AN INFORMATION SYSTEM TO ACCESS CONTEMPORARY ARCHIVES OF ART: CAVALCASELLE, VENTURI, OJETTI, ARGAN, BRANDI

Irene Calloud[1], Andrea Ferracani[2], Vincenzo Lepera[2], Giuseppe Serra[2]

[1]Fondazione Memofonte,
Studio per l'elaborazione informatica delle
fonti storico-artistiche,
Florence, Italy
http://www.memofonte.it/

[2]Media Integration and Communication Center,
University of Florence
Florence, Italy
http://www.micc.unifi.it/

Abstract – The research project aims at the development of a digital library for handwritten documents of artistic and literary culture in XIX and XX centuries. The goal is to provide access to contemporary archives of documents related to well-known art historians: Giovan Battista Cavalcaselle, Adolfo Venturi, Ugo Ojetti, Giulio Carlo Argan and Cesare Brandi. All the sources, constituted by unedited and edited archival documents, are stored in a relational database mapped to an ontology. The system has been designed and implemented by Media Integration and Communication Center, Fondazione Memofonte, Scuola Normale Superiore of Pisa, University of Florence and University of Udine.

INTRODUCTION

The project aims to provide an innovative search and browsing system for textual resources, handwritten or printed, connected to the main figures of the Italian historiography and art criticism in XIX and XX centuries. The research focus is on a group of documents, written by Giovan Battista Cavalcaselle (1819-1897), Adolfo Venturi (1856-1941), Ugo Ojetti (1871-1946), Giulio Carlo Argan (1909-1992) and Cesare Brandi (1906-1988), that can be investigated and analyzed to reveal the relationships between art critics, intellectuals, artists and public during those years.

Funded by MIUR (Ministero dell'Istruzione, dell'Università e della Ricerca), the Project (2009-2011) is developed by MICC together with the Fondazione Memofonte onlus (project coordinator), the Scuola Normale Superiore di Pisa, the University of Florence and the University of Udine [1].

The professional background and the experience of the research units in digital technologies for cultural heritage led to the definition of a method to analyze the huge amount of heterogeneous sources as whole (manuscripts, correspondences, bibliographic references, historic travel notes etc.), and each one document for its own peculiarities. Searching and browsing these documents it will be possible to understand the authors' interest for contemporary artists. All the units have contributed to the definition of methodologies and criteria for critical edition and for the creation of lexicon.

The activities have been developed and distributed amongst partners as follows: the Fondazione Memofonte [2], thanks to the specific knowledge in on-line publication of rare and inedited textual and figurative sources in the artistic and literary historiography, is responsible for the management of the Ugo Ojetti materials (Biblioteca Nazionale Centrale di Firenze). The operation will improve the fruition of the documents, considering the fact that Ojetti's material in the National Library have not been yet organized in a systematic way. The main goal is the creation of a wide archive of all the art exhibitions, national and international,

curated by Ojetti, to the purpose of discovering the interrelations between the author and the contemporary artistic panorama.

The analysis of the University of Udine [3] focuses on how to better evaluate the so far neglected textual and literary aspects of two important historical productions which are important in the Italian art criticism: the documents by G.B. Cavalcaselle (Biblioteca Marciana di Venezia), specially focusing on drafts and notes for the edition of "History of Painting in North Italy", and the letters and manuscripts by A. Venturi (held at Scuola Normale di Pisa), studied in order to understand the genesis and the change of style and descriptive techniques of the "Storia dell'arte italiana".

The research unit of the University of Florence [4] conducts a deep archive research for the rearrangement and cataloguing of the Giulio Carlo Argan's private archive in Rome. Moreover, the transcription and the metadata annotation of the documents will permit the analysis of the resources pertaining to the scientific portrait of the academic. On the other hand, this work will also shed light on the complex artistic, political, social plots of the Italian twentieth-century that emerge strongly between the lines of this correspondence, in which appear the names of major artists, critics and intellectuals of the century.

Finally, the Scuola Normale Superiore of Pisa [5], works on Cesare Brandi's edited production. Its articles on journals and newspapers allow to draw up its initiatives for divulging and organizing art exhibitions. Using the browsing and search system, the unit develops studies on linguistic and lexical interpretation of Brandi's works, especially focusing on the creation of lexical indexes, covering various typologies of writings, both edited and inedited.

The project aims mainly at finding in the archives material still largely inedited, like documents, letters, drafts, different versions of essays, papers and notes. In the meanwhile, the archival research activity provides an advancement in the rearrangement, inventory and cataloguing of archives themselves. The search on the documents provides the users an overall view of the popularity of artists and artworks in Italy and abroad and facilitates the identification of the relationships between art commerce, collecting and exhibitions in private setting or in public museums. It also permits to understand how art historians and critics managed to publish their essays or pamphlet in journals, newspapers or in other editing initiatives.

Resources is stored in the database indexing the records on the basis of the cataloguing standard "Union List of Artist Names", issued by Getty Institute [6], which seems to be the most suitable. Events (exhibitions and travels) and sources (manuscripts, correspondence, bibliographic references, historic notebook etc.) are the main concepts outlined in the digitized documents. Controlled vocabularies have been included for many cataloguing fields, selecting entries from national or international vocabularies and thesauri, or adding specific entries.

THE SYSTEM

The proposed system is a web application, based on Model View Controller architecture, and has been completely developed by MICC [7] using the Symfony framework [8]. Currently the application is implemented in an experimental version and is under test by the institutes belonging to the consortium.

The system is mainly composed by two components: the content management system and the search engine. The homepage of the web application presents the artists with an essential biography, a general bibliography and an image gallery. The figure 1 shows the homepage of the application.

Figura 1 – Home page of the system.

After identification, the user can access the application backend for insertion / modification of the data. Sources and events are the two main categories the documents belonging to the archive can be classified in. Sources are original documents, usually manuscripts, for which it is possible to a identify producer, a sender, a date, a content and the references. Events consist of documents relating to exhibitions that are directly related to the artists and the sources. For these can be specified the name, the place, a start and end date, the curator and committee.

Furthermore, the application allows to include multimedia objects, such as images, for both the sources (e.g. an image from a notebook) and events (e.g. a poster for an event).

The query system has been implemented not only as a general purpose search engine but also to meet the particular needs of researchers in art. For this reason, it is composed by two components: a full-text search and a advance search.

The full-text search allows the user to search using keywords. The engine indexing of the data is based on Lucene [9]. Lucene is a free/open source information retrieval software library extremely flexible and adaptable to every need of search. In fact, at the core of Lucene's logical architecture is the idea of a document containing fields of text. This flexibility allows Lucene's API to be independent from the file format.

The interface for full-text search allows the user to select one or more archives that constitutes the data model and provides a simple 'Google like' search. More complex queries can be composed using Boolean operators (AND, OR, NOT) and the wildcard ? (one character) and * (*n* characters) operators.

The advanced search provides some additional capabilities to perform queries adding a set of specific filters for each category (sources and events) and also relating the different typologies of documents (sources and events in relation).

The search by 'source' allows to filter the documents with respect to the archive, date, typology, producer, sender, summary, transcription and title; whereas, the search by 'event' allows to filter with respect to the archive, data, typology, name and place.

Finally, the search by 'sources and events in relation' allows to perform cross searches between sources and events. Thus, the user can search, for example, all the letters related with a specific event. In particular, the filters are the union of the filters of the two components defined above.

Archivio informatizzato delle testimonianze di cultura artistica e letteraria in fondi manoscritti tra Ottocento e Novecento: **da Cavalcaselle a Brandi**

B A C K E N D

Home   Lista archivi   Ricerca   Profilo   Logout

Ricerca Libera

archivio:
☑ Tutti gli Archivi   ☐ Argan   ☐ Brandi   ☐ Cavalcaselle - Venturi   ☐ Ojetti

Testo: Dante   [cerca]

la ricerca ha restituito 23 fonti e 2 eventi

Archivio	Fonti	Eventi
Argan	Biglietto (1) Lettera (6)	
Brandi	Catalogo (1) Prefazione (1)	Mostra (1)
Cavalcaselle-Venturi	Taccuino (1)	
Ojetti	Articolo (10) Lettera (2) Scritto (1)	Mostra (1)

Figura 2 - The full-text search interface.

Semantic annotation and visualization

The system provides an automatically semantic annotation of the documents. This feature is essential in order to give the cataloguer a cloud of suggestions from which to choose while performing the categorization of the texts, task extremely time consuming for sources such as letters or notebooks that can be very long. Text mining algorithm aims to provide and to extract some meaningful representation of the semantic content of the knowledge base.

Not all words have the same importance, of course, and the frequency is not the only factor determining the weight of a term in a text. Even the words that occur once may be very important. Many of the most frequent words are 'stop words' such as for example, "e", "di", "da", "il" (corresponding to "and", "of", "from", "the" English respectively).

For this reason, in the first step of the algorithm removes such words. Then follows the stemming operation. This process allows the reduction of the inflected form of a word to its root form, called the theme. The theme does not necessarily correspond to the morphological root (lemma) of the word. It is usually sufficient to map related words to the same theme.

For example "manoscritto", "manoscritti" ("manuscript", "manuscripts", in English) map to the theme "manoscritt". We use the TreeTagger tool [10] for grammatical analysis to the text. The tool is able to distinguish nouns, names, adjectives and verbs, so the algorithm proceeds to discard the verbs, no useful in the annotation process.

Once completed the procedure of removing stop words and stemming, the algorithm extracts the keywords using a corpus based approach known as Term Frequency - Inverse

Document Frequency (TF-IDF). This approach allows to calculate a weight for each word in the document as follows:

- highest when term occurs many times within a small number of documents (thus lending high discriminating power to those documents);
- lower when the term occurs fewer times in a document or occurs in many documents (thus offering a less pronounced relevance signal);
- lowest when the term occurs in virtually all documents.

The words classified as having an high weight by the system are identified as keywords and proposed to the cataloguer on the interface. These metadata and keywords selected and added by the annotators are stored in a relational database. A program that performs the mapping of the database to an ontology is then used to create a knowledge base that describes the semantics of the documents. An ontology consists of concepts, concept proprieties, and their interactions to provide a formal description of a domain and provides a shared vocabulary that overcome the heterogeneity of information [11]. Several tools [12,13,14] have been developed to make the data contained in relational databases accessible as virtual RDF ontology graphs that can be navigated or queried by SPARQL endpoints.

The system uses the D2RQ tool [12] to expose database data as a semantic knowledge base. It provides a mechanism for mapping the tables and columns of a relational database to the classes and properties of an ontology. Here an example from the mapping file used in the system:

```
# Table evento                                map:evento_nome a d2rq:PropertyBridge;
map:Evento a d2rq:ClassMap;                       d2rq:belongsToClassMap map:Evento;
    d2rq:dataStorage map:database;                d2rq:property iswc:eventTitle;
    d2rq:uriPattern "evento/@@evento.oggetto_id@@";   d2rq:property rdfs:label;
    d2rq:class iswc:Event;                        d2rq:column "evento.nome";
    d2rq:classDefinitionLabel "evento";           d2rq:propertyDefinitionLabel "label";
    d2rq:classDefinitionComment "Un evento"
```

The example shows a d2rq:ClassMap instance with the URI map:Evento that contains information used by D2RQ to map the 'event' table from the database to a class in the ontology and a label property mapped to the column 'evento.nome' of the database.

Through this mechanism semantic queries can be performed on the system and output responses can be formatted as JSON data. An advanced visualization of the knowledge base has been integrated in the system using Simile Exhibit [15], a JavaScript framework that enables a data set to be visualized, sliced and embedded in a web page. It provides several views such as basic table, timeline, map, gallery. Figure 3 shows an example of visualization with a map, a timeline and a faceted browsing menu.

CONCLUSIONS

In this paper we present an innovative search and browsing web system for handwritten documents of artistic and literary culture in XIX and XX centuries written by well-known art historians: Giovan Battista Cavalcaselle, Adolfo Venturi, Ugo Ojetti, Giulio Carlo Argan and Cesare Brandi. The system provides a traditional search engine combined with semantic features performing automatic keywords extraction, semantic annotation andviews for advanced geolocalized and time based visualizations.

References

[1] Partners of the project: Fondazione Memofonte. Studio per l'elaborazione informatica delle fonti storico-artistiche (coordinator); Scuola Normale Superiore di Pisa, Laboratorio Lartte; Università degli Studi di Firenze, Facoltà di Lettere e Filosofia, Dipartimento di Storia delle Arti e dello Spettacolo; Università degli Studi di Firenze, Facoltà di Ingegneria, Centro per la Comunicazione e Integrazione dei Media (MICC); Università degli Studi di Udine, Dipartimento di Storia e Tutela dei Beni Culturali.
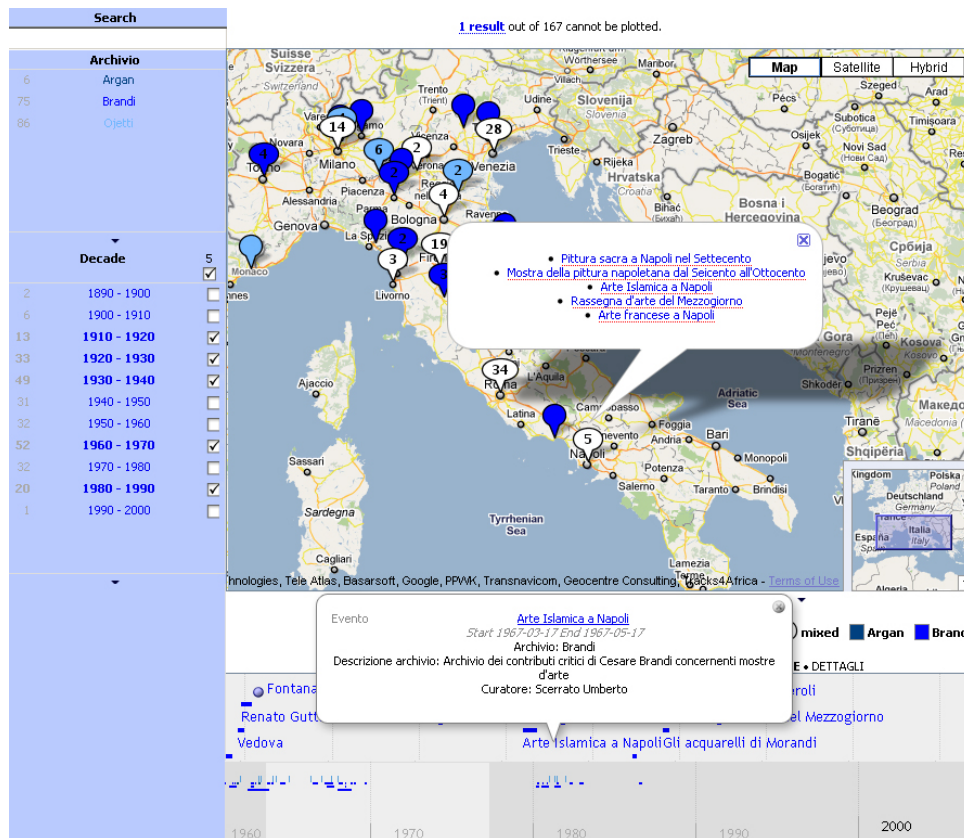
Figura 3 - Advanced visualization of the geolocated events and timeline.

Associated investigators: Alberto Del Bimbo (Micc, Univ. Firenze), Paola Barocchi and Miriam Fileti Mazza (Fondazione Memofonte), Benedetto Benedetti (SNS), Antonio Pinelli (Univ. Firenze), Donata Levi (Univ. Udine). Researchers: Paola Bonani, Irene Buonazia, Irene Calloud, Martina Dei, Annamaria De Santis, Andrea Ferracani, Claudio Gamba, Nadia Marchioni, Giorgia Marotta, Elena Miraglio, Emanuele Pellegrini, Katiuscia Quinci, Giuseppe Serra

[2] Associated investigators: Paola Barocchi, Miriam Fileti Mazza; Researchers: Irene Calloud, Martina Dei, Elena Miraglio

[3] Associated investigator: Donata Levi, Emanuele Pellegrini

[4] Associated investigator: Antonio Pinelli; Researchers: Paola Bonani, Claudio Gamba, Nadia Marchioni Katiuscia Quinci

[5] Associated investigator: Benedetto Benedetti; Researchers: Irene Buonazia, Annamaria De Santis, Giorgia Marotta

[6] ULAN, http://www.getty.edu/research/tools/vocabularies/ulan/index.html.

[7] Associated investigator: Alberto Del Bimbo; Researchers: Andrea Ferracani, Vincenzo Lepera, Giuseppe Serra

[8] Symfony, web PHP Framework, http://www.symfony-project.org/

[9] M. McCandless and E. Hatcher and Otis Gospodnetić, "Lucene in Action", Manning Publications, 2011.

[10] Helmut Schmid, "Probabilistic Part-of-Speech tagging using decision trees". In *Proc. of the International Conference on New Methods in Language Processing*, pages, 1994.

[11] T. Gruber, "Principles for the design of ontologies used for knowledge sharing," *Intational Journal of Human-Computer Studies*, vol. 43, no. 5-6, pp. 907–928, 1995.

[12] D2RQ, http://www4.wiwiss.fu-berlin.de/bizer/d2rq/

[13] SquirrelRDF, http://jena.sourceforge.net/SquirrelRDF/

[14] OpenLink Virtuoso, http://virtuoso.openlinksw.com/

[15] SIMILE: http://simile.mit.edu/