This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870811

SPICE

Social cohesion, Participation, and Inclusion through Cultural Engagement

D4.4: Linking and Discovering Digital Assets (v1.0)

Deliverable information	
WP	WP4
Deliverable dissemination level	PU Public
Deliverable type	R Document, report
Lead beneficiary	OU
Contributors	UNITO, MAIZE, UH
Date	Release date of current version
Authors	Enrico Daga (OU), Agnese Chiatti (OU), and Antonio Lieto (UNITO)
Date	31/10/2022
Document status	Final
Document version	v1.0

PAGE INTENTIONALLY BLANK

Project Information

Project Start Date: 1st May 2020
Project Duration: 36 months
Project Website: https://spice-h2020.eu

Project Contacts

Project Coordinator

Silvio Peroni

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
Department of Classical Philology and Italian Studies – FICLIT

E-mail: silvio.peroni@unibo.it

Project Scientific Coordinator

Aldo Gangemi

Institute for Cognitive Sciences and Technologies of the Italian National Research Council

E-mail: aldo.gangemi@cnr.it

Project Manager

Adriana Dascultu

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
Executive Support Services

E-mail: adriana.dascultu@unibo.it

SPICE Consortium

No.	Short name	Institution name	Country
1	UNIBO	ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA	Italy
2	AALTO	AALTO KORKEAKOULUSAATIO SR	Finland
3	DMH	DESIGNMUSEON SAATIO - STIFTELSEN FOR DESIGN-MUSEET SR	Finland
4	AAU	AALBORG UNIVERSITET	Denmark
5	OU	THE OPEN UNIVERSITY	United Kingdom
6	IMMA	IRISH MUSEUM OF MODERN ART COMPANY	Ireland
7	GVAM	GVAM GUIAS INTERACTIVAS SL	Spain
8	PG	PADAONE GAMES SL	Spain
9	UCM	UNIVERSIDAD COMPLUTENSE DE MADRID	Spain
10	UNITO	UNIVERSITA DEGLI STUDI DI TORINO	Italy
11	FTM	FONDAZIONE TORINO MUSEI	Italy
12	MAIZE	MAIZE SRL (previously CELI SRL)	Italy
13	UH	UNIVERSITY OF HAIFA	Israel
14	CNR	CONSIGLIO NAZIONALE DELLE RICERCHE	Italy

Executive Summary

SPICE is an EU H-2020 project dedicated to research on novel methods for citizen curation of cultural heritage through an ecosystem of tools co-designed by an interdisciplinary team of researchers, technologists, museum curators engagement experts, and user communities. This technical report D4.4 presents the results of Task 3 of Work Package 4: “Linking and Discovering Digital Assets“.

This task develops novel methods for supporting the interlinking and discoverability of digital assets. Specifically, the aim is to automatically generate semantic metadata that would allow to link between distributed assets originating from different data management systems. The methods developed leverage taxonomies of concepts and combines symbolic and sub-symbolic methods for automatically generating metadata annotations, to complement the annotations derived by museums’ curators. Linking is performed by automatically classifying artworks under two complementary dimensions: (a) descriptive features of the artworks (e.g. nature, abstract, people, etc...), and (b) emotions and sentiments (e.g. joy, fear, shame, ...). First, we focus on **content-based descriptions**, specifically, we consider the problem of automatically classifying artworks within a taxonomy of descriptive artwork features. Crucially, we explore how neuro-symbolic learning, combining image features, textual metadata, and Knowledge Graph embeddings, could help in mitigating the problems derived from data sparsity in cultural heritage image archives. Second, we focus on **sentiment-based descriptions**. Specifically, we consider the problem of automatically classifying artworks within a taxonomy of common-sense, affective descriptions. We introduce the novel DEGARI 2.0 system, an affective reasoner that exploits the logic $\mathbf{T}^{\mathsf{CL}}$ in order to generate and classify content according to the circumplex theory of emotions devised by the cognitive psychologist Robert Plutchik.

Document History

Version	Release date	Summary of changes	Author(s) - Institution
v0.1	16/09/2022	Prepared template and outline	Enrico Daga (OU)
v0.2	20/09/2022	Chapter 2 completed	Enrico Daga (OU), Agnese Chiatti (OU)
v0.3	07/10/2022	Chapter 2 completed	Antonio Lieto (UNITO)
v0.4	11/10/2022	Introduction and Conclusions	Enrico Daga (OU)
v0.5	12/10/2022	Version submitted to internal reviewers	Reviewers: Jason Carvalho (OU) and Guillermo Jiménez Díaz (UCM)
v0.6	25/10/2022	Feedback from internal reviewers integrated	Enrico Daga (OU)
v1.0	31/10/2022	Submission to EU	Coordinator

Table of contents

1 Introduction

SPICE is an EU H-2020 project dedicated to research on novel methods for citizen curation of cultural heritage through an ecosystem of tools co-designed by an interdisciplinary team of researchers, technologists, and museum curators and engagement experts, and user communities. In the SPICE project, we are researching on an intelligent system that classifies artworks to support several tasks such as metadata curation and linking across image collections.

This technical report D4.4 presents the results of Task 3 of Work Package 4: "Linking and Discovering Digital Assets". The task develops novel methods for supporting the interlinking and discoverability of digital assets. Specifically, the aim is to automatically generate semantic metadata that would allow to link between distributed assets originating from different data management systems. The methods developed leverage taxonomies of concepts and combine symbolic and sub-symbolic methods for automatically generating metadata annotations, to complement the annotations derived by museums' curators.

Linking is performed by automatically classifying artworks under two complementary dimensions: (a) descriptive features of the artworks, and (b) emotions and sentiments.

First, we focus on **content-based descriptions**, specifically, we consider the problem of automatically classifying artworks within a taxonomy of descriptive artwork features. Deep Learning (DL) methods have proved to be very successful for many image classification tasks. However, applying DL methods to real-world cultural heritage collections for the task of artwork subject classification is problematic. Objects in this domain are characterised by different levels of heterogeneity: of media and techniques, of categories, of time-periods, just to mention a few. This heterogeneity makes the related training features sparsely distributed. In Chapter 2, we report on an empirical investigation where we apply Neuro-Symbolic Deep Learning techniques to a paradigmatic case of cultural heritage archive: the Tate Gallery collection open data. We pose the question of what type of feature engineering could help in reducing the impact of data sparsity in this domain. Crucially, we explore how neuro-symbolic learning, combining image features, textual metadata, and Knowledge Graph embeddings, could help in mitigating the problems derived from data sparsity in cultural heritage image archives.

Second, we focus on **sentiment-based descriptions**. Specifically, we consider the problem of automatically classifying artworks within a taxonomy of common-sense, affective descriptions. DEGARI is one of the sensemaking tools developed in coordination with WP6 (see Deliverable 6.3 and Deliverable 6.6) used to both enrich, via reasoning mechanisms, and suggest novel affective-based connections among items within a collection [1]. Here we report how a novel version of the system, called DEGARI 2.0 [2], has been exploited in the project to discover and create semantic connections between different items within a collection. Interestingly, the same mechanisms can also be used to collect connections also between different museum collections.

The rest of the deliverable is structured as follows. Chapter 2 is dedicated to studying how neuro-symbolic learning can automatically generate descriptive, content-based annotations. Next, Chapter 3 reports on the DEGARI system for automatically generating affective-based annotations, before concluding the report in Chapter 4.

2 Neurosymbolic learning for generating content-based descriptions

2.1 Introduction

Deep Learning (DL) methods have expedited the advancement on image classification tasks [3]. However, image classification through DL is still an open challenge in domains characterised by a high variance, for example, of data samples and labels [4, 5]. The negative impact of noisy labels, in particular, has been sufficiently acknowledged in the literature as one unavoidable problem in many real-world settings [6].

In the SPICE project, we are researching an intelligent system based on DL that classifies artworks to support several tasks in the domain such as metadata curation or knowledge linking and discovery across cultural heritage archives. Crucially, in this domain, datasets are characterised by different types of heterogeneity - e.g., diversity of media and techniques, and of time-periods. Due to this variance, training features are sparsely distributed across the categories of interest.

In this chapter, we explore the application of Deep Learning (DL) techniques to a paradigmatic case of cultural heritage archive: the Tate Gallery collection open data [7]. We interrogate on (a) what type of data preparation strategies could be applicable to these data and on (b) how neuro-symbolic learning, combining image features, textual metadata, and Knowledge Graph (KG) embeddings, could help in mitigating the problem of data sparsity.

To answer these questions, we devise a layered set of experiments. First, we aim at evidencing the negative impact of data sparsity on standard DL approaches and start by only considering visual features. Secondly, we look into how metadata could help in partitioning the training space and mitigating some effects of the sparsity of image features. Third, we incrementally introduce new features from textual metadata and background knowledge, including Knowledge Graph embeddings, and explore how they improve the classification performance.

The chapter is structured as follows. After introducing the related work (Section 2.2), we provide the background context of this research (Section 2.3). Concurrently, we characterise the problem of data sparsity in artwork subject classification and present our research questions. In Section 2.4, we illustrate the approach and system architecture, which is based on current state of the art methods applied to this domain. Section 2.5 reports on the implementation of the experiments and results. Findings from these experiments are instrumental in deriving the lessons learnt and future directions of this work, as further discussed in Section 2.6.

2.2 Related work

Deep Learning (DL) is applied to a wide variety of problems in the context of cultural heritage applications [4, 5]. The tasks can range from the identification of artworks from noisy Web pictures (NoisyArt [8]), to the classification of artistic media [4], stylistic, and genre-specific artwork traits [9]. In this work, we focus on the problem of learning subject classifications from an heterogeneous cultural heritage archive. The problem of *classifying artwork subjects* is unique in its own respect compared to the types of image classification tasks that are typically tackled in the Cultural Heritage literature, which are thoroughly reviewed in [10]. The most relevant state of the art approach which we have identified to model the case of artwork subject classification is ContextNet [5], which focuses on learning a set of tasks such as Genre, Period, and School from an homogeneous set of paintings. A limitation of this approach is that only attributes in the target dataset are considered in the learning, disregarding other potential sources of artistic knowledge. To harness this potential, Castellano and Vessio proposed an extension of ContextNet, where properties gathered from Wikidata and DBpedia are used to construct a dedicated ArtGraph [11]. Inspired by the work in [5, 11] we propose to reuse the knowledge which has been previously distilled from DBpedia in the form of KG embeddings, through the RDF2Vec model [12]. Differently from prior works, we intend to explore the integration of off-the-shelf KG embeddings, as an alternative method to curating ad-hoc artistic Knowledge Graphs.

Moreover, we propose to adopt different types of embeddings to qualify different artistic features. Specifically, we test the integration of KG embeddings produced on the artist metadata with a linguistic model (distilBERT [13]), as a feature preparation from artwork titles. Combining different types of features and embeddings to leverage the strengths of the various approaches is common in recent research in neuro-symbolic learning [14]. However, no work so far has explored the data sparsity problems that emerge when DL methods are applied to cultural heritage image collections.

2.3 Background and research questions

In the SPICE project [15], a team of researchers and museum professionals are developing novel methods for citizen participation and engagement, focused on the method of *slow looking*. This approach is based on designing scripts made from a set of prompts or questions about selected artworks [16]. Prompts are designed based on properties of the artworks that are *factual* (e.g. abstract, landscape, people, objects) rather then *contextual* (e.g. genre, period, author). A Deep Learning system should classify artworks according to a given subject list (e.g. abstract, landscape, people, objects) and use these metadata to link images across collections, thus helping curators in reusing scripts across similar artworks. However, cultural heritage collections are characterised by a heterogeneity of images and subject metadata. This characteristic hinders the successful application of state of the art DL approaches, which are typically developed on homogeneous samples (e.g. only on paintings), and optimised for categories that are not related to the actual content of the artwork (e.g. "Genre", "Century" [5]).

The Tate Gallery archive is a paradigmatic case of cultural heritage image collection, characterised by artworks with a significant heterogeneity of *factual* properties. The dataset provides metadata and image urls of collection items summarised in two CSV files with general metadata (artworks and artists), and detailed metadata distributed in approximately 100k JSON files[1]. The collection includes artworks from more than 3000 artists spanning 142 genres over a period of approximately 500 years. Metadata was manually annotated by expert curators, including a taxonomy of more than 16,000 distinct *subjects*, organised in 11 top-level subjects covering key concepts relevant to the slow looking application scenario: *abstraction*, *architecture*, *nature*, *people*, etc.. The Tate Gallery collection demonstrates dimensions of heterogeneity that are typical of cultural heritage image archives:

- **Image heterogeneity**: images represent artworks produced with different mediums and techniques
- **Sample heterogeneity**: the data distribution is very unbalanced
- **Semantic heterogeneity**: subject annotations are based on the content of the assets but are produced incrementally over a large time-span by an unspecified number of annotators.

This makes the labelled data incomplete, messy, and sparsely distributed. In this chapter, we explore how to address data sparsity from different perspectives, that we illustrate.

Tackling data sparsity by configuring the learning space. On the one hand, real-world subject taxonomies are overly heterogeneous both semantically and in terms of class population. This characteristic of real-world collections significantly complicates the learning of robust classification models. For example, learning to differentiate a collie from a spitz is, in principle, more difficult than learning to tell dogs and cats apart, especially given the lack of sufficient examples representing different dog breeds. On the other hand, the few high-level subjects that have sufficient population (e.g., macro-classes such as people, nature, society, etc.), are also too generic and therefore difficult to abstract from heterogeneous training samples. Thanks to the fact that subjects are organised taxonomically, we entertain the idea that we could use such sparsity to our advantage. Specifically, we make the hypothesis that learning low-level subjects can help the classification of high level subjects. That is, we ask: *[RQ1] Can we use the knowledge of the subject taxonomy structure to help with the categorization?*

[1] The following summaries are produced with SPARQL Anything [17] on the original data sources from the Tate Gallery Collection open data project on GitHub. Data and queries can be reviewed and reproduced [18]

Tackling data sparsity by partitioning data by the means of key features. We observe how metadata could provide a useful input in understanding the reasons of the visual heterogeneity of artwork collections. Specifically, we partition the learning task dividing the data by technique/medium used. For example, we compare the task of learning subjects on artworks of any medium with the results obtained when subjects are learned only on photography. Thus, we pose the following research question: *[RQ2] Does splitting the set by artwork medium (e.g., photography, graphite, sculpture,...) help?*

Tackling data sparsity with neuro-symbolic learning. Finally, we explore the impact of different combinations of features on the learning performance, including both visual and non-visual features, such as the textual embeddings from the artwork title and the knowledge graph embedding from the artist entity on DBpedia. In other words: *[RQ3] Does combining text embeddings, KG embeddings, and visual embeddings improve the categorization performance?*

2.4 Methodology

To devise an architecture for subject classification from artwork images, we take as a reference ContextNet [5], a state of the art architecture for neuro-symbolic learning on artwork classification tasks. Namely, we treat the different artwork subjects - e.g., nature, people, architecture, as tasks to learn jointly in a Multi Task Learning (MTL) fashion. To generate visual embeddings from the input images, we maintain the same Convolutional Neural Network (CNN) backbone as ContextNet, i.e., a ResNet50 [19] from which the last fully-connected layer is removed.

However, differently from [5], our aim is to classify artwork subjects which can express a variety of factual elements about the artwork - e.g, physical, social, or abstract concepts. Therefore, we introduce a few modifications to the ContextNet framework to accommodate the task of subject artwork classification. The resulting pipeline is illustrated in Figure 2.1.

First, different annotations which concern the same task (or subject) can co-exist in an artwork. For instance, J.M.W. Turner's "Nant Peris, looking towards Snowdon" in Figure 2.1 depicts a mountain peak and a river view, both overlooked by a cloudy sky. These elements all fall under the *nature* subject. Thus, we configure the Network for multi-label classification. Specifically, in the last layer of the Network, Softmax activation is replaced by a Sigmoid activation. Indeed, while Softmax activations, which are typically interpreted as classification probabilities, are distributed across neurons, sigmoid outputs are computed independently on each class node. As a result, Sigmoids can predict more than one class with high probability. Similarly, we rely on a binary cross-entropy loss function, so that multiple subject predictions can be generated for an artwork.

In ContextNet, the visual embeddings are projected to the vector representation of the artwork context in the broader painting set [5]. This representation is derived from a KG where paintings are grouped by author and also annotated with attributes such as timeframe and medium. In the pipeline of Figure 2.1, we adopt a similar approach to ContextNet and infuse the Network with background knowledge through an encoder module, optimised through a smooth $\ell 1$ loss function. However, we test a different combination of embeddings to represent non-visual artwork features. In particular, we embed the title and artist metadata in the visual representation of the artwork. For the artwork title, we capitalise on a pre-trained linguistic model which has been shown to provide compact textual embeddings: DistilBERT [13]. To model the authorship information, instead, we apply the off-the-shelf RDF2Vec model [12] to the DBpedia entities which represent each artist. We can further concatenate the linguistic and KG embeddings, to derive a unified representation for the injected background features. Nonetheless, because the individual features are maintained as separate modules (Figure 2.1), we can also test the effects of incrementally adding new features to the learning process.

Ultimately, different components contribute to the overall training loss of the Network. We use the same notation as [5] to characterise these contributing factors through a set of parameters. First, the visual classification is influenced by the different learning tasks. Formally, the contribution of the t-th task to the binary crossentropy loss (ℓ_c) is weighted with respect to a λ_t so that $\sum_{t=1}^{T} \lambda_t = 1$. Similarly, because the classification and encoder modules are optimised through different loss functions, the relative contribution of each function to the overall loss is weighted

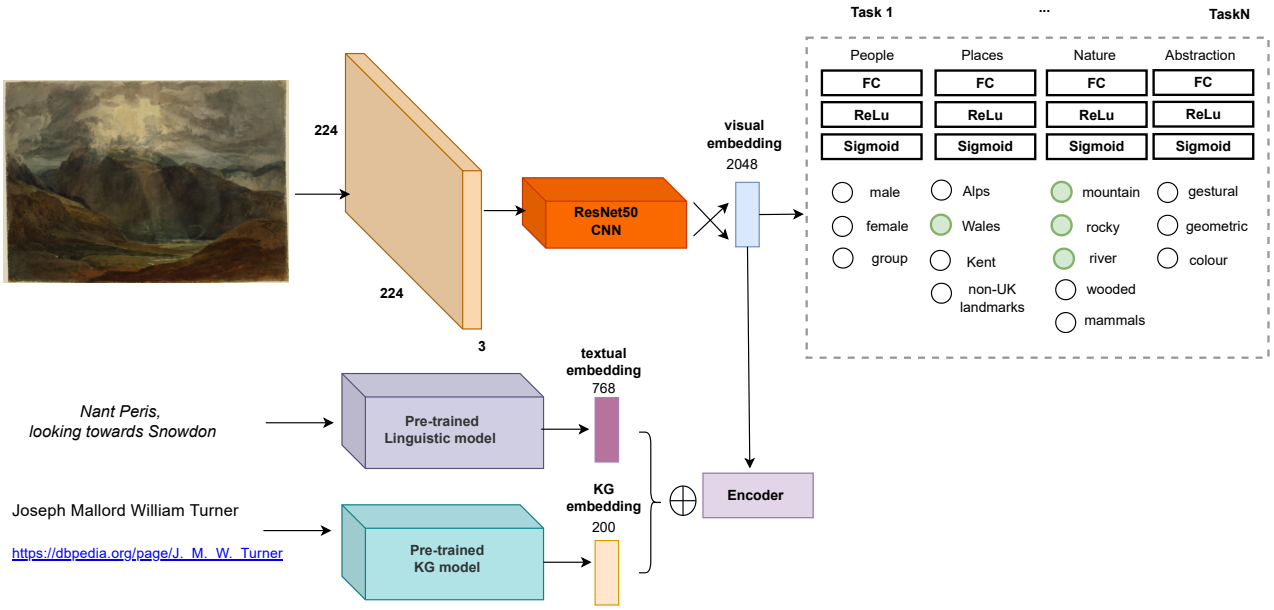


Figure 2.1: The proposed architecture for artist subject classification, where visual embeddings are combined with textual and KG embeddings, to jointly train the Network on multiple tasks - e.g., recognising people, places, nature, and so forth.

through different parameters. Let these weights be λ_c for the classifier loss, and λ_e, i.e., the complement to 1 of λ_c, for the encoder loss. By proxy, these two parameters allow us to leverage the degree to which the visual and non-visual components of the embeddings influence the training.

2.5 Experiments

2.5.1 Data preparation

In our experiments, we focus on the Tate Gallery Open Data as our reference cultural heritage archive. The published data includes two summary CSVs and more then 10k JSON files with detailed metadata on artworks and artists. The SPARQL Anything framework [17] provides a means to filter and integrate the required features from the broader CSVs and JSON documents provided in this collection [18]. In addition to retrieving the JPEG image files that are available from the Tate website, we extracted the following data fields: (i) the unique artwork identifier, (ii) the title, (iii) the subject annotations, as well as (iv) the medium, or material, of each artwork. We also queried DBpedia to retrieve the entities, marked through a Uniform Resource Identifier (URI), which match a certain artist name. Because a string can match multiple DBpedia entities and to account for homonyms, we manually validated the collected artist entities.

We start by considering the 11 top-level concepts that provide abundant training examples (i.e., at least 3'000 examples per class), for the purpose of supervised Deep Learning. These are listed in Table 2.1.

Moreover, we want to reduce the sparsity of the example distribution across the sub-categories of a subject. With the term sub-category, we refer to the children of a subject, in the Tate taxonomy - e.g., “female figure" is a sub-category of “people". Thus, we focus on the six subjects which provide the highest number of sub-concepts. As highlighted in blue in Table 2.1, these are: nature, architecture, places, people, objects, and abstraction.

With this premise, we further prune the space of sub-concepts with a two-fold objective. First, because samples should ideally overlap across the top-level subjects that are learned jointly, we select sub-categories which are worth

Table 2.1: Statistics of the top-level subjects in the Tate collection.

Top subject	Artworks	Subjects
Nature	36,477	24
Architecture	29,787	19
Places	23,842	12
People	20,798	14
Society	13,991	6
Objects	12,381	10
Abstraction	8,503	8
Emotions, Concepts, Ideas	8,248	4
Work and Occupations	5,133	3
Symbols and Personifications	5,022	3
Leisure and Pastimes	3,129	2

at least 700 examples. Second, for each subject, we want to select non-overlapping sub-categories which identify distinct concept groups. Thus, if two child categories of the same node are selected at the previous step, only the more specific one is retained. For instance, we prioritise annotations of male and female portraits over the more generic portrait label.

After retaining only records which are annotated with respect to the target categories and sub-categories, we are left with 54,494 records[2], 99.6% of which (54,293 records) have a non-corrupted image file associated.

Ultimately, we want to ensure that examples are balanced across categories when forming our training, validation, and test splits. Therefore, we resort to the multi-label stratified sampling strategy proposed in [20, 21], which is conveniently provided with the scikit-multilearn package[3]. At this stage, we apply a 80/10/10 ratio to split the data into training, validation, and test sets.

Additionally, because we are also interested in grouping artworks by artistic medium (RQ2), we prepared a dedicated subset for each medium. The metadata which describe the different artistic media are sparsely annotated. Thus, we ought to apply a series of basic Natural Language Processing (NLP) steps to converge towards coherent groups. Specifically, we reduced the raw text to lowercase, and derived a set of tokens which excludes the standard English stopwords, and which is free from alphanumeric characters and spurious white spaces. After deriving word stems through the Snowball method, we also filtered out any duplicated tokens - e.g., "*paper* graphite, on *paper*".

Thanks to the availability of a reference glossary of art terms on the Tate website[4], we could derive a set of keywords to canonicalise heterogeneous medium annotations. Specifically, we merged semantically-related keywords (e.g., ink and pen) to gather a sufficient number of data points per medium. In sum, we converged towards ten subsets that are representative of different materials. These are: *painting*, *sculpture*, *graphite*, *etching*, *screenprint*, *watercolour & gouache*, *ink & pen*, *lithograph*, *engraving & intaglio*, and *photography*.

2.5.2 Experimental setup

To address our main research questions (Section 2.3), we configure three distinct experiments. Across all experiments, the performance on each task (i.e., subject) is evaluated in binary terms. Specifically, in addition to the overall classification accuracy on a task, we also track the Precision (P), Recall (R), and F1 achieved on the retrieval of positive examples of a task. For instance, in the context of the *people* task, the correct recognition of a person in a portrait increases the P, R, and F1 metrics, irrespective of the system's ability to recognise that a different artwork does not depict any people.

[2] This number is lower than the sum of the figures in Table 2.1, as the same artwork can be annotated with more than one subject.
[3] http://scikit.ml/
[4] https://www.tate.org.uk/art/art-terms/

Experiment A. The objective of the first experiment is to assess whether or not configuring the learning space on the basis of the Tate subject taxonomy improves the classification performance (RQ1). Thus, we start by considering a simplified version of the architecture presented in Section 2.4, where only the visual embeddings extracted from a CNN are considered. In this context, we compare two training configurations. The first configuration only relies on the macro-categories which represent each task, whereas the second configuration considers finer-grained annotations for each task. In the former case, the network is optimised to generically classify people, places, objects, natural, abstract elements, and architectural components. In the latter configuration, the goal is to learn sub-classes of each subject - e.g., to recognise mountains, rivers, and beaches, as opposed to classifying “nature" generically - as in the example of Figure 2.1.

Experiment B. The second experiment is conceived to test the effects of splitting training examples by artwork medium - e.g., watercolour, sculpture, intaglio, to further reduce the sparsity of the learning space (RQ2). Therefore, in this setup, we start by evaluating the performance results obtained when the complete image collection is considered, without discriminating by artistic medium. We then repeat the performance assessment across the ten sub-samples which we have prepared for different artistic media, as described in Section 2.5.1.

Experiment C. The last experiment is a study of the impact of visual, textual, and KG embeddings when learning artwork subjects (RQ3). In particular, we explore the integration of numeric features representing the title and artist of an artwork. Therefore, at this stage, we contrast the performance of the following pipelines or ablations:

(img) The first method considers only visual embeddings to classify artworks, thus following the same methodology of Experiments A and B.

(img + text) In this pipeline, the visual embeddings are also optimised with respect to the linguistic embeddings extracted from a pre-trained DistilBERT model. Specifically, the linguistic model is fed with the title of each artwork. To derive a single vector for each input sentence, we follow a series of transformations which are standard practice in NLP. First, the hidden states produced by the last four layers are summed together to derive a word vector for each input token. Then, we average the second to last hidden layers of each token to form the sentence embedding.

(img + KG) We also test a neuro-symbolic variation of the “img" pipeline, where the visual embeddings are projected onto the 200-dimensional embeddings returned by a RDF2Vec model [12] which was pre-trained on DBpedia entities. Specifically, if the DBpedia URI associated with an author is found in the RDF2Vec feature space, the related KG embedding is retrieved to guide the optimisation of the visual embedding, through the encoder module (Section 2.4).

(img + text + KG) Lastly, we consider the scenario where the textual embedding and the KG embedding are concatenated, to contribute to the learning routine. In other words, this ablation models the methodology of Section 2.4 (Figure 2.1).

2.5.3 Results

Experiment A. The results obtained on the test set when training the model only on macro-subject labels are reported in Table 2.2a. With the exception of the nature task, the model is incapable of detecting the presence of any artwork elements. The non-zero accuracies indicate that the model has learned to produce only negative predictions for the tasks (except for nature). In fact, for the *nature* class, which contributes the highest number of training examples, the model outputs mostly positive predictions. Hence, it has simply learned to replicate the imbalanced distributions of the training data.

However, when finer-grained subject annotations are introduced, the performance improves, particularly in terms of Precision (Table 2.2b). Nevertheless, the overall performance remains dramatically low across all tasks. Indeed, while introducing finer-grained categories may help discriminating different subjects, it also makes the distribution of training examples for each specialised category more sparse.

Table 2.2: Performance comparison when we change the hierarchical level (granularity) of the subject labels used for training.

(a) Training on subject level one

Task	Acc	Pre	Rec	F1
nature	0.64	0.64	1	0.78
people	0.65	0	0	0
architecture	0.47	0	0	0
objects	0.8	0	0	0
places	0.58	0	0	0
abstraction	0.86	0	0	0

(b) Training on subject level two

Task	Acc	Pre	Rec	F1
nature	0.39	0.45	0.01	0.02
people	0.65	0.29	0	0.01
architecture	0.52	0.37	0.01	0.01
objects	0.84	0.14	0	0
places	0.63	0.21	0	0
abstraction	0.86	0.75	0.01	0.02

Experiment B. Based on findings from the previous experiment, here we consider finer-grained subject categories for training our models. However, in this experiment, data points are further sampled by artistic medium. As shown in Figure 2.2, in the majority of cases where the model is trained only on a specific medium, the F1 score is higher than in the baseline scenario, where all materials are considered. A marginal performance decay was instead recorded when classifying *places* from *sculptures*, where the already low F1 dropped to zero. In the scenario where only *graphites* were considered, all results are equivalent or only marginally higher than the baseline curve, except for the *abstraction* task, were the improvement was most pronounced. In the remaining scenarios, splitting examples by artistic medium significantly benefited the performance. In particular, the highest F1 scores were achieved by training only on:

- photographies to classify *nature* and *people*
- engraving & intaglio examples to classify *people* and *places*
- ink & pen works on the *architecture* task
- sculptures to classify *objects*
- screenprints to classify *abstraction*.

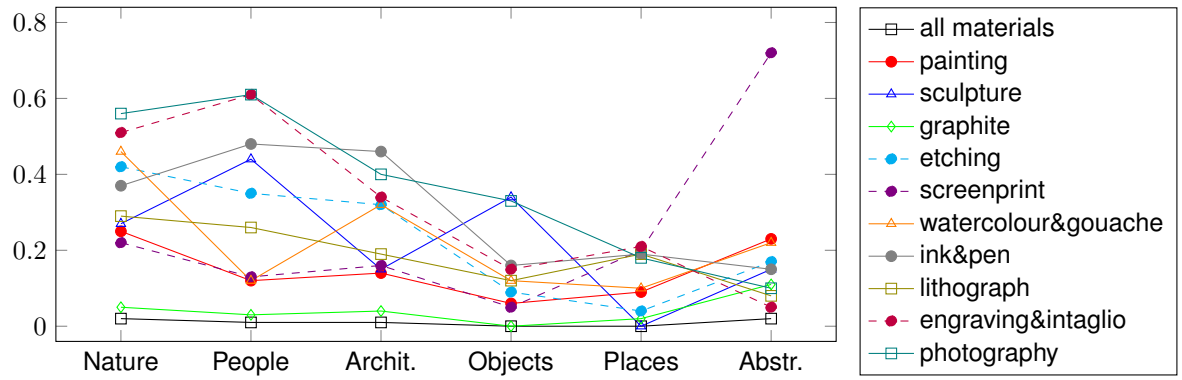


Figure 2.2: Comparison of F1 scores before (black curve) and after (coloured curves) splitting the dataset by artwork media.

Experiment C. Figures 2.3 and 2.4 illustrate the results obtained with the top-performing methods of Experiment B (i.e., those trained solely on *photography*, *engraving & intaglio*, *sculpture*, *ink & pen*, and *screenprint*), through different combinations of features. Crucially, the integration of non-visual features enhanced the baseline DL performance across the majority of tasks and media. However, different performance trends can be observed that are medium-specific and task-specific.

To classify photography, linguistic embeddings were relatively more beneficial, in terms of performance increase, than KG embeddings (Figure 2.3a). However, combining different embedding types led to highest F1 on the classification of nature, people, and objects. Similarly, on the ink & pen sample, the integration of all tested features produced the highest performance when classifying nature, people, places, and abstraction (Figure 2.3d).

In the case of sculptures, the largest margin of improvement is associated with the introduction of linguistic embeddings, for the majority of tasks (Figure 2.3c). Screenprints, instead, exhibit an opposite trend: overall, integrating KG embeddings was preferable, in terms of performance, to only relying on linguistic features (Figure 2.3e). Interestingly, the top performance achieved through the baseline on the abstraction task was unmatched, even after integrating both the title and the artist features (Figure 2.4a). Indeed, the abstraction subjects explored in this evaluation mostly encode colour and geometric traits, which are best learned through visual features.

On the engraving & intaglio set, the effects of applying neuro-symbolic learning differs from task to task (Figure 2.3b). On the nature, architecture, places, and abstraction tasks, the introduction of text and KG embeddings ensured a significant performance increase. By contrast, the improvement is only marginal when classifying objects. The case of the people task is interesting because the integration of the linguistic embeddings led to a performance degradation, and the introduction of KG embeddings only matched the baseline F1. However, leveraging both types of embeddings improved the performance by 5%.

Overall, different tasks are learned most efficiently through a different combination of features, on different mediums. Specifically, the highest F1 scores are observed:

- **for nature**: *img+text* on *engraving* (Figure 2.4b)
- **for people**: *img+text+KG* on *photography* (Figure 2.4d),
- **for architecture**: *img+text* and *img+KG* on *engraving & intaglio* (Figures 2.4b,2.4c)
- **for objects**: *img+text+KG* on *photography* (Figure 2.4d), *img+text* on *sculpture* (Figure 2.4b)
- **for places**: *img+text* on *engraving & intaglio* (Figure 2.4b)
- **for abstraction**: *img* on *screenprint* (Figure 2.4a).

2.5.4 Implementation details

The experiments discussed in this section were conducted on Google Colaboratory. In our training configuration, we initialised the ResNet50 module with weights pre-trained on ImageNet. Weights of the classification heads were instead initialised through the Xavier method [22]. Consistently with [5], parameters were updated via stochastic gradient descent. In particular, we set the learning rate to 0.0001, with a weight decay of 0.00001 and a momentum of 0.9. Through preliminary experiments to this paper, we found that updating parameters across the entire model is preferable to fine-tuning only the last classification layer, likely due to the marked differences between the ImageNet benchmark and the Tate collection. All tested models were trained for up to 150 epochs, with an early stopping condition whenever the validation loss did not decrease for 30 successive epochs.

Each task contributed equally to the classification loss, i.e., for the six tasks explored in this paper, λ_t was set to 0.165. After testing different weight configurations for the loss, we set $\lambda_c = 0.9$ and $\lambda_e = 0.1$ across all experiments. That is, we observed empirically that giving higher importance to the visual embeddings at training time ensures a higher performance, on average.

Input images were resized to 224×224 pixels and normalised with respect to the ImageNet mean and standard deviation. We relied on the Pytorch and transformers Python libraries to implement the proposed architecture. The RDF2Vec embeddings, which we downloaded locally to speed up the processing time, are conveniently exposed through the KGVec2Go resource [23].

The code, data, and pre-trained models which reproduce these experiments are available at: https://bit.ly/3p3WV3M.

2.6 Discussion

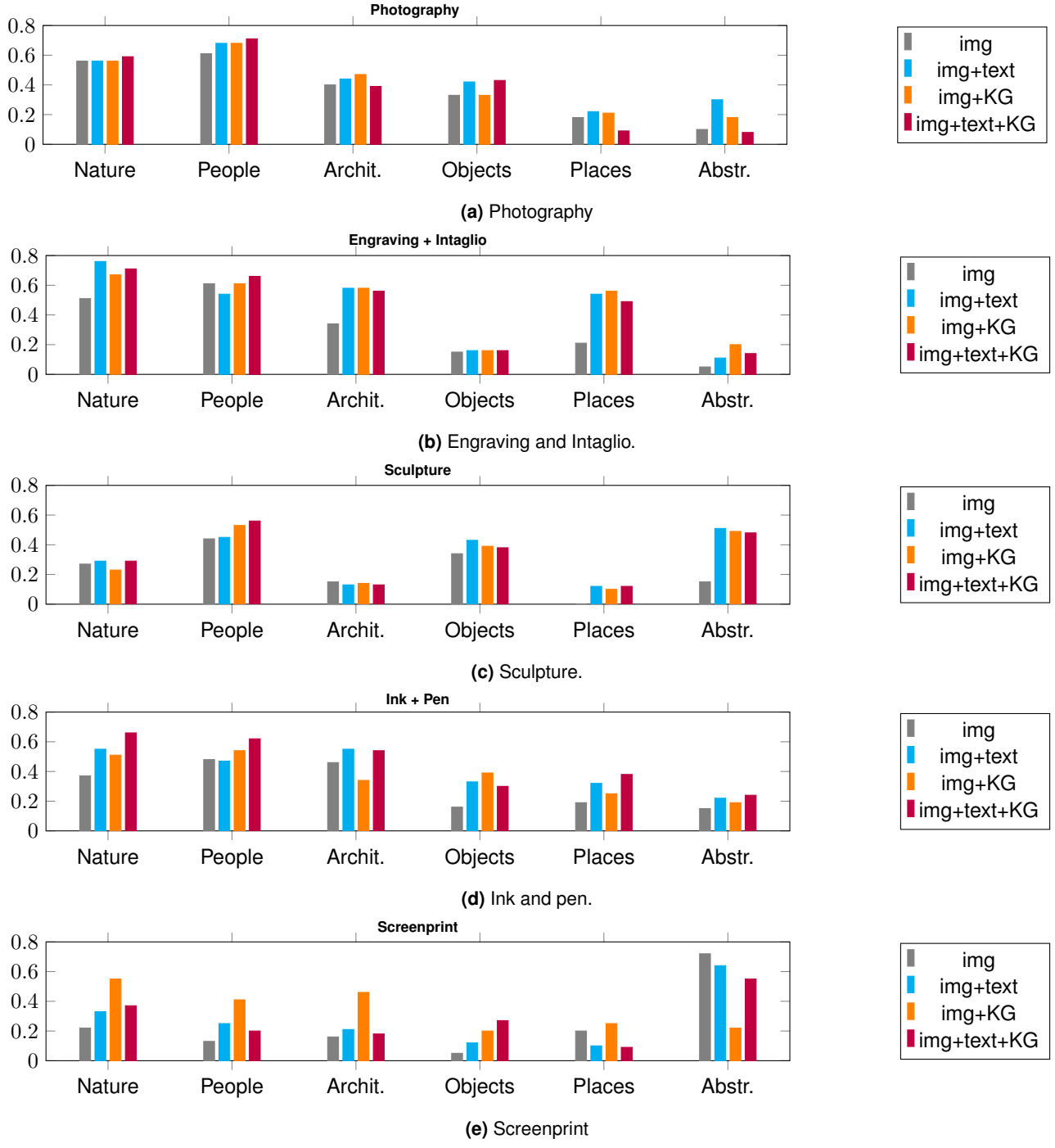
We conducted experiments with the purpose of exploring strategies for handling the data sparsity that characterises cultural heritage collections. We asked whether the taxonomical structure of the labels can help in learning top level categories (RQ1). Indeed, organising the training space according to the semantic hierarchy of objects helped improve the classification performance. However it is not sufficient, alone, to handle sparsity. Next, we explored the idea that image features may vary depending on the artistic medium (RQ2) and therefore learning should be performed separately for each medium. We demonstrated how splitting by artistic medium helped significantly improve the performance across the majority of tasks (with only one caveat: *places* on *sculptures*). Finally, we conducted extensive experiments to study the impact of different feature sets on the various subjects, learnt on the different media (RQ3). Here, we observe that different features are helpful for learning different tasks on different media, and that there is no feature set that performs systematically better on all media and subjects.

On the basis of these findings, we elaborate on possible future work. First, learning what combinations are performing well was a costly operation. This evidence poses the question of how to autonomously learn the feature set that is most representative for each subject. Hence, these results spark an important meta-learning task: **To what extent can we automatically devise the appropriate learning strategy depending on the three dimensions of *feature modality*, artistic *medium*, and *subject*?**

Second, we found that certain tasks are better learned on certain media. Could we use this behaviour as an opportunity for transfer learning? In other words, could we use a model trained on a specific medium to recognise the same subject on another medium? Even further, could we use the learned model on a different image collection, to compensate for the imbalance (or even scarcity) of training examples? In the context of the SPICE project, we plan to apply this method for characterising subjects on citizen-curated collections such as the Irish Museum of Modern Arts (IMMA) archive[5], within the Ireland case study.

Finally, in this work, we have attacked the classification problem as a Multi Task Learning (MLT) setting, following relevant priors in the state of the art. Future work includes exploring the correlation between different tasks, i.e., studying which tasks are best learned jointly and which ones should be learned separately.

[5] http://imma.ie



(a) Photography

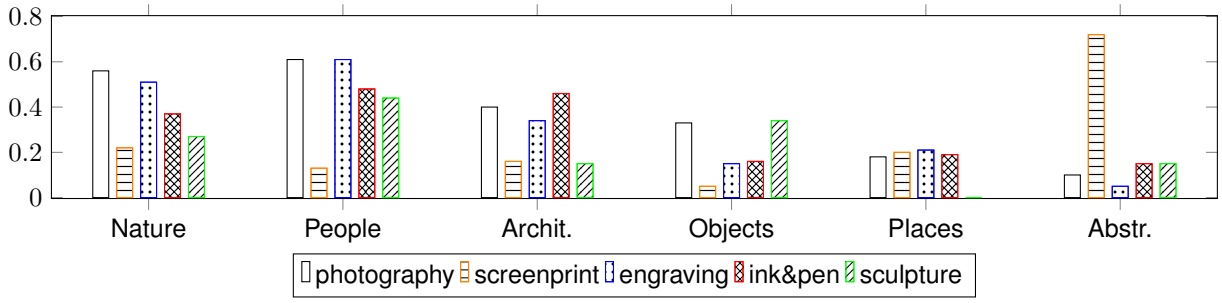
(b) Engraving and Intaglio.

(c) Sculpture.

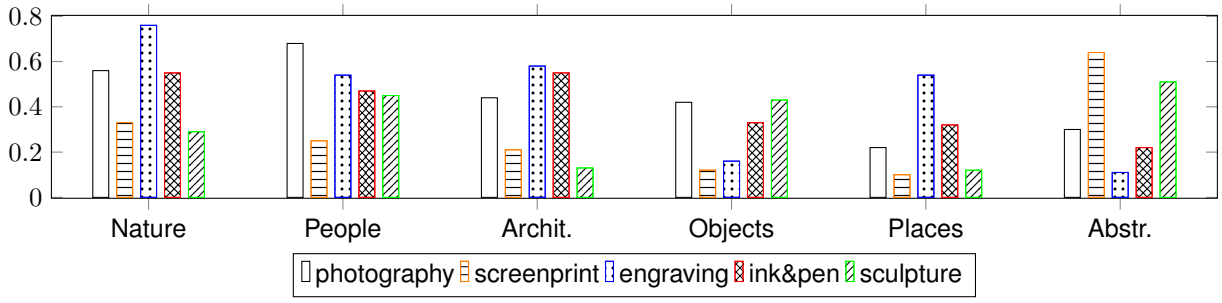
(d) Ink and pen.

(e) Screenprint

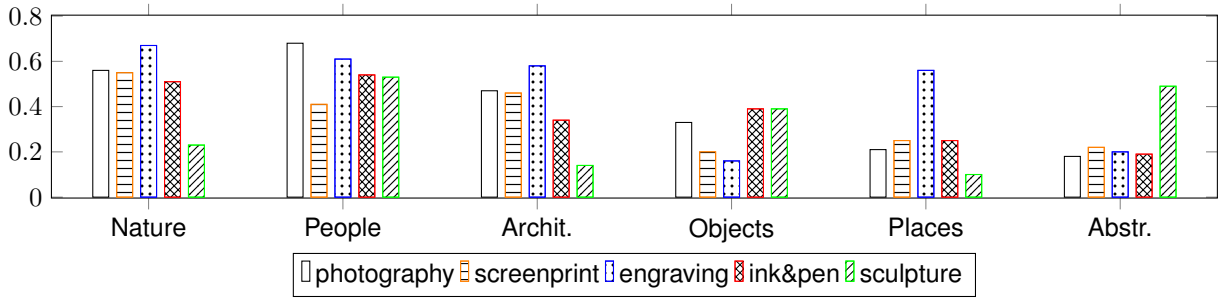
Figure 2.3: Experiment results by *medium*: comparison of F1 scores after incrementally combining visual embeddings (img) with textual embeddings on the artwork title (text) and Knowledge Graph embeddings on the artist feature (KG).



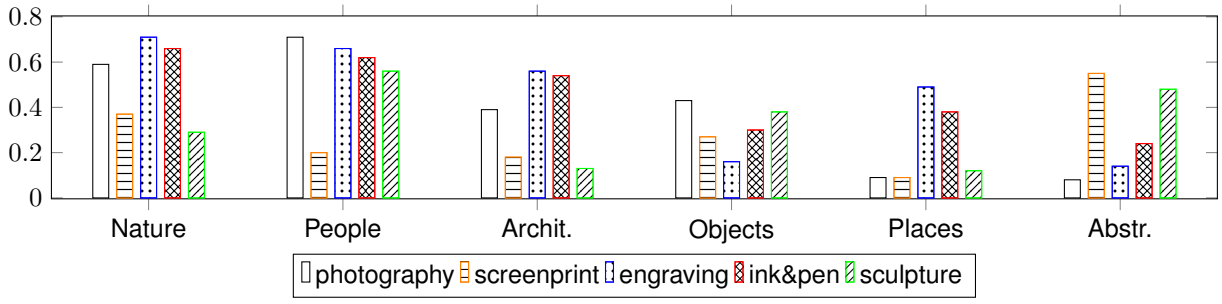
(a) img



(b) img + text



(c) img + KG



(d) img + text + KG

Figure 2.4: F1 results for the experiments with different features (img, text, KG), grouped by medium.

3 The DEGARI system for generating affective-based descriptions

3.1 Introduction

DEGARI 2.0 is an affective reasoner, one of the two developed in the context of SPICE (the other one is the Semantic Annotator, developed in WP3: see Deliverable 3.4), that exploits the logic $\mathbf{T}^{\mathrm{CL}}$ [24] in order to generate and classify content according to the circumplex theory of emotions devised by the cognitive psychologist Robert Plutchik [25, 26][1]. According to this theory, emotions, and their interconnections, can be represented on a spatial structure, a wheel (as reported in the left of the Figure 3.1), in which the affective distance between different emotional states is a function of their radial distance. DEGARI 2.0 exploits the Plutchik's ontology (see Deliverable 6.2), formalizing such a theory. In particular, by following Plutchik's account, complex emotions are considered as resulting from the composition of two basic emotions (where the pair of basic emotions involved in the composition is called a dyad). The compositions occurring between similar emotions (adjacent on the wheel) are called primary dyads. Pairs of less similar emotions are called secondary dyads (if the radial distance between them is 2) or tertiary dyads (if the distance is 3), while opposites cannot be combined[2]. An illustrative example showing the rationale used by DEGARI 2.0 to generate the compound emotions (in this case, the emotion Love as composed by the basic emotions Joy and Trust, according to Plutchik's theory) is reported in Figure 3.1. In the next sections, we frame the contribution of such a system in the context of affective-based item linking and suggestions.

3.2 Related work

Emotions have been acknowledged as a primary component of the artistic experience for centuries; recently, their role in art has been demonstrated through physiological experiments showing how correlates of emotions, such as brain response and facial expressions, are affected by the experience of art [28, 29]. In addition to their role in defining the way people experience artistic expression [30], from paintings and musical works to movies and novels, emotions also provide a universal language through which people convey their experience of art, well beyond words. Despite the differences in the expression of emotions across languages, and the influence of cultural factors, in fact, emotions own an universal origin [31]: rooted in evolution, they provide the basis for intercultural communication, as effectively demonstrated by the advancements in facial expression recognition [32, 33]. In this sense, emotions can provide a suitable means for connecting people belonging to different groups, intended as culture, age, education, and different sensory characteristics. Pervasive in human communication, emotions are expressed through multiple channels, ranging from facial expression and body posture to spoken and written language. The expression of emotions through language, in particular, lies at the basis of several models of emotions, including Shaver's [34] and Plutchik's [26], and has prompted the creation of a number of resources for sentiment analysis in language [35, 36, 37]. The application of these resources to art is straightforward: for example, the WikiArt Emotions project [38] has collected the emotional response to the WikiArt online art collection, yielding a dataset of 4,105 artworks with annotations for the emotions evoked in the observer. Experiments such as WikiArt Emotions have paved the way to the extraction of emotions from text and tags to create affective art recommenders, like ArsEmotica [39, 40] or the first version of DEGARI [41], able to classify and group artistic items well beyond the standard 6 basic emotions of

[1] The reasons leading to the choice of this model as grounding element of the DEGARI 2.0 system is twofold: on the one hand, this it is well-grounded in psychology and general enough to guarantee a wide coverage of emotions, thus giving the possibility of going beyond the emotional classification and recommendations in terms of the standard basic emotions suggested by models like the Ekman's one (widely used in computer vision and sentiment analysis tasks). This affective extension is aligned with the literature on the psychology of art suggesting that the encoding of complex emotions, such as *Pride* and *Shame*, could give further interesting results in AI emotion-based classification and recommendation systems [27]. Second, as anticipated above, the Plutchik wheel of emotions is perfectly compliant with the generative model underlying the $\mathbf{T}^{\mathrm{CL}}$ logic.

[2] The ontology is available here: https://raw.githubusercontent.com/spice-h2020/SON/main/PlutchikEmotion/ontology.owl.

Ekman's theory [31] and embracing richer, finer-grained models. DEGARI 2.0 is framed within this global challenge and has been used to link and suggest museum items sharing different types of emotional stances.

3.3 Background and research questions

State of the art affective recommenders for the artistic domain are not yet able to deal with the 'echo chamber' problem (i.e. they are only able to suggest and aggregate items evoking the same emotion) and, in addition, they have never been tested on the grounds of diversity and inclusion. On the other hand, here we show how DEGARI 2.0 can be used as an inclusive, explainable and diversity-seeking affective art recommender, aimed at bridging the differences in the experience of art between different communities. In particular, such system aims at overcoming the limitations of traditional recommendation approaches by exploiting the Plutchik's ontological structure to suggest museum items not only labeled with the same emotions, but - as mentioned - also to group and recommend artworks evoking *similar* (but not exactly the same) emotions or *opposite* emotions. This kind of alternation in the content suggestion mechanism aims at leading to more comprehensive exploration and fruition of museum collections. Indeed, suggesting museum items evoking different emotions from the ones already experienced via the fruition of other artworks, is based on the notion of perspective taking [42], i.e. seeing the world (e.g. an exhibition in this case) from other perspectives. Since this approach is used to promote empathy, cohesion and inclusion across social groups, reaching this goal would represent a huge advancement with respect to the current technologies (e.g. like social media or standard recommender systems) that often lead people toward content that fits their own viewpoint, promoting fragmentation and fostering confirmation biases, instead of cohesion, inclusive reflection, and critical thinking.

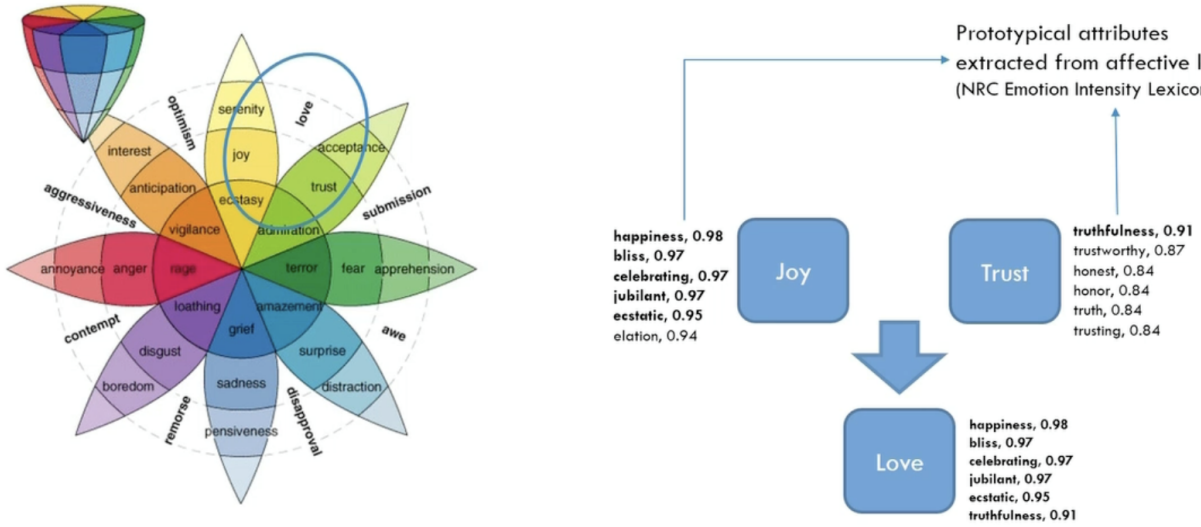


Figure 3.1: Generation of novel Compound Emotions with DEGARI 2.0 by exploiting the Plutchik's ontology (e.g. Love as composed by Joy and Trust in the picture). The features and the probabilities characterizing each basic emotion are obtained from the NRC lexicon. The Plutchik's wheel of emotion in this figure reports only the compound emotions representing the primary dyads, but our system works on the entire spectrum of dyads.

3.4 Methodology

In order to connect associated lexical features (used as typical properties) to the classes of the Plutck's ontology recreated in the T $\mathbf{T}^{\text{CL}}$ language, we used the NRC lexicon [35].

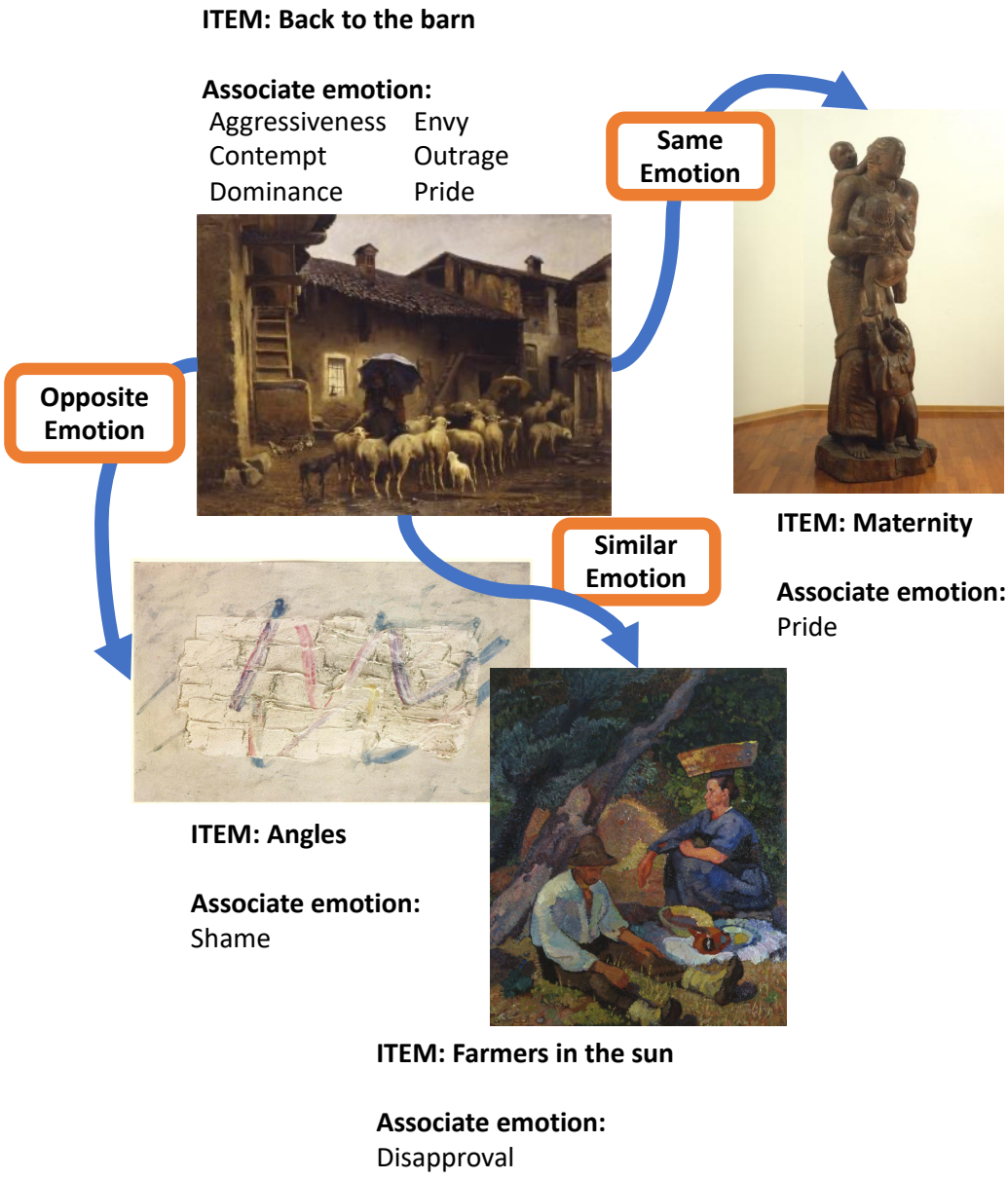


Figure 3.2: Example of Same, Similar and Opposite emotion recommendations of DEGARI 2.0 from the GAM dataset. This figure shows how the system is able not only to generate new compound emotions (see e.g. Figure 1) but also to group and suggest cultural items according to their obtained Plutchik's-based affective classification. The entire dyadic structure of the Plutchik's model is exploited to recommend items evoking different emotional stances with the aim of providing a more inclusive and affective-based interpretations of cultural content.

Once the prototypes of the compound emotions are generated, the system is able to reclassify museum items taking the new, derived emotions into account. As a consequence, such a reclassification allows the system to group and recommend museum items based on the novel assigned labels and, as mentioned, a novel prerogative of DEGARI 2.0 consists in the possibility of delivering also diversity-seeking recommendations.

The Figure 3.2 reports an example of these different kinds of suggestions for the artefact entitled "Ritorno alla stalla"

("Back to the barn") of the GAM museum (Galleria Arte Moderna, in Turin). Based on the system's output, this item is emotionally linked to "Maternità" ("Maternity": a statue of the GAM collection classified as evoking the same emotional content as the original painting: "Pride"), to "Contadini al sole" ("Farmers in the sun", a painting labelled with the similar emotion "Disapproval"; note that this emotion is considered "similar" according to the Plutchik's model since it is the one spatially adjacent to the category "Outrage" that is one of the categories, a "tertiary dyad" in the Plutchik's theory, to which the system has assigned the original item) and, finally, to the abstract painting "Angles", labelled with the opposite emotion "Shame" (in this case the opposition concerns the label "Pride"). Overall, the system tries to categorize and link the items with respect to any of the original emotional categories found.

As anticipated, a final crucial feature of the DEGARI 2.0 classification system is represented by the fact that the rationales of its classifications are entirely transparent and explainable (see Deliverable 6.3).

3.5 Experiments

DEGARI 2.0 has been evaluated in a number of ways (see Deliverable 6.6). Here we report an evaluation focused on the user acceptability of the received inclusive-based affective recommendations for the GAM collection in Turin. This measure is important since it is an indicator of how good the system is in detecting and suggesting the discovered connections between items that are affectively related. The reported data extends a previous evaluation of the same type reported in Deliverable 6.3. The evaluation consisted in a user study where 74 deaf participants, after having been exposed to a number of affective-based recommendations based on their original selection, were asked to compile an online questionnaire about the received suggestions. Here they had to rate, on a 10-point scale (from 1 to 10), the received recommendations based on the 'same-emotion' 'similar-emotions' and 'opposite emotions' categories . Overall they rated 91 recommendations.

Below, we report the results of the prototype applications of DEGARI 2.0 to the datasets provided by Gallery of Modern Art (GAM).

Mean score:	5,79			
Median total score:	6			
		Same Emotion	**Similar Emot.**	**Opposite Em.**
mean		5,78	6,23	5,25
median		6	6	5
standard deviation		0,61	0,71	1,06

Table 3.1: Results of the ratings of the deaf group in GAM on the DEGARI recommendations

The overall obtained results about the ratings are shown in Table 3.1. The users showed a moderate acceptance of the received content suggestions. The average rating assigned to the total set of emotion categories proposed by DEGARI was 5.79 with a median value of 6/10. Table 3.1 shows the mean, median and standard deviation values for each emotion recommendation group (same, similar and opposite emotions). The recommendations that received a better rating were the ones suggesting items linked to the original one through the property "similar emotion". The recommendations of items evoking opposite emotions (with respect to the original item selected in the game) were the ones that received the worst rating.

3.6 Discussion

Overall, this experiment shows that the effort of tackling diversity-seeking, affective-based and explainable museum recommendations received a moderate, improvable, acceptance. This datum also suggests that there are mechanisms of cognitive resistance that prevent a full acceptance of connections going in a different direction from one's own preferences. From here, a first guideline that can be extracted for the improvement of diversity-seeking affective

recommenders concerns the opportunity to adopt presentation devices for the **mitigation of cognitive resistance effects**. Although the search for mitigation measures that wrap diversity into some meaning frame is an open research area, the effectiveness of narrative formats [43, 44] and of ethically-driven digital nudging techniques [45, 46] is worth exploring. A more immediate strategy that could be adopted in our system is also represented by the progressive recommendation of items evoking emotions that are gradually more distant from the starting one (where the distance can still rely on the radial structure of the Plutchik's wheel encoded in the ontology).

4 Conclusions

In this deliverable, we developed novel methods for supporting the interlinking and discoverability of digital assets across museum collections, for supporting citizen curation applications. In order to do that, we focused on automatically generating semantic metadata that would allow linking between distributed assets originating from different data management systems, and complement the annotations derived by the curators. Next, we plan to integrate these methods within the workflow of SPICE pilots, specifically the Ireland and Turin case studies.

Bibliography

[1] A. Lieto, G. L. Pozzato, S. Zoia, V. Patti, and R. Damiano, "A commonsense reasoning framework for explanatory emotion attribution, generation and re-classification," *Knowledge-Based Systems*, p. 107166, 2021.

[2] A. Lieto, G. L. Pozzato, M. Striani, S. Zoia, and R. Damiano, "Degari 2.0: A diversity-seeking, knowledge-based, explainalable, and affective art recommender for social inclusion," *Cognitive Systems Research*, 2022.

[3] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self-and unsupervised learning for image classification," *IEEE Access*, vol. 9, pp. 82 146–82 168, 2021.

[4] H. Yang and K. Min, "Classification of basic artistic media based on a deep convolutional approach," *The Visual Computer*, vol. 36, no. 3, pp. 559–578, 2020.

[5] N. Garcia, B. Renoust, and Y. Nakashima, "Contextnet: representation and exploration for painting classification and retrieval in context," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 1, pp. 17–30, 2020.

[6] G. Algan and I. Ulusoy, "Image classification with deep learning in the presence of noisy labels: A survey," *Knowledge-Based Systems*, vol. 215, p. 106771, 2021.

[7] The Tate Gallery, "Tate Collection metadata," 2014. [Online]. Available: https://github.com/tategallery/collection

[8] R. Del Chiaro, A. D. Bagdanov, and A. Del Bimbo, "Noisyart: A dataset for webly-supervised artwork recognition." in *VISIGRAPP (4: VISAPP)*, 2019, pp. 467–475.

[9] S. Liu, J. Yang, S. S. Agaian, and C. Yuan, "Novel features for art movement classification of portrait paintings," *Image and Vision Computing*, vol. 108, p. 104121, 2021.

[10] G. Castellano and G. Vessio, "Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview," *Neural Computing and Applications*, vol. 33, no. 19, pp. 12 263–12 282, 2021.

[11] G. Castellano, G. Sansaro, and G. Vessio, "Integrating contextual knowledge to visual features for fine art classification," in *DL4KG'21: Workshop on Deep Learning for Knowledge Graphs*. CEUR, 2021.

[12] P. Ristoski, J. Rosati, T. Di Noia, R. De Leone, and H. Paulheim, "Rdf2vec: Rdf graph embeddings and their applications," *Semantic Web*, vol. 10, no. 4, pp. 721–752, 2019.

[13] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019.

[14] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, "Neuro-symbolic artificial intelligence: Current trends," *arXiv preprint arXiv:2105.05330*, 2021.

[15] E. Daga, L. Asprino, R. Damiano, M. Daquino, B. D. Agudo, A. Gangemi, T. Kuflik, A. Lieto, M. Maguire, A. M. Marras *et al.*, "Integrating citizen experiences in cultural heritage archives: requirements, state of the art, and challenges," *ACM Journal on Computing and Cultural Heritage (JOCCH)*, vol. 15, no. 1, pp. 1–35, 2022.

[16] P. Mulholland, E. Daga, M. Daquino, L. Díaz-Kommonen, A. Gangemi, T. Kulfik, A. J. Wecker, M. Maguire, S. Peroni, and S. Pescarin, "Enabling multiple voices in the museum: challenges and approaches," *Digital Culture & Society*, vol. 6, no. 2, 2020.

[17] E. Daga, L. Asprino, P. Mulholland, and A. Gangemi, "Facade-X: an opinionated approach to SPARQL anything," in *Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands*, vol. 53. IOS Press, 2021, pp. 58–73.

[18] E. Daga, "SPARQL Anything showcase: open data from the Tate Gallery," May 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6518424

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[20] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," *Machine Learning and Knowledge Discovery in Databases*, pp. 145–158, 2011.

[21] P. Szymański and T. Kajdanowicz, "A network perspective on stratification of multi-label data," in *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, ser. Proceedings of Machine Learning Research, L. Torgo, B. Krawczyk, P. Branco, and N. Moniz, Eds., vol. 74. ECML-PKDD, Skopje, Macedonia: PMLR, 2017, pp. 22–35.

[22] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[23] J. Portisch, M. Hladik, and H. Paulheim, "Kgvec2go–knowledge graph embeddings as a service," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 5641–5647.

[24] A. Lieto and G. L. Pozzato, "A description logic framework for commonsense conceptual combination integrating typicality, probabilities and cognitive heuristics," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 32, no. 5, pp. 769–804, 2020.

[25] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[26] ——, "The nature of emotions," *American scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[27] P. Silvia, "Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 3, pp. 48–51, 2009.

[28] N. N. Van Dongen, J. W. Van Strien, and K. Dijkstra, "Implicit emotion regulation in the context of viewing artworks: ERP evidence in response to pleasant and unpleasant pictures," *Brain and Cognition*, vol. 107, pp. 48–54, 2016.

[29] H. Leder, G. Gerger, D. Brieber, and N. Schwarz, "What makes an art expert? Emotion and evaluation in art appreciation," *Cognition and Emotion*, vol. 28, no. 6, pp. 1137–1147, 2014.

[30] I. Schindler, G. Hosoya, W. Menninghaus, U. Beermann, V. Wagner, M. Eid, and K. R. Scherer, "Measuring aesthetic emotions: A review of the literature and a new assessment tool," *PloS one*, vol. 12, no. 6, p. e0178899, 2017.

[31] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[32] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures." *Emotion*, vol. 18, no. 1, p. 75, 2018.

[33] I. M. Revina and W. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, 2021.

[34] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor, "Emotion knowledge: further exploration of a prototype approach." *Journal of Personality and Social Psychology*, vol. 52, no. 6, p. 1061, 1987.

[35] S. Mohammad, "Word affect intensities," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. European Language Resources Association (ELRA), 2018. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2018/summaries/329.html

[36] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok, "Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis," in *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, 2020, pp. 105–114.

[37] Y. Susanto, A. G. Livingstone, B. C. Ng, and E. Cambria, "The hourglass model revisited," *IEEE Intelligent Systems*, vol. 35, no. 5, pp. 96–102, 2020.

[38] S. Mohammad and S. Kiritchenko, "WikiArt emotions: An annotated dataset of emotions evoked by art," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani,

H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga, Eds. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1197

[39] V. Patti, F. Bertola, and A. Lieto, "Arsemotica for arsmeteo. org: Emotion-driven exploration of online art collections," in *The Twenty-Eighth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2015)*, I. Russell and W. Eberle, Eds., Association for the Advancement of Artificial Intelligence. AAAI Press, 2015, pp. 288–293. [Online]. Available: http://www.aaai.org/Library/FLAIRS/flairs15contents.php

[40] F. Bertola and V. Patti, "Ontology-based affective models to organize artworks in the social semantic web," *Information Processing & Management*, vol. 52, no. 1, pp. 139–162, 2016. [Online]. Available: https://doi.org/10.1016/j.ipm.2015.10.003

[41] A. Lieto, G. L. Pozzato, S. Zoia, V. Patti, and R. Damiano, "A commonsense reasoning framework for explanatory emotion attribution, generation and re-classification," *Knowl. Based Syst.*, vol. 227, p. 107166, 2021.

[42] T. Pedersen, A. Wecker, T. Kuflik, P. Mulholland, and B. Diaz-Agudo, "Introducing empathy into recommender systems as a tool for promoting social cohesion," in *Joint Proceedings of the ACM IUI 2021 Workshops, April 13-17, 2021, College Station, USA*. CEUR Workshop Proceedings, 2021.

[43] A. Wolff, P. Mulholland, and T. Collins, "Storyspace: a story-driven approach for creating museum narratives," in *Proceedings of the 23rd ACM conference on Hypertext and social media*, 2012, pp. 89–98.

[44] R. Damiano, V. Lombardo, A. Lieto, and D. Borra, "Exploring cultural heritage repositories with creative intelligence. the labyrinth 3d system," *Entertainment Computing*, vol. 16, pp. 41–52, 2016.

[45] A. Augello, G. Città, M. Gentile, and A. Lieto, "A storytelling robot managing persuasive and ethical stances via act-r: an exploratory study," *International Journal of Social Robotics*, pp. 1–17, 2021.

[46] C. Gena, P. Grillo, A. Lieto, C. Mattutino, and F. Vernero, "When personalization is not an option: An in-the-wild study on persuasive news recommendation," *Information*, vol. 10, no. 10, p. 300, 2019.