Patents editor in order to automatically construct terminological databases

Chihebeddine Ammar, Kais Haddar

Patents editor in order to automatically construct terminological databases

Chihebeddine Ammar — Kais Haddar

Laboratoire MIRACL, Université de Sfax, Pôle technologique de Sfax : Route de Tunis Km 10 B.P. 242, 3021 Sfax

ammarchihebeddine@hotmail.com, kais.haddar@fss.rnu.tn

RÉSUMÉ. Les demandes de brevet ont une structure similaire dans le monde entier. Elles comprennent une page de couverture, un mémoire descriptif, des revendications, des dessins (si nécessaire) et un abrégé. Au monde Arabe, il n'y a pas de collection numérique uniforme de document de brevets et donc pas de collection XML. Dans ce contexte, nous présentons la création d'un modèle de document de brevet standard pour les brevets Arabes et le développement d'un générateur de collection XML de brevets ayant une structure uniforme est simple à utiliser. Cette collection va nous être utile pour la construction d'une base de données terminologique pour les domaines scientifiques et techniques.

ABSTRACT. Patent applications are similarly structured worldwide. They consist of a cover page, a specification, claims, drawings (if necessary) and an abstract. In the Arabic world, there is no uniform digital collection of patent documents and therefore no XML collections. In this context, we aim to create a standardized document model for Arabic scientific patents and develop a generator of XML patent collection having a uniform and easy to use structure. This collection will be useful for us to build a terminological database for scientific and technical fields.

MOTS-CLÉS : Normalisation, modèle de document de brevets, Bases de données terminologique.

KEYWORDS: Normalization, Document patent model, Terminological databases.

1. Introduction

One of the very rich in terminology work streams are the scientific patents. They are similar, for example, to a scale repository. They also cover several scientific and technical fields, while offering rich interdisciplinary relations. That is why we will need several terminological databases, one for each field.

Indeed, standardized modeling patent allows us to maintain a standard for the representation of texts in digital form, so that we protect patents data by bringing them in digital databases. It will provide a single common data model for all terminological data regardless of the data's language, source, field, etc. Also, we will be able to build collections of uniform patents which facilitate the extraction and the exploitation of patents data and the extraction of links between valid terms. Standardized modeling patent ensure also interoperability between applications. Finally, it will allow us to easily enrich other terminological databases.

Patent information, in the Arabic world, remains almost the preserve of patent agents or lawyers versed in research needed to be done before any patent application filing or preparation a lawsuit. That's why, the development of computerized databases on patents, will open access to all categories of users: businessmen, economists, researchers, etc., and make them aware of the potential value of the information that patents contain.

Patents are available in different formats: Full text, PDF document, set of images, XML, etc. They have heterogeneous components that require different modeling. Also, patents have linguistic structures like text and titles, and nonlinguistic structures like figures, citations,

tables and formulas. In fields such as mechanics, automatic extraction based only on the text will fail.
In addition to the text, figures and citations information, all patent publications contain a relatively rich set of well-defined metadata. These metadata are often found in the cover page of patents and titles of figures and tables. To cope with the large volume of data and metadata, we will develop a patents terminological editor to generate terminological databases. This will allow us to develop heuristics, based on metadata such as the applicant(s) name(s), the inventor(s) name(s) or priority documents, etc., for finding interesting documents.
The structure of the XML documents may be used for the processing performed to differentiate various elements according to their semantic. Thus, a section title, a summary, bibliographic data, or examples can be used to identify different aspects of the text. Indeed, scientific patents can be easily processed as XML documents. So we can treat their structures[1] as a source of information.
The work presented in this paper is the continuation of studies (Ammar et al., 2014) on the standardization for Arabic patents. We propose a standardized model for Arabic patents and create a patents editor in order to generate a patent collection having similar structure. It is an original idea because nobody treated terminology in Arabic patents in previous works.
This article is organized as follows. In section 2, we present previous works. We present, in section 3, our Arabic standard patent document editor. Section 4 is devoted to the evaluation and discussion and we conclude and enunciate some perspectives in section 5.

2. Previous works

Previous works on patents (Lopez and Romary, 2009; Lopez and Romary, 2010b; Magdy et al., 2009) were mainly based on purely statistical approaches. They used standard techniques of information retrieval and data extraction.
Some of the previous works use machine learning tools to extract header metadata from patents using support vector machines (SVM) (Do et al., 2013), hidden Markov models (HMM) (Binge, 2009), or conditional random fields (CRF) (Lopez, 2009). Others use machine learning tools to extract metadata of citations (Hetzner, 2008), tables (Liu et al., 2010), figures (Choudhury et al., 2013) or to identify concepts (Rao et al., 2013). All these approaches rely on previous training and natural language processing.
In (Lopez et al., 2010a), the authors developed multilingual terminological database called GRISP covering multiple technical and scientific fields from various open resources.
The European Patent Office, EPO[2], offers inventors a uniform procedure of application, and a register of multilingual patents (English, French, and German).
The MAREC (MAtrixware REsearch Collection) database is formed of patent documents from the European Patent Office. It is a standard corpus of patent data available for research purposes. It consists of 19 million of patent documents in different languages (English, French, and German) in a standardized XML schema highly specialized.
The ePCT[3] is a WIPO (World Intellectual Property Office) online service that provides secure electronic access to the files of international applications filed under the international patent system as maintained by the International Office. It is also possible to file international applications using ePCT-Filing.
In the literature, works on Arabic patents are missing. This is caused by the fact that in the Arabic world, there is no North African or Arabian Intellectual Property Office and therefore

[1] Remind that an XML document is structured as a tree consisting of hierarchical elements which may have one or more attributes, the leaf nodes have information.
[2] EPO: European Patent Office, http://www.epo.org/
[3] https://pct.wipo.int/LoginForms/epct.jsp

no uniform collections of Arabic patents. In Tunisia, for example, the INNORPI[4] (National Institute for Standardization and Industrial Property) does not propose a digital collection of patent documents and therefore no XML collections. As a result, Arabic patents have no unique structure. For the Tunisian patents, as illustrated in Table 1, the cover page doesn't have abstracts and patent documents could be in one of the three languages (Arabic, English or French). In the regional office[5] for the Gulf Cooperation Council (GCC Patent Office[6]), there are only Arabic patents and there is an Arabic abstract in the cover page. The layout of the description part varies also from place to place. For example, the summary and the background of the invention could not exist in some patent descriptions. The Tunisian patents themselves have no unique structure in that some of them have no abstract, have missing bibliographic data and even no cover page. For these reasons, a normalization phase for Arabic patents is necessary.

	Tunisian patents	**GCCPO patents**
Language	Arabic, French, English	Arabic
Digital document	No	Yes
International patent classification	No	Yes
Search report	No	No
Citations	No	Yes
International publication	Yes	No

Table 1. *Comparison between Tunisian and GCCPO Patents.*

3. Arabic standard patent document editor

Patent applications are similarly structured worldwide. They consist of a cover page, a description, claims (Hong, 2013), drawings (if necessary) and an abstract. The cover page of a published patent document usually contains bibliographic data such as the title of the invention, the filing date, the priority date, the names and addresses of the applicant(s) and the inventor(s). It also has an abstract, which briefly summarizes the invention, and a representative drawing. Bibliographic data are extremely useful for identifying, locating and retrieving patent documents. The patent description must describe the claimed invention and give technical information. The claims determine the patentability and define the scope of the claimed invention.

Patent documents are often difficult to understand and have a variety of structures. So, we propose an Arabic patent document model and develop a patent editor which automatically generates a collection of XML patent documents having a similar structure. It will facilitate the task of terms and keywords extraction.

In the following, we will present our UML based patent document model for bibliographic and application data. The structure of the patent can be divided into two parts: bibliographic data taken from the cover page and application data from the rest of the patent document.

[4] INNORPI: National Institute for Standardization and Industrial Property, http://www.innorpi.tn

[5] Certificates of Patents granted by the GCC Patent Office secure legal protection of the inventor's rights in all Member States.

[6] GCC Patent Office: Gulf Cooperation Council Patent Office, http://www.gccpo.org/

Figure 1 shows the class diagram of the patent bibliographic data in which all associations are strong composition associations. It contains Bibliographic Data class which includes the Filing Number and Date, the Publication Date and Language and Classification of the patent. Bibliographic Data object is associated with one or more Title of Invention in different languages, zero or more Priority patent applications, one or more Inventor(s) and Applicant(s), zero or one Representative (agent) and zero or more International Publications (PCT).

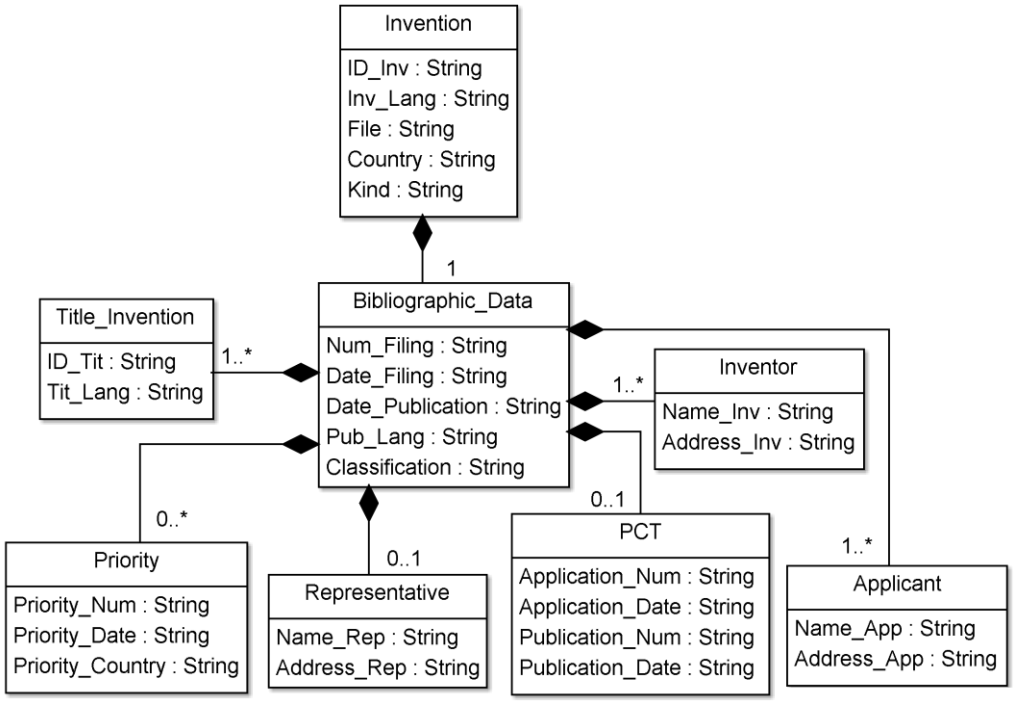

Figure 1. *The class diagram of patent bibliographic data.*

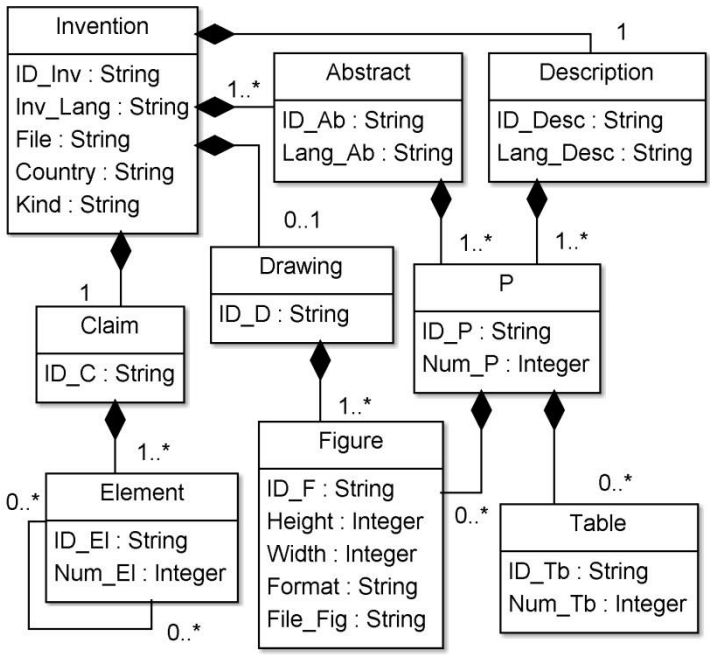

Figure 2. *The class diagram of patent application data.*

The Figure 2 shows the class diagram of the patent application data in which all associations are also strong composition associations, because, if a composite is removed, all of its component parts will be removed with it. It presents the association of the Invention class with one or more Abstract in different languages, one Claim and Description parts and zero or one Drawing part.
The two above presented diagrams allow us to introduce a DTD (Document Type Definition) for scientific Arabic patents, as shown on Figure 3. The role of the DTD is to precisely define the unique structure of Arabic patent documents, no matter the patent offices.

```
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT Invention (Bibliographic_Data, Description, Claim, Drawing?, Abstract+) >
<!ATTLIST Invention
ID_Inv CDATA #REQUIRED
Inv_Lang (AR|FR|EN) "AR"
File CDATA #REQUIRED
Country CDATA #REQUIRED
Kind CDATA #REQUIRED >
<!ELEMENT Bibliographic_Data (Title_Invention+, Priority*, PCT?, Applicant+, Inventor+, Representative?) >
<!ATTLIST Bibliographic_Data
Num_Filing CDATA #REQUIRED
Date_Filing CDATA #REQUIRED
Date_Publication CDATA #REQUIRED
Pub_Lang (AR|FR|EN) "AR"
Classification CDATA #REQUIRED >
<!ELEMENT Title_Invention ( #PCDATA ) >
<!ATTLIST Title_Invention
ID_Tit CDATA #REQUIRED
Tit_Lang (AR|FR|EN) "AR" >
<!ELEMENT Priority EMPTY >
<!ATTLIST Priority
Priority_Num CDATA #REQUIRED
Priority_Date CDATA #REQUIRED
Priority_Country CDATA #REQUIRED >
<!ELEMENT PCT EMPTY>
<!ATTLIST PCT
Application_Num CDATA #REQUIRED
Application_Date CDATA #REQUIRED
Publication_Num CDATA #REQUIRED
Publication_Date CDATA #REQUIRED >
<!ELEMENT Applicant EMPTY>
<!ATTLIST Applicant
Name_App CDATA #REQUIRED
Address_App CDATA #REQUIRED >
<!ELEMENT Inventor EMPTY>
<!ATTLIST Inventor
Name_Inv CDATA #REQUIRED
Address_Inv CDATA #REQUIRED >
<!ELEMENT Representative EMPTY>
<!ATTLIST Representative
Name_Rep CDATA #REQUIRED
Address_Rep CDATA #REQUIRED >
<!ELEMENT Description (P+) >
<!ATTLIST Description
ID_Desc CDATA #REQUIRED
Lang_Desc (AR|FR|EN) "AR" >
<!ELEMENT P (#PCDATA|Figure|Table)* >
<!ATTLIST P
ID_P CDATA #REQUIRED
Num_P CDATA #REQUIRED >
<!ELEMENT Figure EMPTY >
<!ATTLIST Figure
ID_F CDATA #REQUIRED
Height CDATA #REQUIRED
Width CDATA #REQUIRED
File_Fig CDATA #REQUIRED
Format (jpg|tif) #REQUIRED >
<!ELEMENT Table (#PCDATA) >
<!ATTLIST Table
ID_Tb CDATA #REQUIRED
Num_Tb CDATA #REQUIRED >
<!ELEMENT Claim (Element+) >
<!ELEMENT Element (#PCDATA) >
<!ATTLIST Element
ID_El CDATA #REQUIRED
Num_El CDATA #REQUIRED >
<!ELEMENT Drawing (Figure+) >
<!ATTLIST Drawing
ID_D CDATA #REQUIRED >
<!ELEMENT Abstract (p+) >
<!ATTLIST Abstract
ID_Ab CDATA #REQUIRED
Num_Ab CDATA #REQUIRED >
```

Figure 3. *A DTD for scientific Arabic patents*

4. Evaluation and Discussion

We did not have a collection of document in digital form because it is not the official in Tunisia for example. So we created our small collection of multilingual patents from various fields using our patents editor.

To cope with the large volume of patents data and metadata, we developed a patents terminological editor to generate TMF (ISO 16642:2003 – Terminological Markup Framework) terminological databases. This will enable us to facilitate the extraction and information retrieval tasks from the cover pages (metadata), and the other parts (data) of patents. Our terminological database contains terms of different technical and scientific fields and various patents with different structures. We can distinguish two categories of terms: the scientific and technical terms and the other terms. Scientific and technical terms in their turn were divided according to their technical and scientific fields.

The results of our terminological database are presented in Table 2 (only technical and scientific terms are counted). It concerns Tunisian and Gulf Arabic patents and it can be easily merged with other terminological databases. We hope that our terminology database will improve patent search.

Collection	**Number of patents**	**Number of terms**		
		Full text	**Cover page**	**Abstract**
INNORPI	28	236	25	63
GCCPO	30	302	312	96

Table 2. *Over view of the number of technical and scientific terms in our terminological database.*

5. Conclusion

Our main obstacle is that the structure of patents differs from an intellectual property office or institute to another in the Arabic world. The cover page of a Tunisian patent differs from the Egyptian or Moroccan patent cover page. We conducted a standardized modeling for Arabic patents based on the forms of patents published in the Arabic world. It provides us a single common patent document model. We developed a patent document editor to create a collection of Arabic patents having unified structure. In future, we plan to enlarge our patents collection and then our terminological database.

We will merge several terminology databases of patents. We aim to better extract information from a collection of multilingual scientific patents and to combine onomasiological and semasiological models. We are also developing a new annotation procedure, to annotate our learning and test collections.

We aim also to evaluate the relevance of our terminological database with an information retrieval system for patent documents, and to realize a prior art search for the collection of patents.

Information retrieval technics in multilingual patents are not lacking in previous works, we will test whether the results of this works remain valid if one expands the collection by documents into other languages (Arabic, for example), and if they will be affected by changing the type of the document collection, calculating noise, redundancy, cost, precision, recall, silence, etc.

References

Chihebeddine Ammar, Kais Haddar and Laurent Romary. 2014. A standard TMF modeling for Arabic patents. Terminology and Knowledge Engineering 2014, Jun 2014, Berlin, Germany.

Cui Binge. 2009. Scientific literature metadata extraction based on HMM. Cooperative Design, Visualization, and Engineering. Springer Berlin Heidelberg, 2009. 64-68.

Erik Hetzner. 2008. A simple method for citation metadata extraction using hidden markov models. In Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries (JCDL '08). ACM, New York, NY, USA, 280-284.

Huy Hoang Nhat Do, Muthu Kumar Chandrasekaran, Philip S. Cho, and Min-Yen Kan. 2013. Extracting and Matching Authors and Affiliations in Scholarly Documents. In Proceedings of the Thirteenth Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL'13), Indianapolis, ACM. 2013.

ISO 16642:2003. Computer applications in terminology: Terminological markup framework

Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. Proceedings of the 13th European Conference on Digital Library (ECDL), Corfu, Greece, 2009.

Patrice Lopez and Laurent Romary. 2010a. GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains. 7th international conference on Language Resources and Evaluation LREC'10, La Valette, Malte 2010.

Patrice Lopez and Laurent Romary. 2009. Multiple retrieval models and regression models for prior art search. 10th Workshop of the Cross-Language Evaluation Forum CLEF'09, Corfu, Greece, September 30 - October 2, 2009.

Patrice Lopez and Laurent Romary. 2010b. Experiments with Citation Mining and Key-Term Extraction for Prior Art Search. 11th Workshop of the Cross-Language Evaluation Forum CLEF'10, Padua, Italy, 2010.

Pattabhi RK Rao, Sobha Lalitha Devi and Paolo Rosso. 2013. Automatic Identification of Concepts and Conceptual relations from Patents Using Machine Learning Methods. Proceedings of the 11th international conference on Natural Language Processing, ICON-2013, Noida, India, December, 18-20.

Sagnik Ray Choudhury, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones and C. Lee Giles. 2013. Figure Metadata Extraction from Digital Documents. ICDAR 2013, 135-139.

Soonwoo Hong. Claiming what counts in business: drafting patent claims with a clear business purpose, SMEs Division, WIPO.

Walid Magdy, Johanne Levelin and Gareth J. F. Jones. 2009. Exploring Standard IR Techniques on Patent Retrieval. 10th Workshop of the Cross-Language Evaluation Forum CLEF'09, Corfu, Greece, September 30 - October 2, 2009.

Ying Liu, Kun Bai, Prasenjit Mitra and C. Lee Giles. 2010. Tableseer: automatic table metadata extraction and searching in digital libraries. Proceeding of the 7th annual international ACM/IEEE joint conference on Digital libraries - JCDL '07, 91-10.