Clustering over the Cultural Heritage Linked Open Dataset: Xlendi Shipwreck

Mohamed BEN ELLEFI[1], Mohamad Motasem NAWAF[1], Jean-Christophe SOURISSEAU[2], Timmy GAMBIN[3], Filipe CASTRO[4], and Pierre DRAP[1]

[1] Aix Marseille University, CNRS, ENSAM, University of Toulon, LIS UMR 7020, 13397 Marseille, France;
firstname.lastname@univ.amu.fr

[2] Aix Marseille Univ, CNRS, Ministère de la Culture et de la Communication, CCJ UMR 7299, 13094 Aix En Provence, France;
jean-christophe.sourisseau@univ-amu.fr

[3] Archaeology Centre (Car Park 6), University of Malta, Msida MSD 2080, Malta;
timmy.gambin@um.edu.mt

[4] Ship Reconstruction Laboratory 4352 TAMU, Texas A-M University, College Station, Texas 77843, USA;
fvcastro@tamu.edu

Abstract. Cultural heritage (CH) resources are very diverse, heterogeneous, discontinuous and subject to possible updates and revisions in nature. The use of semantic web technologies associated with 3D graphical tools is proposed to improve the access, the exploration, the mining and the enrichment of this CH data in a standardized and more structured form. This paper presents a new ontology-based tool that allows to visualize spatial clustering over 3D distribution of CH artifacts. The data that we are processing consists of the archaeological shipwreck "Xlendi, Malta", which was collected by photogrammtry and modeled by the Arpenteur ontology. Following semantic web best practices, the produced CH dataset was published as linked open data (LOD).

1 Introduction

The study of the history of seafaring is the study of the relations of humans with rivers, lakes, and seas, which started in the Paleolithic. An understanding of this part of our past entails the recovery, analysis, and publication of large amounts of data, mostly through non-intrusive survey methods. The methodology proposed in this paper aims at simplifying the collection and analysis of archaeological data, and at developing relations between measurable objects and concepts. It builds upon the work of J. Richard Steffy, who in the mid-1990s developed a database of ship components. This shipbuilding information, segmented in units of knowledge, tried to encompass a wide array of western shipbuilding traditions which developed through time and space and establish relations between conception and construction traits in a manner that allowed comparisons between objects and concepts. Around a decade later Carlos Monroy transformed Steffy's database into an ontological representation in RDF-OWL, and expanded its scope to potentially include other archaeological materials [12]. After establishing a preliminary ontology, completed through a number of interviews with naval and maritime archaeologists, Monroy combined the database with a multi-lingual glossary and built a series of relational links to textual evidence that aimed at contextualizing the archaeological information contained in the database. His work proposed the development of a digital library that combined a body of texts on early modem shipbuilding technology, tools to analyze and tag illustrations, a multi-lingual glossary, and a set of informatics tools to query and retrieve data [3].

Our approach extends these efforts into the collection of data, expands the analysis of measurable objects, and lays the base for the construction of extensive taxonomies of archaeological items. The applications of this theoretical approach are obvious. It simplifies the acquisition, analysis, storage, and sharing of data in a rigorous and logically supported framework. These two advantages are particularly relevant in the present political and economic world context, brought about by the so-called globalization and the general trend it entailed to reduce public spending in cultural heritage projects. The immediate future of naval and maritime archaeology depends on a paradigm change. Archeology is no longer the activity of a few elected scholars with the means and the power to define their own publication agendas. The survival of the discipline depends more than ever on the public recognition of its social value. Cost, accuracy, reliability (for instance established through the sharing of primary data), and its relationship with society's values, memories and amnesias, are already influencing the amount of resources available for research in this area. Archaeologists construct and deconstruct past narratives and have the power to impact society by making narratives available that illustrate the diversity of the human experience in a world that is less diverse and more dependent on the needs of world commerce, labor, and capital.

In the context of semantic web works toward the development of culture heritage applications, we cite recent projects that among others, provide multimedia access to distributed collections of CH resources: *(i)* data portals like ADS[5], ARIADNE[6], EUROPEANA[7] and STITCH[8], *(ii)* vocabularies like the CIDOC-CRM[9] and the Getty vocabularies[10]. A different approach is adopted by [11], where authors present a framework that relies on the Ontology-Based Data Access (OBDA) paradigm to allow for virtual integration based on rewriting SPARQL queries over the EPNet ontology to SQL queries over distributed data sources.

[5] http://data.archaeologydataservice.ac.uk/query/
[6] http://www.ariadne-infrastructure.eu/
[7] https://www.europeana.eu/portal/fr
[8] https://www.cs.vu.nl/STITCH/
[9] http://www.cidoc-crm.org/
[10] http://vocab.getty.edu/

This work is centered on the Xlendi shipwreck, named after the place where it was found off the Gozo coast in Malta. The shipwreck was located by the Aurora Trust, an expert in deep-sea inspection systems, during a survey campaign in 2008. The shipwreck is located near a coastline known for its limestone cliffs that plunge into the sea and whose foundation rests on a continental shelf at an average depth of 100 m below sea level. The shipwreck itself is therefore exceptional; first due to its configuration and its state of preservation which is particularly well-suited for our experimental 3D modeling project. The examination of the first layer of amphorae also reveals a mixed cargo, consisting of items from Western Phoenicia and Tyrrhenian-style containers which are both well-matched with the period situated between the end of the VIII and the first half of the VII centuries BC. The historical interest of this wreck, highlighted by our work, which is the first to be performed on this site, creates a real added-value in terms of innovation and the international reputation of the project [5].

This paper is a continuity for a previous work published in [5] where we developed tools combining photogrammetry and knowledge representation that provide new analysis of the visible part of the cargo. We have also developed an ontology that models both the photogrammetric process and the measured objects, as detailed in [1]. The focus of this paper is to publish the produced CH dataset as linked open data following the semantic web best practices. Furthermore, we introduce a new GUI tool for clustering over the distribution of different artifacts in the published LOD dataset. In 2001 the UNESCO Convention for the Underwater Cultural Heritage established the necessity of making all archaeological data available to the public[11].

The rest of the paper is organized as follow: first, section. 2 will presents the adopted photogrammetrical process during data gathering. Further, section. 3 discusses the motivation behind our conceptual model then introduces the newly published dataset with an illustrative example. Next, section. 4 presents our GUI clustering tool that provides a 3D visualization of the resources density distribution in th published dataset. Finally, we conclude and give some future direction in the last section.

2 Photogrammetry Survey

Data acquisition and processing using photogrammetry allow the capture of an impressive amount of underwater site features and details [13]. In the in Xlendi shipwreck, the aim of deploying a photogrammetry framework is to perform survey and produce a complete 3D model and overall orthophoto. The acquisition system used for the photogrammetric survey was installed on the Rmora 2000 submarine made by COMEX[12]. This two-person submarine has a depth limit of 610 m with a maximum dive time of 5 hours, which provides more than enough time for the data acquisition phase of the photogrammetry survey. What is of crucial importance to us are the three high-resolution cameras that are

[11] http://vww.unesco.org/new/en/culture/themes/underwater-cultural-heritage/

[12] http://comex.fr/

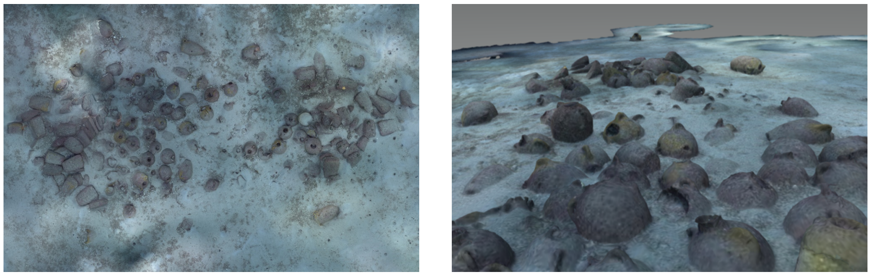


Fig. 1. Example of obtained models for the underwater site Xlendi. An Overall orthphoto (left) and a close-up view of the generated 3D model (right)

synchronized and controlled by a computer. All three cameras are mounted on a bar located on the submarine just in front of the pilot. Continuous lighting of the seabed is provided by a Hydrargyrum medium-arc iodide lamp (HMI) powered by the submarine. The continuous light is more convenient for both the pilot and the archaeologist who can better observe the site from the submarine. The high frequency acquisition frame rate of the cameras ensures full coverage whereas the large scale of acquired images gives the eventual 3D models extreme precision (up to 0.005 mm/pixel for the orthophoto). Briefly, the deployed procedure consists of three phases, the first two are done in real-time while the third is achieved in a later step. Starting with image orientation phase, it is possible to know the exact pose of the camera at each image acquisition. On the other hand, contrary to PMVS, our developments directly use the images produced by the cameras, without any distortion correction nor rectification. We refer to [5] for more details, see Figure1. Deployed in this way, the acquisition system entails zero contact with the archaeological site making it both non-destructive and extremely accurate.

The next section will introduce our method for modeling the photogrammetry data using ontologies in order to facilitate data sharing between researchers with different backgrounds, such as archaeologists and computer scientists. The ontology-based model can be particularly useful to improve and expand data analysis and to identify patterns or to generate different statistics using a simple query language that is close to natural language.

3 Xlendi As Linked Open Dataset

3.1 Ontology Conceptualization

Cultural heritage data is very heterogeneous and can have different ambiguous descriptions. Hence, the most challenging problem for metadata designers and cultural heritage experts is to provide a common conceptualization of the data. This conceptualization provides a common way of representing knowledge about some domain and a way to share a common understanding of information structure. Once we have common understanding, we can try to reason/query over this information, i.e. inference, consistency checking, etc. To develop a transversal data mining techniques and adapted systems, conceptualization must provide an intelligible description that allows a better understanding for experts manipulating the data. By organizing this information in an ontology, the conceptualization can be used to cover different terminologies and to represent a clear specification of the different meanings. In this way, the ontology model can guide the design of the knowledge bases to store the various experimental data as well as the measurement process in a knowledge manner. In the remainder of this paper, we adopt the computational meaning of ontology which can be seen as a structured system of fundamental concepts and relationships and of an agreed epistemology, i.e. clearly defined rules of evidence and reasoning, which do not privilege individual experiences or beliefs that cannot be argued against, and which at the same time include clear evaluation mechanisms for the credibility of research conclusions [9].

In a collaborative work between archaeologists and ontology designers, we developed a common ontology that models cultural heritage artifacts in term of their typologies, photogrammetric process and spatial representation, as in [1], where we presented our model for profiling archaeological amphorae. We serialized our ontology using the Web Ontology Language OWL2[13], and we made available on[14]. Following the linked data best practices [2], metadata designers reuse and build on, instead of replicating, existing ontologies and vocabularies. Motivated by this observation, we linked our ontology to the CIDOC-CRM ontology [4] and GeoSPARQL[15] in order to allow more integrity cross different cultural heritage datasets using different ontologies, i.e. enabling to perform federated queries cross multiple datasets. The ontology was modeled closely linked to the Java class data structure in order to be able to manage the photogrammetric process as well as the measured items. Note that each concept or relationship in the ontology has a counterpart in Java (the opposite is not necessarily true). Finally, our ontology has been integrated in the linked open vocabularies for better terms reuse, see[16].

3.2 Xlendi LOD Dataset

We draw the reader intention that the data from Xlendi shipwreck was processed by photogrammetry in a previous work [5]. The focus of this paper is to publish this dataset as linked open data following the semantic web best practices. Hence, the dataset processed by photogrammetry is stored in ABox OWL file and we made it available as open data on the datahub under the name "Xlendi Amphorae", see[17]. For better understanding of the dataset, we detail in the following the two sample files "XlendiApmhoraeSample" and "PhotographSample":

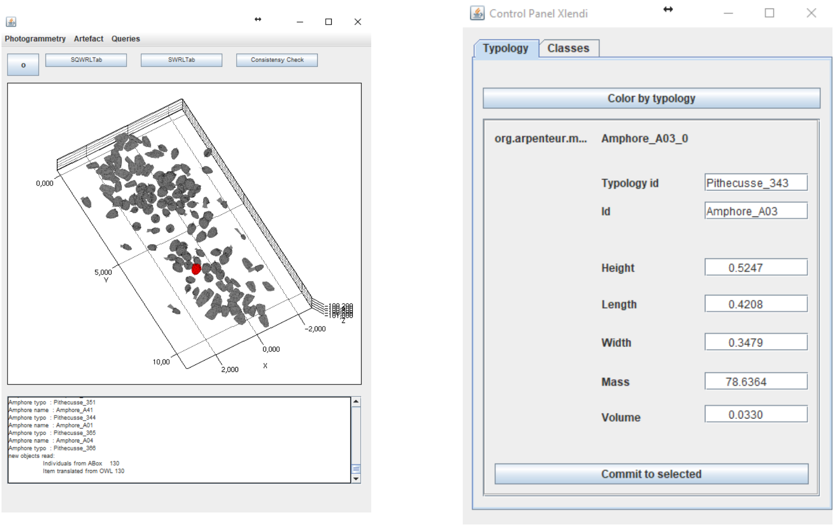
[13] W3C Consortium recommendation, see https://www.w3.org/TR/owl2-overview/

[14] http://www.arpenteur.org/ontology/Arpenteur.owl

[15] http://www.opengeospatial.org/standards/geosparql

[16] http://lov.okfn.org/dataset/lov/vocabs/arp

[17] https://datahub.ckan.io/dataset/xlendiamphorae



```
Amphore typo  : Pithecusse_351
Amphore name  : Amphore_A41
Amphore typo  : Pithecusse_344
Amphore name  : Amphore_A01
Amphore typo  : Pithecusse_365
Amphore name  : Amphore_A04
Amphore typo  : Pithecusse_366
new objects read:
            Individuals from ABox    130
            Item translated from OWL 130
```

org.arpenteur.m... Amphore_A03_0

Typology id: Pithecusse_343

Id: Amphore_A03

Height: 0.5247

Length: 0.4208

Width: 0.3479

Mass: 78.6364

Volume: 0.0330

Fig. 2. A 3D visualization of amphorae stored in the ABox by our GUI tool. (a) Amphora_A03 spatial position in the Xlendi site. (b) the typology dimensions corresponding to the amphora selected in the 3D spatial site

- We start with an example of the amphora instance *Amphore_A03* in the RDF file "XlendiApmhoraeSample". The spatial description of this amphore is represented through the *hasTransformation3D* relation which points to the *Transfo1003825059*, which provides connections to the corresponding RotationMatrix and the IPoint3D, i.e. respectively *Mat1743553655* and *IPoint3D635001030* that together provide information about the shape and the localization of *Amphore_A03*.
- The RDF file "PhotographSample" in the XlendiAmphorae dataset depicts an example of a photograph instance *Photograph_13*. This photograph instance is connected to a camera and a 3D transformation. The camera is described by a set of camera settings properties and enriched by a distortion specifications, which is particularly crucial for the photogrammetry measuring. The 3D transformation describes the photograph with a set of 3D points and a set of rotation matrix.

Finally, the complete set of CH artifacts, amphorae and grinding-stones, are made available in the "XlendiArtifacts" OWL file.

3.3 Xlendi Data Linking

Following semantic web best practices, we need to provide links to further candidate datasets that may contain similar instances in order to join the LOD cloud[18]. However, we provide the only available RDF data that represents am-

[18] http://lod-cloud.net/

phorae collected from the Xlendi shipwreck. Hence, we looked into DBpedia, being the most obvious target in the LOD. The only similarity that we found within this multi-domain dataset consists on the instances of the widely used concept "Camera". However, in our dataset the distinguishing criterion between different cameras is the setting (calibration, distortion, etc) not the camera type i.e. different instances refers to the camera Nikon D700 but with different setting. For this purpose, the identity link is not able to be adopted in this case ("owl:sameAs" to DBpedia camera Nikon_D700[19]), according to [10]. Other broader links such as skos:broadMatch[20] might be semantically more appropriate since they indicate a broader matching links.

3.4 Xlendi Data Visualization

For better visualization of the stored dataset, we developed a graphical user interface tool that loads the published dataset to a 3D graphical visualization. Figure. 2 shows a view of our GUI tool demonstrating the 3D density distribution of amphorae and grinding-stones in the Xlendi shipwreck. In this way, the user can graphically depict information about different artifacts based on their 3D spatial representation. The Figure. 3.2 shows the case of *Amphore_A03* and its localization in the shipwreck while the corresponding information about the artifact typology is depicted in Figure. 3.2. Note that our GUI tool offers further services which are currently in a work in progress statue from which we can cite [6], where we demonstrated a prototype of implemented spatial queries using SQWRL (a SWRL[21]-based query language) in our tool. For example, the operator "isCloseTo" built-ins which allows to select artifacts present in a sphere centered regarding to a specific one. In the next section we will introduce a new tool that offers a 3D clustering functionalities over the ABox part of the dataset.

4 Spatial clustering for Xlendi Dataset

Within the vast domain of data mining, spatial data mining is an important field of research and has always been of particular interest for archaeological community. Spatial data mining can be seen as the process of extracting potentially useful and previously unknown information from spatial datasets. One of the most fundamental tasks in spatial data mining is spatial clustering which has been steadily gaining importance over the past decade. Clustering algorithms are attractive for the class identification tasks. There are many spatial clustering methods available and each of them may give a different grouping set of a dataset. Here, we focus on density-based clustering algorithms where the idea is that objects which form a dense region should be grouped together into one cluster. These algorithms search for regions of high density in a feature space

[19] http://dbpedia.org/page/Nikon_D700

[20] http://www.w3.org/2004/02/skos/core

[21] https://www.w3.org/Submission/SWRL/

that are separated by regions of lower density. Thus, density-based methods can be used to filter out noise, and discover clusters of arbitrary shape.

In our tool, we implemented two well known clustering algorithms K-Means++ [8] and the DBSCAN [8] (i.e. as a density-based algorithm for discovering clusters in large spatial databases with noise). The main intuition behind our choice is to provide the user multiple choices to address users needs. For example, if the user knows in advance the number of clusters, K-means++ will be the more appropriate choice. Otherwise, DBSCAN clustering can be performed without knowing the number of clusters. Furthermore, we give the user the choice to select properties on which the clustering will be based, i.e. clustering Xlendi artifacts based on their typology, volume, length or height.

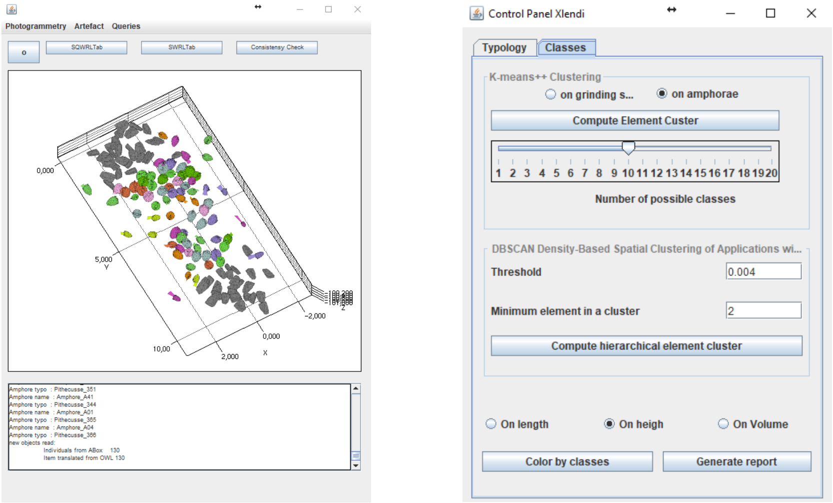
In the following, we detail our implementation of DBSCAN. This algorithm is mainly used to cluster point objects, which is perfectly in line with our model where any spatial object can be represented as a point (as detailed in Section.2). The main intuition is that, within each cluster, there is a typical density of points which is considerably higher than outside of the cluster. Subsequently, the density within the areas of noise is lower than the density in any other area of the clusters. Two important parameters are required for DBSCAN: a distance threshold - ϵ, and a minimum number of points - $MinPts$. The parameter ϵ defines the radius of neighborhood around a point A. It's called the ϵ-neighborhood of A. The parameter $MinPts$ is the minimum number of neighbors within ϵ radius.

Following the main concepts defined in DBSCAN [8], let's consider the set of amphorae in the Xlendi shipwreck as a set of n points $\{A_0, .., A_i, .., A_n\}$, that DBSCAN will cluster as follow:

1. For each point A_i, the algorithm computes the distance between A_i and the other points. Finds all neighbor points within distance ϵ of the starting point (A_i). Each point, with a neighbor count greater than or equal to $MinPts$, is marked as core point or visited.
2. For each core point, if it's not already assigned to a cluster, creates a new cluster. Finds recursively all the density points connected to it and assigns them to the same cluster as the core point.
3. Iterates through the remaining unvisited points in the data set.

Note that points not belonging to any clusters are treated as outliers or noise.

Figure. 4 depicts the setup panel within our GUI tool where user can define the setup parameters of the clustering methods. The setup panel provides access to: *(i)* K-Means++ setup parameters by selecting the appropriate number of clustering; *(ii)* DBSCAN parameters: the minimum number of neighbors (i.e. $MinPts = 2$) and the threshold ϵ (i.e. $\epsilon = 0.004$); *(iii)* the artifact property to cluster on. The clustering result is represented by different distribution of colors within the site, as shown in Figure 4. Furthermore, our GUI tool generates a report describing the deviation ratio of the generated clusters in term of: average, median, minimum, maximum, median absolute deviation (MAD) and root mean square (RMS).



```
Amphore typo  : Pithecusse_351
Amphore name  : Amphore_A41
Amphore typo  : Pithecusse_344
Amphore name  : Amphore_A01
Amphore typo  : Pithecusse_365
Amphore name  : Amphore_A04
Amphore typo  : Pithecusse_366
new objects read:
            Individuals from ABox    130
            Item translated from OWL 130
```

○ on grinding s... ● on amphorae

Threshold 0.004

Minimum element in a cluster 2

○ On length ● On heigh ○ On Volume

Fig. 3. A 3D visualization of Xlendi amphora distribution by our GUI tool. (a) 3D visualization of colored clusters of amphorae. (b) A view of the clustering setup panel

Finally, our clustering tool is integrated into our 3D geographic information system that merges photogrammetry and ontologies with an aim to the automatic production of 3D (or 2D) models through ontological queries: these 3D models are in fact at the same time a graphic image of the archaeological knowledge and the current interface through which the user can edit the dataset. Further clustering functionalities can be integrated in our GUI tool, as we cite the work on [7] where we proposed a clustering model for the Montreal Castle in Shawbak, Jordan.

5 Conclusion

In this paper, we introduced the Xlendi shipwreck dataset that was published as linked open data. We developed tools combining photogrammetry and knowledge managements to provide a 3D virtual survey of the cargo. The tool allows to load the LOD dataset and to visualize the spatial distribution of the different artifacts in the shipwreck. Based on this distribution, the user is able to extract different information about the artifacts dimension typologies. Different clustering methods are implemented and can be processed over the artifacts distribution aiming to be exploited according to cultural heritage tasks and users preferences.

Future directions can go towards the extension of our tool with a new interface allowing to assist CH users in building semantic queries and rules. Also, we are currently looking for potential candidate datasets that may contain similar artifacts as the Xlendi dataset in order to produce a 5-stars linked open data[22],

[22] http://5stardata.info/en/

i.e. connecting Xlendi amphorae to the ones having similar typologies in the ADS[23].

Acknowledgments

The authors would like to thank the iMareCulture project for partially funding this work, URL: http://imareculture.weebly.com.

References

1. Ben Ellefi, M., Papini, O., Merad, D., Boi, J.M., Royer, J.P., Pasquet, J., Sourisseau, J.C., Castro, F., Nawaf, M.M., Drap, P.: Cultural heritage resources profiling: Ontology-based approach. In: Companion of the The Web Conference 2018. pp. 1489–1496. WWW '18 (2018), https://doi.org/10.1145/3184558.3191598
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
3. Cobar, C.A.M.: A digital library approach to the reconstruction of ancient sunken ships. Texas A&M University (2010)
4. Doerr, M.: The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. AI magazine 24(3), 75 (2003)
5. Drap, P., Merad, D., Hijazi, B., Gaoua, L., Nawaf, M.M., Saccone, M., Chemisky, B., Seinturier, J., Sourisseau, J.C., Gambin, T., et al.: Underwater photogrammetry and object modeling: a case study of xlendi wreck in malta. Sensors 15(12), 30351–30384 (2015)
6. Drap, P., Papini, O., Pruno, E., Nucciotti, M., Vannini, G.: Ontology-based photogrammetry survey for medieval archaeology: Toward a 3d geographic information system (gis). Geosciences 7(4) (2017), http://www.mdpi.com/2076-3263/7/4/93
7. Drap, P., Papini, O., Pruno, E., Nucciotti, M., Vannini, G.: Ontology-based photogrammetry survey for medieval archaeology: Toward a 3d geographic information system (gis). Geosciences 7(4) (2017)
8. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. vol. 96, pp. 226–231 (1996)
9. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Handbook on ontologies, pp. 1–17. Springer (2009)
10. Halpin, H., Hayes, P.J.: When owl: sameas isn't the same: An analysis of identity links on the semantic web. In: In: International Workshop on Linked Data on the Web (LDOW) (2010)
11. Kharlamov, E., Kotidis, Y., Mailis, T., Neuenstadt, C., Nikolaou, C., Özçep, Ö., Svingos, C., Zheleznyakov, D., Brandt, S., Horrocks, I., et al.: Towards analytics aware ontology based access to static and streaming data. In: International Semantic Web Conference. pp. 344–362. Springer (2016)
12. Monroy, C., Furuta, R., Castro, F.: Synthesizing and storing maritime archaeological data for assisting in ship reconstruction. Oxford Handbook of Maritime Archaeology; doi: 10.1093/oxfordhb/9780199336005.013.0015 (2011)
13. Scaradozzi, D., Sorbi, L., Zoppini, F., Gambogi, P.: Tools and techniques for underwater archaeological sites documentation. In: Oceans-San Diego, 2013. pp. 1–6. IEEE (2013)

[23] http://data.archaeologydataservice.ac.uk/query/