ARTO: An Artwork Object Ontology for Descriptive and Contextual Captioning

Can Yang[1 [0000−0002−0223−659X]], Bernardo Pereira Nunes[1 [0000−0001−9764−9401]], Sergio Rodríguez Méndez[1 [0000−0001−7203−8399]], Yige Chen[1 [0009−0006−9952−9773]], Rubén Manrique[2 [0000−0001−8742−2094]], and Marco Antonio Casanova[3 [0000−0003−0765−9636]]

[1] Australian National University, Canberra ACT 2612, Australia
Can.Yang@anu.edu.au, Bernardo.Nunes@anu.edu.au, Sergio.RodriguezMendez@anu.edu.au, Aylin.Chen@anu.edu.au
[2] Universidad de los Andes, Bogotá 111711, Colombia
rf.manrique@uniandes.edu.co
[3] PUC-Rio, Rio de Janeiro 22451-900, Brazil casanova@inf.puc-rio.br

Abstract. Existing artwork ontologies focus mainly on artwork management, lacking detailed content and contextual representation. To fill this gap, we propose the ARTwork object Ontology (ARTO), which is designed to provide a comprehensive representation of artwork from both descriptive and contextual perspectives. The descriptive model represents the artistic expression of artwork, including visual elements, scenes, and emotions, while the contextual model captures background information, such as historical framework and related events. The ontology uses a data-driven approach, building on available data and incorporating concepts from established cultural heritage ontologies such as CIDOC-CRM and EDM. We evaluate ARTO using the OntOlogy Pitfall Scanner! and conducte interviews with 10 art experts to validate and refine our model. The results demonstrate the adequacy and comprehensiveness of ARTO through its application in artwork captioning to enhance automated artwork description and understanding.

1 Introduction

With the development of digital technologies, cultural heritage institutions have evolved how they catalogue and manage artworks [17,35], which relies heavily on accurate and consistent metadata. Building upon a theoretical foundation [38] that established a comprehensive conceptual framework for artwork representation, this research advances from theory to practice. The previous work highlighted the critical need for integrating descriptive, contextual, and interpretive aspects of artworks. While that framework offered valuable theoretical insights, implementing it in practice remains a significant challenge. Despite advances in artwork management, artwork captioning remains a difficult task. Our research seeks to bridge this gap between theory and practice by developing a concrete implementation through a novel ontology that captures and models the complex

and comprehensive knowledge in the art domain, providing a solid foundation for generating captions in a broad scope.

It is worth noting that existing ontologies often fail to capture multiple perspectives and focus primarily on basic information and broad contexts of artworks, such as title, format, size, and time of creation [12]. The lack of detailed and insightful content representations of artwork ontologies also affects the accuracy and completeness of artwork captions, as they may not objectively, comprehensively, and completely reflect the hidden meanings of the artwork, thus limiting the viewer's understanding of the artwork and the story behind it. Thus, it is needed to develop ontologies that incorporate multiple, integrative perspectives. Besides, established captioning approaches ([1,10,19,22,33,37]) are mainly for realistic images. Although there have been some attempts at image captioning for artworks, such as [9, 16, 18, 20, 21, 29, 39], most of them directly apply existing methods used for recognition tasks to train machine learning models using art-related data often overlooking the significant differences between artworks and realistic images. Captions for artworks highlight more cultural and historical information and interpretation of images, rather than simply listing their contents. Thus, this paper introduces a novel artwork caption ontology - ARTwork Object Ontology (ARTO) - that aims to provide comprehensive and holistic information about artworks. While we demonstrate ARTO's capabilities using paintings as our primary example, its foundational structure is designed to accommodate various art forms. The hierarchical representation of elements, objects, and scenes provides a versatile framework that naturally extends to other artistic mediums. For instance, sculptures can be analysed through the same scene-object relationships with additional three-dimensional properties, while performance arts can be conceptualised as sequences of interconnected scenes. We chose paintings for our initial implementation due to their rich visual complexity and well-documented contextual information, which allows us to thoroughly validate the ontology's expressive capabilities. Given the complexity and challenge of generating captions for paintings in the artwork captioning domain, demonstrating the applicability of the ontology through this medium provides a solid theoretical foundation for its broader utility across different forms of artistic expression. The proposed ontology is structured in two parts: (i) the *Artwork Descriptive Model*, emphasising artistically specialised content in artworks, including visual elements, objects, scenes, and connotations; and, (ii) the *Artwork Contextual Model*, providing a comprehensive collection of artwork-related information, including historical context and related events.

Our validation process for ARTO combined its data-driven development with qualitative evaluation through Think-Aloud Protocol (TAP) interviews [5]. This approach enabled us to validate both the structure and content of the model from a practical perspective, ensuring alignment with the requirements and objectives of artwork captioning tasks. The logical side of ARTO was also evaluated using the OntOlogy Pitfall Scanner! (OOPS!) [25]. Unlike existing artwork ontologies such as CIDOC-CRM [4], EDM [12], and Linked Art[4], ARTO specifically focuses

[4] `https://linked.art/`

on artwork captioning through enhanced content representation and contextual information. Our main contribution is an ontology comprising the *Artwork Descriptive Model* and *Artwork Contextual Model*, validated through a systematic protocol that can be extended to broader arts domains.

The paper is organised as follows: Section 2 reviews artwork ontologies and knowledge graphs (KGs) in artwork and general image captioning. Section 3 presents ARTO's components - the Artwork Descriptive Model and Artwork Contextual Model - with an RDF-star[5] application example. Section 4 evaluates the ontology through automatic assessment and expert interviews. Section 5 concludes with our contributions and future work on KG construction and integration with Large Language Models (LLMs) for artwork captioning.

2 Related Work

Existing ontologies and vocabularies have made significant contributions to the organisation and representation of knowledge in the art domain.

The CIDOC Conceptual Reference Model (CIDOC-CRM) [4] stands as a cornerstone in this field, offering a widely adopted standard for describing concepts and relationships in cultural heritage. While CIDOC-CRM is good at representing art-related relationships and events, it lacks detailed visual descriptions and emotional content capture, which are crucial for artwork captioning. In the realm of bibliographic records, the Functional Requirements for Bibliographic Records (FRBR) [13] model has been adapted to describe cultural heritage objects. FRBR's clear hierarchical structure enhances information retrieval and interoperability. However, its application in the visual arts is limited. Building upon CIDOC-CRM, the Visual Representation (VIR) Ontology [7] extends its capabilities to focus more specifically on visual heritage. VIR offers a structured framework for capturing visual elements of artworks. However, it lacks a detailed representation of visual elements and symbols, and lacks a representation of contextual semantics, such as the symbolic meanings of elements, relationships between characters, etc. Therefore, the artwork description has limitations. The ArCo [8] KG primarily focuses on cataloguing Italian cultural heritage. It emphasises the description of cultural and historical contexts, but has some limitations in supporting artwork image captioning. As ArCo is designed specifically for Italian cultural heritage, it may not apply to a wider type of artworks. Moreover, its focus on cataloguing rather than captioning results in a lack of visual description of artworks. The Europeana Data Model (EDM) [12] has made significant progress in facilitating interoperability and data integration among diverse sources in the digital cultural heritage domain. Its support for rich multilingual data and Linked Open Data principles enhances the connectivity of cultural heritage information. Nevertheless, EDM's primary focus is metadata management rather than detailed artwork captioning. The Visual Resources Association Core (VRA CORE) [34] is a widely used metadata standard for describing visual

[5] `https://w3c.github.io/rdf-star/cg-spec`

resources, including artworks and cultural heritage objects. Its primary objective is to facilitate data interoperability and sharing among cultural heritage institutions. However, VRA Core lacks descriptions of visual elements, making it inadequate for capturing the visual features and artistic techniques of artworks. The ICON Ontology [28] brings a specialised focus to the description of visual artworks, particularly in terms of iconographic and iconological content. While ICON excels in representing symbolic meanings and objects in artworks, it has limitations in comprehensively representing visual content. Specifically, it lacks multi-level representations of artworks that are crucial for a complete understanding of an artwork's visual structure. Linked Art is a data model for describing artworks and other cultural heritage objects. The goal of Linked Art is to provide a standardised model to facilitate data interoperability and sharing. Although Linked Art can describe multiple aspects of artworks, it may not be detailed enough to capture the visual details important for caption generation.

Getty Research Institute has developed the Art & Architecture Thesaurus (AAT) [14] and the Union List of Artist Names (ULAN) [15] vocabularies that provide a foundation for knowledge in the art domain. The AAT is a structured vocabulary that covers concepts in art and related domains. The ULAN focuses on standardising artist names and has been employed to construct visualisations of artist social networks [31].

Although existing ontologies are widely used in the art domain, they lack comprehensive representation across descriptive, contextual, and interpretive dimensions needed for artwork captioning. Existing approaches often excel in one aspect, while struggling to balance multiple information requirements simultaneously. Building upon the framework proposed by Yang et al. [38], our ARTO ontology addresses these limitations through two integrated components: the Artwork Descriptive Model for visual content-based interpretations and the Artwork Contextual Model for historical and cultural aspects. ARTO focuses specifically on paintings, capturing rich visual details, relationships, and contextual information necessary to generate comprehensive artwork captions.

3 The ARTwork object Ontology (ARTO)

ARTO captures the multi-perspective of artworks through two interconnected models: the Artwork Descriptive Model and the Artwork Contextual Model. The Artwork Descriptive Model focuses on the expressive content of the artwork itself, addressing the need for detailed visual and emotional content representation that is often lacking in existing ontologies. The Artwork Contextual Model details the metadata, creation background, provenance, and artists' experience. By combining these two models, ARTO provides a structured framework that enables the generation of comprehensive, multi-layered artwork captions.

3.1 Methodology

The construction of the artwork ontology follows the NeOn methodology [30], which emphasises the reusing and refactoring of existing ontologies. Firstly, we

collected and analysed data from online galleries and museums to build the conceptual structure of the ontology. Then, we identified the core concepts, such as artists, events, styles and creation techniques, and defined the relationships between them, such as "belongs to", "has event" and "created by", forming a semantic network. Throughout this process, we consider reusing existing ontologies and vocabularies by analyzing established standards such as CIDOC-CRM [4], EDM [12] and the Event ontology [27] and establishing mappings between ARTO's elements and corresponding concepts in these ontologies through semantic relationships. This approach promotes interoperability and leverages established conceptualizations.

As for ontology design evaluation and validation, we employed OOPS! [25] to assess the logical consistency and completeness of the ontology. We also invited 10 art experts to review the ontology's structure and content, incorporating their feedback to refine and improve the ontology. Finally, we have published our ARTO [6] where we provide human-readable and machine-readable documentation, describing its structure and content. To accommodate different users and systems, the ontology is available in multiple serialisation formats, such as RDF/XML and Turtle. We use the MIT license agreement to specify the usage rights and permissions. We have established version control and update mechanisms to incorporate new knowledge and address any identified issues, ensuring the ontology remains up-to-date and relevant.

3.2 Ontology Design Rationale

Table 1. Data Sources Statistics

Data Source	Number of Artworks	Has Schema?	Number of Attributes	Copyright
Europeana	2,224,268	Yes	17	Fully Public
Wikiart	172,394	No	17	Partially Public
Artic	93,836	No	9	Partially Public
American art	39,388	No	13	Partially Public
Louvre	512,613	No	9	Public Metadata
Nation Museum	82,463	No	6	Partially Public
NGA.gov	139,723	No	26	Fully Public
MET museum	218,930	No	12	Partially Public
Brooklyn Museum	67,529	No	13	Partially Public
Getty	86,848	Yes	10	Partially Public
British Museum	2,123,229	No	16	Partially Public
Art bank	10,727	No	9	For research and study

Designing an ontology requires both specialised expertise in the field of art and a deep understanding of the existing data intended to populate the on-

[6] ARTO is available at: `https://w3id.org/arto`.

tology. The ontology not only needs to capture the broad categories and relationships inherent in the art but also be fine-tuned to the specific characteristics, techniques and contexts that are present in the collection. Our design approach combines data-driven analysis with theoretical foundations in art. We adopted a methodology that integrates the analysis of artwork data, reuse of established ontologies, and validation through expert interviews. This approach ensures that ARTO is both practically grounded and theoretically sound. We began by collecting artworks from well-known art websites listed in Table 1 [7], such as Wikiart, Art Institute of Chicago and the US National Gallery of Art. These websites typically provide two types of information: basic artwork details (e.g., title, artist, location, year) and their contextual descriptions. From the collected artworks, we filtered to retain only those with available images, since our ontology mapping requires visual content analysis. We first extracted and organised the basic artwork information according to their meanings, as different websites often use varying terminologies. The other part is about the overall description of the artwork, which varies greatly. Popular artworks usually have a lot of descriptions, including specific visual elements of the artwork and interpretations, the artist's creative background and even the artist's biography. Most lesser-known artworks, however, have little to no description. We further analysed the content of these descriptions and classified them. Using the collected data, we constructed a preliminary ontology. ARTO encompasses 31 classes covering artwork-related concepts (such as Artwork and its subclasses Painting/Sculpture/Photograph), scene elements, objects, visual elements, and contextual aspects (including Agent, Time, and Location). The ontology defines 25 object properties establishing relationships between entities, and 10 datatype properties capturing specific attributes. Regarding constraints, ARTO implements 14 mandatory properties (minCardinality) ensuring essential information capture, 20 unique properties (cardinality 1) maintaining data integrity, and incorporates domain/range constraints for all properties. The ontology demonstrates strong semantic interoperability through its integration with 8 external ontologies including CIDOC-CRM, Schema.org, and FOAF, and establishes semantic alignments through equivalent class mappings. Our analysis of multiple data sources revealed the need for more detailed artwork representation. ARTO addresses this by defining a structured framework (Figure 1) for comprehensive, context-rich image captioning.

3.3 Artwork Descriptive Model

A review of existing artwork-related ontologies reveals that most of them do not attempt to model the content of artworks, leaving a significant gap. The content is an essential part of an artwork, as artists convey specific themes and emotions through elements such as colour, texture and composition, all of which are key components of the artwork. Due to the complexity of artwork content, traditional classification methods struggle to capture their rich connotations.

[7] Please see details in `https://w3id.org/arto`

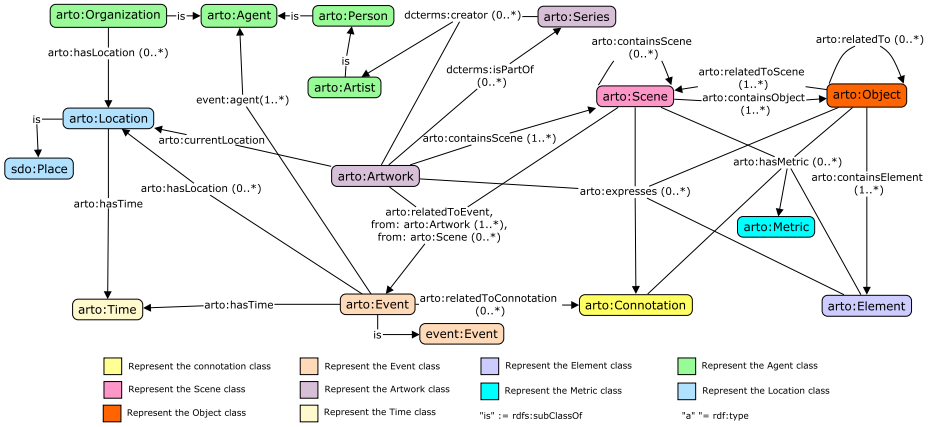

Fig. 1. Core Model of the Artwork Object Ontology. *The colours are used consistently in this paper to represent different classes and their instances.

Similarly to the Document Object Model (DOM) [36], which neutrally captures the inner structure of documents, we propose the Artwork Descriptive Model aiming at describing the content of artworks. We note that this model currently targets artworks that contain distinguishable real-world objects, such as "landscape paintings" or "portrait paintings". Figure 2 presents the concept map of the Artwork Descriptive Model, which decomposes artwork content into four levels: basic element, object, scene and connotation:

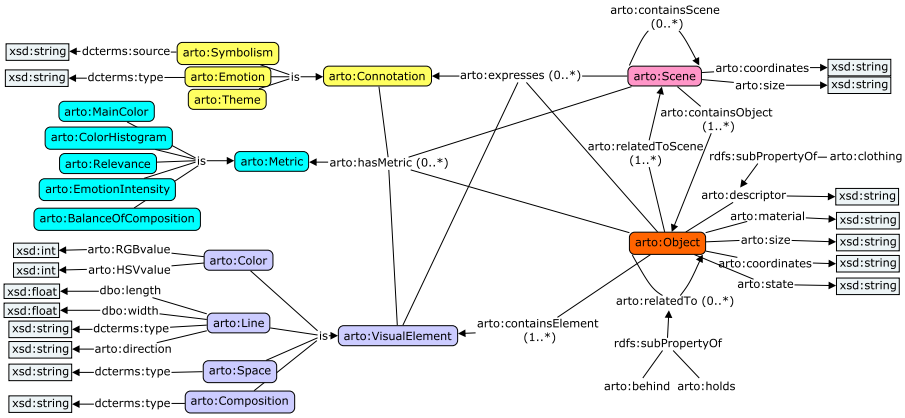

Fig. 2. Artwork Descriptive Model.

– **Basic Elements**: Basic elements contain colour, line, shape, texture, space and composition. Colour and line form the foundation, as their combination

can create shape and texture. Space pertains to the representation of objects and their three-dimensional relations within the artwork and composition involves the arrangement and layout of these objects.
- **Object**: Objects are the main components of the artwork. Various objects interrelate and collectively form the entire artwork. Each object has its specific state and attributes that reflect the details and depth of the artwork.
- **Scene**: A scene represents meaningful visual content within the artwork, which is composed of a group of objects or visual elements that convey a specific meaning or sentiment. A complex artwork is typically not constituted by only one scene, but rather by multiple meaningful scenes that jointly articulate the overall theme of the artwork. The visual content of the entire artwork can be considered as a macro-level scene (for static artworks). Each scene plays an indispensable role in the whole artwork.
- **Connotation**: From the high-level perspective to analyse the symbolism, theme and emotion of the artwork. While the theme of most artworks can be discerned directly from their content, some works, especially those that are implicit or abstract, cannot be recognised directly. For such artworks, understanding the background of its creation, and the artist's experiences becomes the key to truly understanding the connotation of the artwork.

To map the Artwork Descriptive Model to the ontology, each layer of the model is represented as a class. To facilitate quantitative analysis, the `Metric` class represents measurable aspects of the content as we explain in what follows.

The `VisualElement` class represents the fundamental visual components of an artwork, with `Colour`, `Line`, `Space` and `Composition` as its key subclasses for paintings. Each subclass of `VisualElement` has its own unique properties. For instance, the `Colour` class includes properties of `RGBvalue` and `HSVvalue`, while the `Line` class has properties like type, direction, length, and width. These properties enable a detailed description of the basic elements within an artwork.

The `Object` class represents the identifiable and meaningful entities depicted in an artwork, which are composed of various `VisualElement`s. Objects can be further described by their specific properties, such as `size`, `material` and `state`, which indicates the condition or appearance of the object. The properties of an `Object` vary depending on its type, therefore, the `descriptor` property allows for extendable subproperties to describe any aspects of different objects. Figure 6 illustrates the extensibility through the "clothing" subproperty.

An artwork's entire visual content can be viewed as a macro `Scene`, thus each artwork must `containsScene` with at least one scene. The `Scene` class represents a meaningful aggregation of objects and visual elements, where objects are linked to their respective scenes through the `relatedToScene` relationship, allowing an object to be assigned to one or more scenes. Additionally, a `Scene` can contain other `Scene` instances, creating a hierarchical structure. Since scenes in artworks may depict either real or fictional events, a `relatedToEvent` property connects the `Scene` and `Event` classes, establishing a relationship between visual representations and their narrative context.

The `Connotation` class is at the top of the hierarchy, capturing the high-level semantics and implicit meanings of an artwork. It has three subclasses: `Symbolism`, `Emotion` and `Theme`. The `Symbolism` represents the underlying message of the objects or scenes. The `Emotion` describes the emotional expression of the artwork. The `Theme` subclass encapsulates the central idea or subject of the artwork. In artwork, the `Scene`, `Object` and `Element` are all relevant to the expression of `Connotation`. Moreover, many connotations actually originate from specific events, such as mythology or historical stories. For instance, "Achilles' heel" is used to describe a fatal weakness, which comes from the mythological story of Achilles, a hero in the Trojan War. Similarly, "Waterloo" was originally just a place name in Belgium, but because Napoleon was defeated there, it is now commonly used to describe a major failure or downfall.

For structural clarity and better analysis of the content of the artwork, we designed a `Metric` class to represent the various possible metrics, including some statistical information about the basic elements such as colour statistics, emotion intensity and metrics of other visual elements. These classes jointly build a multi-level ontology for representing the content of artworks.

3.4 Artwork Contextual Model

While the Artwork Descriptive Model focuses on the intrinsic elements and content of the artwork, the Artwork Contextual Model provides necessary contextual information to situate the artwork within its historical, cultural and artistic context. This model: (i) characterises artworks by capturing essential properties, such as the genre, medium and style; (ii) brings out underlying events or facts that are represented by the characters or objects in the artwork; (iii) captures events related to the artwork, such as exhibitions, provenance, or events related to the artist, such as where they studied, and travelled, as well as identifies the time and location for each event; (iv) establishes connections between artworks and various agents involved in their life cycle—from creation to presentation.

The Artwork Contextual Model builds upon and extends existing ontologies in the cultural heritage domain, such as CIDOC-CRM [4] and EDM [12]. These ontologies provide a solid foundation for representing cultural heritage information. Our model aims to narrow down the context of artworks and simplify the representation of concepts. We borrowed concepts representing classes and properties from CIDOC-CRM and EDM, such as `E5_Event`, `E53_Place` and `E52_Time-Span`, to build our Artwork Context Ontology.

Our model is also inspired by the W3C Provenance (PROV-O)[8] and Event Ontologies [27]. PROV-O offers a way to represent the provenance information of things, which is crucial for tracing the history and ownership of artworks. The Event Ontology provides a framework for describing events and their relationships to entities, which fits well with our goal of capturing events related to artworks and artists. By leveraging these ontologies' modelling structure and

[8] https://www.w3.org/TR/prov-o/

approach, we built our event modelling, which effectively represents historical context, events behind creation, provenance and others.

Furthermore, our model incorporates elements from other art-related ontologies, such as the Art and Architecture Thesaurus [14] for representing artistic styles, techniques and materials, and the Union List of Artist Names [15] for identifying and disambiguating artists. By reusing and extending existing ontologies, the Artwork Context Model (Figure 3) provides a more detailed representation of artworks while maintaining interoperability with other data sources. The Artwork Context Model covers properties and the context of artworks. The ontology comprises five main classes: `Artwork`, `Event`, `Location`, `Time` and `Agent`, emphasising extensibility and offering flexible design for different classes.

Although ontologies for time, location, and similar concepts exist, we developed custom classes tailored to the specific requirements of the art domain. Existing ontologies offer detailed representations but are often too complex for contextualising artworks in our use case. For instance, many artworks lack a precise time point and instead exist within an ambiguous time period, making standard time representations difficult to apply. Following the KISS principle ("Keep It Simple, Stupid"), we designed custom classes (`Time`, `Location`, and `Event`) that simplify the model structure while retaining core functionality. This enhances usability and maintainability while maintaining semantic compatibility and interoperability with established ontologies through subclass relationships.

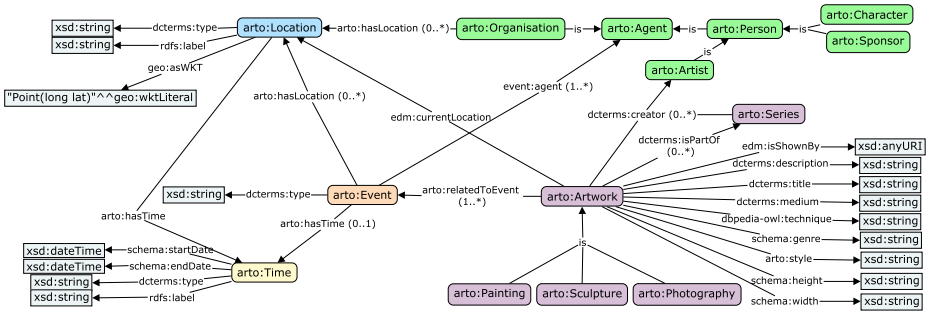

Fig. 3. Artwork Contextual Model.

While the foundational structure of our ontology is designed to accommodate various art forms, the `Artwork` class primarily encompasses visual art, including categories such as Painting, Sculpture, Photography, Printmaking, and Digital Art. For this paper, we focus on `Painting` as the primary subclass of `Artwork` due to its rich visual complexity and well-documented contextual information. Although our illustration in Figure 3 explicitly demonstrates `Painting`, `Sculpture`, and `Photography`, the ontology's categories are designed to be open-ended and flexible, allowing for extension to embrace the diverse and evolving nature of artistic expression in the art world.

Additionally, to maintain relevance and applicability, the `Series` class represents a collection of artworks. The ontology includes the fundamental properties of an artwork, including title, theme, description, elements, style, medium, format and image resource. It also allows for the representation of more specific properties relevant to different types of artworks.

The `Time` class represents one of the most relevant aspects of an artwork. Time potentially encodes events that influenced its creation. The temporal data provide the chronology of artworks and also imply their historical, cultural, or personal contexts. Representing time is not a straightforward task, especially for historical data. The time data might be annotated with precise dates or be associated with broader epochs, seasonal timelines, or even vague indications like 'late Renaissance'. These diverse temporal expressions mandate a flexible representation model. Therefore, we use `dcterms:type` to categorise temporal expressions, which can accommodate specific years, months, days, seasons as well as broader, more qualitative time descriptions such as 'early/late 19th century'. All instances of time are treated as intervals meaning each time instance has a `startDate` and an `endDate`. This interval-based methodology provides a structured and consistent means of representing time. Therefore, the `Time` class is adaptable and extensible. A direct adoption of the Time Ontology[9] would introduce unnecessary complexity to our model.

Regarding the `Location` class, we have implemented a strategy similar to the time type, considering the frequent correlation between location and time in historical contexts. The discernment of specific locations within historical records poses considerable challenges due to the dynamism and variability inherent in geographical delineations over time. To streamline this intricate process, we also use `dcterms:type`, which could encompass a wide range of geographical entities such as countries, kingdoms, states, provinces, cities, districts, etc., allowing the representation of diverse spatial granularities. Moreover, we have integrated geographical coordinates to serve as precise locational indicators, facilitating more accurate spatial analysis. Therefore, our ontology adopts the `geo:wktLiteral` datatype from GeoSPARQL [23] standard to represent the specific coordinates of the location. Where historically existent nations or regions lack unequivocal coordinates or where boundary definitions have evolved, we endeavour to align them with contemporary geographical demarcations to deduce an approximate scope. This alignment process is meticulous, ensuring that the approximated locations maintain as much historical accuracy as feasible, respecting the historical integrity of the regions involved. To connect and represent historical location names, we set an `hasTime` property that links the location with a time limitation. Our approach maintains flexibility to accommodate updates and refinements in location data. As new information becomes available or as historical understandings evolve, the `Location` class can be adjusted to reflect more accurate or nuanced geographical understandings, ensuring the continual relevance and accuracy of the spatial representations within our framework. This multifaceted

[9] https://www.w3.org/TR/owl-time/

approach to location allows for a comprehensive and adaptable representation of geographical entities and their interrelations over time.

In the given context, an `Agent` emerges as the main participant in any event, primarily comprising organisations and persons. `Organisation` can represent a myriad of institutional forms, such as museums, galleries and art schools. The `Person` class can be broken down into more specific classifications, such as artists, sponsors and characters. Moreover, this classification is dynamic and inclusive, accommodating the incorporation of additional roles as various events are incorporated. It is open-ended, allowing for the integration of new participant types as they are identified, ensuring the `Agent` class remains comprehensive and reflective of the diversity and complexity of participation in events.

The structuring of the `Event` class is fundamentally anchored in three cardinal elements: time, location, and agents. The concept modelling is inspired by PROV-O, Event ontology and DOLCE [6] which provides a comprehensive framework for capturing and representing real-world events. Our approach ensures the capture of various background information pivotal to understanding the artworks. For instance, it encompasses events occurring during the creation of the artwork, such as the artist's experiences, interactions, and the prevailing historical and cultural environment, all of which intricately weave into the fabric of the resultant artwork. These elements offer insights into the motivations, influences, and conditions that shaped the artwork. It is also designed to integrate events post-creation, encompassing aspects like provenance history and exhibition history. These elements collectively contribute to the evolving narrative of the artwork, marking its journey through time and space and its interactions with various entities. To maintain an optimal balance between flexibility and comprehensiveness, we introduced the `dcterms:type` attribute. This captures the evolving and diverse events without compromising on specificity. We employed RDF-star to represent the `Event` class for its enhanced capability to represent more complex relationships and events compared to traditional RDF, which is inherently constrained to binary relationships. When we use RDF to describe events, we need an additional event entity to serve as an intermediate entity to connect all the entities related to the event, such as people, time, and location. However, with RDF-star, different entities are directly associated and it can better handle complex events and model their direct relationships. This approach results in cleaner graph structures and more intuitive querying, making it easier to capture temporal, spatial, and contextual information. RDF-star not only extends our ability to encapsulate intricate events associated with artworks but also provides a more concise and easy-to-understand representation.

3.5 Artwork Object Ontology Example

Figure 4 depicts "The Shepherdess" by William-Adolphe Bouguereau. We represent its contextual information using RDF-star format in Figure 5[12], while Figure 6 illustrates the artwork's content through a multi-level deconstruction.

[11] https://en.wikipedia.org/wiki/The_Shepherdess

[12] Example details can be found at `https://w3id.org/arto`.

Fig. 4. "The Shepherdess"[11], also known as "The Little Shepherdess", is a painting by William-Adolphe Bouguereau completed in 1889. The title is taken from the Southern French dialect. The painting depicts an idyllic, pastoral scene of a lone young woman in peasant attire posed for the artist, balancing a stick (likely her crook) across her shoulders, standing barefooted in the foreground. In the background are oxen grazing in a field.

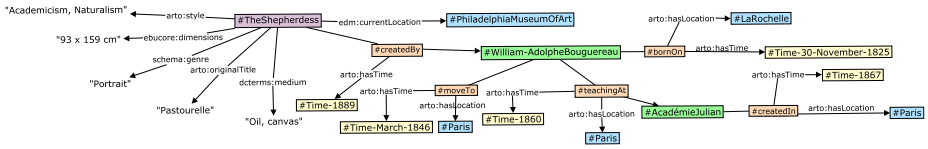

Fig. 5. ARTO Contextual Model of "The Shepherdess".

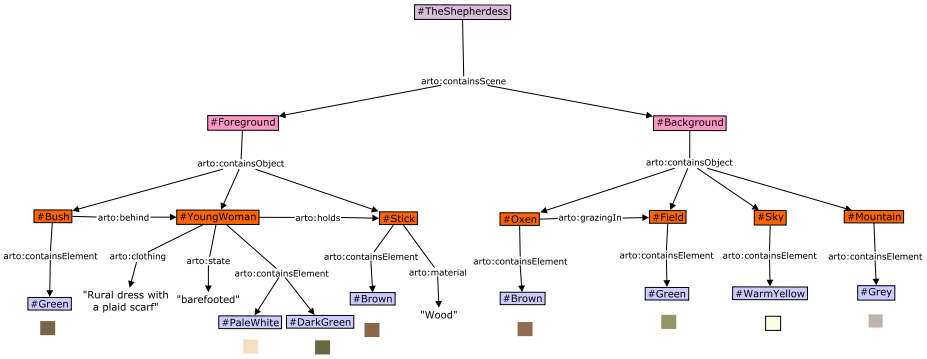

Fig. 6. ARTO Descriptive Model of "The Shepherdess".

4 Evaluation

4.1 OOPS! Assessment

To validate the ontology's quality and completeness, we used OOPS! [25], an online tool that detects and resolves potential errors during ontology creation. OOPS! can detect 40 common design flaws, including class loops, missing annotations, and incorrect data types. By uploading the ontology, we systematically identified and resolved issues such as inappropriate class relationships, inconsistent attribute use, and logical inconsistencies.Through this validation process, we also identified and addressed several structural improvements: we added inverse relationship declarations where appropriate, enhanced the ontology with equivalent class definitions to improve reasoning capabilities, and optimized the documentation with comprehensive class descriptions and usage examples. This thorough validation and improvement process ensured our ontology maintains high quality and usability standards.

4.2 Think-aloud Protocol (TAP)

TAP is a widely used method that can be considered a form of synchronised verbal reporting [5, 26, 32]. This protocol was applied[13] to gather valuable input and guidance from art experts, with a focus on identifying the key features of artwork caption, validating the feasibility of the proposed ontology, and providing feedback for improvement. In the TAP, subjects are asked to verbalise their immediate thoughts, feelings, and cognitive processes while performing a specific task [11]. Through this method, we were able to identify and record the complete thought process of the artwork captioning task.

Protocol Design: The general procedure of the TAP[14] is: firstly, the art experts were expected to select two of their most familiar and two least familiar paintings from the given ten examples. Subsequently, they were required to verbally describe the four works and then label the captions they generated according to the categorisation of captions (descriptive, contextual and interpretive) [2, 3]. Afterwards, they listed and ranked the aspects they considered most important when describing the artwork. The experts were asked to share their approach to describing the artwork, and their insights on describing subjective aspects of the artwork, such as emotion and subject matter. Next, they outlined the elements that they found most challenging to describe and their thoughts on how to verify the accuracy of the description. Finally, after presenting the experts with the initial design of the ontology, the experts needed to apply it to make captions for four artwork examples and evaluated whether the generated content was comprehensive and structured well.

Participants: We recruited 10 art experts who have advanced proficiency in painting or the visual arts, including professors at art colleges or practising artists. Participation was voluntary, lasted approx. one hour, and participants' information was kept confidential throughout the study.

[13] The ANU Human Research Ethics Committee approved this research (2023/1399).
[14] Interview details and related information can be found at `https://w3id.org/arto`.

Results: We analysed interviewees' processes, content, and challenges in generating artwork captions to assess the ontology's alignment with practical needs and comprehensiveness. Feedback from art experts further guided its refinement. Key findings from the interviews and their impact on validating and improving the ontology are discussed below.

In caption creation, 90% emphasised the artist's background and environment, 60% focused on artwork content (composition, colour, characters), and 40% highlighted meaning and interpretation. Familiar artworks prompted more background details, while unfamiliar ones lacked context. These findings validate ARTO's comprehensive coverage of key captioning elements.

The categorisation of captions into descriptive, contextual, and interpretive types received unanimous approval, verified through Ground Theory analysis [24] of participants' caption examples. Participants also positively evaluated the ontology examples, confirming ARTO's comprehensiveness.

Art experts varied in their views on emotional analysis in artworks. Most emphasised the importance of understanding both artwork content and artists' personal experiences for emotional interpretation. However, one expert challenged the necessity of objective emotional analysis, arguing that emotions can only be understood through the artist's experience. Based on these expert perspectives on emotional expression and interpretation, ARTO's design specifically addresses these concerns through its Connotation class and its subclasses (Symbolism, Emotion, and Theme). While some experts emphasized the subjective nature of emotional interpretation in artworks, others highlighted the need for structured representation of emotional content. Our arto:Connotation class bridges these viewpoints by capturing both objective elements (through the Emotion subclass's structured attributes) and subjective interpretations (through the relationships between emotions, symbols, and themes). The model allows for multiple interpretations while maintaining connections to the artwork's visual elements and the artist's documented experiences. This approach aligns with experts' emphasis on understanding both the artwork's content and the artist's personal context for emotional interpretation. The experts highlighted several key challenges in caption creation. Half found professional-level description difficult, while 30% struggled with understanding creator intent and another 30% with verifying contextual information. ARTO addresses these challenges by providing professionally analysed content and verified contextual information, supporting both background understanding and individual interpretation.

Beyond directly evaluating the ontology design, the interview results offered insights for future work. Experts varied in their captioning methods and priorities: 30% addressed content, background, and interpretation; 30% focused on content and background; 20% on content and interpretation; and 20% solely on content. While content was universally prioritised (mentioned by 100% of experts), the inclusion of background and interpretation differed significantly.

To further investigate the importance of different aspects, we conducted a detailed survey where participants ranked 11 aspects from most relevant (1) to least relevant (11). The results showed that "Artist" had the highest average

importance rank, followed by "Content", "Composition", and "Background about the Artist". On the other hand, "Objects", "Emotion", and "Visual Elements" were considered less important on average.

Experts highlighted the importance of including background events in artwork captions, noting how the artist's experiences, social and historical context, and creation motivations influence technique, style, and thematic expression. Regarding interpretation, half of the participants advocated for objectivity, while the other half favoured comprehensive information and publicly recognised interpretations.

5 Conclusion and Future Work

We proposed a comprehensive artwork ontology comprising two components: the Artwork Descriptive Model, which represents artwork content (visual elements, scenes, emotions), and the Artwork Contextual Model, which captures background information (metadata, historical context). This integrated approach enables deeper artwork understanding and representation. Our validation process combined automated error detection using OOPS! with expert interviews involving 10 art specialists. While experts unanimously agreed on ARTO's comprehensiveness and the importance of descriptive and contextual information, they debated the role of subjective interpretation in captions. Some advocated for purely objective facts, while others argued for including personal perspectives. Responding to concerns about limited subjective interpretations, our approach acknowledges that machine learning models can only draw from existing human perspectives. ARTO therefore focuses on providing objective information about artwork elements, historical connections, and established interpretations, while creating space for viewers to develop their own subjective understanding.

Future work is to explore LLMs for generating comprehensive and contextually-rich artwork captions from an ARTO-based Knowledge Graph populated with extensive artwork data from various sources. We also plan to incorporate additional types of artwork, such as sculpture, photography, and digital art.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 6077–6086 (2017), https://api.semanticscholar.org/CorpusID:3753452
2. Anderson, T.: Defining and structuring art criticism for education. Studies in Art Education **34**(4), 199–208 (1993), http://www.jstor.org/stable/1320404
3. Barrett, T.M.: Criticizing art : understanding the contemporary / Terry Barrett. Mayfield Pub. Co Mountain View, Calif (1994)
4. Bekiari, C., Bruseker, G., Doerr, M., Ore, C.E., Stead, S., Velios, A.: Definition of the cidoc conceptual reference model v7.1.1. The CIDOC Conceptual Reference Model Special Interest Group (2021). https://doi.org/10.26225/FDZH-X261
5. Bernardini, S.: Think-aloud protocols in translation research: Achievements, limits, future prospects. Target. International Journal of Translation Studies **13**(2), 241–263 (2001)

6. Borgo, S., Ferrario, R., Gangemi, A., Guarino, N., Masolo, C., Porello, D., Sanfilippo, E.M., Vieu, L.: Dolce: A descriptive ontology for linguistic and cognitive engineering. Applied Ontology **17**, 45–69 (2022). https://doi.org/10.3233/AO-210259, https://doi.org/10.3233/AO-210259, 1
7. Carboni, N., de Luca, L.: An ontological approach to the description of visual and iconographical representations. Heritage **2**(2), 1191–1210 (2019). https://doi.org/10.3390/heritage2020078, https://www.mdpi.com/2571-9408/2/2/78
8. Carriero, V.A., Gangemi, A., Mancinelli, M.L., Marinucci, L., Nuzzolese, A.G., Presutti, V., Veninata, C.: Arco: The italian cultural heritage knowledge graph. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web – ISWC 2019. pp. 36–52. Springer International Publishing, Cham (2019)
9. Cetinic, E.: Towards generating and evaluating iconographic image captions of artworks. Journal of Imaging **7**(8) (2021). https://doi.org/10.3390/jimaging7080123, https://www.mdpi.com/2313-433X/7/8/123
10. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10575–10584 (2019), https://api.semanticscholar.org/CorpusID:219635470
11. Ericsson, K.A., Simon, H.A.: Verbal reports as data. Psychological review **87**(3), 215 (1980)
12. Europeana Foundation: Europeana data model (edm) v5.2.8, https://pro.europeana.eu/page/edm-documentation
13. on FRBR/CRM Dialogue, I.W.G., Bekiari, C., Doerr, M., Le Boeuf, P., Riva, P.: Definition of frbroo: A conceptual model for bibliographic information in object-oriented formalism. Tech. rep., International Federation of Library Associations and Institutions (IFLA) (Mar 2017), https://repository.ifla.org/handle/123456789/659
14. Institute, G.R.: Art architecture thesaurus. Online (2024), available from: http://www.getty.edu/research/tools/vocabularies/aat/index.html [Accessed 13th March 2024]
15. Institute, G.R.: Union list of artist names. Online (2024), available from: http://www.getty.edu/research/tools/vocabularies/ulan/index.html [Accessed 13th March 2024]
16. Ishikawa, S., Sugiura, K.: Affective image captioning for visual artworks using emotion-based cross-attention mechanisms. IEEE Access **11**, 24527–24534 (2023). https://doi.org/10.1109/ACCESS.2023.3255887
17. Korro Bañuelos, J., Rodríguez Miranda, Á., Valle-Melón, J.M., Zornoza-Indart, A., Castellano-Román, M., Angulo-Fornos, R., Pinto-Puerto, F., Acosta Ibáñez, P., Ferreira-Lopes, P.: The role of information management for the sustainable conservation of cultural heritage. Sustainability **13**(8), 4325 (2021)
18. Liu, F., Zhang, M., Zheng, B., Cui, S., Ma, W., Liu, Z.: Feature fusion via multi-target learning for ancient artwork captioning. Information Fusion **97**, 101811 (2023). https://doi.org/https://doi.org/10.1016/j.inffus.2023.101811, https://www.sciencedirect.com/science/article/pii/S1566253523001203
19. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3242–3250 (2016), https://api.semanticscholar.org/CorpusID:18347865

20. Lu, Y., Guo, C., Dai, X., Wang, F.Y.: Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training. Neurocomputing **490**, 163–180 (2022). https://doi.org/https://doi.org/10.1016/j.neucom.2022.01.068, https://www.sciencedirect.com/science/article/pii/S092523122200087X
21. Lu, Y., Guo, C., Dai, X., Wang, F.Y.: Artcap: A dataset for image captioning of fine art paintings. IEEE Transactions on Computational Social Systems **11**(1), 576–587 (2024). https://doi.org/10.1109/TCSS.2022.3223539
22. Mokady, R., Hertz, A.: Clipcap: Clip prefix for image captioning. ArXiv **abs/2111.09734** (2021), https://api.semanticscholar.org/CorpusID:244346239
23. Nicholas J. Car, Timo Homburg, Matthew Perry, John Herring, Frans Knibbe, Simon J.D. Cox, Joseph Abhayaratna, Mathias Bonduel: OGC GeoSPARQL - A Geographic Query Language for RDF Data. OGC Implementation Standard OGC 22-047, Open Geospatial Consortium (2023), http://www.opengis.net/doc/IS/geosparql/1.1
24. Noble, H., Mitchell, G.: What is grounded theory? Evidence-Based Nursing **19**(2), 34–35 (2016). https://doi.org/10.1136/eb-2016-102306, https://ebn.bmj.com/content/19/2/34
25. Poveda-Villalón, M., Gómez-Pérez, A., Suárez-Figueroa, M.C.: OOPS! (OntOlogy Pitfall Scanner!): An On-line Tool for Ontology Evaluation. International Journal on Semantic Web and Information Systems (IJSWIS) **10**(2), 7–34 (2014)
26. Prokop, M., Pilař, L., Tichá, I.: Impact of think-aloud on eye-tracking: A comparison of concurrent and retrospective think-aloud for research on decision-making in the game environment. Sensors **20**(10), 2750 (2020)
27. Raimond, Y.: The event ontology, https://motools.sourceforge.net/event/event.html
28. Sartini, B., Baroncini, S., van Erp, M., Tomasi, F., Gangemi, A.: Icon: An ontology for comprehensive artistic interpretations. J. Comput. Cult. Herit. **16**(3) (aug 2023). https://doi.org/10.1145/3594724, https://doi.org/10.1145/3594724
29. Sheng, S., Moens, M.F.: Generating captions for images of ancient artworks. In: Proceedings of the 27th ACM International Conference on Multimedia. p. 2478–2486. MM '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3343031.3350972, https://doi.org/10.1145/3343031.3350972
30. Suárez-Figueroa, M.C., Gómez-Pérez, A., Fernández-López, M.: The NeOn Methodology for Ontology Engineering, pp. 9–34. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-24794-1_2, https://doi.org/10.1007/978-3-642-24794-1_2
31. Szekely, P., Knoblock, C.A., Yang, F., Zhu, X., Fink, E.E., Allen, R., Goodlander, G.: Connecting the smithsonian american art museum to the linked data cloud. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) The Semantic Web: Semantics and Big Data. pp. 593–607. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
32. Veenman, M.V., Elshout, J.J., Groen, M.G.: Thinking aloud: Does it affect regulatory processes in learning? Tijdschrift voor Onderwijsresearch (1993)
33. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 3156–3164 (2014), https://api.semanticscholar.org/CorpusID:1169492

34. VRA Core: Vra core schemas, https://www.loc.gov/standards/vracore/schemas.html
35. Whitaker, A., Bracegirdle, A., De Menil, S., Gitlitz, M.A., Saltos, L.: Art, antiquities, and blockchain: new approaches to the restitution of cultural heritage. International Journal of Cultural Policy **27**(3), 312–329 (2021)
36. World Wide Web Consortium, https://dom.spec.whatwg.org/: Document Object Model (DOM) — Living Standard (2024), accessed: 2024-03-15
37. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 2048–2057. PMLR, Lille, France (07–09 Jul 2015), https://proceedings.mlr.press/v37/xuc15.html
38. Yang, C., Pereira Nunes, B., Rodríguez Méndez, S., Chen, Y., Manrique, R., Casanova, M.A.: Towards Comprehensive Artwork Representation: Motivations and Challenges in Capturing Multi-Dimensional Art Descriptions. In: The 2024 ACM/IEEE Joint Conference on Digital Libraries. p. 5. JCDL '24, ACM, New York, NY, USA (December 2024). https://doi.org/10.1145/3677389.3702517
39. Zheng, B., Liu, F., Zhang, M., Zhou, T., Cui, S., Ye, Y., Guo, Y.: Image captioning for cultural artworks: a case study on ceramics. Multimedia Systems **29**(6), 3223–3243 (Dec 2023). https://doi.org/10.1007/s00530-023-01178-8, https://doi.org/10.1007/s00530-023-01178-8