A standard TMF modeling for Arabic patents

Chihebeddine Ammar, Kais Haddar, Laurent Romary

A standard TMF modeling for Arabic patents

Chihebeddine Ammar[1], Kais Haddar[1], and Laurent Romary[2]

[1] Department of Computer Science, University of Sfax,
Laboratory MIRACL, Route de Soukra, B.P. 1171, Sfax, 3000, Tunisia
ammarchihebeddine@hotmail.com kais.haddar@fss.rnu.tn
[2] INRIA & Humboldt Universität zu Berlin,
Institut für Deutsche Sprache und Linguistik
laurent.romary@inria.fr

Abstract. Patent applications are similarly structured worldwide. They consist of a cover page, a specification, claims, drawings (if necessary) and an abstract. In addition to their content (text, numbers and citations), all patent publications contain a relatively rich set of well-defined metadata. In the Arabic world, there is no North African or Arabian Intellectual Property Office and therefore no uniform collections of Arabic patents. In Tunisia, for example, there is no digital collection of patent documents and therefore no XML collections. In this context, we aim to create a TMF standardized model for scientific patents and develop a generator of XML patent collections having a uniform and easy to use structure. To test our approach, we will use a collection of XML scientific patent documents in three languages (Arabic, French, and English).

Keywords: TMF, Operability, Patents, Terminological databases

1 Introduction

Works on how to define a database's standard models are abundant in literature, especially in fields such as data warehouse. A standardized modeling terminological databases consists in integrating, homogenizing and giving terminology data a unique sense understandable by all users. It provides us a tool to integrate and merge terminology data from multiple source systems while improving terminology data quality and maintaining maximum interoperability between different applications. There are several standardized modeling terminological databases: TMF [4] and [5], TEI [11], etc.

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.

In fact, scientists inventors are the best to present the technical words of a field. Since, when drafting their patent applications, they will carefully choose words and named entities of a specific domain. In addition, patents may contain extreme examples of noise, deliberately vague and misleading wording for the

title, abstract and claims while maintaining relatively standard technical terminologies in the body description of the patent. But on the other side, in the same field, there is a risk that terms will be represented in different ways from one patent to another.

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.

Another motivation was to decide which standard will we choose to model our terminological databases, which standard will best represent patents terms and which approach to use, onomasiological or semasiological approach?

Patents are available in different formats: Full text, PDF document, set of images, XML, etc. They have heterogeneous components that require different modelings. Also, patents have linguistic structures like text and titles, and non-linguistic 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 allows 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.

We propose in this paper to treat the problem of extraction and operating information from a collection of Arabic scientific patents. In order to achieve this, we will propose a standardized model for multilingual patents and generate terminological databases from patent collections. It is an original idea because nobody treated terminology in Arabic patents in previous works.

Our learning collection includes a small number of multilingual patent documents. Each patent is associated with one of the three languages: Arabic, English, and French. In this paper, we will focus on the Arabic patents. Some of them have their translation in one (or two) of the other languages. Others have a translation of technical words or keywords of the invention and even a literal

[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.

translation in the same paragraph. These translations are usually of a very high quality because they are made by professional human translators.

This article is organized as follows. Section 2 is devoted to the presentation of the previous works. In section 3, we present our TMF standardized model for patents. Section 4 is devoted to the evaluation and discussion and we conclude and enunciate some perspectives in section 5.

2 Previous works

Information retrieval technics in multilingual patents are not lacking in previous works, the question is 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 documents collection, calculating noise, redundancy, cost, precision, recall, silence, etc.

2.1 CLEF initiative

The most recent previous works have been performed under the CLEF initiative on non-Arabic patents. The CLEF [12] initiative 2000-2014 (Conference and Labs of the Evaluation Forum, formally known as Cross-Language Evaluation Forum) promotes research and stimulates development of multilingual and multimodal IR systems for European languages. It provides tracks to evaluate the performance of systems for: for example, from 2009, the intellectual property: The aim is to encourage and facilitate research in the field of data mining in patents by providing a large database of experimental data. This database is formed of patent documents from the European Patent Office and it is called MAREC (MAtrixware REsearch Collection) which is a standard corpus of patent data available for research purposes. It consists of 19 million of patent documents in different languages (English, French, German) in a standardized XML schema highly specialized.

Previous works [2], [3] and [8] were mainly based on purely statistical approaches. They used standard techniques of information retrieval and data extraction. But there are others who have worked on purely linguistic or hybrid approaches. [6] and [7] used claims section as a bag of words and information source. In [9], the authors developed multilingual terminological database called GRISP covering multiple technical and scientific fields from various open resources.

2.2 Comparison between various Intellectual proprety offices

Patent applications are similarly structured worldwide. They consist of a cover page, a description, claims [1], 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 informations. The claims determine the patentability and define the scope of the claimed invention.

The European Patent Office (EPO) [14] offers inventors a uniform procedure of application, and a register of multilingual patents (English, French, German). In the Arabic world, there is no North African or Arabian Intellectual Property Office and therefore no uniform collections of Arabic patents. In Tunisia, for example, the INNORPI [13] (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, 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[2] for the Gulf Cooperation Council (GCC Patent Office) [15], there is only Arabic patents and there is an Arabic abstract in the cover page. The layout of the desciption 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 (e.g. Tunisian patents) is necessary.

3 TMF standardized model

The onomasiological terminological resources, usually go from one sense to the various embodiments of the term in different languages. Classic examples of onomasilogic dictionaries are thesaurus, synonym dictionaries, etc. The terminology is interested in what the term means: notions, concepts, and words or phrases that nominate. This is the notional approach. Ideally, a concept corresponds to one term and a term corresponds to a concept. Motivated from the industrial practice terminology , the Terminology Markup Framework (TMF, ISO 16642) was developed as a standard for these resources. This Standard also allows the modeling of lexical resources, but also contains a serialization, in this case, an XML format. So if we have onomasiological resources such as glossaries, thesaurus or **words networks**, we refer to TMF (ISO 16642) modelling.

In this paper, we treat patents as **networks of terms** and citations. So we consider that TMF is the most appropriate for patent modeling.

The meta-model (Fig. 1) of TMF is defined by a logical hierarchical levels. It thus represents a structural hierarchy of the relevant nodes in linguistic description. Each structural node or level can be described using basic or complex unit of information. The meta-model describes the main structural elements and

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

their internal connections. It is combined with data categories [10] from a data category selection (DCS). And finally, this model is matched with a model defined by the user. So we can appropriate the model according to our needs. The

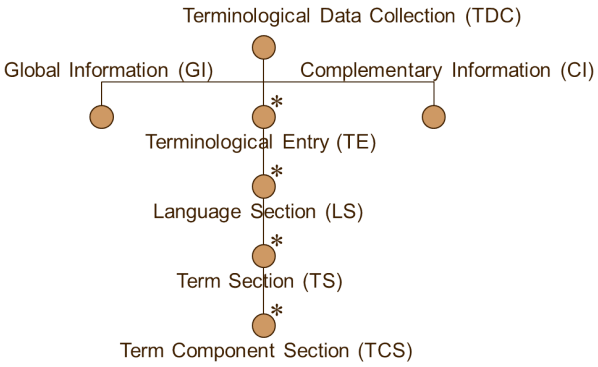

Fig. 1. TMF model.

aim of the meta-model is to act as a reference regarding possible interoperability requirements. So it defines a principle of interoperability between two Terminological Mark-up Languages which guarantees equivalence since they are based on the same set of data categories. This principle is guaranteed thanks GMT (Generic Mapping Tool) (Fig. 2) advocated by TMF to allow passage between two TMLs.

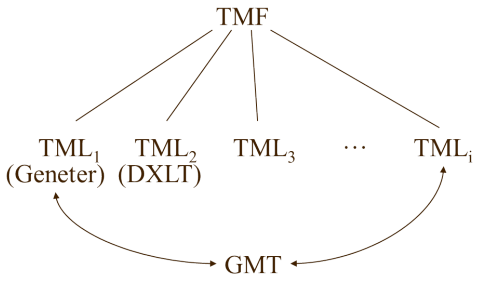

Fig. 2. GMT: Generic Mapping Tool.

In the following, we will present our TMF standardized model for bibliographic and application terminology. 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. Fig. 3 shows the class diagram of

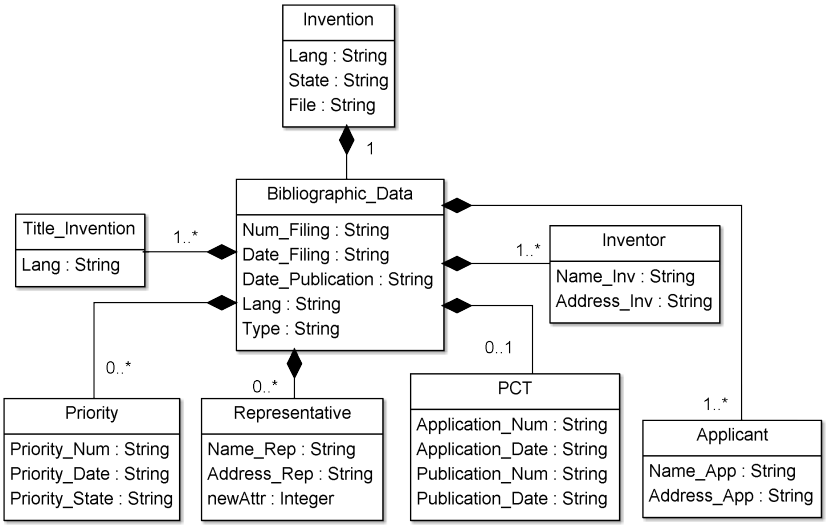

Fig. 3. The class diagram of patent bibliographic data.

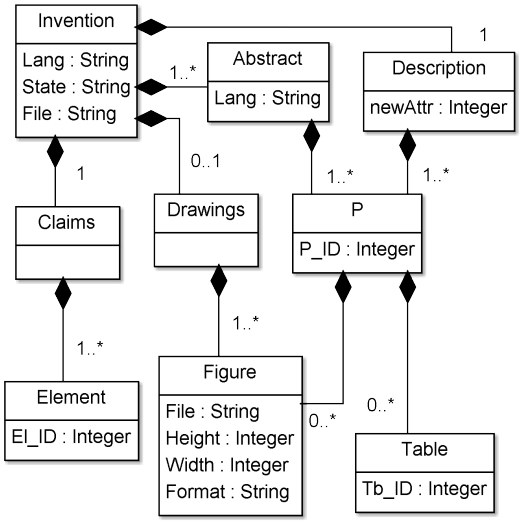

Fig. 4. The class diagram of patent application data.

the patent bibliographic data in which all associations are a strong composition associations. It contains, *Bibliographic Data* class which includes the *Filing Number* and *Date*, the *Publication Date* and the *Language* and *Type* of the patent. *Biblioghraphic 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* and zero or more *Internation Publications (PCT)*.

The Fig. 4 above shows the class diagram of the patent application data in which all associations are also a 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 *Claims* and *Description* parts and zero or one *Drawings* part. The two above presented diagrams allow us to introduce a DTD for scientific Tunisian patents. The DTD given in the following is much more simple and basic than European patents DTD. In fact, European patents contain more bibliographic data.

```
<!ELEMENT Invention (Bibliographic_Data, Description, Claims,
 Drawings?, Abstract+) >
<!ATTLIST Invention
  State CDATA #IMPLIED
  Lang (AR|FR|EN) "AR"
  File 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
  Lang (AR|FR|EN) "AR"
  Type CDATA #REQUIRED >
<!ELEMENT Title_Invention ( #PCDATA ) >
<!ATTLIST Title_Invention Lang (AR|FR|EN) "AR" >
<!ELEMENT Priority EMPTY >
<!ATTLIST Priority
  Priority_Num CDATA #REQUIRED
  Priority_Date CDATA #REQUIRED
  Priority_State  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+) >
<!ELEMENT p (#PCDATA|Figure|Table)* >
<!ATTLIST p P_ID CDATA #IMPLIED >
<!ELEMENT Figure EMPTY >
<!ATTLIST Figure
  Height CDATA #REQUIRED
  Width  CDATA  #REQUIRED
  File CDATA  #REQUIRED
  Format  (jpg|tif)  #REQUIRED >
<!ELEMENT Table (#PCDATA) >
<!ATTLIST Table Tb_ID CDATA #IMPLIED >
<!ELEMENT Claims (Element+) >
<!ELEMENT Element (#PCDATA) >
<!ATTLIST Element El_ID CDATA #IMPLIED >
<!ELEMENT Drawings (Figure+) >
<!ELEMENT Abstract (p+) >
```

A terminological record is a structured presentation that allows us to provide all information relating to a term in a clear and orderly way. Our terminological record includes: the entry identifier, subject field, definition, term, sub-term, synonym, example and abbreviation. Fig. 5 shows an example of bibliographic terminological entry (Title of invension) in the form of an XML document conforming GMT in the three languages (French, Arabic and English).

4 Evaluation and Discussion

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 TMF modeling for multilingual patents based on the forms of patents published in the Arabic world in general and precisely in Tunisia.

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 to generate our terminology database.

To cope with the large volume of patents data and metadata, we developed a patents terminological editor to automatically generate terminological databases.

```
<struct type="TE">
    <feat type="EntryIdentifier">601</feat>
    <feat type="SubjectField">Page de couverture</feat>
    <feat type="Definition">L'élément qui représente les données
        biblioghraphiques d'un brevet</feat>
    <struct type="LS">
        <feat type="Lang">Francais</feat>
        <struct type="TS">
            <feat type="Term">Titre de l'invention</feat>
            <feat type="Synonym">Intitulé du brevet</feat>
            <feat type="Example">Voiture écolo européenne</feat>
        </struct>
    </struct>
    <struct type="LS">
        <feat type="Lang">Arabe</feat>
        <struct type="TS">
            <feat type="Term">عنوان الإختراع</feat>
        </struct>
    </struct>
    <struct type="LS">
        <feat type="Lang">Anglais</feat>
        <struct type="TS">
            <feat type="Term">Title of invention</feat>
        </struct>
    </struct>
</struct>
```

Fig. 5. Terminological entry in the form of an XML document conforming GMT.

This will enable us to facilitate the extraction and information retrieval tasks from the cover pages (metadata), and the other parts (data) of patents. The results of our terminological database are presented in Table 1. 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.

Table 1. Over view of the number of terms in our terminological database

Collection	**Number of Patents**	**Number of terms**		
		Full text	Cover page	Abstract
INNORPI	28	6924	25	238
GCCPO	30	7632	312	224

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.

5 Conclusion

In this paper, we have proposed a TMF modeling for Arabic patents which provide us a single common data model for all terminological data and we developed a patents terminological editor to automatically generate terminological databases. In future, we plan to enlarge our patents collection and then our terminological database. Patent documents are often difficult to understand and have a variety of structures. So, we aim to 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. 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.

References

1. Hong, S.: Claiming what counts in business: drafting patent claims with a clear business purpose, SMEs Division, WIPO
2. Lopez, P., Romary, L.: Multiple retrieval models and regression models for prior art search, In CLEF 2009 Workshop. Corfu, Greece (2009)
3. Lopez, P., Romary, L.: Experiments with Citation Mining and Key-Term Extraction for Prior Art Search, In CLEF 2010 Workshop. Padua Italy (2010)
4. Romary, L.: An abstract model for the representation of multilingual terminological data: TMF Terminological Markup Framework, TAMA (2001)
5. ISO 16642:2003. Computer applications in terminology – Terminological markup framework
6. Verberne, S., D'hondt, E.: Prior art retrieval using the claims section as a bag of words, In CLEF 2009 Workshop. Corfu, Greece (2009)
7. Szarvas, G., Herbert, B., Gurevych, I.: Prior Art Search using International Patent Classification Codes and All-Claims-Queries, In CLEF 2009 Workshop. Corfu, Greece (2009)
8. Magdy, W., Leveling, J., Jones, G.J.F.: DCU @ CLEF-IP 2009: Exploring Standard IR Techniques on Patent Retrieval, In CLEF 2009 Workshop. Corfu, Greece (2009)
9. Lopez, P., Romary, L.: GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains, In Seventh international conference on Language Resources and Evaluation (LREC) 2010. La Valette, Malte (2010)
10. ISO 12620:1999. Computer applications in terminology Data categories
11. TEI: Text Encoding Initiative, `http://www.tei-c.org`
12. CLEF: Conference and Labs of the Evaluation Forum, `http://www.clef-initiative.eu`
13. INNORPI: National Institute for Standardization and Industrial Property, `http://www.innorpi.tn`
14. EPO: European Patent Office, `http://www.epo.org/`
15. GCC Patent Office: Gulf Cooperation Council Patent Office, `http://www.gccpo.org/`