Big Data Perspective with Otological Modeling for Long Term Traceability of Cultural Heritage

Muhammad Naeem, Muhammad Fahad, Néjib Moalla, Yacine Ouzrout, Abdelaziz Bouras

Big Data Perspective with Ontological Modeling for Long Term Traceability of Cultural Heritage

[1]Muhammad Naeem, [1]Muhammad Fahad, [1]Néjib Moalla,
[1]Yacine Ouzrout, [2]Abdelaziz Bouras
[1]Université Lumière Lyon 2, DISP Laboratory France
{FirstName.LastName}@univ-lyon2.fr
[2]Qatar University, ICT Qatar Chair, College of Engineering, Doha, Qatar
abdelaziz.bouras@qu.edu.qa

Abstract. The safeguarding of cultural heritage has brought forward the generation of heterogeneous, complex, diversified and irreplaceable digital data. It becomes difficult for an object with missing characteristics to perform the premises identification, object identification as well as its movement recording. Therefore, it is an imperative need of traceability of the cultural digital objects. In this study, we have proposed an expert system to address the issues of achieving and maintaining traceability of cultural objects. The system has employed big data technologies as well as ontological modeling capabilities to semantically trace objects. We have designed *Cultural Heritage Ontology (CHOnt)* that captures all the semantics for inference mechanisms. We have shown that the proposed system is capable of sound expressiveness with an immense potential in offering a scalable solution as a common vehicle through which archaeologists, IT specialists and even a non-professional can trace, evaluate, enhance, analyze and exchange all types of information.

Keywords: Big Data, DBSCAN, Cultural Heritage Ontology, Traceability

1 Introduction

Cultural heritage of a civilization bears numerous artifacts, such as folk music, dance, language, folk lore, seer axiom, oral literature, manners, games, etiquette, handicraft, traditional medicine, architecture arts and other preservation methods used for the expressivity of a region. Each artifact mentioned above is either tangible or intangible but strictly continuous in its nature. This intrinsic nature of dynamic behaviors of a culture has motivated the artists, poets, writers and painters to preserve the heritage in the shape of tangible formats. With the advent of modern technology, the concept of *Digital Preservation (DP)* or knowledge retention was emerged [1]. The core purpose of DP is the process of ensuring communication between future and past via innovative artifacts. In other words, it points out the persistence of digital resources in such a way that enables their rendering with easiness, availability and comprehensibility for the ancillary contemporary reuse. A big advantage of DP is that it can ensure the process of protecting the continuance concerns of forthcoming generations. There is an imperative motivation to maintain universal knowledge

repositories addressing digital museums, communication conduits, digital archives and other type of digital memory systems. The museums and archivists are constantly collecting cultural heritage. According to The World Museum Community, there are more than 55,000 museums in 202 countries [2]. Here a question arises, if a museum receives a digital resource of an artifact, then it may or may not have all of its related information. The completion of a set comprising all relevant information can helps in investigation of the authenticity of antique paintings. The experts usually adopt verification mechanism through stylistic evaluation, objective tests of the ageing of the underlying material or with the help of modern scientific tools. Some researcher have highlighted the problem of retrieval as important issue in the cultural heritage domain [3]. It has also been reported that the unstructured data handling in the cultural heritage data is a problem [4]. Another perspective was how to find out the semantics of the cultural heritage [5] . We have argued in this study about the scalability of all of such techniques handling the cultural data in the wake of mixture of structured and unstructured data. We have formulated a research question: *How can the traceability of a new or existing but questionable object can be performed given large amount of structured and unstructured data in the domain of cultural heritage.* The proposed system has used various technologies including Big Data, Natural Language Processing (NLP) and Ontological Modeling. We in this study have proposed an expert system which can assist an expert to conclude more precisely and with improved confidence.

The reaming of the paper is organized as below. The section 2 is related the overview of the techniques describing the problem in this context. In section 3, we have introduced the architecture and then discussed it in detail. This section is further divided into numerous sub sections in which we have discussed each component, experimental work carried out and its justifications. The last section is concerned with the conclusive remarks in which we have highlighted how our proposed architecture is useful for the traceability of the cultural heritage domain.

2 Literature Review

We have reviewed the literature related to the provision of traceability of the cultural data in all of the possible ways. We noticed that most of the research work has been carried out by means of applying “Semantic engineering” on the data itself [5] [1] . The problem with such an approach is that developing a semantic network on varied level of information produces “less concise” ontologies. We have argued in such situations that prior to feeding the data in the semantic engineering, one must converge the plethora of information into a fine grained dataset. More the data set is precised, better the output of the final semantic expressivity of cultural heritage.

As the information era dawned with the advent of modern computational tools, the vast amount of digital data has been reorganized in the structured and unstructured format. This data eventually culminated into the databases grouping the cultural heritage knowledge [6]–[8].

Apart from these specialized sources, other online knowledge repositories such as wikis and weblogs could also be observed for the source of information; albeit such

sources are less explicit and void of systematic traceability of a query related to newly arrived artifact in a museum.
The research question is that how can we introduce new functionalities and opportunities to improve the quality of cultural data, whether using innovative semantic web techniques alone are sufficient? [9], [10] introduced work employing intelligent engineering, however their approaches were limited to only provision of facilitation of better accessibility to end users of their systems.
The usefulness of NLP has also been highlighted in this domain [4]. They have used the NLP for the purpose of identifying essential information earlier. They highlighted that most of the systems deal only with the relational data. However, they highlight that there are numerous situations when there is only unstructured data. They proposed "WissKI" tools for the semantic annotations using controlled vocabularies as well as formal ontologies. Their focus was mostly concerned with the recognition of events with the free text of documents.
Some previous work related to the digital cultural preservation was related to approached with the aim of proposing techniques for improving the retrieval, organization, and understanding of non-homogenous cultural knowledge through the cross-analyzing multi sources information [3].

3 Methodology

Motivated from the discussion in the previous sections, we have introduced an architecture which can handle the research problem addressed in the previous section. Figure 1 shows the flow of the components and the interaction between them. It also illustrates showing the input and output details of each of the three components. We shall discuss each one of them in detail in the following subsections.

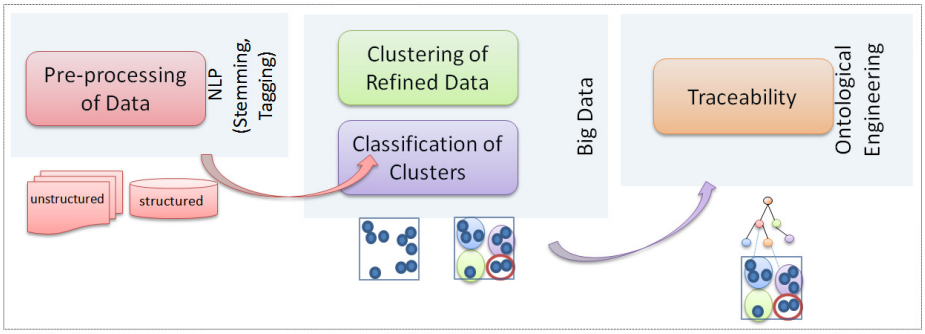

Fig. 1. Proposed architecture for the traceability of cultural heritage

3.1 Natural Language Processing

We collected data from the online resource (Museum of Archaeology and Anthropology, University of Pennsylvania) [11]. The original data consists of 346,474

object records. Although the data is in comma separated format, however, there is a problem with this large dataset. In many cases, the data is in semi-structured nature with various free text fields. This requires that we should perform natural language initial processing such as stemming, lemmatization, etc. However, the application of natural language processing cannot provide good results unless the semantic annotation is performed using the controlled vocabulary.

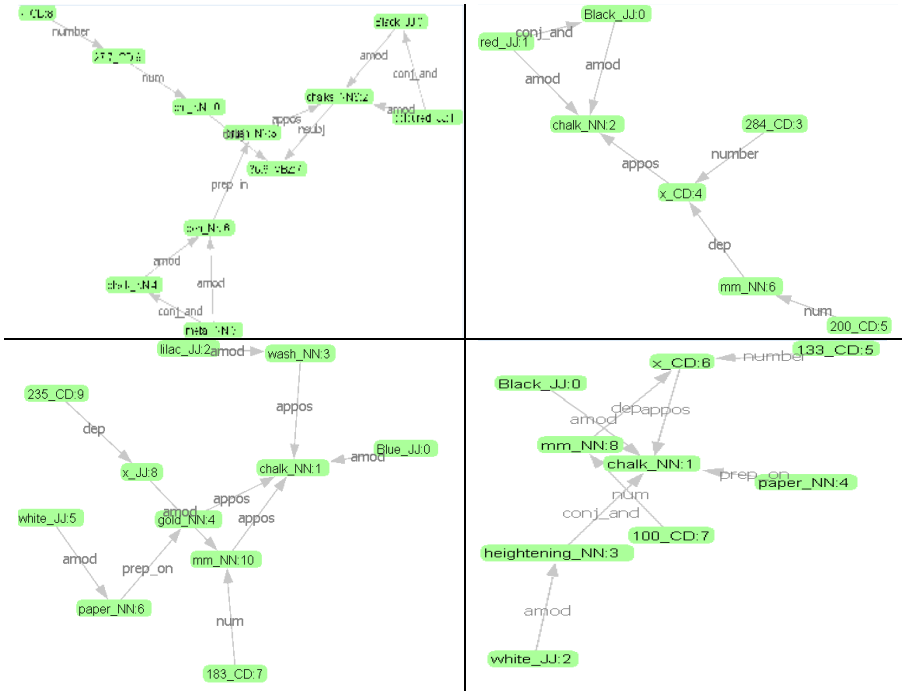

Fig. 2. A sample from the forest of semantic graph

In the column data, we faced problem such as "Tempera on wood" available in various dimensions. As the dimensions vary, it gives a unique state. The second problem was the hierarchy problem. "Oil on panel" and "Oil on canvas" both fall under the same category of Oil Painting. We need to provide the hierarchy in such a way that for grouping the hierarchy is useful, but for in depth tracing the more detail granularity is required. In the first part of this step, we take whole of the free text of the column, tagging it into pieces of Nouns, Adjectives, Adverbs and Verbs separated using NLP Library [12]. In the second part, we develop a forest of semantic network which connects all of these concepts realized into the previous part. This semantic network is used as a base of the controlled vocabulary. For example, Figure 2 is showing some of graphs in the forest in which the common technique is "chalk". This word is derived from graphs which have semantics relevant to this concept. Figure 1 is showing only four of such graphs; the conclusive concept out of these graphs is derived by the means of semantic graph matching. In the third part of this step, we perform the semantic similarities between graphs in the semantic forest. This process

is useful to identify all of the related concepts in a hierarchy. The outcome of this operation provides us a data by the means of which we have tuned up the granularity level of the distinct states in the column.

3.2 Clustering on Big Data

Previously, we discussed that the cultural heritage data is also increasing on the same pace as that of other domain of knowledge. It is a known fact that the extraction of significant structures out of arbitrary high dimensional data has always been a challenging task. Although stratified sampling also serves the purpose of preparing the samples for the input data. However stratified sampling is more or less a random grouping performed on the strata. The clustering technique employs the objective function and refines the data according to the objective function [13] [14]. In the case of cultural data, we are more focused on aggregating data. Hence, in this case the clustering is more helpful for applying our methodology. Clustering techniques capable of running on Hadoop platform can give the second level solution. We have employed the DBSCAN clustering algorithm using MapReduce approach [15], [16]. The added advantages of using DBSCAN over MapReduce are following:

1. DBSCAN possesses the capability to produce irregular shape clusters which are more close to the distribution nature of data.
2. MapReduce ensures the scalability of the executing functions. This aspect is useful in case of high dimensional data as described earlier in our case.

In DBSCAN, each object is clustered given two core parameters. Mathematically, we can define it by equation 1.

$$N_{\varepsilon}(p):\{q \mid d(p,q) \leq \varepsilon\} \tag{1}$$

Where N denotes the number of objects between two given objects p and q (both are inclusive); ε is radius. On each point scale (as for p in the equation 1), it gives a circular cluster. However as lots of circles individually grows up, collection of tiny circles (sub-clusters) are realized into a dense regions in the data space which is separated by regions of lower object density. This identifies a maximal set of density connected points. DBSCAN is useful for such situations where data is distributed in numerous small density zones. The cultural data by virtue of its nature bears density oriented distribution. The underlying reason is that given a specific type or format, specific cultural information are aligned in a peculiar way. One can notice that DBSCAN is sensitive to external parameters of radius size and number of observations. These parameters are usually found out by means of k^{th} nearest neighbor.

The MapReduce can be mathematically defined by the following definitions.

Definition 1: MapReduce MR is a function of three alternating function such that:

$$MR = f(M, \Lambda, \Omega) \tag{2}$$

where M is a mapper, Λ is a reducer and Ω is a sorting function.

Definition 2: Given a list of key value pairs $\left\langle k_i, v_i \right\rangle_{i=1}^{n}$ which are comprised of a string $v \in \{v_1, v_2, v_3, ...v_m\}$ composed of m number of features; The mapper can be defined as:

$$\left\langle k_i^{'}, \left\{ v_{i1}^{'}, v_{i2}^{'}, v_{i3}^{'}, ... \right\} \right\rangle \leftarrow M \left\langle k_i, v_{im} \right\rangle_{i=1}^{n} \quad (3)$$

Notice that the set mapped values for each key is unbounded. This set may be an empty set or a set with arbitrary length.

Definition 3: Given a list of key value pairs $\left\langle k_i^{'}, \left\{ v_{i1}^{'}, v_{i2}^{'}, v_{i3}^{'}, ... \right\} \right\rangle$; The reducer Λ generates the same key with a new set of value list. This value list is unbound; The equation 4 defines it as below.

$$\left\langle k_i^{'}, \left\{ v_{i1}^{''}, v_{i2}^{''}, v_{i3}^{''}, ... \right\} \right\rangle \leftarrow \Lambda \left\langle k_i^{'}, \left\{ v_{i1}^{'}, v_{i2}^{'}, v_{i3}^{'}, ... \right\} \right\rangle \quad (4)$$

The result of this step ends up in producing cluster samples which can be easily classified. Certainly the data has been reduced but still this is not converged enough to give a precise meaningful notion.

3.3 Classification on Cluster Objects

In the previous step, we obtained clustering samples. The clustering of sample serves to reduce data problem complexity by providing users with groups of similar entities. However, clusters are unable to highlight relationships among various features, especially in the case of data analysis on the high dimensional data sets. Therefore, we *classify cluster samples* in order to reduce a data set. This also bears special relevance in this case as we are motivated to reduce the data at the minimum loss of information. Generally, there are two types of classification (*Hard* and *Soft* classification) based on the type of class assignment to each of the clusters while grouping them according to their similar features. With the *Hard classification* mechanism, one can determine whether an instance can either be or not to be in a particular class. With the *Soft classification*, one can extrapolate whether an instance can be predicted to be in some specific class with some likelihood and often a probability distribution across all of the classes. We apply soft classification as it is suitable to have a probability distribution that depicts the level of confidence depending on the similarity of their features.

3.4 Ontological Modeling and Traceability

In the previous step, we performed classification of clusters which results into grouping the individual clusters characterized by their similarity features. However, the completeness of classification is still arguable; because this classification lacks semantics to be better queried and searched for the traceability of any cultural object. Therefore, we built a *Cultural Heritage Ontology (CHOnt)* that provides enriched

model for the inference of cultural objects as illustrated in Figure 3. Cultural Heritage includes tangible and intangible cultural objects which belong to societies or groups that are inherited from past generations.

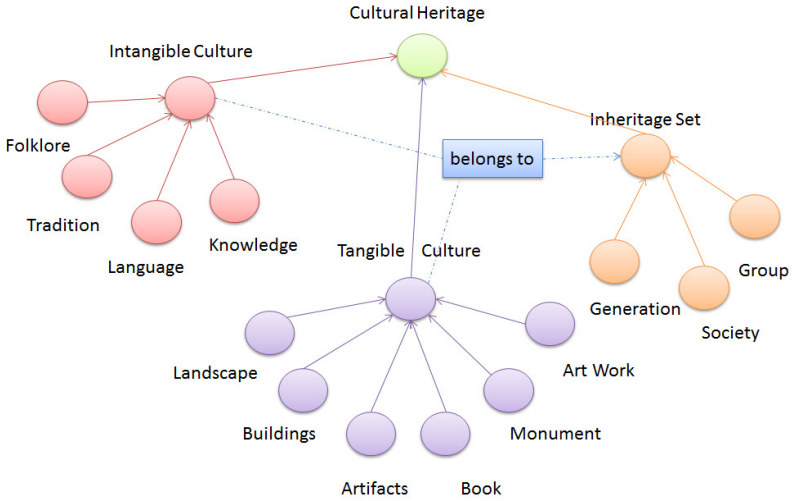

Fig. 3. Top level view of *Cultural Heritage Ontology (CHOnt)*

In *CHOnt* ontology, each cultural object that belongs to the cultural heritage museum is tracked by its *ID*, *Name* and *Description,* captured by the *Cultural-Object* concept. In addition, *Cultural-Object* concept is associated with the *Object_Property* via a *has_features* property (illustrated by Figure 4). Each object possesses several properties, such as *Form, Type, Period, School, and Technique*. Figure 4 below also illustrates the expanded view of *Type* and *Form* concepts. We have checked the consistency of this ontology by using Description Logic (DL) Reasoner to avoid any type of inconsistencies [16]. This ontological model is the fundamental building block for the traceability of the cultural heritage objects. When an anonymous object with certain known features arrives at a museum, the question for its traceability appears as a challenging task. We brought forward a user interface to describe its known features and provide some more descriptive information if available. This information is captured inside the *CHOnt*. Next step is to perform the *semantic matching making* to trace that object. For the match matching we are using our previously build ontology mapper *DKP-AOM* [17] to detect the aggregated semantic similarity between the anonymous object features and the underlying classified clusters of objects. DKP-AOM provides different strategies, such as string matching, synonym matching, etc. to find the level of confidence for traceability of cultural object. String matching strategy enables direct matching of anonymous object properties with the concepts of *CHOnt.* Synonym matching strategy uses WordNet lexical database to find all possible synonyms to able semantic traceability when different terminologies are used for the same type of object properties. Based on the level of confidence aggregated by different similarity measures, we trace cultural object and illustrate its likelihood based on its semantic similarity.

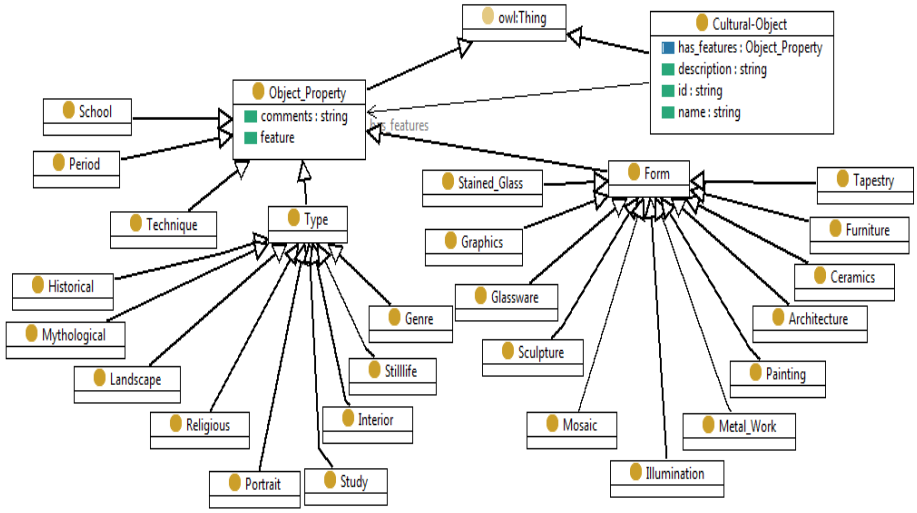

Fig. 4. *Cultural-Object in Cultural Heritage Ontology (CHOnt)*

4 Results and Discussion

In this section, we shall analyze the results obtained from the experiment executed on the proposed architecture (see Figure 1 for the detail). The cultural data usually contains a lot of free text. If we apply the machine learning techniques, the learning model found large number of distinct states. Table 1 is showing the weighted average result of Naïve Bayes classification obtained before and after application of forest of semantic network. Certainly, one can argue that reducing large number of unique states in feature variables as well as reducing classes benefit in reducing error in classification. However, at this point, the strength of the system appears when by means of ontological modeling and forest of semantic network can pin point the missing information accordingly.

Table 1. Performance of the Proposed Architecture

	TPR	FPR	Precision	Recall	F-Measure
Without NLP	0.479	0.063	0.457	0.479	0.46
With NLP	0.95	0.042	0.954	0.932	0.945

For example some techniques such as “Brush, brown ink and oil on blue primed paper”, “Ceiling painting in oil”, “Charcoal and oil on canvas”, “Charcoal and oil on cardboard”, “Charcoal and pastel on paper” etc. have conceptually alike (although not equivalent). These terms have semantic meanings. Unless, numerous concepts are not clustered, the classification is always prone to give poor results. There was significant

margin to increase the classification accuracy and the same was acquired by the proposed hybrid approach in this study. Figure 5 is providing a visualization which is a reflection of prime sections of the data. These include Form, Type, School, Period, and Technique. Every concept has a lot of instances. The challenge in the visualization is to organize the maximum information into a significant shortest description. A glance view over the figure informs about the possible inclusion of any unidentified object. However, still the problem is that this is a manual technique; this might be helpful to validate the ontological model but still we need an ontological based expert system to enable automatic traceability.

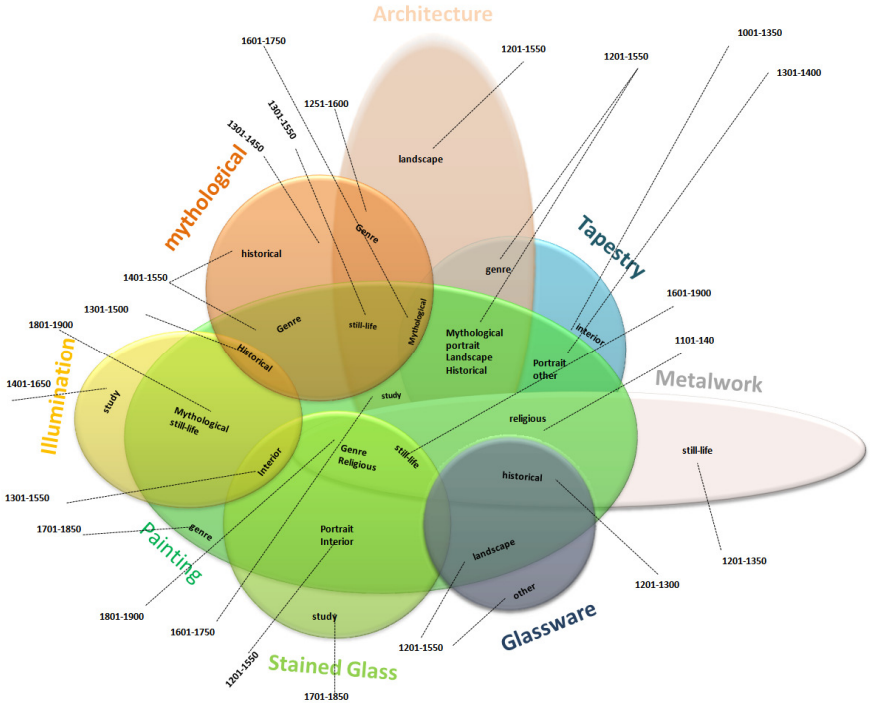

Fig. 5. Visualization of aggregated set of cultural heritage data

5 Conclusion

The study of culture heritage and its preservation is interesting as well as important for variety of domains including anthropology, psychology, archaeology, museology, sociology, communication, management and business. This research presents an expert system based on heterogeneous architecture with the purpose of traceability and estimation of missing information for a newly arrived artifact in a museum. The technique can be used to eliminate the risk of inclusion of possible inconsistencies, and preserve only significant concise information. We introduced the layout mechanism of the essential functionalities to validate the architecture and the

interoperability of various context aware technological modules. The proposed architecture can find answers to interesting research problems by modeling structured and unstructured data for the purpose of traceability.

References

[1] M. Naeem, N. Moalla, Y. Ouzrout, and A. Bouaras, "An ontology based digital preservation system for enterprise collaboration," in *Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference on*, 2014, pp. 691–698.

[2] T. W. M. Academy, "The World Museum Academy." 2014.

[3] C. Cesarano, A. Picariello, D. R. Recupero, and V. Subrahmanian, "The OASYS 2.0 Opinion Analysis System.," *ICWSM*, vol. 7, pp. 313–314, 2007.

[4] G. Goerz and M. Scholz, "Adaptation of nlp techniques to cultural heritage research and documentation," *CIT J. Comput. Inf. Technol.*, vol. 18, no. 4, pp. 317–324, 2010.

[5] L. Hardman, L. Aroyo, J. van Ossenbruggen, and E. Hyvonen, "Using AI to Access and Experience Cultural Heritage," *Intell. Syst. IEEE*, vol. 24, no. 2, pp. 23–25, 2009.

[6] NADB, "National Archeology DataBase." 2009.

[7] K. Delevan, "Library Research Guides. Finding Primary Sources. Visual Materials.," 2011.

[8] B. Museum, "Database of the British Museum." 2015.

[9] E. Hyvönen, "Semantic portals for cultural heritage," in *Handbook on ontologies*, Springer, 2009, pp. 757–778.

[10] O. Stock and M. Zancanaro, "Personalized active cultural heritage: The PEACH experience," *Handb. Res. Cult.-Aware Inf. Technol. Perspect. Models Perspect. Models*, p. 446, 2010.

[11] M. of A. and A. University of Pennsylvania, "University of Pennsylvania Museum of Archaeology and Anthropology." 17-May-2015.

[12] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.

[13] M. Naeem and S. Asghar, "KDSSF: A Graph Modeling Approach," *Int. J. Comput. Appl.*, vol. 33, no. 4, 2011.

[14] M. Naeem, S. Gillani, M. A. Qadir, and S. Asghar, "gSemSim: Semantic similarity measure for intra gene ontology terms," *Int. J. Inf. Technol. Comput. Sci. IJITCS*, vol. 5, no. 6, p. 32, 2013.

[15] H.-P. Kriegel and M. Pfeifle, "Density-based clustering of uncertain data," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005, pp. 672–677.

[16] L. A. Barroso, J. Clidaras, and U. Hölzle, "The datacenter as a computer: An introduction to the design of warehouse-scale machines," *Synth. Lect. Comput. Archit.*, vol. 8, no. 3, pp. 1–154, 2013.

[17] M. Fahad, N. Moalla, A. Bouras, M. A. Qadir, and M. Farukh, "Towards Classification of Web Ontologies for the Emerging Semantic Web.," *J UCS*, vol. 17, no. 7, pp. 1021–1042, 2011.