. Among the most frequently referenced materials were marble, wood, metal, fabric, stone, and canvas.

Paintings on canvas—a significant portion of the database—were consistently described in terms of their oil-based medium and canvas texture. The oil paint was noted for its smooth, glossy, and slightly raised surface, resulting from the layers of paint and varnish applied to the canvas. The underlying canvas was occasionally described as textured or grainy, particularly in areas where the paint application was thinner or where the natural weave of the material was visible. These descriptors provided a nuanced understanding of the interplay between the medium and the surface, highlighting both the visual and tactile qualities of the paintings. It is worth noting that for the paintings, a significant portion of the descriptions focused on the depicted elements within the compositions rather than solely the physical materials of the paintings themselves. These descriptions provided insights into the tactile and material qualities of the objects, textures, and scenes portrayed in the artworks, offering an additional layer of sensory engagement. Many paintings featured detailed depictions of clothing, with descriptions emphasizing their tactile qualities. Terms such as soft, flowing, woven, silky, and smooth were frequently used to characterize garments, while materials like furs and velvets were specifically noted for their warmth and textured surfaces. In most instances, clothing materials in the dataset were described as soft, reflecting their pliable and delicate nature. However, certain categories of attire, such as work clothes, were distinctly characterized by heavily textured qualities. Other than clothes, metals, such as armor, jewelry, or decorative elements, were described with terms like hard, cold, shiny, and polished, highlighting their rigid and reflective qualities. In addition, elements such as natural features—including trees, rocks, and water—were described with tactile specificity. Trees and wooden objects were often characterized as grainy or rough, while rocks and stones were depicted with descriptors such as hard and textured. Water, when included in the paintings, was frequently associated with cool sensations, capturing its fluid and flowing nature. Another category of depicted elements included everyday objects, such as furniture, vessels, and household tools. These items were annotated with haptic descriptors related to their presumed materials, such as polished wood for chairs or hard, smooth metal for tools and utensils.

Regarding artefacts and objects, the dataset revealed a rich variety of materials, each accompanied by detailed descriptions that emphasized their tactile and sensory properties. Marble emerged as one of the most frequently documented materials, often described as smooth, polished, and cold to the touch. Wood, in

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

contrast, was characterized by descriptors such as grainy, carved, and burnished. These terms emphasized the material's natural texture and the intricate craftsmanship evident in its shaping and finishing. The grainy texture of wood often reflects its organic origins, while carved details speak to the creative expression and precision of the artisan. Metal artefacts were predominantly described with terms like hard, smooth, and cold. In some instances, metals were also noted for their shiny or reflective surfaces, suggesting additional finishing techniques such as polishing or plating that enhance their visual and tactile appeal. Furthermore, stone artefacts were often noted for their rough and textured qualities, emphasizing their unrefined, robust nature. The rough surfaces of stone artefacts often contrasted with the smoother materials in the dataset, offering a tactile diversity that adds depth to the overall collection.

Beyond the material descriptors, the records often highlighted crafting techniques that directly influenced the tactile and aesthetic qualities of the artefacts. Techniques such as carving, polishing and weaving were frequently mentioned, offering valuable context for understanding how artists transformed raw materials into culturally significant objects. For example, carving was described concerning materials like wood and marble, emphasizing the precision required to create intricate designs or detailed figures. Polishing, particularly on marble and metal, played a key role in enhancing sensory appeal. Descriptors like hand-polished marble referred to a smooth, reflective surface. While the local dataset provided a foundation, the analysis also revealed certain limitations. Some artefacts, despite their significance, lacked detailed descriptions of their tactile properties. Moreover, certain sensory experiences, such as nuanced thermal sensations or the interplay of textures, were underrepresented. These gaps highlighted the need for supplementary data from external repositories, to address these shortcomings and further enhance the descriptive quality of the dataset.

Findings

Sensory data can inform the curation of CH artefacts, focusing on the tactile and material dimensions, and enriching the interpretative and experiential aspects of cultural assets. By adopting a structured, CIDOC-CRM as a conceptual framework, was systematically analyzed, the sensory metadata was validated through Europeana datasets, and the haptic descriptions from the SHIFT database were refined. This approach ensured that sensory attributes—such as material composition, texture, and production techniques—were extracted, categorized, and interlinked in a way that supports multi-sensory engagement and accessibility-driven curation. The findings highlight that explored materials were consistently documented across datasets, with frequent references to marble, wood, stone, metal, fabric, and ceramic. However, notable variations emerged in how their tactile attributes were described. Europeana’s curated datasets provided more nuanced sensory details, often specifying surface texture, temperature perception,

Funded by the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement no 101060660.

and production techniques that were missing or less detailed in local records. For example, while both datasets described a sculpture as "marble," the Europeana dataset further classified it as having a "polished, cold-to-the-touch surface," enriching its haptic representation. Similarly, textile artefacts in local databases were broadly labeled as "fabric," whereas Europeana sources distinguished between "silky, woven textures" and "heavy, rough woolen materials," adding greater specificity to their sensory categorization.

Beyond refining material classifications, the findings underscore the potential of sensory-focused curation to reshape how CH is organized and presented. By structuring haptic descriptors within CIDOC-CRM, artefacts can be categorized by shared tactile experiences. This enables curators to explore new exhibition formats that emphasize sensory dimensions, allowing visitors to engage with cultural artefacts through thematic groupings such as "smooth surfaces," "textured reliefs," or "woven and embroidered fabrics." The structured documentation of sensory metadata within CIDOC-CRM and its extensions (CRMsci for observed tactile attributes) further ensures semantic consistency and interoperability across CH datasets. While the integration of sensory data significantly enriched the dataset, it also revealed areas for further development. Certain materials and haptic characteristics remain underrepresented, either due to gaps in existing descriptions or inconsistencies in how tactile properties are documented across sources. Differences in terminology further highlight the need for standardized vocabularies to ensure that sensory descriptors are consistently applied across datasets and curation efforts. Despite these challenges, the enriched metadata provides a strong foundation for the development of tools and systems that incorporate sensory information into CH curation. The insights and data generated through this process directly contribute to the SHIFT information corpus, reinforcing sensory-based categorization as an innovative curatorial method. By enabling curators to organize artefacts based on their sensory properties, this approach facilitates the design of thematic exhibitions that highlight tactile experiences, making CH more accessible and engaging for diverse audiences. The structured integration of haptic metadata within curatorial practices bridges the gap between materiality and experience, fostering a deeper connection between visitors and artefacts.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

5. CONCLUSION

In this deliverable, we have presented the final developments and outcomes of the SHIFT project, focusing on the pre-processing, feature extraction, and curation of CH assets. By leveraging state-of-the-art ML techniques and integrating external knowledge repositories, we have built a robust framework that enables the automated analysis, enrichment, and interlinking of cultural content across multiple modalities, including text, image, audio, and sensory attributes.

The tools and methodologies developed within WP4 highlight the significant advancements made in the field of CH curation, ensuring that assets are not only preserved but also made more accessible, discoverable, and interpretable by a wide range of stakeholders. From improving contextual understanding to enabling more inclusive and engaging storytelling, the outcomes of the SHIFT project contribute to a deeper, more dynamic relationship with CH.

A key part is the integration of the concept of memory twins – rich digital counterparts of cultural assets that capture both their factual and experiential dimensions. By integrating feature extraction techniques that span objective metadata (e.g., artist, materials, historical context) and subjective aspects (e.g., emotions, sensory perceptions), we enable more nuanced representations of CH. These digital surrogates facilitate deeper engagement by preserving not only the physical and historical characteristics of artefacts but also their affective and interpretative significance.

Moreover, the integration of feature extraction techniques, such as OCR, and the development of novel tools for linking artefacts across visual, textual, and auditory dimensions, represent a critical step toward ensuring the sustainability and scalability of CH preservation efforts in the digital age. The concept of memory twins further strengthens these efforts by fostering dynamic interconnections between cultural assets, allowing them to evolve within digital repositories through enriched metadata, contextual associations, and user interactions. These tools can serve as a foundation for future projects, extending beyond the SHIFT platform and contributing to global efforts in digitizing, curating, and making CH more accessible.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

REFERENCES

[achlioptas21] Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., & Guibas, L. J. (2021). Artemis: Affective language for visual art. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11569-11579). [bamman19] Bamman, D., Popat, S., & Shen, S. (2019). An Annotated Dataset of Literary Entities. NAACL 2019.

[bamman20] Bamman, D., Lewke, O., & Mansoor, A. (2020). An Annotated Dataset of Coreference in English Literature. LREC.

[bies12] Bies, A., Mott. J., Warner, C., & Kulick, S. (2012). English Web Treebank. Web Download. Philadelphia: Linguistic Data Consortium.

[bogdanov19] Bogdanov, D., Won M., Tovstogan P., Porter A., & Serra X. (2019). The MTG-Jamendo Dataset for Automatic Music Tagging. Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019).

[burkhardt24] F. Burkhardt, B.T. Atmaja, A. Derington, F. Eyben, B.W. Schuller, "Check Your Audio Data: Nkululeko for Bias Detection", In Proc. of Conference of COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques, 2024.

[carneiro12] Carneiro, G., Da Silva, N. P., Del Bue, A., & Costeira, J. P. (2012). Artistic image classification: An analysis on the printart database. In Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12 (pp. 143-157). Springer Berlin Heidelberg.

[fontaine07] Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. Psychological science, 18(12), 1050-1057.

[galves17] Galves, C., de Andrade, A. L., & Faria, P. (2017, December). Tycho Brahe Parsed Corpus of Historical Portuguese.

[garcia18] Garcia, N., & Vogiatzis, G. (2018). How to read paintings: semantic art understanding with multi-modal retrieval. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

[gerlach20] Gerlach, M., & Font-Clos, F. (2020). A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. Entropy, 22(1), 126.

[goncalves24] Goncalves, L., Salman, A. N., Naini, A. R., Velazquez, L. M., Thebaud, T., Garcia, L. P., ... & Busso, C. (2024). Odyssey 2024-speech emotion

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

recognition challenge: Dataset, baseline framework, and results. Development, 10(9,290), 4-54.

[gonthier18] Gonthier, N., Gousseau, Y., Ladjal, S., & Bonfait, O. (2018). Weakly supervised object detection in artworks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

[jocher23] Glenn, J., Chaurasis, A., & Qiu, J. (2023) Ultralytics YOLO8, Version = 8.0.0, https://github.com/ultralytics/ultralytics

[jocher24] Jocher, G., & Qiu, J. (2024) Ultralytics YOLO11, Version = 11.0.0, https://github.com/ultralytics/ultralytics

[kounadis-bastian24] Kounadis-Bastian, D., Schrüfer, O., Derington, A., Wierstorf, H., Eyben, F., Burkhardt, F., & Schuller, B. (2024). Wav2Small: Distilling Wav2Vec2 to 72K parameters for Low-Resource Speech emotion recognition. arXiv preprint arXiv:2408.13920.

[kroch20] Kroch, A. (2020) Penn Parsed Corpora of Historical English. Web Download. Philadelphia: Linguistic Data Consortium.

[li23a] Li, Y. A., Han, C., Raghavan, V., Mischler, G., & Mesgarani, N. (2023). Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. Advances in Neural Information Processing Systems, 36, 19594-19621.

[li23b] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., ... & Wei, F. (2023, June). Trocr: Transformer-based optical character recognition with pre-trained models. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 11, pp. 13094-13102).

[liao22] Liao, P., Li, X., Liu, X., & Keutzer, K. (2022). The artbench dataset: Benchmarking generative models with artworks. arXiv preprint arXiv:2206.11404.

[lombardi20] Lombardi, F., & Marinai, S. (2020). Deep Learning for Historical Document Analysis and Recognition-A Survey. Journal of imaging, 6(10), 110.

[marcus93] Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational linguistics 19, 313-330. Reprinted in Susan Armstrong, ed., 1994, Using large corpora. Cambridge, MA: MIT Press. 273-290.

[mensink14] Mensink, T., & Van Gemert, J. (2014, April). The rijksmuseum challenge: Museum-centered visual recognition. In Proceedings of international conference on multimedia retrieval (pp. 451-454).

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

[milani21] Milani, F., & Fraternali, P. (2021). A dataset and a convolutional model for iconography classification in paintings. Journal on Computing and Cultural Heritage (JOCCH), 14(4), 1-18.

[morgan19] Morgan, S. D. (2019). Categorical and dimensional ratings of emotional speech: Behavioral findings from the Morgan emotional speech set. *Journal of Speech, Language, and Hearing Research*, *62*(11), 4015-4029.

[pares22] Pares, F., Arias-Duart, A., Garcia-Gasulla, D., Campo-Frances, G., Viladrich, N., Ayguade, E., & Labarta, J. (2022). The MAMe dataset: on the relevance of high resolution and variable shape image properties. Applied Intelligence, 52(10), 11703-11724.

[popovici11] Popovici, D., Bogdan, C., Polceanu, M., & Querrec, Ronan. (2011). Applying of an Ontology based Modeling Approach to Cultural Heritage Systems. Advances in Electrical and Computer Engineering. 11. 105-110.

[redmon16] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[ruta22] Ruta, D., Gilbert, A., Aggarwal, P., Marri, N., Kale, A., Briggs, J., ... & Collomosse, J. (2022, November). StyleBabel: Artistic style tagging and captioning. In European Conference on Computer Vision 2022 (pp. 219-236). Springer.

[schuller18] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM, 61(5), 90-99.

[sims19] Sims, M., Park, J. H., & Bamman, D. (2019). Literary Event Detection. ACL 2019.

[schlosberg54] Schlosberg, H. (1954). Three dimensions of emotion. Psychological review, 61(2), 81.

[schrufer24] O. Schrüfer, M. Milling, F. Burkhardt, F. Eyben, B. Schuller, "Are you sure? Analysing Uncertainty Quantification Approaches for Real-world Speech Emotion Recognition", in Proc. INTERSPEECH 2024.

[smith07] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.

[soleymani13] Soleymani, M., Caro, M. N., Schmidt, E. M., Sha, C. Y., & Yang, Y. H. (2013, October). 1000 songs for emotional analysis of music. In Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia (pp. 1-6).

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

[strezoski17] Strezoski, G., & Worring, M. (2017). Omniart: multi-task deep learning for artistic data analysis. arXiv preprint arXiv:1708.00684.

[zinnen23] Zinnen, M., Hussian, A., Tran, H., Madhu, P., Maier, A., & Christlein, V. (2023). SniffyArt: The Dataset of Smelling Persons. Proceedings of the 5th Workshop on analySis, Understanding and proMotion of heritAge Contents.

[wagner23] Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9), 10745-10759.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.