

DESIGN OF PATENT DOCUMENT AUTOMATIC PROCESSING SYSTEM BASED ON SCHEMA

ZEYAO LI AND YAO LIU*

Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Beijing 100038, P. R. China
*Corresponding author: liuy@istic.ac.cn

Received March 2018; accepted June 2018

ABSTRACT. *The number of patent documents is large and wide, and it is of great significance to the standardization of patent documents. At present, the large-scale patent institutions around the world are in accordance with the structure of their patent data to deal with patented standards. However, they are not compatible with patent document in China, and the disclosure of patent documents is not profound enough. In view of that, this paper, based on the improvement of the defects of existing patented structured documents, designs the schema and function of patent document specification based on Schema, and applies it to the automatic processing of patent documents.*

Keywords: Patent document, Structure, Schema, Automatic processing

1. **Introduction.** XML Schema is a language used to describe and regulate the logical structure of an XML document. It describes the structure of an XML document based on an extensible markup language and defines a legal component group of an XML document. With its unique grammatical advantages, it allows for more rigorous and clear rules on XML documents. As the most effective carrier of technical information, the patent covers over 90% of the world's latest technical intelligence. The patent has a standard format and is easy to store, and how to use Schema effectively to express patent data standards has become an important issue. At the same time, in order to support the data exchange within the publishing agency, the publishing agency and the data processing unit, to meet the need for deep processing of literature, it is necessary to establish a unified data exchange platform: data exchange pool system, to realize the structured and unified processing of book, paper, patents and other scientific literature. The purpose of designing a patent document automatic processing system in this paper is to generate a standardized Schema document of patent, which can be conveniently stored and read, improve work efficiency, promote the informatization and modernization of patent management, and then provide support for constructing data exchange pools.

2. **Research Status.** At present, many institutions have promulgated and implemented the structured standards for patent data. The major patent agencies in the world generally adopt DTD to define the structure of patent documents, for example, the European Patent Office (EPO) in Europe and the United States Patent and Trademark Office (USPTO).

In the European Document Database (DOCDB), the root node of a patent document is named <exch:exchange-document>, which includes basic attribute information, the element about bibliographic data: <exch:bibliographic-data>, the element about summary: <exch:abstract> and the element about simple patent family information: <exch:patentfamily> [1]. The element patent document structure is shown in Figure 1.

The USPTO divides the U.S. patent database into two parts: the application public database and the approval announcement database. Correspondingly, two DTDs are also

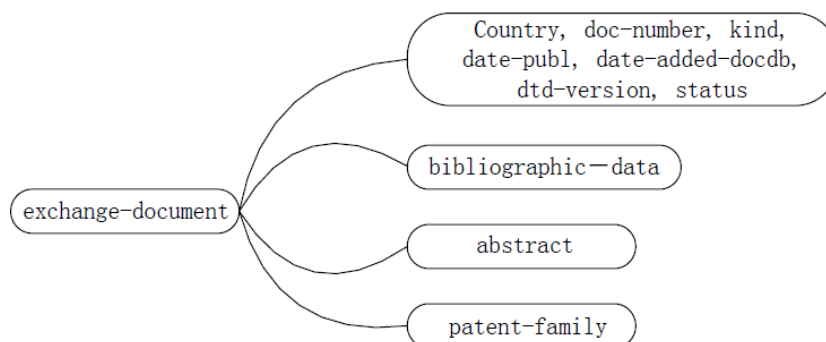


FIGURE 1. The structure of DOCDB patent

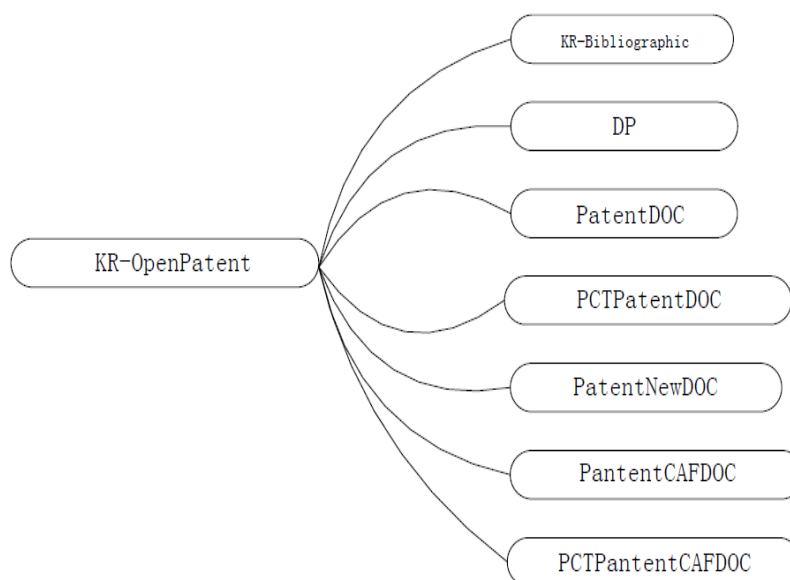


FIGURE 2. The structure of KIPO application invention patent

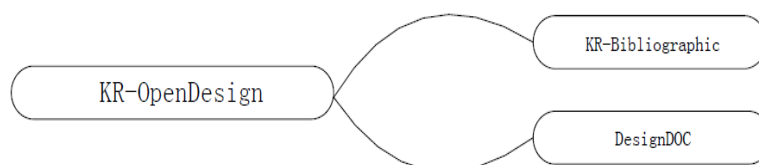


FIGURE 3. The structure of KIPO industrial design patent

provided: “application.DTD” and “approved.DTD”. The element about bibliographic data in the “approved.DTD” not only includes the elements in “application.DTD”, but also the corresponding elements for design patents, such as Locarno classification, cited patent information, cited non-patent information, field of patent retrieval, late or incapacitated inventor information, examiner information, publication information on international patent, and botanical information on plant patents [2].

In addition, the patent agencies of some other countries adopt XML Schema to define their own structure of patent documents, such as Korea’s KIPO (Korea Intellectual Property Office). Figure 2 and Figure 3 show the structures of the application invention patent and industrial design patent. In the Korean patent document, specifically, the sub-elements of the element about bibliographic data (KR-Bibliographic) include the publication information, registration information, application information, IPC classification information on patent, and the type, the name of patent document. The elements

about full-text data of the invention include the abstract, the application body, the claims, the accompanying drawings, and the like; the elements about full-text data of the industrial design patent include stereoscopic design drawings, graphic design drawings, design of charting and fonts and design drawings; specifically, the stereoscopic design drawings contain various view information, such as the left view, the right view, the front view, the back view, the elevation view, and the overlook view [3].

Through the analysis of the patent structure defined by the above major patent agencies, they have some common defects:

- (1) Single data type, cannot meet the comprehensibility of documents and the role of data exchange;
- (2) Although the DTD supports the sequential description of the element nodes, the DTDs must be implemented in various possible order of the elements. This method is not only cumbersome but sometimes even unrealistic;
- (3) DTD does not support the use of the namespace, the inheritance and reusability of documents are unsatisfactory;
- (4) The restriction definition of attribute is relatively vague, and it is difficult for ordinary users to understand;
- (5) DTD documents cannot be directly processed using XML tools, need to develop the processing tools and increase the work burden;
- (6) The division of data items is not detailed enough, the availability of full-text data is ignored, the text can be further parsed synthetically [4], semantic annotation [5], and providing knowledge product services [6].

In order to effectively solve the above problems, this paper designs a new structural model of patent documents based on Schema, describes the changes of data elements in the process of patents from the production to the failure, and parses the full text of the patent in more detail, and it can meet the needs of patent documents processing and service from two aspects of both form and content. At the same time, it allows the existence of multiple namespaces, has a wealth of embedded data types and powerful function of data structure definition, which can effectively achieve data inheritance and reuse, also easy to extend.

3. Design Idea. The design idea of this article is to construct a new structural model of patent documents based on the patent DTD of the State Intellectual Property Office using XML Schema. Based on this model, users can freely customize the Schema through system interaction. This customized model of patent documents based on Schema is not only standard and legal, but also meets the different needs of users for handling and displaying patent documents. According to the user's personalized Schema, the system parses the patent documents on demand and generates a structured XML file to store the patent information, which lays the foundation for subsequent storage and retrieval. The research framework is shown in Figure 4.

Through research, the State Intellectual Property Office has given some patent elements' normative definitions through DTD, and some did not specify [7]. The Schema proposed in this paper is partly converted from DTD of the element, and the other part is designed according to the standard of the State Intellectual Property Office for the bibliographic items of patent documents [8].

4. Key Technologies. On the basis of improving the defects of the existing patent structured documents, the standard specification of patent documents based on Schema is designed through the technical paradigms of element transformation, attribute transformation and entity declaration transformation.

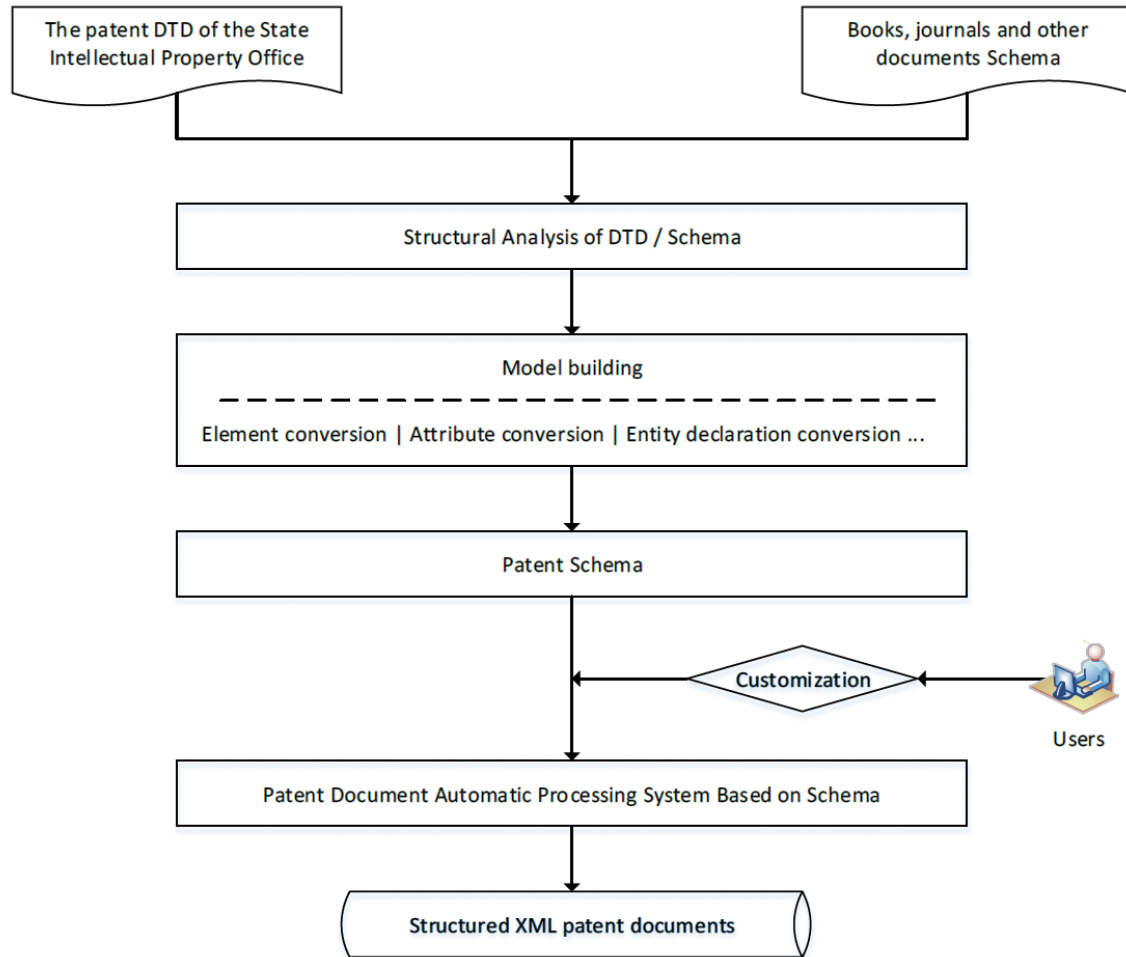


FIGURE 4. The research framework

4.1. **DTD to Schema.** (1) Element conversion. XML Schema uses `<element>` to define elements. In `<element>`, other sub-elements can be defined, and attributes can also be added to elements, such as name, type, minOccurs, and maxOccurs. Use `<sequence>` instead of “,” in the DTD, and `<choice>` in place of “|” in the DTD. Use the combination of minOccurs and maxOccurs to represent “?”, “+”, and “*” in the DTD.

As defined in the patent DTD published nationally, the definition of applicant data elements is as follows:

```
<!ELEMENT cn-applicants (cn-applicants-name, cn-applicants-address)>
```

In the Schema proposed in this article, the above information is equivalently expressed as:

```
<xsd:element name="cn-applicants">
  <xsd:complexType>
    <xsd:sequence>
      <xsd:element name=cn-applicants-name type="xsd:string"/>
      <xsd:element name=cn-applicants-address type="xsd:string"/>
    </xsd:sequence>
  </xsd:complexType>
</xsd:element name="cn-applicants">
```

(2) Attribute conversion. The DTD uses the keywords `#IMPLIED`, `#FIXED`, and `#REQUIRED` to specify whether attributes appear, and supports the definition of attribute defaults. XML Schema provides more explicit markup to achieve a clear and understandable representation. Schema obsoletes `#IMPLIED` and no longer supports the

implicit state of the attribute. Instead, the requirement must specify an explicit state, and the disabled attribute is used to indicate that the attribute is disabled. The expression of the default value is more intuitive and is given directly by default [9]. As in the definition of the patent element DTD published by the state, the definition of the element attributes of the preservation information of the microbiological sample is as follows:

```
<!ATTLIST bio-deposit
  id ID #IMPLIED
  num CDATA #REQUIRED
  url CDATA #IMPLIED
  dnum CDATA #IMPLIED
>
```

In the Schema proposed in this article, the above information is equivalently expressed as:

```
<xsd:attribute name="id" type="xsd.string" use="optional"/>
<xsd:attribute name="num" type="xsd.string" use="prohibited"/>
<xsd:attribute name="url" type="xsd.string" use="optional"/>
<xsd:attribute name="dnum" type="xsd.string" use="optional"/>
```

The above attributes are described in the document as shown in Table 1.

TABLE 1. Element attribute description table

Chinese name	English name	Type	Restriction	Repeatability
Identifier	id	String (50)	Optional	Non-repeatable
Numbering	num	String (3)	Disabled	Non-repeatable
Uniform Resource Locator	url	String (50)	Optional	Non-repeatable
Patent number	dnum	String (50)	Optional	Non-repeatable

(3) Entity declaration conversion. Schema does not support entities and therefore requires processing of entity declarations in the DTD. For parameter entities, they can be expanded and then converted into XML Schema; general entity declarations can be converted to element declarations in the Schema.

4.2. Schema description of patent documents. In the Schema model designed in this paper, the patent document element is represented by the node <cn-patent-document>, which includes two child elements <bibliographic-data> and <application-body>. An XML element can have properties in addition to other elements, text, or a mixture of the two. In order to describe the patent document elements more fully, three mandatory attributes are set: id, lang, and correction-code; and seven optional attributes: country, file-reference-id, doc-number, kind (invention, utility model, or design), date-produced, date-publ, and the status (public, authorized, substantive examination Effective, etc.).

The <bibliographic-data> contains 17 sub-elements, and their element names and meanings are listed in Table 2. Most of the elements in the table also contain child elements of the next level, for example, child elements of <publication-reference> include <publication-number>, <publication-date> and <publication-institution>; child elements of <partie-s> are <cn-applicants>, <cn-inventors>, <designers> (design patents only), <cn-paten-tee> and <agents>. Even these child elements can continue to nest their own child elements. For example, <cn-applicants> contains two child elements: <cn-applicants-name> and <cn-applicants-address>.

It should be noted that the element <classification-locarno> of number 6 is a design-specific element; the numbers 8, 10, 12, 13, and 17 are unique to the invention/utility model. In the patent application process, it is necessary to go through the following stages in order: application, review, authorization, invalidation, termination, and pledge. Under different legal conditions, there will be corresponding changes in the description of

TABLE 2. The child elements of <bibliographic-data>

Num	Element name	Meaning
1	<publication-reference>	Announcement/announcement data
2	<application-reference>	Application data
3	<correction>	Correct the data
4	<priority-claims>	Priority data
5	<classification-ipc>	International Patent Classification Data
6	<classification-locarno>	International Classification of Designs
7	<title>	name
8	<references-cited>	Citation
9	<division>	Divisional original application data
10	<cn-related-publication>	Announcement/announcement data related to the same application
11	<date-pct-article-22-39-fulfilled>	Enter the national phase date
12	<pct-or-regional-filing-data>	PCT International Application Related Data
13	<pct-or-regional-publishing-data>	PCT International Application International Published Data
14	<cn-domestic-priority-claim>	National priority data
15	<parties>	Parties in patent affairs
16	<cn-related-documents>	Other domestic application data related to this application
17	<bio-deposit>	Microbial sample deposit information

patent documents. For example, the patent-specific elements after authorization include the patentee data in 1 and 15.

<application-body> consists of four sub-elements: <abstract>, <claims>, <description>, and <design-figures> (design only). Similarly, they also contain their own children. For example, <abstract> includes <abstract-text> and <abstract-drawing>; <description> includes five child elements: <technical-field>, <background-art>, <disclosure>, <description -of-drawings> and <mode-for-invention>.

5. Platform Design. The document automatic processing system adopts a three-tier client/server architecture, namely a client layer, a service layer, and a data layer. The system is compatible with mainstream operating systems, and has good openness, ease of use, extensibility, and security. The system architecture is shown in Figure 5.

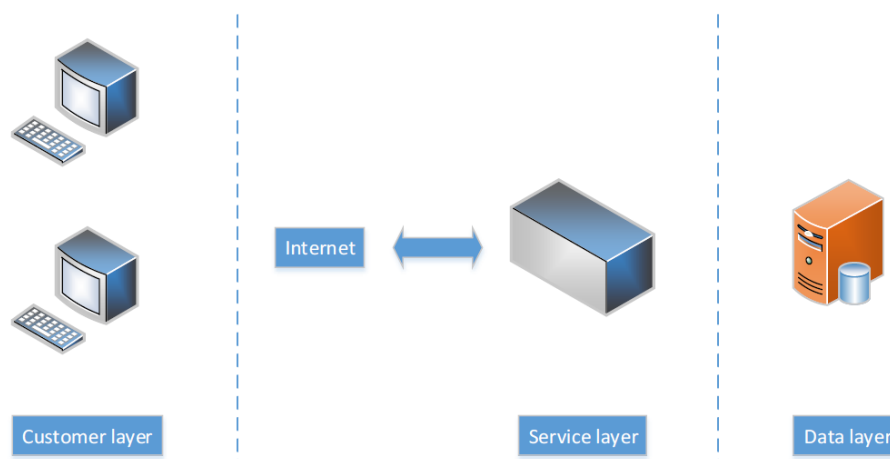


FIGURE 5. Three-tier architecture of the system

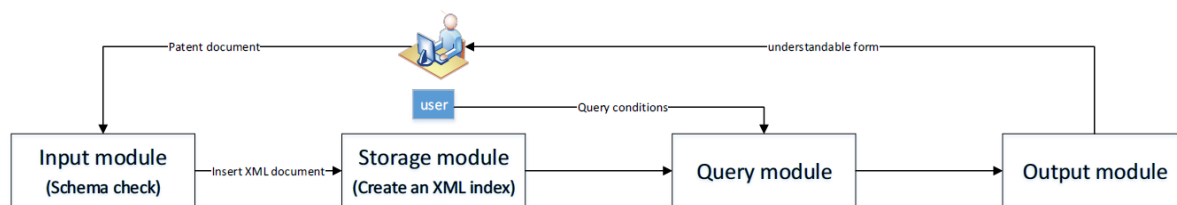


FIGURE 6. Basic module diagram

The role of the customer tier is to provide users with a friendly service interface, so that users can implement queries, edits, uploads, and downloads of system data through their browsers. After the server responds, it will be required in a way that the user can easily understand. Content is output to the client.

The service layer is the business logic implementation layer of the system. After receiving the request for user access, the data layer provides calculation, and finally returns the processing result to the user.

The data layer is used to store and manage patent XML documents, using the distributed document storage database MongoDB, the server provides the data management and query engine.

From the perspective of the system's modules, the system mainly includes four important parts: input module, storage module, query module, and output module.

The system has the following basic functions:

(1) Input of patent documents

The user uploads the patent documents to be processed through the interactive interface.

(2) Formatting of patent documents

In the input module, the user can customize the patent template and modify the patent document based on the Schema. Elements that conform to the definition of the Schema file are tagged according to the element name specified in the Schema of the patent document; items that do not match or lack are missing. After the patent document is converted into an XML document, it is stored in the database as a node.

(3) XML document storage

The processed document is indexed and stored as a text file in the database. An element index table is created with element nodes which have the same element name. In the index table, each record identifies a seven-tuple – (pre, post, parent_pre, depth, type, min, max) – as the representation of the node, where “pre” represents the sequence traversal number of the node, “post” is the sequence traversal sequence number, “parent_pre” represents the sequence traversal number of the parent element, “depth” represents the number of layers the element is in, “type” is the data type of the node, “min” and “max” represent the minimum and maximum number of occurrences in the document; in addition, a document total index table is created to store the relevant information such as the document identifier (id) and the document location (url) [10].

(4) Patent inquiries

According to the search formula, a patent document matching the <XML> node is searched in the database, and the query result is returned to the output interface in a way that the user can understand.

(5) Query result output

Choose the most suitable output form to display the result in the output interface, and also provide functions such as viewing and downloading. In order to realize the structured processing of scientific literature and facilitate data exchange, Schema parsing, storage, and data exchange functions for books, papers, and other resources have been realized.

6. **Conclusions.** In this paper, a new patent document structure model based on Schema is designed to improve the existing patented structured documents in view of the inconsistency of patent data structuring standards, the lack of disclosure of the content of patent documents and the inability to effectively implement data exchange and other issues. Based on the defects, through the technology paradigm of element quasi-commutation, attribute conversion, and entity statement conversion, Schema-based patent document specification standards were completed, user Schema was customized, and the basic framework and functions of the system were designed. Standards are used in the automatic processing of patent documents.

REFERENCES

- [1] X. Chen and W. Zhu, European Patent Office data collection processing and maintenance, *Chinese Inventions and Patents*, no.7, pp.71-73, 2006.
- [2] P. Qi, C. Huo and H. Liu, Characteristics of the U.S. patent system and element analysis of DTD bibliographic project, *Digital Library Tribune*, no.12, pp.44-50, 2013.
- [3] B. Li, L. Gao, F. Zhang et al., Research on Korean patent literature, *Journal of Standard Science*, no.9, pp.35-38, 2012.
- [4] Y. Liu, Y. Li and Y. Huang, Research on semantic and syntactic analysis of patent literature, *ICIC Express Letters*, vol.10, no.2, pp.471-477, 2016.
- [5] Y. Liu, H. Shi, D. Zheng and Y. Huang, Study on semantic annotation for professional literature, *ICIC Express Letters, Part B: Applications*, vol.5, no.5, pp.1383-1389, 2014.
- [6] Y. Liu, Y. Huang and Y. Wang, Research on the key technologies of pyrios knowledge service platform, *ICIC Express Letters, Part B: Applications*, vol.6, no.5, pp.1323-1328, 2015.
- [7] State Intellectual Property Office, *Interim Measures for Processing Patent Data of Inventions and Utility Models in China with XML*, 2007.
- [8] State Intellectual Property Office, *Patent Documents Bibliographic Criteria*, 2006.
- [9] H. Chen and N. Wang, Comparison and application of XML Schema and DTD, *Microcomputer Development*, vol.14, no.1, pp.66-68, 2004.
- [10] X. Xu, *Schema-Based XML Indexing Technology Research*, Master Thesis, Chongqing University, 2006.