Knowledge-based access to art collections: the KIRA system

Flora Amato, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlì

DIETI - University of Naples "Federico II", via Claudio 21, 80125, Naples, Italy
CINI - ITEM National Lab, Via Cinthia, 80126, Naples, italy

{flora.amato,vmoscato,picus,giancarlo.sperli}@unina.it

Abstract. This discussion paper represents an extended abstract of a recent publication where we presented KIRA (*Knowledge-based Information Retrieval from Art collections*), a system to query, browse and analyze cultural digital contents from a set of distributed and heterogeneous repositories. KIRA relies on a Big Data infrastructure with the following features: capability to gather information from different data sources; advanced data management techniques and technologies; ability to provide useful and personalized data to users based on their preferences and context. KIRA thus provides retrieval and presentation functionalities to search information of interest and present it to the users in a suitable format and according to their needs. Using ad-hoc APIs, our system can also support several applications: mobile multimedia guides, web portals to promote the Cultural Heritage, multimedia recommender and storytelling systems and so on. We discuss the main ideas that characterize the system, showing its use for several applications.

Keywords: Knowledge Management, Big Data, Information Retrieval

1 Introduction

The enhancement and promotion of worldwide Cultural Heritage (CH) using Information and Communication Technologies (ICT) represents nowadays an important research issue, with a variety of potential applications. ICT have radically changed the modern CH scenery: simple traditional Information Systems for the management of cultural artifacts have left the place to complex systems that expose rich information extracted from heterogeneous data sources (e.g. Digital Libraries and Open Archives, Multimedia Art Collections, Social Media, Web Encyclopedias, etc.). A large number of proposals, which focus on how ICT solutions should be applied to the CH domain for different purposes, has been presented in the literature [1]. Indeed, several recent European projects (e.g. Ariadne, Europeana, etc.) have already suggested a set of methodologies/technologies together with the best ways and practices to manage and organize the cultural knowledge for different contexts and applications.

In spite of the great effort, some research problems have to be still faced during the design of a modern Cultural Heritage Information System, especially if we consider high change rate, large volume, and intrinsic heterogeneity of cultural data: i) the adoption of architectural models for *Big Data* management [2]; ii) the access, retrieval, integration and analysis of information from distributed and very heterogeneous art repositories [3]; iii) the transformation of the captured data into useful knowledge and the related management in according to the different "views" of a cultural item exploiting the LD/LOD (*Linked Data/Linked Open Data*) paradigm [4]; iv) the access to the knowledge based on the user profile and the *context*.

This paper represents an extended abstract of the work [5], where we describe KIRA (*Knowledge-based Information Retrieval from Art collections*), a system to query, browse and analyze cultural digital contents from a set of distributed and heterogeneous art repositories. In particular, the system prototype has been developed within the Cultural Heritage Information Systems (CHIS) National project, promoted by DATABENC[1]. KIRA is able to manage all the digital contents related to *Cultural Items*. More in details, in our vision each *Cultural Heritage environment* (e.g. museums, archaeological sites, old town centers, etc.) is grounded on a set of cultural *Points of Interest* (PoI), which correspond to one or more cultural items (e.g. specific ruins of an archaeological site, sculptures and/or pictures exhibited within a museum, historical buildings and famous squares in an old town center and so on). In order to meet variety, velocity and volume of the managed information, KIRA is characterized by the following technical features that are typical of a *Big Data* platform:

- capability to gather information from distributed and heterogeneous data sources (e.g. Social Media , Digital Libraries and Open Archives, Multimedia Collections, Web Encyclopedias, Web Data Services, etc.);
- advanced data management techniques and technologies;
- advanced information retrieval services and ability to provide useful and personalized data to users based on their preferences and context.

The paper is organized as follows. Section 2 describes the proposed data model. Section 3 presents a system description with several implementation details. Section 4 reports a possible application of our system and discusses some conclusions and the future work.

2 Data Model for Cultural Items

The introduced data model relies on the concept of "Cultural Item" (CI): examples of CIs are specific ruins of an archaeological site, sculptures and/or pictures exhibited within a museum, historical buildings in an old town center and so on.

[1] The High Technology District for Cultural Heritage (DATABENC) management of the Campania Region, in Italy (www.databenc.it).

In the CH domain, a CI can be opportunely described with respect to a variety of annotation schemata, for example the archaeological view, the architectural perspective, the archivist vision, the historical background, etc., which usually exploit different sets of "metadata" and possibly domain taxonomies or ontologies [6]. In a simplified way, we consider a *ontology* $O = (V, E)$ as a network of concepts belonging to the CH domain, where a node $v \in V$ represents a "concept" and an edge $e \in E$ a relationship between two concepts. Thus, we define an *annotation schema* and a *semantic annotation* for a CI.

Definition 1 (Annotation Schema). *Given a set of ontologies $\mathcal{O}$, an* Annotation Schema *is a tuple $\lambda_{\mathcal{O}} = (A_1, \ldots, A_n, B_1, \ldots, B_m)$, where $A_1, \ldots, A_n$ are attributes for which $\forall i \in [1, n], \exists O = (V, E) \in \mathcal{O}$ s.t. $dom(A_i) \subseteq V$, and $B_1, \ldots, B_m$ are attributes for which $\forall j \in [1, m], \nexists O = (V, E) \in \mathcal{O}$ s.t. $dom(B_j) \subseteq V$.*

The attributes $A_1, \ldots, An$ are *Ontological Attributes* (OAs) and correspond to concepts that are relevant for the specific domain(s) being modeled. In turn, *Non-Ontological Attributes* $B_1, \ldots, B_m$ (NOAs) can contain other useful information, such as multimedia items (e.g. audio, video, images, texts and 3D models, etc.) characterized by a set of low-level features and other metadata. In particular, we can adopt both "literals" or a set of URIs (*Uniform Resource Identifiers*), which allow to access the related cultural information according to the LD/LOD paradigms, as values of the annotation attributes. In addition, a CI may be associated with a specific "Point of Interest" (POI), defined by a set of geographic coordinates, and corresponding either to a single point or to a set of lines and more complex polygons of the considered environment.

Definition 2 (Semantic Annotation). *Given a set of ontologies $\mathcal{O}$, an annotation schema $\lambda_{\mathcal{O}}$ and cultural item* CI*, a* Semantic Annotation *of* CI *is a tuple $\lambda_{\mathcal{O}}(\text{CI}) = (a_1, \ldots, a_n, b_1, \ldots, b_m)$, where $\forall i \in [1, n], a_i \in dom(A_i)$ and $\forall j \in [1, m], b_i \in dom(B_i)$*

Using various sets of ontologies and semantic annotations [7], we can thus describe a cultural item from different points of view supporting several applications. A large set of relationships can also be instantiated among cultural items and the entire system *Knowledge Base* (KB) can be modeled as a *graph*[2].

Definition 3 (Knowledge Base). *The* Knowledge Base *is a graph $G = (C, R)$: each node $c \in C$ can be a cultural item or an ontological attribute (concept) [8], while each edge $r \in R$ represents a relationship derived from a semantic annotation or established between two cultural items.*

Figure 1 shows how a portion of knowledge related to the Paestum ruins can be easily represented in our model.

[2] All possible relationships in the model are opportunely defined "a-priori" and the related meaning can be found in a proper thesaurus.

Leveraging different annotation schemes and ontologies, our model allows achieving interoperability goals. The KB content can be easily exported in the most used formats (e.g. XML, RDF, OWL) and according to the most diffused harvesting standards for CH applications (e.g., EDM, Italian ICCD, etc.). On the other hand, the LD/LOD paradigm permits us to deal with several problems related to data consistency and copyright constraints[3].

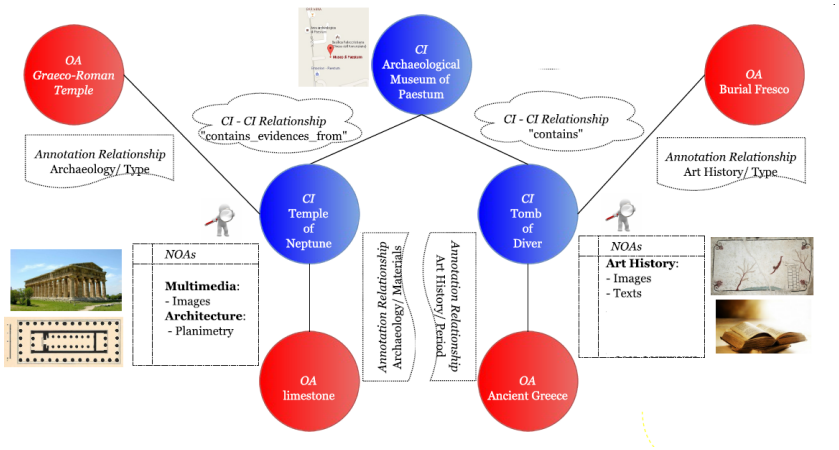

Fig. 1: KIRA - Data Model for CIs

3 System Description

3.1 Architecture

KIRA has to deal with the large and heterogeneous amount of information: annotations and descriptions provided by cultural heritage foundations, web encyclopedias and open archives, multimedia contents coming from social media networks and digital libraries, opinions and comments of users from common online social networks, etc. For this reason, KIRA presents a layered architecture typical of a Big Data platform [2], exploiting the related stack of technologies (see [5] for more details). In the *data source layer*, each data source is properly "wrapped" in order to extract the information of interest that is then represented as required by the described data model. In particular, each Wrapper is specialized for a particular kind of source (i.e. Social Media Networks, Digital Repositories, Open Archives) and must address all the interoperability issues, providing a

[3] Some cultural items descriptions are accessible only using URI, thus the data management issues are in charge to the related source.

set of functionalities to access data sources and gather all the desired data, possibly leveraging the available APIs[4]. In the *data storage and management layer*, data are stored in the Knowledge Base in compliance with the above-described data model, and managed also exploiting the LD/LOD paradigm. In addition, specific semantics to be attached to the data is provided using the annotation schemes, including ontologies, vocabularies, taxonomies, etc. related to the Cultural Heritage domain. The KB leverages different Data Repositories realized by advanced data management technologies (e.g. Distributed File Systems, NoSQL and relational systems) and provides a set of basic APIs to read/write data by an Access Method Manager. As a basis for the *data processing layer* our system provides a Query Engine that can be invoked by user applications to search data of interest using information retrieval facilities. In particular, our system supports all the basic functionalities for multimedia and semantic information retrieval by means of proper Information Filters. The *data analytics layer* is based on different Analytics Services allowing to create personalized "dashboards" for a given cultural environment. In addition, it provides basic data mining, graph analysis and machine learning algorithms useful to infer new knowledge and provided mechanisms for personalized and *context-aware* access to data.

3.2 Functionalities and Implementation Details

One of the most important functionalities provided by KIRA consists in the capability of gathering the different kinds of data from different sources: User Data, Social Data, Digital Repository and Multimedia Data.

User Data basically include preferences and needs that are useful to define the related *profiles*: data on users (e.g. favorite artistic genre and artists) constitutes the *Personally Identifiable Information* (PII) that is stored in the Knowledge Base and can be used as additional filter in the retrieval. As to *Social Data*, the current prototype only considers information coming from Twitter. In particular, KIRA retrieves user comments and posts information about a given CI by exploiting the related APIs. Social data can be used in several applications requiring a "social vision" of cultural items. We collect *Digital Repositories' Data*, information describing cultural items from on-line digital repositories (e.g. museums, libraries, open archives, multimedia collections, etc.). The wrapper for this kind of sources can import such data descriptions and convert them into a JSON format. *Multimedia data* (e.g. images, texts, video, etc.) related to a given cultural item can be similarly collected using the wrapper facilities. In particular, the descriptions in terms of basic metadata are captured and stored within

[4] Data integration problems for heterogeneous data sources are addressed by means of classical schema mapping techniques, record linkage and data fusion techniques [3], according to the specific data source. Eventually, data stream management problems have to be considered.

KIRA, while raw multimedia data can be opportunely linked and, in other cases, temporarily imported into the system for content-based analysis [9][5].

The data gathered by the Wrappers are then stored and managed by the Knowledge Base. One of the basic functionalities of the KB is to export the related content into the *Europeana Data Model*[6] (EDM) format (see [5] for more details). Metadata semantics is provided by the set of annotation schemes (in XML, RDF or OWL formats). All the data can be represented as sequences of triples ($\langle$ subject, predicate, object $\rangle$) in according to the described data model. The KB is based on several technologies that are briefly described in the following[7].

The data describing basic properties of CIs (e.g. name, short description, etc.) and basic information on users profiles are stored into a *key-value data store* (i.e. *Redis*). The complete description in terms of all the metadata of CIs using the different annotation schemes are in turn saved using a *wide column data store* (i.e. *Cassandra*). We use a table for each kind of CIs having a column for each "metadata family"; column values can be literals or URIs. The *document store* technology (i.e. *MongoDB*) is used to deal with JSON messages, complete user profiles and descriptions of internal resources (multimedia data and textual documents, etc.) associated with a cultural item. All the relationships among cultural items within a cultural environment and interactions with users (behaviors) are managed by means of a *graph database* (i.e. *Titan*). The entire cartography related to a cultural environment together with POIs is managed by a GIS (i.e. *PostGIS*), which provides the functionalities to filter and visualize on a map the geographic area around a given $\mathcal{PoI}$. Multimedia data management is realized using the *Windsurf* library [9]. We exploit an *RDF store* (i.e. different *Allegrograph* instances) to memorize data views in terms of triples related to a given cultural environment and useful for specific applications, providing a SPARQL endpoint for the applications. All system configuration parameters, internal catalogs and thesauri are stored in a relational database (i.e. *PostegreSQL*). Finally, semantics of data can be specified by linking values of high-level metadata to some available internal (managed by Sesame) or external ontological schemes.

This heterogeneous Knowledge Base provides basic *Restful APIs* to read-/write data and further functionalities for importing/exporting data in the most common diffused Web standards. The search of data useful for the applications can be eased by using different information filters that implement the right queries to the various databases. The implementation of such filters is based

[5] Note that multimedia data that are managed by the system are suitably filtered before the storing process. The number and kinds of multimedia data required by the application are tuned by means of configuration parameters.

[6] http://pro.europeana.eu/edm-documentation

[7] We chose to adopt such heterogeneous technologies in order to meet the specific requirements of the applications dealing with the huge amount of data at stake. For example, Social Networking applications typically benefit of graph database technologies because of their focus on data relationships. In turn, more efficient technologies (key-value ore wide-column stores) are required by Tourism applications to quickly and easily access the data of interest.

on *Apache Spark*, and we distinguish four kinds of query on the KB: i) *query by keywords/tags*: through such a query a user/application can search a set of CIs using keywords (as in Google search engine) or specific tags (the query is then "expanded" with similar search terms leveraging the system thesauri); ii) *query by metadata*: by this query a user/application can search a set of CIs using specific metadata of internal or external annotation schemes (the query for OAs is semantically "expanded" with the concepts of managed ontologies that are similar to the target one); iii) *query by example*: through such a query a user/application can search a set of multimedia contents – related to CIs – that are similar to a target one (the query processing is base on the *Windsurf multimedia libraries*); iv) *query by user preferences*: user profiles are exploited to find the set of cultural items that are more similar to user preferences using co-clustering techniques [10].

4 System Running Example and Conclusions

We describe a possible application of our system to support the development of a multimedia guide for the *Paestum archeological site*. The ancient buildings, together with the museum and its main artifacts, constitute the set of cultural items for our case study. Tourists, both from their places and while visiting ruins, can browse these cultural items and enjoy a useful multimedia guide describing them, or be recommended other nearby places, comments of other users and other information of interest. When users search a specific cultural item, as

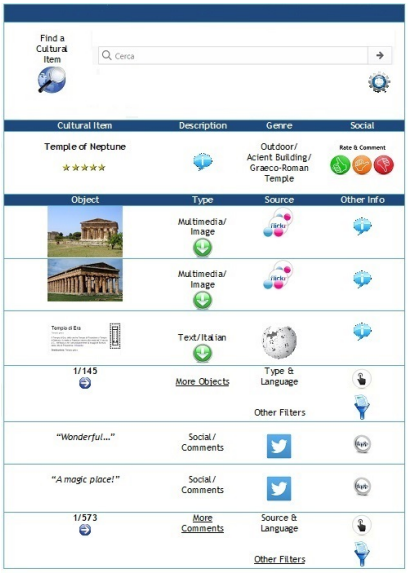

Fig. 2: KIRA - System Query Interface

an example the Temple of Neptune, our system provides a basic description with the related multimedia objects (i.e. audio, images, video and texts) and detailed users' comments. The list of proposed cultural items with descriptions and multimedia objects depends on the user's preferences and system settings: images rather than voice, or expert-level rather than layman-level descriptions of the art pieces; specific metadata and annotation schemes. In addition, *query by example* facilities can be exploited to determine other images that are similar to a given multimedia object. In addition, a *semantic search* can be performed on specific ontological attributes to find other cultural items of the same type. At the same time, users can choose to retrieve some interesting information, to read comments, opinions and ratings about the visited cultural items and to express their own ratings and opinions. Figure 2 shows a running example (obtained by assembling different screenshots) concerning the search of $\mathcal{CI}$s related to the Paestum ruins. Users can browse the data by means of an appropriate GUI; they can filter objects belonging to a given $\mathcal{CI}$ using different criteria: type of multimedia data, language, etc. Future work will be devoted to collect the huge amount of data related to all the different cultural objects of the Campania region and to experiment our system from the efficiency and effectiveness points of view with respect to the information retrieval and filtering tasks providing a comparison with other systems.

References

1. C. Guccio, M. F. Martorana, I. Mazza, and I. Rizzo, "Technology and public access to cultural heritage: the italian experience on ict for public historical archives," in *Cultural Heritage in a Changing World*. Springer, 2016, pp. 55–75.
2. C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.
3. X. L. Dong and D. Srivastava, "Big data integration," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*. IEEE, 2013, pp. 1245–1248.
4. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web*, vol. 7, no. 1, pp. 63–93, 2016.
5. F. Amato, V. Moscato, A. Picariello, and G. Sperlì, "Kira: a system for knowledge-based access to multimedia art collectionsl," in *Proceedings of the IEEE 11th International Conference on Semantic Computing, (ICSC 2017)*. IEEE, 2017.
6. M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. van de Sompel, "The europeana data model (edm)," in *World Library and Information Congress: 76th IFLA general conference and assembly*, 2010, pp. 10–15.
7. F. Amato, A. De Santo, V. Moscato, F. Persia, A. Picariello, and S. Poccia, "Partitioning of ontologies driven by a structure-based approach," 2015, pp. 320–323.
8. F. Amato, A. Mazzeo, A. Penta, and A. Picariello, "Using nlp and ontologies for notary document management systems," 2008, pp. 67–71.
9. I. Bartolini and M. Patella, "Multimedia queries in digital libraries," in *Data Management in Pervasive Systems*. Springer, 2015, pp. 311–325.
10. I. Bartolini, V. Moscato, R. Pensa, A. Penta, A. Picariello, C. Sansone, and M. L. Sapino, "Recommending multimedia visiting paths in cultural heritage applications," *Multimedia Tools and Applications*, vol. 75, no. 7, pp. 3813–3842, 2016.