RESEARCH PAPERS OF THE WROCŁAW UNIVERSITY OF ECONOMICS
No. 25 2008
Knowledge Acquisition and Management

Stefan du Château, Danielle Boulanger, Eunika Mercier-Laurent

MODEME, IAE Research Center Lyon University, Lyon, France
stefan.du-chateau@orange.fr; db@univ-lyon3.fr; eunika@innovation3D.fr

KNOWLEDGE MANAGEMENT SYSTEM FOR CULTURAL PATRIMONY INCLUDING VOICE INTERFACE FOR KNOWLEDGE ACQUISITION

Abstract: This paper presents our research on a first part of hybrid Knowledge Management System for Cultural Patrimony – knowledge acquisition and ontological indexing. It describes our experimentation of a voice interface for the on field inventory of cultural heritage. This acquisition system includes signal processing, natural language techniques and knowledge modeling tools for future retrieval. We discuss the first results and raise some challenges of future work.

Key words: cultural heritage, voice interface, knowledge acquisition, natural language processing, ontology.

1. Introduction

The inventory of the cultural heritage is usual practice in the research on historical cities. Before going to specific cities or villages inventory researchers conduct a preliminary study about a given place or topic. Once at place they collect the information available writing, drawing, making plans, taking pictures or videos. The study documents and the collected information are registered in a data base which could be general or personal. When back to the office a researcher can improve the gathered information on a given object or add some elements from archives to update the content of the data base. This base can contain a description of the masterpieces still existing, preserved as vestiges, destroyed or disappeared but known through documents [Verdier et al. 1999].

All categories of masterpieces are concerned, such as religious, civil, military, in a perimeter as large as are the human activities. Each masterpiece has its own spatio-temporal context – history, past and present, it can be moved from one historical context to the other or can be modified. This kind of information and related knowledge is impossible to represent just in a classic data base.

Collecting of information in form of paper files and transferring them into a laptop is laborious and time consuming. The amount of the completed and corrected information is still very large and heterogeneous. Our objective is to design and experiment a new collecting support system to help the cultural heritage inventory researchers to perform their work better and in a more efficient way. It is also to help indexing and retrieving information and knowledge on a given masterpiece and its context. This paper describes our work on such a hybrid system using a voice interface (signal processing), natural language processing and knowledge modeling for the information gathering, management and retrieval. Section 2 presents relative work, section 3 describes our experiments and in section 4 we discuss directions for future work.

2. State of the art knowledge modeling for cultural heritage

This section presents some relative research work in the area of cultural heritage inventory and in speech recognition and "translation" of audio records into knowledge models.

2.1. The inventory of the cultural heritage

Main projects in the field of the inventory of the cultural heritage use data bases. Each base is conceived for a specific application and is not easily extensible. The relative lack of flexibility makes these systems incompatible with the notion of knowledge based systems, which has to be flexible. The known existing systems contain a lot of incompatible data recorded in various data bases using several languages. In such situations an intelligent system able to manage this huge amount of data effectively will be very useful.

Among the different European projects we can quote MICHAEL[1] the purpose of which is to valorize Europe's cultural heritage. This project provides a multilingual interface to encourage the interoperability of different national heritage data bases. HEDD[2], conducted by the English Heritage Committee brings together 22 museums, 3 libraries and deposits of archives and uses ontologies to model common knowledge distributed in heterogeneous data.

Other projects such as those of the national Gallery of Finland[3], the University of Queensland in Australia[4] and SCULPTOR[5], use the ontology CIDOC-Conceptual Reference Model[6] (CRM) [Doerr 2006] as a tool for knowledge modeling. They unite big galleries and European cultural institutions.

[1] http://www.michael-culture.eu/project.html.

[2] http://www.fish-forum.info.

[3] http://www.fng.fi/fng/rootnew/en/vtm/etusivu.htm.

[4] http://www.metadata.net/harmony/MW2002_paper.pdf.

[5] http://www.sculpteurweb.org/html/approach.htm.

[6] http://cidoc.ics.forth.gr/index.html.

2.2. Automatic knowledge acquisition and speech recognition

For years many artificial intelligence researchers have been working on such topics as automatic knowledge acquisition and speech comprehension, mainly with signal processing techniques. The first voice interface was probably this of a workstation called Buroviseur, built in INRIA in 1981 [Kayak 1982], [Mercier-Laurent 1980]. The voice interface was also used for knowledge acquisition for expert systems [Balaram 1988] or for human-machine dialog in machine learning systems [Michalski et al. 1983]. This technology is now mature and can be integrated in applications using a large vocabulary with more than 60 000 words [Haton et al. 2006].

The quality of voice acquisition systems depends on many parameters such as external acoustic environment (noisy or silent) and the quality of the equipment employed. The main specialists of the field state that these performances provide 90% of a correct recognition [Veronis 2000]. This performance can seem insufficient in a system with full automatic transcription; however it is acceptable in the half automatic system, where the results are validated by an expert, especially when it is a question of not validating the whole of the re-transcribed text, but only a part corresponding to the predefined information.

The outlines of the extraction of information systems were defined during several Message Understanding Conferences (MUC) which took place between 1987 and 1998. It can be stated that between the first conference in 1987 and the last one in 1998, the initial ambitions – the understanding of a text by computers – were revised to finally become systems of information extraction.

The goal of the information extraction is to produce a structured representation of unstructured texts by searching for given patterns in the texts which are relevant to an application [Ibekwe-SanJuan 2007]. The text mining information extraction systems are based mainly on two technologies: the one uses automatic learning, the other one uses natural language processing (NLP).

The techniques of machine learning provide the possibility to automatically extract dictionaries and specialized grammar, as well as annotations. They allow reducing the time needed to construct linguistic resources. Their main disadvantage is that they need an important text corpus for each application domain. Information extraction techniques based on NLP use morphosyntactic analyses of text documents. This technique splits a text in sentences and terms. The tagging is based on external resources and grammar, defined by the user for a given field.

While the voice recognition and text mining are not new, the association of both is, based on our knowledge, not really deployed. Our work links these two domains and applies them to knowledge modeling.

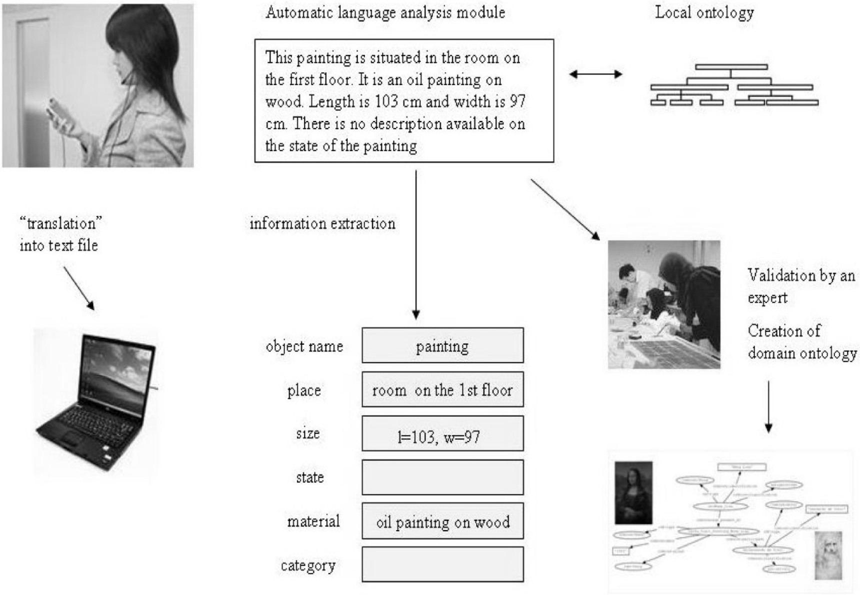
There are only a few publications on knowledge modeling in the field of cultural heritage. The main known contribution is the domain ontology CIDOC-CRM, based on object knowledge representation which is flexible and convertible into various formats such as RDF, XML, DAML+OIL, OWL. The CRM covers all information required for the scientific documentation in the field of cultural heritage.

In terms of concepts and relations between concepts and the construction of ontologies, text mining has been described in numerous works. Among them we quote [Charlet 2002] and [Bourigault, Aussenac-Gilles 2003], who work on the construction of ontologies from texts in the medical domain.

3. Our work – voice acquisition system

Our voice acquisition system is presented in Fig. 1. It follows four steps:

1. Voice acquisition of a given masterpiece description.
2. Automatic transcription of the voice file into the text file by Dragon[7].
3. Extraction of concepts and relations between concepts.
4. Validation of the extracted concepts found in the previous stage by expert.


Automatic language analysis module
Local ontology
"translation" into text file
information extraction
Validation by an expert
Creation of domain ontology

This painting is situated in the room on the first floor. It is an oil painting on wood. Length is 103 cm and width is 97 cm. There is no description available on the state of the painting

object name	painting
place	room on the 1st floor
size	l=103, w=97
state	
material	oil painting on wood
category	

Fig. 1. Voice acquisition system helping assistant in knowledge acquisition

The validated descriptors are registered in a data base and will be used to update the existing ontologies. The acquired voice information is distributed in fields of the data base such as: DENOMINATION, CATEGORY, MATERIAL, DESCRIPTION, and INSCRIPTION without constraining the speaker to say the name of the descriptive field. These fields constitute the descriptive system defined by the heritage inventory department [Ver-

[7] http://www.nuance.fr/naturallyspeaking/.

dier et al. 1999]. Some of these fields are compulsory, the others optional. The contents of certain fields are defined by a lexicon, the contents of the other fields remaining free.

Usually the acquisition of the data is made using a keyboard and needs to strictly respect a data acquisition model. In the case of a voice acquisition, there is no structure to guide the acquisition. The involved people are specialists in the given field, thus we can expect a coherent and well structured text.

3.1. Robust syntactic analysis

Despite the good performances of the re-transcription software, some syntactical and semantic errors can occur in the re-transcribed files. The origin of its errors can be directly connected to the way the speaker dictates the text (waiting time, hesitation, back on sentences or words). The transcription process itself may also cause errors.

We started the acquisition without any text archives what made the applying of machine learning methods impossible. We have chosen the robust incremental syntactic analyzer [Hagège, Roux 2003]. Such an analyzer always insures good results even with a badly structured or erroneous input text.

Incremental means that the rules of disambiguation, category, construction of constituents and the extraction of syntactical dependencies are applied one after the other. The specific and reliable rules are first to filter the rare or exceptional configurations, while the more general rules are at the end of grammar [Hagège, Roux 2003]. We use the XIP[8] analyzer created by XRCE[9] for this experimentation.

3.2. From the data entry form to the extraction patterns[10]

As we mentioned before, the information to find is defined by the descriptive system of the inventory [Verdier et al. 1999]. It indicates the type of information to be looked for, but also controls, in certain cases, the vocabulary to be used. The terms have to correspond with the entry of a lexicon.

The descriptive system of the inventory will partially guide the conception of the extractions patterns and local grammar.

The collected information on the field can be split in two categories:

Physical aspects: material of manufacturing, structure, place.

All the information relative to the historical, social, ethnographical context.

It is the type of information that can be known only by experts of a given domain. Our system of extraction of information has to be able to take it into account.

[8] XIP (Xerox Incremental Parser) by AïtMokhtar, Chanod et Roux.

[9] Xerox Research Center Europe.

[10] Extraction pattern: set(group) rules allowing to identify the expected, relevant information.

Two scenarios are possible:

The result corresponds exactly to a defined entry of a lexicon. In this case the local grammar must be defined to insure that the analysis and the result of extraction is a word or a constituent, which corresponds exactly to an entry of this lexicon.

The result is an incomplete description of a given place, for example:

"le retable comprend 4 tableaux: Baptême du Christ, Christ au Jardin des oliviers, la Cène et la Résurrection".

The constituent "Baptême du Christ" will be tracked down in the text without problem because it exists in the lexicon, then thanks to an analysis of dependence; it can be associated with the representation. The constituent "Christ au Jardin des oliviers" will not be recognized as representation because it does not exist in this lexicon. The system has to be able to recognize this entry as a constituent, and to suggest it as a possible entry. A local syntactic analysis must be triggered by one of the words of the constituent because they belong to the lexicon, or because the sentence contains a word or a constituent which is associated with the idea of the representation: the representation, are represented.

In our example, the constituent "Jardin des oliviers" and the word "Christ" exist separately in the lexicon representation, which is the condition to propose the constituent Christ in the Garden of olive trees as a possible descriptor of the representation. According to the principle of relations "sort of" the representation of the "Christ au Jardin des oliviers" is a specific case of a representation of Christ.

The identification of the words or the constituents is not the only difficulty which we have to face. The language of the cultural heritage is extremely rich and words can have multiple meanings, which means that the system has to be able to deal with ambiguities. A word or a constituent can be used in various contexts as well as to describe the representation of a masterpiece or a masterpiece itself. In the example a picture representing a chalice the name could be the name of the person represented on the chalice or the artist's name. It frequently happens that the described belongs to a group. The description of this type of objects can hint at the contained or containing elements. We are thus in a situation where several names of a masterpiece are quoted. How can we know which one is the object of the study?

The resolution of ambiguities requires an analysis and the understanding of the local context. Some ambiguities can be decided by using a morphosyntactic analysis of the following or previous words or by searching for linguistic indications according to the given topic.

3.3. The initial position

The study of the organization of descriptors in a text can be of considerable help, notably for the resolution of certain types of ambiguities. The study of the initial position, which leans on the cognitive consideration [Enkvist 1976; Ho-Dac 2007], states that the beginning of a sentence has a great importance, as we place important information in an initial position of sentences. In this perspective, the extraction of the information from the text:

Musée de la société archéologique de Montpellier.

Panneau de Saint Guilhem et Sainte Apolline (87 × 136) en cours de restauration par Anne Baxter.

C'est une peinture à l'huile de très grande qualité, panneau sur bois représentant deux figures à mi corps sur fond de paysage, saint Guilhem et sainte Apolline, peintures enchâssées sous des architectures à décor polylobés; Saint Guilhem est représenté en abbé bénédictin (alors qu'à sa mort en 812 il n'était que simple moine); sainte Apolline tient l'instrument de son martyre, une longue tenaille [...],
will prefer the descriptor Panneau over the descriptor Peinture, to indicate the naming of the studied object.

3.4. Semiautomatic generation of ontology

The collected knowledge on a masterpiece is partial; it is valid only for a lapse of time and cannot be limited to a fix frame defined for a given application.

The knowledge is flexible, the masterpieces of the cultural heritage have a past, a present and maybe a future "life", and they can change in time. As we mentioned before the extraction of information in our case has to correspond to a precise specification.

We have to face two requirements: to fill a data base defined by the descriptive system of the inventory and allow the flexibility of a knowledge management system. For the first the information found by extraction can be adjusted, and validated by an expert if it is necessary. We think that it is also a convenient moment to satisfy the second point; the validated information composed of descriptors and their relation, which describes the material and immaterial aspects of masterpiece, will feed the ontology of a domain in a vaster and more flexible way.

How to define the ontology regarding the problem of modeling, opening, and knowledge sharing? There is a vast variety of definitions of ontology, and that of Gruber [Gruber 1993] seems to correspond the best in case: "ontology is an explicit and formal specification of a conceptualization being the object of a consensus". In other words the ontology of a domain is a set of concepts and relations between these concepts defined by means of a formal language by involved actors and for a particular domain. According to Charlet [Charlet 2002], in an ontology we represent and classify concepts and their characteristics (properties); we also represent relations between these concepts. In our case, we have to describe of what material the object of cultural heritage is made, by whom, when, why, what transformations were done, what is its state of preservation as well as the masterpieces movements. We can say that a certain number of concepts is outlined: time, place, actor(person) and state of preservation. Intuitively, we guess that some of these concepts are connected to each other, as for example the state of preservation and time, transformations and time, movements and place, transformation and person.

The CIDOC-CRM ontology, already quoted in Section 2, presents the necessary formalism allowing reporting relations, which an object can have in time and space. The heart of CRM is constituted by the temporal entity expressing the dependence between time and the various events in the life of the historical object.

If we consider an example of a sculpture described by the inventory system, information such as author, naming, materials… are easily expressed. Because this system is not able to model the various movements of a given object, this information is described using free text and mixed with other type of information in the historic field. The same information can be easily expressed by the CRM ontology, presented in Fig. 2.

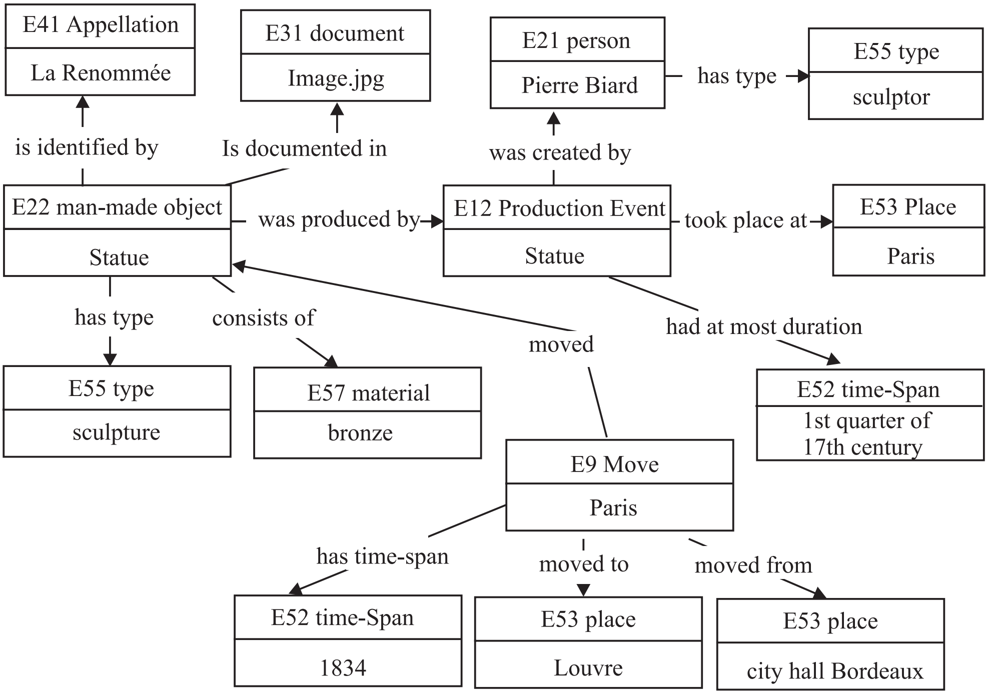


Fig. 2. Example of a sculpture modeling in CIDOC-CRM

The evolution from the model defined by the inventory descriptive system to the CIDOC-CRM ontology is possible by the search for the correspondences between the fields of the descriptive system, in which the content be considered as the instance of one of the classes of the CRM ontology.

For the cases, in which this correspondence could not be found because the information does not exist in the descriptive system, it will be necessary to extract it from the re-transcribed text, under the condition that the speaker registered it. Otherwise it will be necessary to enter it during the validation of the information extracted automatically by the system.

4. Conclusion and perspectives

The originality of our system is the link between three distinctive research domains such as signal processing, ontology and natural language processing. We experimented on field voice knowledge acquisition, “translation” of voice into a text file, the work on text files in order to extract the relative concepts and relation between them in semiautomatic way. The voice interface provides a considerable help and efficiency for an expert working in the field. The knowledge modeling with ontology adds the flexibility to the classic inventory systems and allows future knowledge retrieval.

In the next step we wish to introduce the real dialogue human-machine for knowledge acquisition. So the “knowledge collector” would have a real-time feedback on the understanding by the machine of what he dictates. We believe that the implementation of a transcription system and the extraction of information will be shortly possible on mobile devices. The described voice assistant for knowledge acquisition is a part of larger Knowledge Management System for cultural patrimony allowing the acquisition, modeling and intelligent retrieval of knowledge about objects, their history and contexts.

References

Balaram M. (1998). PC version of a knowledge-based expert system with voice interface. *IEA/AIE*, vol. 2, pp. 1168-1173

Bourigault D., Aussenac-Gilles N. (2003). Construction d’ontologies à partir de texts. *TALN 2003*, Batz-sur-Mer.

Charlet J. (2002). *L’ingénierie des connaissances : résultats, développements et perspectives pour la gestion des connaissances médicales. Mémoire d’habilitation à diriger des recherches*. Université Pierre et Marie Curie, Paris.

Doerr M., Crofts N., Gill T., Stead S., Stiff M. (Eds.) (2006), Definition of the CIDOC Conceptual Reference Model, Produced by the ICOM/CIDOC, October 2006.

Enkvist N.E. (1976). Notes on valency, semantic scope, and thematic perspective as parameters of adverbial placement in English. In: *Reports on Text Linguistics: Approaches to Word Order.* Eds. N.E Enkvist, V. Kohonen, *Åbo*, pp. 51-71

Gruber T.R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, vol. 5, pp. 199-220.

Hagège C., Roux C. (2003). Entre syntaxe et sémantique : Normalisation de la sortie de l’analyse syntaxique en vue de l’amélioration de l’extraction d’information à partir de texts. *TALN 2003*. Batz-sur-Mer, 11–14 juin 2003.

Haton J.-P., Cerisara C., Fohr D., Laprie Y., Smaïli K. (2006). *Reconnaissance automatique de la parole, Du signal à son interpétation*. Dunod, Paris.

Ho-Dac L. (2007). *La position initiale dans l’organisation du discour: une exploration en corpus*. Thèse de doctorat, Université Toulouse le Mirail.

Ibekwe-SanJuan. (2007). *Fouille de textes: méthodes, outils et applications*. Hermès-Lavoisier, Paris–London, 2007.

Kayak J. (1982). *Bureautique et intelligence artificielle, Recueil des conférences*. INRIA.

Mercier-Laurent E. (1980). *Réalisation de communications dans un processeur de consultation de données textuelles*. Thèse Docteur-Ingénieur, INRIA.

Michalski R.S., Carbonell J.G., Mitchel T.M. (1983). *Machine Learning, An Artificial Intelligence Approach.* Vol. I, Tioga.

Verdier H. et al. (1999). *Système descriptif des objets mobiliers*. Editions du Patrimoine, Paris.

Veronis J (2000). Annotation automatique de corpus: panorama et état de la technique. In: *Ingénierie des langues*, Hermes, Paris.