V4Ann: Representation and Interlinking of Atom-based Annotations of Digital Content

Georgios Meditskos, Stefanos Vrochidis, and Ioannis Kompatsiaris

Information Technologies Insititute
Centre for Research and Technology Hellas, Thessaloniki, Greece
{gmeditsk, stefanos, ikom}@iti.gr

Abstract. There is a great potential in creative industries, such as architecture and video game design, for re-using and re-purposing of digital content. Paintings, archival footage, documentaries, movies, reviews or catalogues, and various other forms of artwork can serve as sources of inspiration and design direction towards innovative designs and new concepts. In this paper, we present V4Ann, an ontology-based framework for semantically representing, aggregating and combining annotations (atoms) coming from visual and textual analysis of digital content. The aim is to structure and link data in such a way so as to facilitate the systematic process, integration and organisation of information and establish innovative value chains and end-user applications. The framework is part of the V4Design platform that aims to re-use and re-purpose existing heterogeneous multimedia content by semantically enriching and transforming assets into a 3D representation, so as to inspire and support the design, architecture, as well as 3D and VR game industries.

Keywords: Annotations · Ontologies · Reasoning · Semantic enrichment · Multimodal data

1 Introduction

Vast amounts of multimedia content is being produced, archived and digitised, resulting in great troves of data of interest. Examples include user-generated content, such as images, videos, text and audio posted by users on social media and wikis, or content provided through official publishers and distributors, such as digital libraries, organisations and online museums. This digital content can serve as a valuable source of inspiration to the cultural and creative industries to produce new assets or to enhance and (re-)use the already existing ones.

However, the re-use and re-purposing of digital content is mainly realised based on individual designers skills and a variety of non-interlinked heterogeneous tools. To this end, the content remains largely under-exploited, despite its great potential for re-use and re-purpose, due to the lack of appropriate solutions for its retrieval and integration into the design process. For example, existing heterogeneous multimedia content, such as video and images of buildings and

objects, can be collected and transformed (e.g. into 3D models[1]), so as to inspire and support the creation of new content in creative industries. One of the main challenges in this area is to maximise the potential for re-purposing of digital content through the development of innovative technologies to systematically analyse, combine, link and foster searchability and reusability of heterogeneous multimedia content in different contexts.

In this paper we describe V4Ann, an ontology-based framework for capturing and interlinking digital assets and duly annotations at two levels: a) *content analysis level*, during which visual and textual content is analysed to extract labels, called *atoms*; and b) *retrieval and repurposing level*, where the assets (e.g. 3D models and images) are interlinked and contextually enriched to facilitate their discovery. At the content analysis level, V4Ann provides the conceptual structures to capture and interlink multimedia analysis results on digital content, such as video, image and text. During retrieval and repurpose, V4Ann provides practical retrieval capabilities, allowing users, e.g. game designers, to search for assets relevant to their needs. V4Ann is part of the V4Design platform[2], enriching multimedia processing with a semantic annotation layer.

The contribution of our research can be summarised in the following:

- We describe a resource annotation model that implements the W3C standard for defining annotations (Web Annotation Data Model [17]).
- We define a core set of rules that perform valid inferences for annotation propagation and interlinking, as well as for validity checking.
- We propose an atom similarity metric along with a searching algorithm for keyword-based digital asset retrieval.

The rest of the paper is structured as follows: Section 2 presents related work. Section 3 gives an overview of the framework and presents our motivation. In Section 4 we describe the basic concepts of the V4Ann annotation model, while in Section 5 we elaborate on the inference and validation capabilities. Section 6 describes the atom similarity metric and the searching functionality. In Section 7 we present evaluation results and, finally, in Section 8 we conclude our work.

2 Related Work

Annotations are typically used to convey information about a resource or associations between resources. Simple examples include a comment or tag on a single web page or image, video or a blog post about a news article. In 2017, the Web Annotation Data Model (WADM) [17] became the W3C recommendation for defining annotations. It provides an extensible, interoperable framework for expressing annotations, such that they can easily be shared between platforms[3].

In the domain of digital libraries, the Europeana Data Model (EDM) [4] adopts an open and scalable approach that can accommodate the range and level

[1] `https://pro.europeana.eu/project/3d-content-in-europeana`
[2] `https://v4design.eu/`
[3] `https://www.w3.org/TR/annotation-vocab/`

of details of particular standards, such as LIDO for museums, EAD for archives or METS for digital libraries. EDM is not built on any particular standard, however it is conceptually in line with WADM and the ORE[4] initiative.

The Open Provenance Model (OPM) [11] enables to specify what caused "things" to be, i.e., how "things" depended on others and resulted in specific states. In essence, it allows provenance information to be exchanged between systems, by means of a compatibility layer based on a shared provenance model. OPM predates PROV-O [9], and has a very similar approach to modelling provenance by relating agents, artifacts and processes and the concepts of OPM are covered by equivalent PROV-O concepts. PAV [3] extends PROV-O and specifies Provenance, Authoring and Versioning information.

The Dublin Core metadata (DCMI) standard[5] is a simple yet effective general-purpose set of 15 elements for describing a wide range of networked resources. Although DCMI favors document-like objects, it can be applied to other resources as well. The SKOS Core Vocabulary [10] is a model for expressing the basic structure and content of concept schemes. Specifically for multimedia, the Ontology for Media Resources[6] was developed by the W3C Media Annotations Working Group to identify a minimum set of core properties to describe and retrieve information about media resources. VidOnt [18] provides a formally grounded core reference ontology for video representation. Several attempts have been made to map the XML Schema of MPEG-7 to RDFS and OWL [19] and X3D to OWL (OntologyX3D [6]) and the 3D Modeling Ontology (3DMO[7]).

V4DAnn aims to serve as the semantic annotation layer of multimedia processing results for fostering data exchange among analysis services and for human consumption. In order to promote interoperability and extensibility, it implements the WADM pattern, introducing the concept of atoms and providing several annotation entities and properties. In contrast to existing models that mostly focus on metadata defined by data providers and curators, V4Ann aims to capture content analysis results (e.g. visual and textual analysis), serving as a semantic middleware for metadata exchange. For example, EDM views refer to digital representations, whereas in V4Ann a view represents an atom-based interpretation of a content analysis procedure, e.g. aesthetics extraction. However, V4Ann provides alignments to conceptual structures of existing models, such as the EDM, ORE and SKOS (see Section 4 for more details).

As far as semantic enrichment and retrieval are concerned, recent advances in machine learning and especially deep learning have provided us with tools like word representations (e.g. word2vec [20] and Glove [14]), which led to the development of more recent and powerful analysis models [15]. In addition, several approaches have been proposed for question answering over Semantic Web knowledge bases and Linked Data. Most of them generate one or more queries, while others opt for graph-based approaches to mitigate the rigidness often en-

[4] `http://www.openarchives.org/ore/1.0/vocabulary`
[5] `http://dublincore.org/documents/dces/`
[6] `http://www.w3.org/TR/mediaont-10`
[7] `http://3dontology.org`

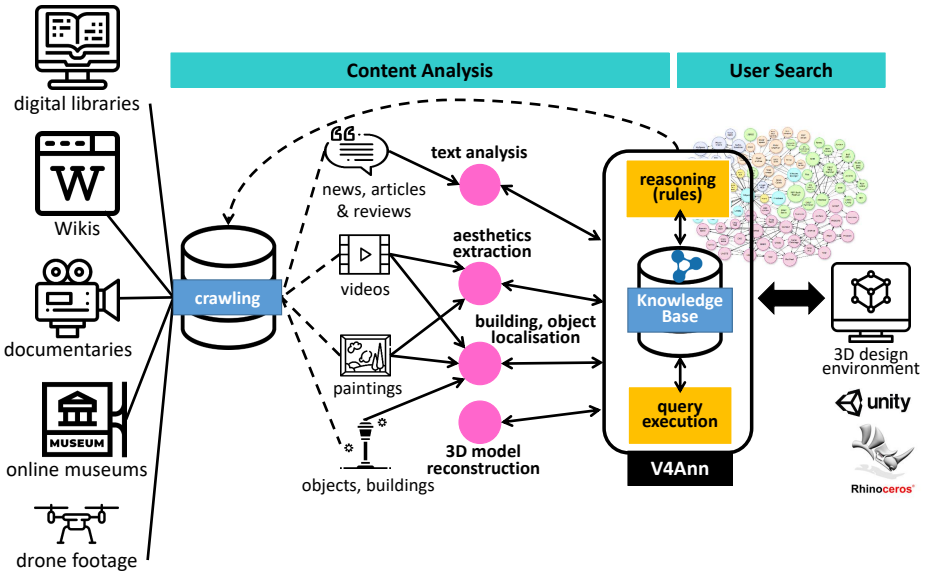

Fig. 1. The position of V4Ann in the integrated V4Design platform.

tailed in formulating appropriate SPARQL queries. Examples include EARL [5] and VoxEL [16]. V4Ann aims to provide a practical context enrichment framework to facilitate basic asset discovery, rather than proposing a fully fledged question answering framework. To this end, it introduces the notions of atom similarity and local contexts.

3 Key Concepts and Motivation

In a world where visual and textual data are in abundance, creative industries need to re-use and re-purpose them so as to remain competitive to other industries and provide to society and creativity a novel financial prism. V4Design is an H2020 project that aims at exploiting state-of-the-art digital content analysis techniques to generate 3D models, extract aesthetic and stylistic information from paintings and videos, localise buildings and objects of interest within visual content, and integrate it with textual information so as to inspire and support the design, architecture, as well as 3D and VR game industries.

V4Ann aims to enrich V4Design with a semantic annotation layer. From one hand, V4Ann acts as the semantic middleware, capturing, interlinking and serving analysis results to multimedia analysis services. On the other hand, it provides the semantic atom-based query infrastructure to retrieve generated assets. The conceptual architecture of V4Design, along with the position of V4Ann, is depicted in Fig. 1. All in all, V4Ann aims to address the following challenges:

- *Annotation propagation and linking:* In a multimodal content analysis setting, like in V4Design, a single media type can be analysed by multiple technologies. For example, an image can be used for extracting building masks, as well as for aesthetics (style) extraction. Also, in many cases, there are interdependencies among the components, e.g. 3D model reconstruction needs as input video frame masks extracted by building localisation. It is important to have an efficient and interoperable way to represent, exchange and further link metadata, both structurally and semantically.
- *Context-aware retrieval*: V4Design aims to create new multimedia content that can be integrated in existing architecture and video game design platforms, such as Unity[8] and Rhino[9]. Therefore, there is a need for practical and efficient retrieval mechanisms on top of the multimodal annotations. For example, to allow users to search for assets with certain styles or with advanced contextual filters, such as "castles near lakes".

In order to address the aforementioned challenges, V4Ann capitalises on and combines existing Semantic Web standards for resource annotation and interlinking, inference and validation. More precisely, the WADM model is used as the core resource annotation pattern, combined with existing structured ontologies and schemata (Section 4). SPIN rules [7] and SHACL shapes [8] are used to derive additional relations among the annotated resources and for validating the generated knowledge graphs (Section 5). Finally, keyword-based context-aware retrieval is facilitated to retrieve assets (Section 6).

4 V4Ann Annotation Model

Fig. 2 illustrates the upper-level concepts of the V4Ann annotation model. The conceptual model revolves around the notions of *annotations*, *media types*, *views* and *atoms*. Annotations serve as resource containers, implementing the annotation pattern of WADM. Each annotation associates a media type (image, video, text, 3D model) with a view, which encapsulates a set of atoms. Each view defines one or more atoms, e.g. entities, tags, styles, etc. that are derived from multimedia content analysis. These atoms describe: a) Aesthetics, i.e. architectural styles and creators that are extracted from images and videos; b) Object and building types that are recognised in images and videos; c) Named entities and concepts that are extracted from textual descriptions, e.g. image captions; d) images and video frames used to reconstruct a 3D model. All atoms derived by aesthetics, localisation and text analysis are disambiguated, i.e. they are already mapped to WordNet, BabelNet or DBpedia resources by the content analysis services. Fig. 2 also presents SKOS mappings to the ORE specification, as well as subclass and subproperty relations to WADM and EDM. In the following we describe in details each key concept.

[8] `https://unity3d.com/`
[9] `https://www.rhino3d.com/`

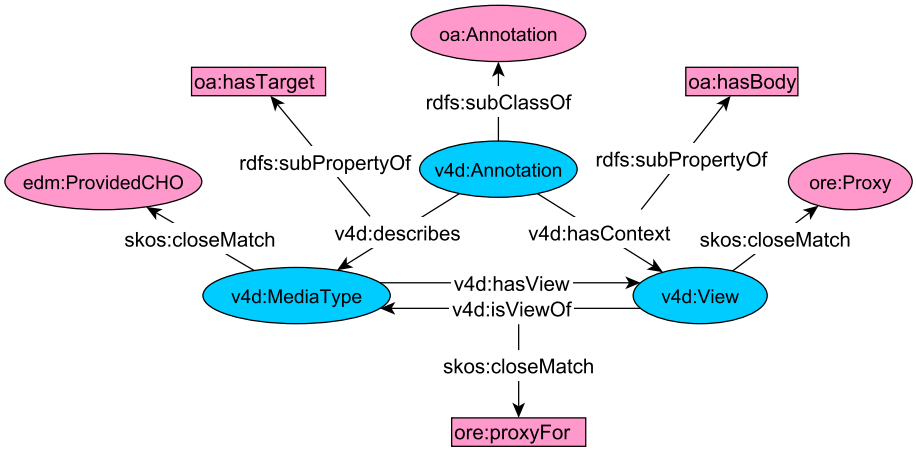

Fig. 2. The core concepts of the V4Ann annotation model defined as specialisation of WADM (`oa` namespace). Mappings to other models are also depicted, such as to Europeana Data Model (EDM) and Object Reuse and Exchange (ORE) initiative.

4.1 Annotation resources

Four domain-specific annotation classes are defined for attaching atom views to media types[10]: `LocalisationAnnotation`, `TextualAnnotation`, `Aesthetics-Annotation` and `3DModelAnnotation`. According to the WADM specification, an annotation has 0 or more *bodies* (`oa:hasBody`), which encapsulate descriptive information, and a 1 or more *targets* (`oa:hasTarget`) that the bodies describe. V4Ann defines two subproperties to restrict the values of these properties, associating the targets (i.e. the media types) with view atoms. Intuitively, a V4Ann annotation has a *context* that *describes* a *media type* using *views*. In terms of OWL 2 semantics, the `hasContext` ($\sqsubseteq$ `oa:hasBody`) property takes as values only instances of the `View` class and the `describes` ($\sqsubseteq$ `oa:hasTarget`) property at least one `MediaType` value. The `Annotation` class is defined as:[11]

$$\texttt{Annotation} \sqsubseteq \texttt{oa:Annotation} \sqcap \exists \texttt{describes.MediaType} \sqcap \forall \texttt{hasContext.View} \quad (1)$$

4.2 Media types

In order to define the targets of annotations (`describes` property assertions), V4Ann provides the `MediaType` upper-level class. There are four media types for annotations: `Video`, `Text`, `Image`, `Mask` $\sqsubseteq$ `Image`, `Texture` $\sqsubseteq$ `Image` and `3DModel`. Each media type can be associated with additional descriptive information, such

[10] In the rest of the paper, we omit the `v4d` namespace.
[11] We use Description Logics [2] to represent the semantics.

as the source of the asset (e.g. the URL), license information, date of retrieval, etc. Intuitively, each media type resource represents a single multimedia asset for which a set of annotation atoms needs to be captured.

4.3 Views and atoms

Views are container classes that encapsulate the annotation metadata (atoms) and they are used in `hasContext` property assertions. Each media type has a different view. For example, the atoms of spatio-temporal building (`BuildingView` $\sqsubseteq$ `View`) and object localisation (`ObjectView` $\sqsubseteq$ `View`) in images and videos specify their type, i.e. whether the image or video contains a building, object or a painting. The semantics of OWL 2 allows us to define useful complex class descriptions to specify further dependencies, as described below. It should be noted that content analysis is not part of the V4Ann framework. As described in Section 3, V4Ann aims to semantically capture the results of content analysis, which is part of the overall V4Design platform [1].

Aesthetics Aesthetics extraction refers to the categorisation of the aesthetics of paintings and images that contain architecture objects and buildings based on their style (e.g. impressionism, cubism and expressionism), the creator (mainly for paintings) and emotion that they evoke to the viewer. Two properties are defined for creators (`v4d:creator` $\equiv$ `schema:creator`) and styles (`v4d:style`), whose domain is the `v4d:AestheticsView` class.

$$\begin{aligned}\texttt{AestheticsAnnotation} \sqsubseteq\ & \texttt{oa:Annotation} \sqcap \\ & \exists \texttt{describes}.\{\texttt{Image} \sqcup \texttt{Video}\} \sqcap \forall \texttt{hasContext}.\texttt{AestheticsView}\end{aligned} \tag{2}$$

$$\texttt{AestheticsView} \sqsubseteq \forall \texttt{creator}.\texttt{Creator} \sqcap \forall \texttt{style}.\texttt{Style} \tag{3}$$

The `Creator` and `Style` classes serve as container classes, allowing the capturing of data-specific properties, such as the classification confidence, as well as contain links to DBpedia and BabelNet. Fig. 3 presents an aesthetics annotation example (left part). The image depicts the Tholos of Delphi that has been given the atom (style) "Greek Architecture".

Object and Building Localisation Building and interior objects localisation on art and architecture-related movies, documentaries and multiple art-images, aims to extract content that can be re-purposed and re-used in a meaningful and innovative way. Examples include buses, trains, as well as statues, buildings, etc.

The extracted atoms (labels) are mapped to the V4Ann annotation model in terms of generated *masks* and *tags*. In videos, the results are also associated with frame(s) to capture the temporal aspects of localisation.

$$\begin{aligned}\texttt{LocalisationAnnotation} \sqsubseteq\ & \texttt{oa:Annotation} \sqcap \\ & \exists \texttt{describes}.\{\texttt{Image} \sqcup \texttt{Video}\} \sqcap \forall \texttt{hasContext}.\texttt{LocalisationView}\end{aligned} \tag{4}$$

$$\texttt{LocalisationView} \sqsubseteq \exists \texttt{hasTag}.\texttt{Tag} \sqcap \forall \texttt{hasFrame}.\texttt{integer} \tag{5}$$

Text Analysis Text analysis provides the atoms that are derived from textual content. For example, in addition to annotating images with building and objects, the assets are further enriched with named entities and concepts extracted from captions, titles and descriptions. V4Ann captures these atoms and associate them with the media type (video or image) that the textual content is relevant to through instantiations of the `TextAnalysisView` class. Example atoms include `name`, `title`, `date`, `creator`, `designer`, `artist`, `location`, etc., defined as subproperties of `Tag`.

$$\begin{aligned}\texttt{TextAnnotation} \sqsubseteq \texttt{oa:Annotation} \sqcap \\ \exists \texttt{describes}.\{\texttt{Image} \sqcup \texttt{Video}\} \sqcap \forall \texttt{hasContext}.\texttt{TextView}\end{aligned} \tag{6}$$

$$\texttt{TextView} \sqsubseteq \exists \texttt{hasTag}.\texttt{Tag} \tag{7}$$

3D Reconstruction 3D reconstruction converts input video and images into 3D point clouds and meshes. Apart from the actual 3D object, this process generates a variety of metadata, such as the number of point clouds, number of faces, textures, etc. The most important atom is the source of reconstruction, i.e. the video or the images the 3D model has been extracted from.

$$\begin{aligned}\texttt{3DModelAnnotation} \sqsubseteq \texttt{oa:Annotation} \sqcap \\ \exists \texttt{describes}.\texttt{3DModel} \sqcap \forall \texttt{hasContext}.\texttt{3DModelView}\end{aligned} \tag{8}$$

$$\texttt{3DModelView} \sqsubseteq \exists \texttt{hasSource}.\{\texttt{Images} \sqcup \texttt{Video}\} \tag{9}$$

A 3D annotation example is depicted in Fig. 3 (right part). The annotation of the 3D model of Tholos is associated ($\texttt{image} \sqsubseteq \texttt{hasSource}$) with the images that have been used for the reconstruction. It is assumed that the example image for aesthetics extraction is part of the set, demonstrating the way multimodal analysis results are interlinked. As we describe in the next section, these links are used to materialise additional relationships in the form of inference rules.

5 Inference and Validation

5.1 Implicit Relationships

Additional inferences are derived by combining native OWL 2 RL reasoning and custom rules. The former is based on the OWL 2 RL profile semantics (OWL 2 RL/RDF rules [12]), which is implemented by state-of-the-art triple stores, such as GraphDB. However, the semantics OWL 2 is limited. For example, only instances connected in a tree-like manner can be modelled [13]. V4Ann implements domain rules on top of the graphs to express richer relations. SPARQL-based `CONSTRUCT` graph patterns are used that identify the valid inferences that can be made on the annotation graphs. It is beyond the scope of the paper to include an extensive coverage of relevant reasoning capabilities. In the following we present the concept of *atom propagation* that illustrates the principle idea.

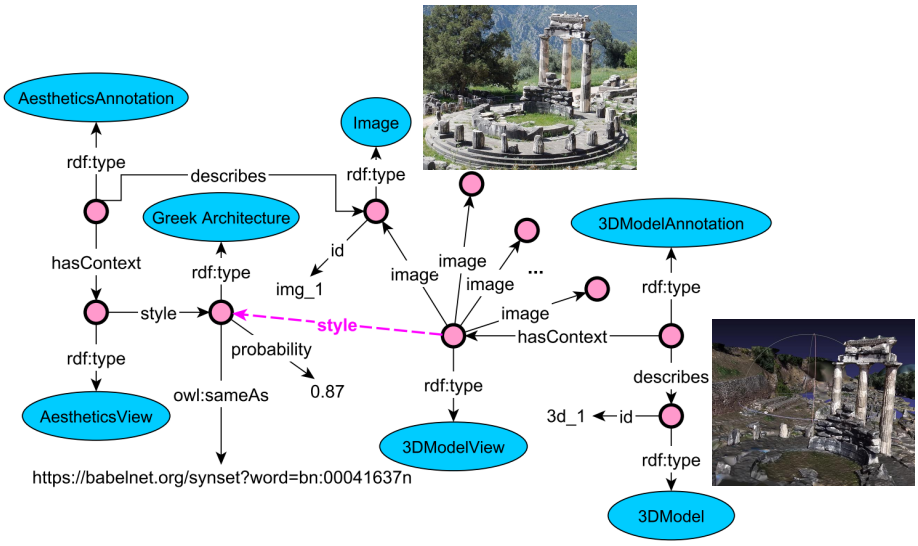

Fig. 3. Example of atom propagation. The dashed arrow illustrates the enrichment of the 3D annotation resource with the aesthetics `style` derived from visual analysis.

Since V4Ann follows a standard-based annotation pattern, additional relations can be further derived. For example, the aesthetics atoms extracted from video frames can be used to annotate the 3D models that have been reconstructed using those frames. The principle idea is that atoms can be *propagated* among one or more views, provided that their annotations are associated.

Fig. 3 illustrates atom propagation between an aesthetics and 3D model annotations. The two annotations are connected at the view level, since the aesthetics annotation describes an image (`img_1`) that has been used to generate the 3D model of Tholos (id `3d_1`). In this case, the view that describes the 3D model inherits the atom (style) of the image (`Greek Architecture`). The corresponding SPARQL graph pattern is given bellow.

```
CONSTRUCT {
   ?view :style ?atom .
} WHERE {
   ?a1 a :AestheticsAnnotation;
      :describes ?img; :hasContext [:style ?atom] .
   ?a2 a :3DModelAnnotation; :hasContext ?view .
   ?view :image ?img .
}
```

5.2 Validation and Consistency Checking

The validation process checks the consistency, structural and syntactic quality of the metadata. We use both native ontology consistency checking (e.g. OWL

2 DL reasoning) and custom SHACL validation rules, following the closed-world paradigm. The former handles validation taking into account the semantics at the terminological level (TBox), e.g. checking class disjointness. The latter detects constraint violations, e.g. missing values and cardinality violations. An example SHACL shape is given below that represents a constraint that all 3D model views should include references to the atoms (images) used to the 3D reconstruction.

```
v4d:3DModelView
   rdf:type sh:NodeShape ;
   sh:property [
      rdf:type sh:PropertyShape ; sh:path v4d:image ;
      sh:class v4d:MediaType ; sh:minCount 1 ;
      sh:name "one or more images" ; sh:nodeKind sh:IRI ;
   ] .
```

6 Context-based Asset Retrieval

In Section 4 we described the process of creating the V4Ann annotation graphs, which involves the representation and further interlinking (e.g. through annotation propagation) of media type atoms. In this section we describe the approach of V4Ann towards enabling keyword-based context-aware retrieval of assets, capitalising on the concept of *local context*.

Definition 1. *The local content l_t of an atom t is defined as the tuple $\langle t, r_t, he_t, ho_t \rangle$, where r_t is the set of conceptually relevant terms, he_t is the set of hypernyms and ho_t is the set of hyponyms of t.*

Intuitively, a local context of an atom constitutes an enriched, pre-constructed semantic signature of this atom, taking into account conceptual and lexical relations from existing semantic networks and datasets, such as WordNet, BabelNet and ConceptNet (Fig. 4). In the case of hypernyms and hyponyms, we use the threshold h to specify the maximum level of relevant atoms. All in all, the retrieval mechanism of V4Ann aims to match incoming local contexts of query atoms (keywords) against local contexts of annotation atoms.

6.1 The $\mathcal{AH}$ Metric

The $\mathcal{AH}$ metric represents the similarity of two atoms taking into account their local context. It depends on a term similarity function S, and on a set F of local context filters. In the following, we assume that $S(A, B)$ denotes the similarity of two atoms A and B, with respect to the function S, and that $S(A, B) \in [0..1]$, with 1 denoting absolute match. We use the notation $A \overset{f}{\sim} B$ to denote that A matches to B, with respect to one of the following filters f:

1. *exact (e).* The two atoms should have either the same URI, or they should be equivalent concepts, that is, $A \overset{e}{\sim} B \Leftrightarrow A = B \vee A \equiv B$.

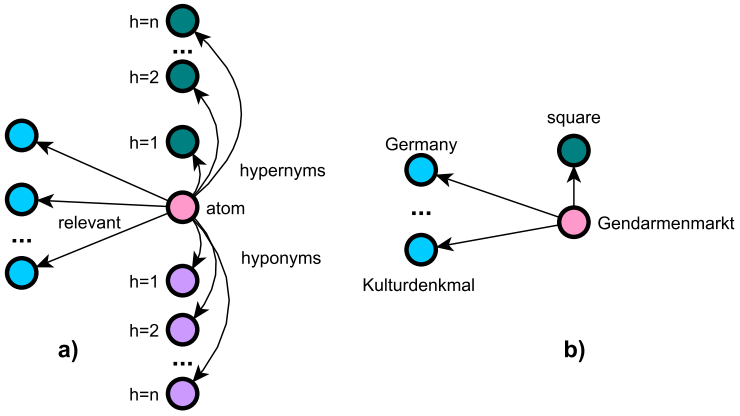

Fig. 4. a) Generic local context of atom: relevant atoms are extracted from ConceptNet and BabelNet properties, hypernyms stem from WordNet and IS-A BebelNet relationships, hyponyms stem from WordNet; b) example local context for "Gendarmenmarkt".

2. *plugin (p)*. The atom B should belong to the set of hypernyms of A (he_A) or to the set of relevant concepts of A (r_A), that is, $A \overset{p}{\sim} B \Leftrightarrow B \in he_A \vee B \in r_A$.
3. *subsume (su)*. The atom B should belong to the set of the hyponyms of A, that is, $A \overset{su}{\sim} B \Leftrightarrow B \in ho_A$.

We generalize the $A \overset{f}{\sim} B$ relation to a set of filters F and we define that the atom A matches the atom B, with respect to a filter set F, if and only if there is at least one filter f in F, such that $A \overset{f}{\sim} B$, that is:

$$A \overset{F}{\sim} B \Leftrightarrow \exists f \in F : A \overset{f}{\sim} B.$$

Definition 2. *The $\mathcal{AH}$ similarity of two atoms X and Y is the normalized value to* [0..1] *that is defined, with respect to a function S and a filter set F, as*

$$\mathcal{AH}(X, Y, F) = \begin{cases} S(X, Y) & \text{if } X \overset{F}{\sim} Y \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

We generalize (10) on two sets S_A, S_B of atoms as

$$\mathcal{AH}_{set}(S_A, S_B, F) = \frac{\sum_{\forall B \in S_B} \max_{\forall A \in S_A} \left[\mathcal{AH}(B, A, F)\right]}{|S_B|} \tag{11}$$

Intuitively, for each atom $B \in S_B$ there should be at least one atom $A \in S_A$ relevant to B, with respect to the filter set F. Otherwise, $\mathcal{AH}_{set}$ returns 0 (absolute mismatch). The overall $\mathcal{AH}_{set}$ similarity is computed as the mean value of the sum of the maximum $\mathcal{AH}$s for each atom B, since each B may have more than one relevant atoms in S_A. In V4Ann, S_A represents the atoms that are associated with an asset, while S_B is the set of keywords.

6.2 Atom Similarity S

As a similarity function $S(A, B)$, V4Ann uses a heuristic function that takes into account the information captured in local contexts of A and B, i.e. in the sets r, he and ho (see Definition 1). The implementation of S is summarised in the following priority rules r_i, where $r_1 > r_2 > r_3 > r_4$.

r_1**:** if $A = B \vee A \equiv B$, then $S(A, B) = 1$.
r_2**:** if $B \in hy_A \vee B \in r_A$, then $S(A, B) = a$.
r_3**:** if $B \in ho_A$, then $S(A, B) = b$.
r_4**:** $S(A, B) = 0$.

Currently, a and b ($a > b$) are defined manually based on domain knowledge regarding the quality of multimedia analysis that produces the atoms (e.g. aesthetics extraction). The empirical definition of these values (currently $a = 0.7$ and $b = 0.3$) aims to promote plugin matches (r_2) over subsumed (r_3).

7 Evaluation and Discussion

7.1 Digital Content

Deutsche Welle (DW) and Europeana are two key content providers in V4Design. DW provides selected parts of their documentary and movie archives so as to localise building structures and objects. Europeana provides their large archive of paintings, pictures of contemporary artwork and related critics, for stylistic and aesthetics extraction and textual analysis. The generated V4Ann annotation graphs contain the atoms that have been extracted from the analysis components, along with interconnections among the annotation resources. Table 1 provides some statistics for the annotation graphs.

7.2 Evaluation

User-centred A user-centred evaluation has been performed with a twofold purpose. First, to collect qualitative feedback on the results, as well as on non-functional aspects, such as query response time. Second, and most important, to generate an annotation dataset and assess the performance of V4Ann.

Participants were invited to evaluate the current implementation by performing keyword-based queries. A list of relevant resources has been provided, such as square names, monuments, building types, etc., in order to help them conduct relevant queries. Users filled in a five-point scale questionnaire (1-completely agree, 5-completely disagree). Sample questions are depicted in Table 2. The feedback can be summarised as it follows:

Table 1. The number of annotations and atoms in the V4Ann annotation graphs, along with the average size of local context for each atom (r + hy + ho).

#annotations	#atoms	avg. local context size
17245	154610	17 per atom

Table 2. Example questions answered by users.

#	**Question**	**Mean (SD)**
Q1	Atoms that are derived from visual analysis are most of the time correct	1.7 ± 0.83
Q2	Atoms that are derived from text analysis are most of the time correct	2.34 ± 0.79
Q3	Many times irrelevant results are top-ranked	3.97 ± 1.29
Q4	There are many irrelevant results	2.43 ± 1.41
Q5	It takes too long for the system to provide a response	4.04 ± 0.99
Q6	There are too many "No results" responses	4.08 ± 0.45

- **Quality of atoms**: The quality and relevance of local contexts depends on the performance of content analysis, e.g. visual and textual analysis. Table 2 shows that visual analysis provides, in principle, better results than text analysis (Q1, Q2).
- **Retrieval results**: According to Q3, the system achieves good top-ranked accuracy, however the complete set of the results contain quite a lot irrelevant entries (Q4). As we explain in the next section, this is mainly relevant to the context provided in the query (i.e. number of keywords). Due to the local context, the system was able to provide a response in most cases (Q6), even partially correct (Q4).
- **Response time:** The response time of the system was positively assessed (Q5). The average response time was 4.1 seconds, which includes query analysis, building of local context and searching algorithm execution.

System evaluation We manually annotated the relevance sets of the performed queries, so as to quantitatively assess performance. Table 3 depicts the average precision and recall achieved for $h = 1$ and $h = 3$ and using different searching filters (Section 6.1). As expected, the stricter the filter is, the more accurate results we obtain (high precision) with low, however, recall. On the other hand, the more relax is the filter, the higher recall is achieved with a negative impact on the precision. This is due to the fact that with a strict filter (i.e. exact), the probability of finding the correct annotation is higher compared to a relaxed filter (i.e. subsume), since in the second case, impartial matches are also allowed.

It should be noted that the overall performance of V4Ann strongly depends on the quality of the atoms, which in turn depends on the quality of the results provided to V4Ann. For example, if the wrong style for a painting is provided by aesthetics, this will affect precision, since V4Ann does not aim at improving the classification of incoming atoms. However, we plan to integrate multimodal data aggregation and fusion techniques to derive the most plausible classification of atoms and help improve the contextual information captured in local contexts.

Another interesting finding involves the threshold h. We observed that for $h = 1$ the framework provides better results than using $h = 3$, i.e. by enriching

Table 3. Average precision and recall (top-20 results).

	$h = 1$		$h = 3$	
	Recall	**Precision**	**Recall**	**Precision**
exact	0.59	**0.77**	0.44	0.51
plugin	0.67	0.69	0.52	0.48
subsume	**0.73**	0.61	0.59	0.42

the local context with additional atoms, up to the third level. Intuitively, h allows to control the amount of contextual information taken into account during the definition of local contexts. A higher h value leads to more generic local contexts that affect precision. For example, the third-level WordNet hypernym of "tower" is "unit", which is too generic, obfuscating the semantics of the atom. The optimal value of h depends on the concreteness of the atoms extracted from content analysis: the more specific is the label/atom, the more room for additional context exists. In our experiments, the labels we get tend to be generic, therefore the best performance is achieved with $h = 1$.

8 Conclusion

In this paper we presented V4Ann, an ontology-based framework for representing, linking and enriching results of multimedia analysis on digital content. V4Ann generates annotation graphs of image, video, textual analysis and 3D model reconstruction, so as to facilitate the systematic process, integration and organisation of information and establish practical repurposing mechanisms.

The annotation model of V4Ann reuses existing standards and schemata, building the atom-based annotations graphs on top of standard ontologies, controlled vocabularies and patterns. The vocabularies are defined in OWL 2 and atoms are associated with assets using the WADM pattern. As such, it promotes interoperability, as well as fosters the use of declarative languages to identify further inferences and ensure the semantic consistency of the knowledge graphs. We also elaborated on the concept of local contexts, as well as on the $\mathcal{AH}$ metric for asset retrieval. We evaluated the framework using actual multimedia content and atoms provided by the V4Design modules and discussed the findings.

V4Ann is accessible through Rhinoceros 3D (Rhino)[12] and Unity plugins developed in the V4Design project through which users (architects and video games designers) can search for assets and import them in the scene. For future work we plan to implement context-aware algorithms to improve the classification accuracy of incoming atoms, as well as to extend the context-aware retrieval algorithm with more sophisticated similarity metrics and functions.

[12] `https://gitlab.com/v4designEU/v4d4rhino`

Acknowledgments This work was supported by the EC funded project V4Design (H2020-779962).

References

1. Avgerinakis, K., Meditskos, G., Derdaele, J., Mille, S., et al.: V4design for enhancing architecture and video game creation. In: IEEE International Symposium on Mixed and Augmented Reality. pp. 305–309 (2018)
2. Baader, F., Calvanese, D., McGuinness, D.L., et al.: The Description Logic Handbook: Theory, Implementation & Applications. Cambridge University Press (2003)
3. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J., Goble, C., Clark, T.: PAV ontology: Provenance, authoring and versioning. Journal of Biomedical Semantics **4**(1), 37 (2013)
4. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., Van de Sompel, H.: The europeana data model (edm). In: World Library and Information Congress: 76th IFLA general conference and assembly. pp. 10–15 (2010)
5. Dubey, M., Banerjee, D., Chaudhuri, D., Lehmann, J.: Earl: Joint entity and relation linking for question answering over knowledge graphs. In: The Semantic Web – ISWC 2018. pp. 108–126. Springer International Publishing (2018)
6. Kalogerakis, E., Christodoulakis, S., Moumoutzis, N.: Coupling ontologies with graphics content for knowledge driven visualization. In: IEEE Virtual Reality. vol. 2006, p. 6 (2006)
7. Knublauch, H., Hendler, J., Idehen, K.: SPIN - Overview and Motivation (2011), `https://www.w3.org/Submission/spin-overview/`
8. Knublauch, H., Ryman, A.: Shapes Constraint Language (SHACL) (2015), `https://www.w3.org/TR/shacl/`
9. Lebo, T., Sahoo, S., McGuinness, D.: PROV-O: The PROV Ontology. W3C Recommendation (2013), `https://www.w3.org/TR/prov-o/`
10. Miles, A.: SKOS Core Vocabulary Specification. English (November 2005), 1–28 (2006), `https://www.w3.org/TR/swbp-skos-core-spec/`
11. Moreau, L., Clifford, B., Freire, J., Futrelle, J., et al.: The Open Provenance Model core specification (v1.1). In: Future Generation Computer Systems (2011)
12. Motik, B., Grau, B.C., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C.: OWL 2 Web Ontology Language Profiles (2012), `https://www.w3.org/TR/owl2-profiles/`
13. Motik, B., Grau, B.C., Sattler, U.: Structured Objects in OWL: Representation and Reasoning. 17th International World Wide Web Conference pp. 555–564 (2008)
14. Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (2014)
15. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Computational Linguistics: Human Language Technologies. pp. 2227–2237 (Jun 2018)
16. Rosales-Méndez, H., Hogan, A., Poblete, B.: Voxel: A benchmark dataset for multilingual entity linking. In: The Semantic Web – ISWC 2018. pp. 170–186 (2018)
17. Sanderson, R., Ciccarese, P., Young, B.: Web Annotation Data Model. W3C pp. 1–56 (2016), `https://www.w3.org/TR/annotation-model/`
18. Sikos, L.F.: VidOnt: a core reference ontology for reasoning over video scenes. Journal of Information and Telecommunication pp. 1–13 (2018)
19. Sikos, L.F., Powers, D.M.: Knowledge-Driven Video Information Retrieval with LOD. Exploiting Semantic Annotations in Information Retrieval pp. 35–37 (2015)
20. Zhao, H., Lu, Z., Poupart, P.: Efficient Estimation ofWord Representations in Vector Space. IJCAI International Joint Conference on Artificial Intelligence (2015)