From Linguistic Linked Open Data to Multimodal Natural Interaction: a case study

Marco Grazioso
Department of Electrical Engineering and Information Technology
University of Naples "Federico II"
Naples, Italy
Email: mar.grazioso@studenti.unina.it

Valeria Cera
Department of Architecture
University of Naples "Federico II"
Naples, Italy
Email: valeria.cera@unina.it

Maria Di Maro
Department of Humanistic Studies
University of Naples "Federico II"
Naples, Italy
Email: maria.dimaro2@unina.it

Antonio Origlia
URBAN/ECO Research Center
University of Naples "Federico II"
Naples, Italy
Email: antonio.origlia@unina.it

Francesco Cutugno
Department of Electrical Engineering and Information Technology
University of Naples "Federico II"
Naples, Italy
Email: cutugno@unina.it

***Abstract*—We present here the conversion of Linguistic Linked Open Data into Semantic Maps to be used to produce contents in a set of technological applications for Cultural Heritage. The paper describes the architectural data collection and annotation procedure adopted in the Cultural Heritage Orienting Multimodal Experiences (CHROME) project (PRIN 2015 funded by Italian University and Research Ministry). Such data will be used in Multimodal Dialogue Systems to obtain precise information about Architectural Heritage, by means of pointing gestures or verbal requests. In particular, we design conversational agents accessing fine-detailed semantic data linked to available 3D models of historical buildings. The starting point of our scientific approach is the Getty Vocabulary on Art & Architecture Thesaurus, integrated with the Getty Thesaurus of Geographic Names (TGN) and the Union List of Artist Names (ULAN). These data are related to 3D mesh of the considered buildings in order to associate abstract concepts to architectural elements. In the field of 3D architectural investigation, a significant amount of research has been conducted to allow domain experts to represent semantic data while keeping spatial references. We will discuss how this will make it possible to support multimodal user interaction and generate Cultural Heritage presentations.**

I. Introduction

Recent advances in graphics hardware, together with the availability of professional video game engines, have increased the chances of developing innovative approaches for Cultural Heritage presentation. The use of game engines has been shown to produce beneficial effects on the interaction quality with systems based on advanced knowledge representation and dialogue-based interaction (e.g. [1]). In particular, the use of embodied conversational agents (ECA), represented as 3D avatars placed in virtual environments, provides a natural way to access to pieces of information. Establishing a dialogue with an artificial character is increasingly becoming one of the most appealing ways of interacting with technological devices. A current trend in managing art information is to make data on art, architecture, and other Cultural Heritage branches available as Linked Open Data (LOD). Following these trends we decided to combine LOD for Cultural Heritage, game engines, and conversational agents, with the aim of building an original Open Data representation to be accessed multimodally in interacting with ECA in a virtual environment.

Concerning the annotation of digital models, scholars associate geometric shapes with semantic descriptors. The most relevant approach to this kind of semantic annotation is presented in [2], where it deals with the geometrical segmentation of architectural digital artefacts, being handled as separate elements organised using part-whole relationships. Each entity is, furthermore, identified by a precise concept in a specialised domain thesaurus. Different geometrical representations (point clouds, nurbs, textured meshes, etc...) are linked to the objects represented by the terms, included in the dic-

978-1-5386-7202-0/18/$31.00 ©2018 IEEE
DOI 10.1109/iV.2018.00060

tionary, depending on the specific descriptive objectives. Each geometrical element can be linked to a single semantic descriptor, while a semantic descriptor may be associated to multiple geometrical elements. More recently, this methodology has been updated [3] and implemented as a cloud-based service called *Aioli*[1]. Using the projective relationship between bidimensional and tridimensional representations, the semantic annotation of digital models, obtained through a set of reference images, is produced by segmenting the same reference images, thus removing the need of a geometrical segmentation. Images sharing the same semantic label may be linked to one or more specific terms in a controlled vocabulary, or they may be characterised with customised attributes. Semantically annotated 3D models contain a significant amount of data that, in promoting Cultural Heritage, may be used to assist non-expert users to navigate cultural contents within interactive technologies. These technologies should be indeed designed to support the exploration of the large amount of information available for Cultural Heritage (texts, images, 3D models, etc...) in an engaging way. To tackle this problem, we pursue the use of ECA, in the form of 3D avatars, immersed in the digital representations of cultural artefacts. Using semantic processing techniques coming from different domains (e.g. Natural Language Processing, Computer Vision, etc...), it is possible to link separate sources of information and generate a consistent presentation exploiting the obtained semantic labels. An interesting work related to the annotation in the field of Cultural Heritage was proposed in [4] where the interaction with users visiting cultural sites is seen as a strategy to provide archaeological information like pictures, descriptions, and metadata, which can potentially empower the work of the architects and provide an enriched user experience.

In this paper, we present the architectural data collection pipeline we adopted to build the 3D meshes representing the relevant structural parts of the Padula Charterhouse "San Lorenzo" (Italy) and how we annotated them with semantic information converting Getty Thesaurus concepts into semantic maps. The obtained data represent a multifaceted documentation of Architectural Heritage describing both geometrical detail and visual experience. We also present the work in progress concerning a software architecture designed to use the semantically enriched 3D data to present relevant information based on user interaction. This architecture will be therefore used to support natural user interaction [5] through the use of Social Signal Processing techniques [6] and game engines.

II. Data collection

The developed pipeline, from the 3D data acquisition to the semantic annotation, was tested on Padula Charterhouse's Church. In order to obtain a 3D reality-based model, in virtual and interactive environments, several surveying techniques were employed, depending on the particular lighting conditions and on the morphological characteristics of the heritage site under consideration. A photogrammetric survey, integrated with terrestrial laser scanning acquisitions (TLS), were conducted to obtain a geometrically accurate 3D model that also delivers a photo-realistic view of the surveyed cultural site.

The record process, based on the state of art techniques, involved a two-step procedure. The first one was performed with a Continuous Wave Faro Focus 3D S120 laser scanner. Starting from the entrance, the laser was positioned differently in each recording to cover the entire volume of the nave, taking into account the geometry, the tangency of surfaces, and shadows. A total number of 10 scans, placing the scanner uniformly along an elliptical path with a spatial resolution of 6 mm at 10 m, was necessary to digitally record the whole area. At a later stage, a terrestrial photogrammetric survey was carried out to integrate the data missing in the laser scanning 3D survey and for texturing purposes. A digital reflex Canon EOS 1300D, coupled with a zoom 18-55 lens set at 24 mm view, was chosen for the acquisition of about 525 images, adopted to assure better colour information for the final texturing of the 3D digital model.

III. The Getty vocabularies-based ontology

The Getty vocabularies contain structured terminology for art, architecture, decorative arts, archival materials, visual surrogates, conservation, and bibliographic materials constructed to allow their use in linked data. The Getty Research Institute[2] released these vocabularies as LOD in order to make their resources freely available, and to consequently supply authoritative information for cataloguers, researchers, and data providers. The Getty vocabularies are compiled resources that grow through contributions from Getty projects and other institutions. In our specific case study, we refer to:

- Art and Architecture Thesaurus (AAT) [7].
- Getty Thesaurus of Geographic Names (TGN).

[1] www.aioli.cloud/

[2] www.getty.edu

- Union List of Artist Names (ULAN) vocabularies.

AAT provides terms, descriptions, and other metadata for generic concepts related to art, architecture, conservation, archaeology, and other Cultural Heritage-related fields. Information on work types, styles, materials, and techniques are also included. TGN provides names, descriptions, and other metadata for modern and historical cities, empires, archaeological sites, and physical features meaningful for art and architecture researches. ULAN provides names, biographies, related people, and other metadata on artists, architects, firms, studios, museums, patrons, sitters, and other people and groups involved in the field of art and architecture. As LOD, data are provided in RDF (Resource Description Framework) format following W3C[3]'s guide lines. This representation make data easily accessible using 'SPARQL Protocol and RDF Query Language' (SPARQL). SPARQL queries contain a set of triple patterns called 'basic graph pattern'. Triple patterns are like RDF triples except that each of the subject, predicate and object may be a variable. Members of the SPARQL family are considered as relational query languages because they have relational or pattern-based operations. Among the advantages of using the three selected resources, we need to mention that they present a shared unique ID for each element representing the same object. This feature makes possible to combine data form different vocabularies in order to obtain a more in-depth information on different objects. Since they are connected and provides complementary information, we decided to combine them. As a result, we were able to generate a larger knowledge base. To represent this knowledge we used a graph database, built with the graph database manager 'Neo4J' [8]. This open source software has been applied to a high number of tasks related to data representation (e.g. [9]). Since it's specialised in exploring graph paths, it helps us to efficiently store concepts information and retrieve data from our knowledge base.

Such ontological base can function as a conceptual hinge for a high-performance multimodal interactive system. Not only is it useful to enrich architectural 3D models, but it is also a source for developing a pointing interpreter capable of understanding the verbal request of information by a user enriched with a possible gesture, that points at an architectural element in the 3D scene. In addition, the knowledge graph can also be used to automatically extend speech recognition grammars with all the terms included in it, providing a wide

[3]www.w3.org

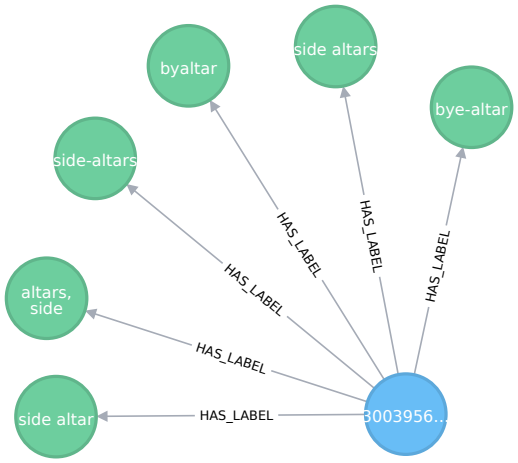

Fig. 1. Example of *side altar* concept showing its different lexical forms in out graph database

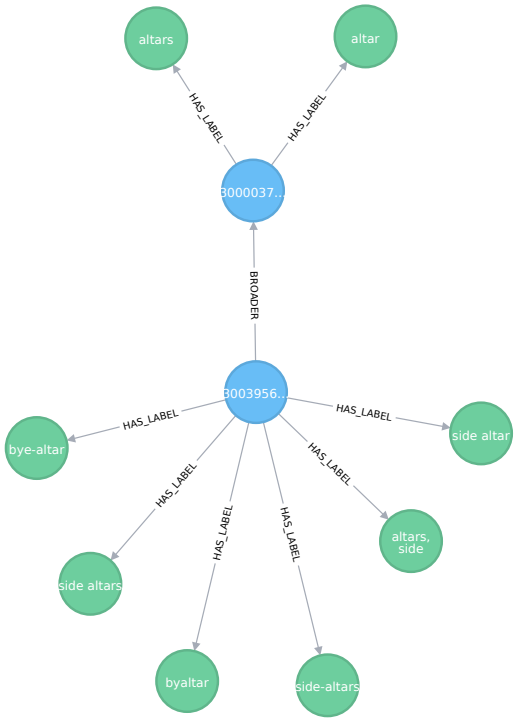

Fig. 2. Example of *broader* relation between *altar* and *side altar* concepts in our graph database

range of possible technical and less technical words that users would pronounce. For instance, AAT describes each term listing its hierarchical-related words: *column* is related to *pillar*, it is contextualised in *colonnade* and it distinguishes from *pier* and *post*. Additionally, translations into several languages are provided for each different lexical form. To better illustrate the information to extract from the developed knowledge base, in Figure 1, the relation between the concept *side altar* and its related lexical forms is shown, whereas in Figure 2 we display the hierarchical relation between *altar* and *side altar*, as they are stored in Neo4J.

IV. Annotation

In order to provide a natural and multimodal way of querying the Getty thesaurus, we propose an original concept representation that consists in semantic annotated maps. The expression 'semantic annotation' is here understood to mean the mapping of abstract concepts to the relevant parts of a 3D mesh. The annotation makes use of maps, describing semantic concepts, applied to 3D models like a texture, thus avoiding the need to geometrically segment the architectural artefact. To achieve this result, we defined an object that extends simple 3D meshes, associating the 3D model not only to the texture but also to the semantic map. For each concept of the thesaurus which was found in the digital architectural model, a semantic map is created an the concept ID is assigned. To improve the previously mentioned approach[3], where the notion of relevance is not provided, we represent the semantic information as a grayscale map: each map records which polygons, in the digital model, are relevant for the concept it represents by using the model's UV map. The level of relevance changes between white, that indicate high relevance (white = 1), and black that indicate no relevance (black = 0). In Figure 3 is presented an example of semantic map of the *altar* concept, while in Figure 4 is presented a normally coloured texture. Using semantic maps and reference IDs for the annotated concepts allows the integration of multiple sources of information (texts, images, audio recordings, etc…) sharing the same annotation scheme. Cross-referencing these sources opens the possibility to produce advanced interfaces to link the descriptions that a specific artefact has in separate domains.

The use of gradients is extremely important in the field of Architectural Heritage. For our maps, annotators refine the quality of the semantic data. The possibility to express more than a binary relevance of each vertex for a given concept is useful in cases where it is not possible to classify an element in a unique and precise way, or when an architectural element cannot be assigned to a specific category (grey = 0,5). The same happens when it is not possible to clearly state the exact boundaries between an architectural element and another. This also makes it possible to consider semantic maps produced by multiple annotators resulting in a final map computed with the mean values for each UV coordinate. This approach has been used in other fields where annotation uncertainty is important, like for emotions [10]. In conclusion, the possibility to express a relevance *level* for a concept makes the map more expressive and adaptable to specific

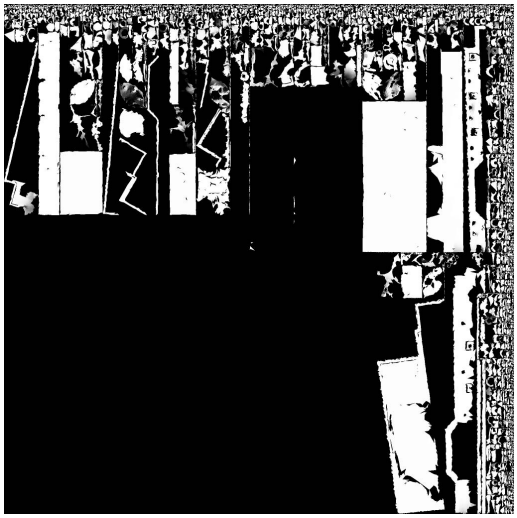

Fig. 3. A semantic map for the *altar* concept

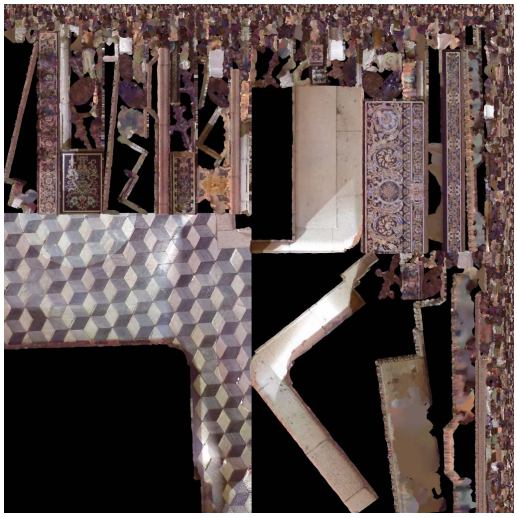

Fig. 4. Colour map of a sample segment of a flattened 3d model

cases.

V. Multimodal Interaction

In order to generate cross-domain presentations, semantically annotated materials can be exploited. These ones are intended to be adopted in the design of a Multimodal Dialogue System, where a virtual character in the 3D scene multimodally interacts with users, making use of different channels of communication, such as natural commands and social signals. We thus decided to take advantage of these possibilities by designing a software architecture that combines specialised modules for knowledge representation and interaction management. The essential components of our software architecture are:

- Neo4J graph database to represent the knowledge.
- a dialog manager to handle interaction (OpenDial).
- a game engine (Unreal Engine 4) to control the virtual character.
- a voice synthesizer (Mivoq).
- a Kinect sensor to collect user's activity data.

The Neo4J graph database contains the knowledge graph obtained combining the Getty vocabularies. OpenDial [11] is a framework for dialogue management that uses probabilistic rules and aims at better merging probabilistic dialogue and rule-based management. OpenDial's probabilistic rules are used to setup and update a Bayesian network consisting of variables that represent the current dialogue state, including uncertainty. The next system action is computed using a utility based approach. The virtual avatar and the real time rendering of the obtained artefacts are controlled using the Unreal Engine 4[4]. Mivoq Voice Synthesis Engine[5] is used to dynamically generate the avatar voice. User gestures and speech are detected with the Kinect sensor and are continuously forwarded to the game engine.

With this approach, the raw user data handling is assigned to a module designed to manage complex and dynamic interfaces that includes video, audio and user control systems, while the dialogue manager processes high-level decisions. This component accesses the encyclopedic knowledge represented in the graph database and selects the most appropriate response to the user input. To better select the response, approaches that use Conceptual Dependency Theory [12], like the one proposed in [13] can be used. The response consists of an abstract *plan* that may include text extracted from the knowledge base, clarification requests or generic action instructions (e.g. *enter another environment*). Because of the dynamic nature of the interaction between users and the avatar, but also between the avatar and the 3D surroundings, the task concerning the action implementation decision was assigned to the game engine. In fact, a reactive behavioural logic is needed to manage interrupts caused by both implicit or explicit user activities. To ensure consistency with the users' behaviour, social signals generation and monitoring must be performed in real time. Lastly, the relative position of the avatar with respect to the concepts that are relevant for the generated utterances must also be evaluated in real time in order to generate pointing gestures.

To support multimodal commands from the users and allow a richer interaction we use the Kinect sensor to detect users' skeletons. We identify shoulder and hand joints and use their coordinates in the virtual space to get a direction vector. Successively, by exploiting the ray-casting system included in the engine, it is possible to emit a single, invisible ray of light that follow the direction vector to capture collision events between the ray and the objects in the scene. From the data included in the collision event, it is, then, possible to extrapolate the UV coordinates of the vertex that is closest to the collision point. These UV coordinates can then be used to query the semantic maps of the object the ray collided with, to extract relevance information for the annotated concepts. When a speech command is detected and multimodal fusion has been performed, relevance information can then be passed to the dialog manager. Further details on interaction management strategies will be formalised on the basis of audiovisual recordings of expert art historians presenting the Campanian Charterhouses to small groups of visitors, which are currently being collected in the framework of the CHROME project.

VI. Conclusion

An original method to semantically annotate the low poly meshes has been developed to allow a direct link between concepts in the Getty thesaurus and geometric parts, introducing the possibility to represent uncertainty. This approach provides a different Open Data visualisation through semantic maps. The semantic data produced with this workflow will allow the development of 3D conversational agents able to refer to the reconstructed annotated environment, capable of being queried by users in a natural and multimodal way.

[4]www.unrealengine.com

[5]www.mivoq.it

REFERENCES

[1] A. Origlia, P. Cosi, A. Rodà, and C. Zmarich, "A dialogue-based software architecture for gamified discrimination tests," in *Proc. of GHItaly*, 2017. [Online]. Available: http://ceur-ws.org/Vol-1956/

[2] L. De Luca, "Relevé et multi-représentations du patrimoine architectural définition d'une approche hybride pour la reconstruction 3d d'édifices," Ph.D. dissertation, Sciences de l'Homme et Société. Arts et Métiers ParisTech, 2006.

[3] T. Messaoudi, P. Véron, G. Halin, and L. De Luca, "An ontological model for the reality-based 3D annotation of heritage building conservation state," *Journal of Cultural Heritage*, vol. 29, pp. 100–112, 2018.

[4] V. Deufemia, V. Mascardi, L. Paolino, G. Polese, and H. de Lumley, "A volunteered geographic information system for collecting and rating petroglyph data," *J. Vis. Lang. Comput.*, vol. 25, no. 6, pp. 963–972, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.jvlc.2014.10.031

[5] D. Wigdor and D. Wixon, *Brave NUI world: designing natural user interfaces for touch and gesture*. Elsevier, 2011.

[6] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[7] T. Petersen, "Developing a New Thesaurus for Art and Architecture," *Library Trends*, vol. 38, no. 4, pp. 644–658, 1990.

[8] J. Webber, "A programmatic introduction to Neo4j," in *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*. ACM, 2012, pp. 217–218.

[9] F. Dietze, J. Karoff, A. C. Valdez, M. Ziefle, C. Greven, and U. Schroeder, "An open-source object-graph-mapping framework for Neo4j and Scala: Renesca," in *International Conference on Availability, Reliability, and Security*. Springer, 2016, pp. 204–218.

[10] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proc. of ICME*, 2010, pp. 1079–1084.

[11] P. Lison and C. Kennington, "Opendial: A toolkit for developing spoken dialogue systems with probabilistic rules," *ACL 2016*, p. 67, 2016.

[12] R. C. Schank, "Conceptual dependency: A theory of natural language understanding," *Cognitive psychology*, vol. 3, no. 4, pp. 552–631, 1972.

[13] P. L. Albacete, S. K. Chang, and G. Polese, *Iconic language design for people with significant speech and multiple impairments*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 12–32. [Online]. Available: https://doi.org/10.1007/BFb0055967