COMPUTER AIDED SYSTEM FOR MANAGING DOMAIN KNOWLEDGE

Application to Cultural Patrimony

Stefan du Château, Danielle Boulanger and Eunika Mercier-Laurent

MODEME, Research Center IAE, University of Jean Moulin-Lyon, 6, av Albert Thomas, F-69008 Lyon, France

Keywords: Knowledge processing, Voice interface, Natural language processing, Domain ontology, Cultural heritage.

Abstract: This paper presents our hybrid system for cultural heritage management. It combines the techniques of signal and natural language processing and knowledge modelling to effectively help a researcher in cultural patrimony in collecting, recording and finding the relevant knowledge. The voice interface serves to describe the artefacts in a given historical place. This audio file is than "translated" into a text file and validated by an expert in the area. The next step is an automatic concept extraction and building specific ontologies for the future processing. After introducing the problem of on field information collecting and managing, we describe the specific work of a researcher in the field of cultural heritage and main difficulties. Furthermore we explain our choice of the architecture of this hybrid system, our experiments and the results. Finally we give some perspective on extending this system to the other domains.

1 INTRODUCTION

A common problem in knowledge engineering is the efficient collection of information and knowledge from sources considered to be scientifically reliable. These can be human experts, written records or computer applications (databases) that cover the domain knowledge. Depending on the situation, treatment and expected outcome, different collection methods can be used.

The work of researchers in the area of cultural heritage consists in one part of gathering of information in the field, in towns and villages in the form of text files, photos, sketches, maps and videos. If necessary, the information gathered for each work is corrected, archived, and finally stored in a database. The storage of information in paper documents or directly on laptops is cumbersome and time consuming. The amount of information collected is very large, the data is heterogeneous and its transformation into a form that can be used for research is not automatic.

The system we propose uses a voice interface that reduces the amount of time used in the process of collection, because the description of the artefacts studied can be voice recorded and saved as an audio file. This is a hybrid system because it relies on technologies of signal processing, knowledge modelling and natural language processing.

2 ARCHITECTURE OF SIMPLICIUS

The architecture of our system takes into account several factors. First, it enables the implementation of three functional steps: the collection of information and knowledge in a specific context, information extraction and semi-automatic generation of a partial domain ontology supervised by a conceptual model. On the other hand, it must respect the constraints imposed by existing: the descriptive system of inventory, lexicons and thesauri and conceptual model CIDOC-CRM (Doerr et al., 2006).

The process leading to the ontology of discourse of an object consists of several steps:

1. The voice acquisition of the description of a artefact.
2. Transcription of audio file into a text file, using Dragon software that we have enriched with a specific vocabulary of cultural heritage.

du Château S., Boulanger D. and Mercier-Laurent E..
COMPUTER AIDED SYSTEM FOR MANAGING DOMAIN KNOWLEDGE - Application to Cultural Patrimony.
DOI: 10.5220/0003111002400245
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2010), pages 240-245
ISBN: 978-989-8425-30-0

3. Display the result text to allow expert correct it if errors.
4. The linguistic analysis and information extraction (Grishman, 1997), (Ibekwe-SanJuan, 2007). This stage leans on the XIP (Xerox Incremental Parser) (Aït-Mokhtar, et al., 2002) software, which we enriched by semantic lexicons and grammatical rules, specific of the domain of the cultural heritage.
5. Validation of information got in the previous stage.
6. Generation of ontology of objects described during the first stage. It is the transfer of an implicit information contained in the SDI (Descriptive System of the Inventory) (Verdier, 1999), defined by the Department of Heritage Inventory, to the explicit knowledge represented by the domain ontology of cultural heritage.

The architecture of our system is shown in the Figure1.

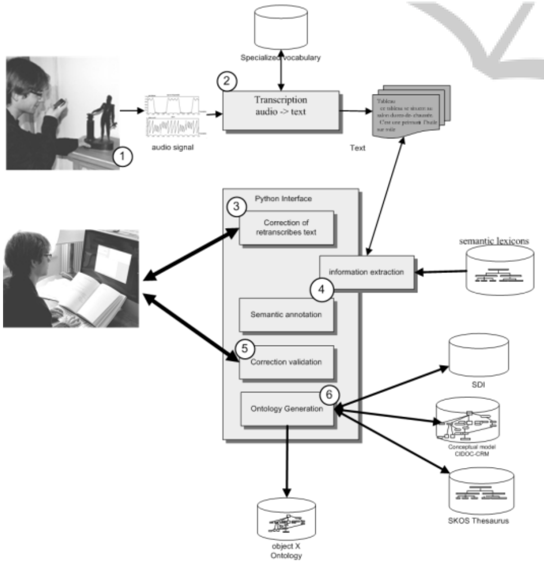

Figure 1: The architecture of Simplicius system.

2.1 From Voice to Text

The audio file is "translated" into text using the Dragon software; we have chosen it for its robustness and its performance in speech recognition.

Text files serve as information retrieval so that information is distributed into fields such as: DESCRIPTION, CATEGORY, MATERIALS, REGISTRATION, NAME (...), without requiring the speaker to specify the description field. The above fields derive from the descriptive system defined by the Department of Heritage Inventory (Verdier, 1999). Some of these fields are mandatory, others optional. The content of certain fields is defined by a lexicon; the contents of other fields remain free.

Currently the data acquisition is done via keyboard and the user has to respect a highly structured data entry form. In the case of voice acquisition, there is no structure required to guide the user, who is usually a specialists in the field; we can therefore assume that the verbal description will be coherent and well structured. This has been proven in our experiments.

2.2 Analysis of Resulting Text and Information Extraction

To analyze the transcribed text (cf stage 4, figure 1), we use the robust XIP parser. This guarantees a result of corpus analysis, even if the text is malformed or erroneous, which can happen if the text is the result of an oral transcript (Hagège, 2003), (Brun, 2009).

As we mention in Section 2.1, the information that has to be identified for extraction is defined by the above-mentioned descriptive system for artwork inventories, which defines not only the type of information that is to be looked for, but also controls, in some cases, the vocabulary to be used. The terms used should match the entry of a lexicon. The descriptive system of the inventory will therefore partially guide the creation of design patterns and of local grammars.

2.2.1 Lexicons

Two types of lexicons have been created: one that contains the vocabulary defined as authorized to fill out fields such as DENO, REPR MATR (...), and other which contains vocabularies for context analysis. Two types of formats are used. For lexicons with extended vocabularies, the term of each has been associated with its infinitive form for verbs and the masculine singular for nouns. In addition, its semantic and morphological trait was added to each term, as shown below

calice
calice
+Denomination+Masc+Sg+Common+Noun
calices
calice
+Denomination+Masc+Pl+Common+Noun

The format of smaller glossaries includes the lemmatised form of the term and the semantic and morphological trait associated with it :

marque : noun += [!insc:+].
cachet : noun += [!insc:+].

2.2.2 Resolution of Ambiguities

The identification of words or phrases is not the only difficulty faced by a system of information extraction. In the context-rich environment of cultural heritage artefact descriptions, the complexity of the language itself and the multiplicity of meanings that can be given to the descriptors used, one of the major problems is the resolution of semantic ambiguity. A word or phrase can be used in different contexts both to describe the characteristics of an artefact as well as the artefact itself, for example *a picture of a chalice,* the name of a person can be that of a person represented, or that of the artist (...). Often, heritage objects that are being described are part of a whole. The description of this type of object can refer to included elements, or to its container. It is therefore in a situation where several artefact names are mentioned. How do we know which is the subject of study ?

In the sentence: ***Calice*** *en argent doré, orné de grappes de raisins, d'épis de blé, de roseaux sur le pied et la fausse coupe, d'une* ***croix*** *et des* ***instruments*** *de la passion dans des* ***médaillons****, sur le pied.*

The terms: **calice**, **croix**, **instruments**, **médaillons** exist in the lexicon DENOMINATION. The term **calice** also exists in the lexicon REPRESENTATION

How can we be sure that, in this case, it is DENOMINATION?

How to choose the term for the DENOMINATION?

Study of the initial position

The study of the ordering of descriptors in a text provides valuable assistance, particularly for solving certain types of ambiguities. The study of the initial position, based on cognitive considerations (Enkvist, 1976), (Ho-Dac, 2007), gives special importance to the beginnings of sentences: the information at the beginning is a given information or at least one that is important.

In this perspective, extracting information from the following text:

Calice *en argent doré, orné de grappes de raisins, d'épis de blé, de roseaux sur le* ***pied*** *et la fausse* ***coupe****, d'une* ***croix*** *et des instruments de la passion dans des* ***médaillons****, sur le pied.*

Will give a preference to the descriptor ***Calice*** compared to other descriptors mentioned above, to designate the name of the object studied.

Local context

Resolving ambiguities requires an analysis and understanding of local context. A morphosyntactic analysis of words surrounding the word whose meaning we seek to identify, as well as searching for linguistic clues in the context of a theme, can resolve some ambiguities.

In the sentence : *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.*

Saint Guilhem can designate a place or a person. Is it a painting that is located in Saint Guilhem, or does it represent Saint Guilhem and Sainte Apolline?

A study of the position and the semantic class of arguments in the relationship: *subject-verb-object,* provides clues for resolving this ambiguity, the principle that the topic is the subject of the sentence, what is known as the word about the phrase, what is said of the theme.

In the above example the verb **representing** contains the feature [Repr: +], which links it with the REPRESENTATION class. In the absence of other significant indices, it can thus be inferred that the purpose of the sentence is "representation" and Saint Guilhem and Sainte Apolline do not designate places, but rather the representation.

2.3 Semi-automatic Generation of a Domain Ontology

The knowledge gathered on an artefact is necessarily partial: it is only valid for a period of time and therefore cannot be limited to a descriptive grid designed for one specific application.

Knowledge is scalable, cultural heritage artefacts have a past, a present and perhaps a future; they undergo transformations over time.

However, we have seen above that the extraction of information in our case must correspond to precise specifications. We are thus faced with two requirements: on the one hand to populate a database defined by a specific inventory description system, on the other hand, to meet the requirements of a knowledge management system.

To satisfy the first requirement, it is essential that the information found by the extraction can be adjusted (if necessary) and validated by an expert.

To satisfy the second item, the validated information, consisting of descriptors and their relationships that describe the tangible and intangible aspects of the artefact, will have to be fed into a domain ontology, which is more extensive and extensible. This provides the necessary openness and sharing of knowledge, as defined by Gruber, *"an ontology is an explicit and formal specification of a conceptualization that is the consensus"* (Gruber, 1993).

In the context of cultural heritage artefacts, which is the one that interests us, the description will focus on how an object was manufactured, by whom, when, for what purpose, it will focus on its transformations and travels, its conservation status and materials used for this purpose. One can see that a number of concepts are emerging such as: Time, Place, Actor (Person), state of preservation. Intuitively, one suspects that some of these concepts can be related to each other, such as conservation status and time, transformations and time, travels and place, transformations and owner.

The ontology CIDOC-CRM presents the formalism required for reporting of relationships that can be implemented in time and space. The heart of CIDOC-CRM consists of the entity expressing temporal dependence between time and various events in the life of the artefact.

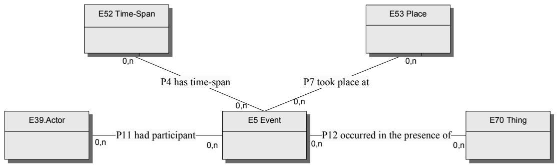

Figure 2: Modelling event in CIDOC-CRM, from (Crofts, 2007).

For clarity and easier reading by the user accustomed to the nomenclature of SDI we have, based on CIDOC-CRM, created a model that defines equivalences between the different fields of SDI and certain classes of CIDOC-CRM (Figure 3).

The transition from the model defined by the inventory descriptive system to the CIDOC-CRM ontology (cf stage 6, figure 1), will be done by searching through the correspondence between the fields of inventory descriptive system, whose contents can be regarded as an instance of one of the classes of the CRM ontology.

For cases where this correspondence can not be made, because the information does not exist in the inventory description system, it will have to be retrieved from the transcribed text, provided that the speaker has record such kind of information. Otherwise it will have to be input when the information extracted automatically by the system is validated.

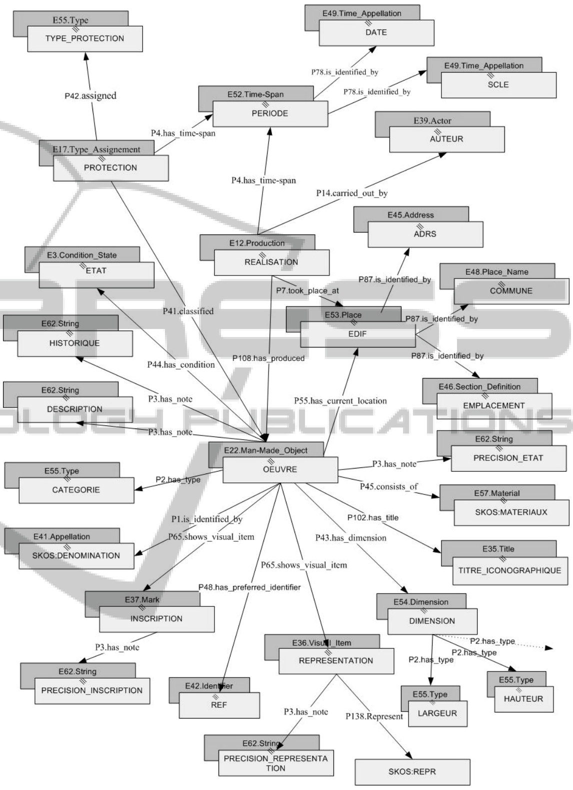

Figure 3: The CIDOC-CRM classes and equivalence with the SDI.

3 EXPERIMENTS AND RESULTS

The application that we propose is still in prototype stage; it is therefore too early to provide a real experience feedback, which would require the operation of our system.

Thus, we present the experiments we have conducted so far with the prototype version of our system and with the help of three researchers familiar with cultural heritage as well as the area of inventory and SDI. Two of the three researchers are female, one of which has a regional accent, while the other speaks with no accent. The third, male researcher speaks with accent.

The dictations were performed in real conditions in a noisy environment. We asked each researcher to verbally describe three objects.

The oral descriptions were transcribed into text. The results are quite satisfactory; the concordance between the original content and the content in the automatically transcribed texts varies between 90 and 98%.

Before presenting them to the module for the extraction of information, the transcribed texts have been corrected by the researchers. For each result of extraction of information, we measured Precision, Recall and F-score, which are presented in the table below.

In order to clarify the presentation we have assigned a letter to designate each speaker: A for the woman speaking with an accent, B for the woman with no accent and the letter C for the man.

Table 1: Results of extraction of information.

Researcher	Precision	Recall	F-score
A	0,898	1	0,943
B	0,854	0,946	0,897
C	0,903	0,94	0,921

Our experiments are not numerous enough to supply a more reliable statistical study, nevertheless the obtained results are sufficiently promising to encourage us to continue developments of our system.

For the moment our system is elaborate for the French language.

Below is an example of the description of a painting performed by a researcher of cultural heritage. The first text is the result from the voice recording transcript. You can see the errors marked in bold.

***Et le Damiani** église Saint-Sauveur. Tableau représentant saint Benoît d'Aniane et saint Benoît de Nursie offrant à Dieu le Père la nouvelle église abbatiale d'Aniane. Ce tableau est situé dans le **coeur** et placé à 3,50 m du sol. C'est une peinture à l'huile sur toile encadrée **et 24 en** bois Doré. **Ça auteure et** de 420 cm sa largeur de 250 cm. Est un tableau du XVIIe siècle. Il est signé en bas à droite **droite** de Antoine Ranc. Est un tableau en mauvais état de conservation un réseau de craquelures s'étend sur l'ensemble de la couche picturale.*

The second is the text after correction. You can consult the translation of this text in English in appendix.

The results outcomes from module of the Extraction of Information are marked in bold.

*Ville d'**COM**{Aniane} **EDIF**{église Saint-Sauveur}. **PREPR{DENO**{Tableau} représentant **REPR**{ saint Benoît d'Aniane] et **REPR** {saint Benoît de Nursie} offrant à **REPR**{Dieu le Père} la nouvelle église abbatiale d'Aniane}. Ce tableau est situé **EMPL**{ dans le choeur et placé à 3,50 m du sol}. C'est une peinture à **MATR**{ l'huile sur toile} encadrée d'un cadre en **MATR**{bois doré}. Sa **DIMS**{hauteur est de 420} cm sa **DIMS**{ largeur de 250 cm}. Est un tableau du **SCLE**{XVIIe siècle}. Il est signé en bas à droite de **AUTR**{Antoine Ranc}. **PETAT** {Est un tableau en mauvais état de conservation un réseau de craquelures s'étend sur l'ensemble de la couche picturale}.*

Where:

COM = Commune, EDIF = Edifice, REPR =Representation, PREPR = Precision on the representation, EMPL = Place, MATR = Materials, DIMS = Dimension, SCLE = Century, AUTR = Author, PETAT = Precision on the state of preservation.

The linguistic analysis, information extraction and ontology creation are done using the second file, as shown schematically in Figure 4.

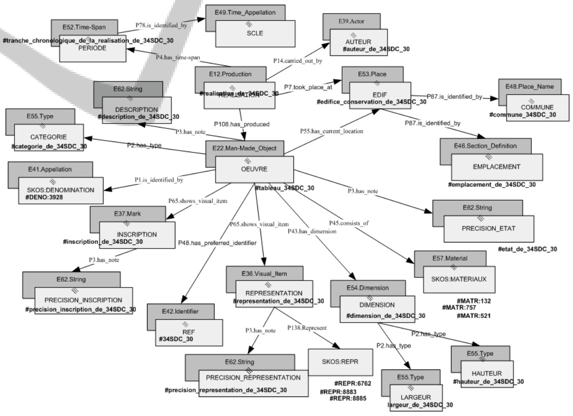

Figure 4: Example of ontology of a work after a description dictation.

4 CONCLUSIONS

The originality of our voice recording system developed to support the acquisition of knowledge of cultural heritage is the link between two areas of research, which were until now developing parallel to each other: signal processing and automatic language processing. Our experiments have been successful and confirm the technical feasibility and usefulness of such applications.

Modelling of knowledge as an ontology and ontological cooperation will provide flexibility and scalability to our system, e.g. extending the scope of the CIDOC-CRM model to model the spatio-temporal knowledge, by adding geospatial information such as topology, directions, distances, location of an artefact relative to reference locations. The recent work of LIG (Laboratoire d'Informatique de Grenoble) and in particular the model ONTOAST (Miron et al., 2007) seem very interesting in this regard. In the context of ontological cooperation arises the problem of coherence among distributed ontologies, who we believe can be resolved by means of the cognitive agents.

In the future, it might be useful to incorporate a speech acquisition control mechanism, in the form of a man-machine dialogue. Thus the speaker would have a real-time feedback on the machine's understanding. This implies in our case the possibility to implement the transcription and information extraction system on a mobile platform.

The OWL format for the creation of the ontology we use ensures its compatibility with the standards of the semantic web. It allows for an easy integration with inference and inquiry systems, thereby facilitating its future use in both scientific and community applications, such as search engines, artefact comparison platforms or the exchange of knowledge with other ontological structures.

REFERENCES

Doerr Martin., N. Crofts, T. Gill, S. Stead and M. Stiff, eds. Definition of the CIDOC Conceptual Reference Model. *ICOM/CIDOC*, October 2006.

Grishman, Ralph. Information Extraction: techniques and challenges. Information Extraction (MT Pazienza ed.), *Springer Verlag (Lecture Notes in computer Science)*, Heidelberg, Germany, 1997.

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

Aït-Mokhtar Salah, Jean-Pierre Chanod, and Claude Roux. Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering, vol. (8/2-3)*, 2002. 121-144.

Verdier Hélène. Système descriptif des objets mobiliers. *Paris: Editions du Patrimoine*, 1999.

Hagège Caroline and Claude Roux. 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.

Brun Caroline and Caroline Hagege. Semantically-Driven Extraction of Relations between Named Entities. CICLing 2009 *(International Conference on Intelligent Text Processing and Computational Linguistics)*, Mexico City, Mexico, March 1-7, 2009

Enkvist, Nils. E. "Notes on valency, semantic scope, and thematic perspective as parameters of adverbial placement in English". *In: Enkvist, Nils E./Kohonen, Viljo (eds.) 1976: Reports on Text Linguistics: Approaches to Word Order*.

Ho-Dac Lydia. La position Initiale dans l'organisation du discours: une exploration en corpus. Thèse de doctorat, Université Toulouse le Mirail, 2007.

Gruber Tom. R. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5, 1993.

Crofts Nick. La norme récente ISO 21127: une ontologie de référence pour l'échange d'infomations de patrimoine culturel, Systèmes d'informations et synergies entre musées, archives, bibliothèques universités, radios et télévisions, Lausanne, 2007.

Miron, Alina., J. Gensel, M. Villanova-Oliver, and H. Martin. "Relations spatiales qualitatives en ONTOAST pour le Web semantique geospatial", *Colloque International de Geomatique et d'Analyse Spatiale* (SAGEO2007), Clermont-Ferrand, France, 18-20 June 2007.

APPENDIX

The Translation of the French Description:

City Aniane church Saint-Sauveur. Painting representing Saint *Benoît d'Aniane* and Saint Benoît of Nursie offering to God the Father the new abbey church of Aniane. This Painting is situated in the choir and placed in 3,50 m above the ground. It is an oil painting on canvas framed in a gilt wood frame. His height is 420 cm width is 250 cm. It is a painting of the XVIIth century. He is signed bottom on the right by Antoine Ranc. It is a picture in a poor state of preservation a cracks network spread throughout the entire painting area.