*International Journal of Knowledge and Language Processing*
Volume 7, Number 4, **2016**

*www.ijklp.org*
KLP International ⓒ2016 ISSN 2191-2734
**pp.1--23**

# Review of XML Based on Patent Documents

Li Zeyao
Institute of Scientific and Technical Information of China
No.15 Fuxing Road, Haidian District
100038, Beijing, China
lizy2016@istic.ac.cn

Received December 2016; revised December 2016

**ABSTRACT**. *As an open standard for data representation, XML language has injected new vitality into web applications even network computing, it plays an irreplaceable role in the standardizing process of patent document data. This paper attempts to give an introduction of XML and it' s description documents -- DTD and XML Schema, and provide a series of contrast. In addition, the paper analyzes the role of XML language in the process of standardization of patent documents and gives a brief description of storage format and description elements of major patent institutions at home and abroad. Finally, the system and application of patent XML are introduced, and the development trend of XML technology in patent field is discussed.*

**Keywords:** XML; DTD, Schema; patent literature, system

1. **Introduction to MXL.** As the HTML language does not have the characteristics of large-scale web applications such as scalability, structure and data verification, it cannot meet the rapid development of Web technology, especially after the emergence of Java in 1995. For this, since 1996, a working group of the World Wide Web Consortium (W3C) has been working on designing a new language that extends beyond HTML. The language was later named Extensible Markup Language (XML).

XML is an optimized subset of SGML. SGML a language standard used to create markup language provided by ISO (International Organization for Standardization) in 1986. SGML provides a representing method for the publishing industry which can separate the data content from the display, making the data independent of the machine platform and the processing program. SGML is useful in many large publishing systems, but its complexity makes it difficult to apply directly to the Internet. HTML is a markup language designed specifically for publishing hypertext on the Web, it is not extensible and cannot meet the needs of many Web applications. In this context, XML has emerged as an optimized subset of SGML used on the Web.

XML is a way to store data across multiple platforms. Similar to SGML, XML is a meta-markup language that allows users to create new markups on demand, which is the extensibility of XML. A tagged element is a building block of an XML document, which can have several attributes and can contain zero or more child elements. These child elements can be text data or tagged elements.

1.1. **XML mode.**

(1) DTD (Document Type Definition)

An XML document can declare a DTD in its Document Type Declaration. A DTD is a syntax constraint on the markup and element structure that appears in an XML document, it can be used to validate an XML document. A DTD is a set of definitions for Element types, Attributes, Entities, and Notations. It defines the tags required for the document, such as the types of elements that can be used in the document and the possible associations between these elements.

(2) Schema

We know the DTD lack of constraint mechanism for the content and semantics of the XML document, which will limit the XML processor for effective type checking, application developers will have to write a special type of test code. Therefore, it is necessary to establish a constraint mechanism which is more comprehensive and more effective for XML, so that the XML processor can do better validity check, thus creating the XML Schema Language.

A schema document written in XML Schema Language defines the rules for the corresponding XML document, in order to constrain its data elements and their relationships. First of all, schema documents constraint XML documents more strictly from two aspects of data structure and data types and can define the rules, but DTD cannot, it limited the constraints of the XML document only from the structure. Second, DTD language has its own grammatical form, but XML Schema Language is actually an application of XML language (similar to the relationship between HTML and SGML language), in fact, the schema document is an XML document, you can use the XML tools to analysis. In this way, schema document can also use the existing DTD language to describe. The relationship between them is shown in the figure.

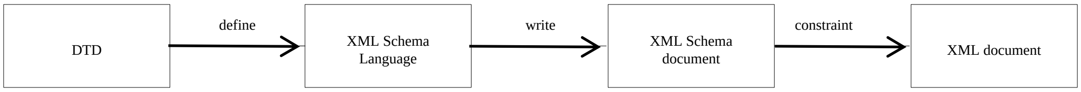

FIGURE1. THE RELATIONSHIP BETWEEN DTD, SCHEME, AND XML

1.2. **XML features.**

(1) Scalability

Just as Java allows users to declare their own classes, XML allows users to create and use their own tags instead of HTML's limited vocabulary. Scalability is vital, companies can use XML to define their own markup language for the applications such as e-commerce and supply chain integration. Even the specific industry can define the special markup language together, as the basis of data sharing and exchange in their field.

(2) Flexibility

HTML is difficult to develop, because it is a hybrid of formatting, hypertext, and graphical user interface semantics, it is difficult to develop these mixed functions at the same time. XML provides a structured data representation that separates the user interface from structured data. In XML, you can use style sheets to render the data to the browser, Such as XSL (Extensible Stylesheet Language, Extensible Stylesheet Language) and CSS2 (Cascading Style Sheets Level 2, Cascading Style Sheets, 2nd Advanced). In addition, hyperlinks between XML documents are supported by a separate XLink (Extensible Linking Language). All of these aspects can be independently improved and developed. Therefore, many of the advanced features that Web users are pursuing can implement easier in an XML environment.

(3) Descriptive

XML documents usually contain a document type declaration, so it is self-describing: not only a human can read XML documents, the computer can handle it, too. The data in the XML document can be displayed in any desired format by applications which can extract, analyzed and processed the XML data. The way that XML used is truly independent of the application system, and these data can be reused, so XML is suitable for open information management. Because of its self-descriptiveness, the data in the document can be created, queried, and updated by applications that use XML, similar to the method of processing data that used in traditional relational databases and object-oriented databases. XML can even be used to represent the data those didn't be regarded as a document but too difficult to deal with for the traditional database. Therefore, XML documents are seen as the database of the document and documentation of the data.

(4) Simplicity

XML is only 20% of SGML’s complexity, but with 80% of it’s function. XML is much simpler, easier to learn, easier to use, and easier to implement than complete SGML. In addition, the birth of XML also absorbed the experience of the people for using the HTML on Web for many years. It is worth noting that XML uses the Unicode character encoding system to support almost all major languages in the world, and text in different languages can be mixed use in the same document, XML-enabled software can handle any combination of all these languages.

All of this makes XML an open standard for data representation, which is independent of the machine platform, provider, and programming language. It will add new vitality to network computing and bring new opportunities for information technology.[1]

## 2. **About Schema.**

2.1. **What is a Schema?** Schema is a language used to describe and standardize the logical structure of an XML document. It is based on Extensible Markup Language (a subset of the

Standard Generalized Markup Language) and is used for the alternative document type definition (DTD). An XML Schema document describes the structure of the Extensible Markup Language document.

It is used to define a legal component group of XML documents, and an XML Schema defines:

The elements that can appear in the document;

The attributes that can appear in the document;

Which elements are child elements;

The order of the child elements;

The number of child elements;

Whether an element should contain text, or be empty;

Element and attribute data types;

The default and fixed values for elements and attributes.

2.2. **Schema syntax.** Schema has its own set of complete grammar, including the following aspects:

(1) The schema element is the first element in the XML schema that declares the XML document to be a schema document. Schema has two properties: name specifies the name of the schema; xmlns specifies the namespace that the schema contains.

(2) Element type is an important content in XML schema, there are two main types: complexType and simpleType. Elements of type complexType can include subelements and attributes, while elements of type simpleType can not contain subelements, nor attributes. Of course, XML also do the expansion of element type, we can more types attributes to do further description about the content and structure of the elements such as simpleContent, emptyContent, mixContent, anyType etc.

(3) The attribute element is used to define the attribute type that appears in the schema document. In the <attribute> include name (attribute name), type (attribute type), use and other attributes. “Use” is used to specify the default value of the attribute, "fixed", "optional" or "required" specifying the property has a certain default value, can have the default value and must be in the element.

(4) The group element is used to group the elements in an XML document. The order of the elements or sub-groups in the group can be specified by the attribute order, and minOccurs and maxOccurs specify the minimum and maximum number of occurrences of the grouping in the XML instance document.

(5) Choice is equivalent to the “|” in DTD. Only one element or group of elements that appears in <choice> ... </ choice> can appear in the corresponding XML instance.

(6) In XML Schema, sequence let the elements included in <sequence> ... </ sequence> become a series, and each member of the family appears in the corresponding instance is in the same order what it was in the series definition.[2]

(7) Type is an important element in XML schema, also a major feature of XML schema. It is used to specify the data type for elements and attributes. XML schema supports two data types: the basic data types defined in the XML 1.0 standard and some extended data types.

(8) Annotation element includes two sub-elements: documentation and appinfo, they are used to describe the information respectively of basic schema information and copyright information and tools, style sheets and other applications.

In short, XML schema adds more content than DTD，it can do more stringent and explicit provisions for XML documents and provides a prerequisite for the complete transformation of relational schemas to XML schemas.[3]

### 2.3. **Schema characteristics.**

2.3.1. **Advantages.** XML Schema is more powerful than DTD. Its syntactic structure is much more complex than DTD, but more expressive than DTD. It is more suitable for the application of various fields and can express the structural and semantic constraints of relational data better. The advantages include the following:

(1) Support a wealth of data types

One of the most important capabilities of XML Schema is the support for data types. It has built-in more than 40 data types, such as long, int, short, double and other common data types, through representing the data type by the value space, lexical space and facet three-part triples to get more flexible.[4]By supporting data types:

Can more easily describe the content what the document allows;

Can more easily verify the correctness of the data;

Can more easily work with the data from database;

Can more easily define data constraints (data facets);

Can more easily define data model (or data format);

Can more easily convert data between different data types.

(2) Use XML syntax

Another important feature of XML Schema is that they are written in XML, which realizes the consistency between the XML document and its description mode, and facilitates the data transmission and exchange. There are a number of advantages to writing XML Schema in XML:

Do not have to learn new languages;

You can use an XML editor to edit a Schema file;

You can use XML parsers to parse a Schema file;

Schema can be handled through the XML DOM;

Schema can be converted by XSLT.

(3) Can describe the element node order

Both XML DTD and XML Schema support the description of the order of subelement nodes, but XML DTD do not provide a description of the unordered case, that is, if an XML DTD is used to describe an out-of-order occurrence of a set of elements, it must use the way that exhaustive all the order that elements may appear to achieve, this method is not only cumbersome, and sometimes even unrealistic. XML Schema provides an <all> tag to describe this situation.

(4) The limitation on the property, default values and enumerations

The XML DTD specifies whether attributes are present with the keywords # IMPLIED, # FIXED, and #RE QUIRED, and supports the definition of attribute defaults. XML Schema provides more explicit markup for easy-to-understand representations. XML Schema obsolete XML DTD’s #IMPLIED, no longer support the implied state of the property, requires the need to give a clear state, and use prohibited to prohibit the property. For the default value of the expression is more intuitive, give it directly by default.

(5) Protect data communications

To protect data communications when data is sent from the sender to the recipient, the point is that both parties should have the same expected value about the content. With XML Schema, the sender can describe the data in a way that the receiver can understand. For example "03-11-2004", was interpreted in some countries as November 3, and in others as March 11. But an XML element with a data type, such as <date type = "date"> 2004-03-11 </ date>, ensures a consistent understanding of the content, because the format of XML datatype "date" is "YYYY-MM-DD".

(6) inheritance and reuse

New Schema can be constructed by taking certain types from existing Schemas, or invalidating any types when they are not needed. At the same time, XML Schema can be divided into separate components, so that when we write Schema, you can correctly reference the components which have been defined already. Inheritance makes the software reuse more effective, helping developers to avoid the situation that every creation must start from scratch, greatly reducing the XML software development process, facilitate the code maintenance, improve the programming efficiency.[5]

(7) Closely connection with namespace

The purpose of introducing namespace in XML is to enable the use of some generic definitions (usually definitions of some elements or data types) defined in other XML documents, and to ensure that no semantic conflicts arise. XML DTD does not support this feature, but XML Schema is very good to meet this point. In addition, XML Schema provides two methods to reference a namespace: include and import.

(8) Capture the error

Even if the document is in good form, there is no guarantee that they will not contain errors, and these errors may have serious consequences. Consider the following scenario: You ordered 5 dozen laser printers instead of 5. With XML Schema, most of these errors are captured by your validation software.

(9) Scalability

XML Schemas are extensible because they are written in XML. With the extensible Schema definition, you can:

Reuse your Schema in other schemas;

Create your own datatypes derived from standard types;

Multiple schemas are referenced in the same document.

2.3.2. **Disadvantages.** Although being written into XML is an advantage, it is also a disadvantage in some respects. The features of W3C XML Schema language may be very lengthy, but DTD is simple and relatively easy to edit.

XML Schema does not have the capabilities that provide most of the data elements to the document, but DTD has.

### 2.4. **Schema development.**

2.4.1. **P-Schema (Physical XML Schema).** In order to solve the problem of mapping between XML Schema to relational schema, Philip Bohannon proposed the concept of P-Schema (Physical XML Schema). P-Schema is an equivalent form of XML Schema because it is transformed by XML Schema. P-Schema extracts the multi-value elements from the original XML Schema in order to processe separately, generates a complex type of the same name, and retains references to the new type in its parent element's type definition.

However, P-Schema's extraction of elements is not complete. XML storage should extract elements that occur multiple times in an XML document, including the elements that recurring multivalued defined by minOccurs / maxOccurs, the elements that appear in a recursive form (nested), and the same element that is referenced more than once. The P-Schema extracts is not perfect because it only simple recurring multivalued elements but doesn’t take the extraction of other repetitive elements into account.[6]

2.4.2. **D-Schema (Deep XML Schema).** Based on the study of Philip Bohannon, puts forward the concept of D-Schema (Deep XML Schema) by analyzing the structure and syntax of XML Schema, which can extract the information of the elements from a deeper level.

①Optional elements: the element or element group in <choice> ... </ choice> can only have one in the corresponding XML instance, we call it optional elements. The elements with the value of MinOccurs is 0 and maxOccurs not more than 1can appear or not in the XML instance document, we also called optional elements. Optional elements can be extracted from the XML Schema, resulting in a new type definition;

②complex elements: complex type (complexType) defined elements known as complex elements. Because the complex elements in XML Schema can not only contain attributes and sub-elements, but also can use complex types to represent nested structures and the relations of recursive reference, so complex elements should be extracted separately;

③multi-valued elements: Additional attributes minOccurs and maxOccurs are used to limit the number of elements in the instance, multi-value element is the element which maxOccurs> 1;

④group elements: mainly refers to the elements defined by the group and attributeGroup, we group these elements as a class of complex elements.

D-Schema inherits the idea of extracting elements from P-Schema. The extraction of elements is mainly concentrated on multi-valued elements and complex elements. The nested structure and elements in P-Schema are still mixed together. D-Schema extends P-Schema to classify elements in the original XML Schema and extract new types such as optional elements, complex elements, multi-valued elements, and group elements. Compared with P-Schema, D-Schema has a deeper extraction of XML Schema, and there is no longer any need to extract components in D-Schema.

## 3. **Overview of patent document specifications.**

3.1. **Introduction to patent literature.** Patent as the most effective carrier of technical information, covering more than 90% of the world's latest technical information, compared to the general technical publications, it provides information as early as 5 to 6 years, it is accurate content, format specifications, easy to store, exchange and sharing. According to the World Intellectual Property Organization estimates, if you can effectively use the patent information, you will make enterprise research and development work to reduce the average technology development cycle of 60%, saving 40% of research funding.[7]

To the word patent, there are three meanings: one refers to the patent, the second refers to the patented invention and creation, the third refers to the patent literature.

Patent literature is the product of the patent system, but also an important basis for the patent system, it plays an important role in the patent examination and international exchanges.

Patent documents include inventions (utility models, designs), patent applications and specifications, as well as other types of documents relating to the invention and various search tools. The patent documents published in China mainly include:

(1) The Patent Gazette, the Utility Model Patent Gazette and the Design Patent Gazette;

(2) Disclosure of patent applications for invention, patent specification;

(3) Utility model patent specification;

(4) Annual Index of Patents.[8]

### 3.2. **Patent data structure.**

### 3.2.1. **Abroad.**

(1) European Patent Office

The European Patent Office (EPO) is the largest intergovernmental organization in the world that collects patent data. By the end of 2013, the European Patent Office has included the patent data from 93 countries and territories. Due to the wide range of information sources, Due to the wide range of information sources, heterogeneous patent data are processed, cleaned and standardized by the European Patent Office (EPO) before being used for DOCDB (Document Management Database), which is a format standard specifications of the patent information database used for data exchange with third parties, the WIPO (World Intellectual Property Organization) and other commercial organizations.[9]

①key attributes

1> data format attributes

DOCDB has a wide range of data sources, but the processing standards and specifications of the data sources are different at all, to maintain the original data characteristics and provide unified patent data, DOCDB retains the multiple expression of the same element. In order to solve these different processing standards, The DOCDB XML adds data format attributes to some of the data elements: data-format, which is used to distinguish the different representations of the same data.

The data format of the key elements is shown in the table: docdb is the standard data generated according to DOCDB normalization rules; docdba is the standard data processed according to the DOCDB level 2 normalization rules; epodoc is the standard data generated according to EPODOC standardization rules; Original indicates data provided by national IP offices or IP offices, and EPO does not guarantee the correctness and formativeness of the data.[10]

TABLE 1. THE DATA FORMAT OF THE KEY ELEMENTS OF DOCDB

| element | element meaning | data format | | | |
|---|---|---|---|---|---|
| | | docdb | docdba | epodoc | original |
| Publication-number | publication number | √ | | √ | √ |
| Application-number | application number | √ | | √ | √ |
| Priority-number | the application number of priority patent | √ | | √ | √ |
| Applicant-name | name of applicant | √ | √ | | √ |
| Inventor-name | name of the inventor | √ | √ | | √ |
| Title | title | | √ | | √ |
| Abstract | abstract | | √ | | √ |

2> Data State Properties

The European Patent Office has continuously adjusted the processing rules for DOCDB patent data, expanded the scope of data processing, resulted in multiple backtracking and revisions of the original patent records, or the deletion of original records due to the withdrawal of patent information, and the addition of new public Of patent records, very easy to lead to patent data processing and exchange difficulties. In order to prevent the above problems, DOCDB XML introduces the data processing status attribute of patent records, which can be used to characterize the status of update, addition or deletion of patent records, so that the data exchange party can distinguish patent records from this attribute. The current state of the patent data attributes are the following:

C: Add a new record representing the new patent record by DOCDB;

A: update the original record information, generally due to changes in processing rules and other reasons, record back and updated the original patent;

D: delete the original record, generally due to the patent records from the database was deleted or withdrawn;

CV: Add an empty record, DOCDB may update the data in future to provide more comprehensive patent information;

DV: Delete an empty record. DOCDB provides the maximum patent information for the record when adding (C) or updating (A) the patent record; and DOCDB XML file containing only the published literature number and literature of the patent record when the data status is D, CV or DV Type, and other simple information.

② key elements

Generally, a single DOCDB XML file contains several pieces of record about the exchange data of patent document. The one-piece patent document record information is defined by the element <exchange: exchange-document>, which includes the basic attribute information of the patent document record and the child elements of descriptive information <exch: bibliographic data>, the abstract element <exch: abstract> and the simple patent family information sub-element <exch: patentfamily>, as shown in the Fig.2.

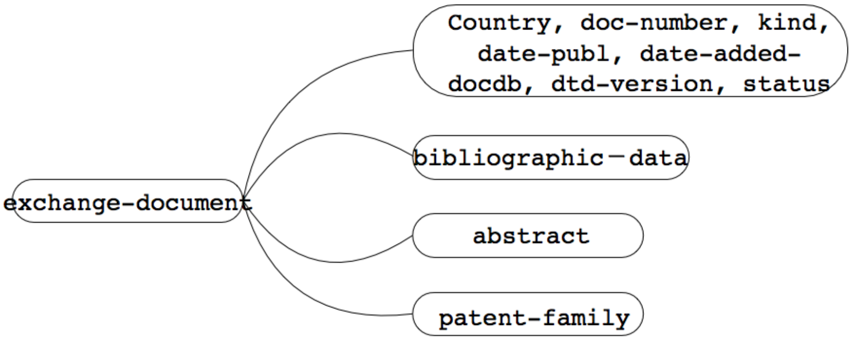
Country, doc-number, kind, date-publ, date-added-docdb, dtd-version, status
bibliographic－data
exchange-document
abstract
patent-family

FIGURE 2. THE BASIC STRUCTURE OF DOCDB XML SINGLE PATENT DOCUMENT

The basic attributes of a patent record include the country of the patent document (country), the publication number(doc-number), the kind of the document(kind), the date of publication (date - publ), the date when the record was included in the DOCDB(date-added-docdb), the DTD version (dtd- version), data status (status) and so on. According to the country, number, type, and date of publication of the patent document, the identification of the patent record can be determined; according to the data state attribute, the data acquisition mechanism can adopt different processing rules in the data processing.

Other key elements include the descriptive information element<exch: bibliographic-data>, the patent classification information elements <exch: patent-classifications>, the citation information elements <exch: references-cited>, the priority information elements <exch: priority-claims> , simple patent family <exch: patent-family> and abstract <exch: abstract>, etc.[11]

(2) the United States

The United States Patent Office was established in 1802, was a department directly under the State Council, to bear the patent-related matters; at the beginning of the 19th century, trademark affairs were also included in the jurisdiction of the Patent Office.In 1975, with the approval of Congress, the US Patent Office was renamed as the US Patent and Trademark Office (USPTO).

①The descriptive information elements of application.DTD

The main elements of the project include the application information, public information, priority information, title, patent classification (including IPC classification and USPC classification), inventor information, applicant information, international patent application information, patent information, agent or agency information and plant patent information. Its structure is shown in fig.3. [12]

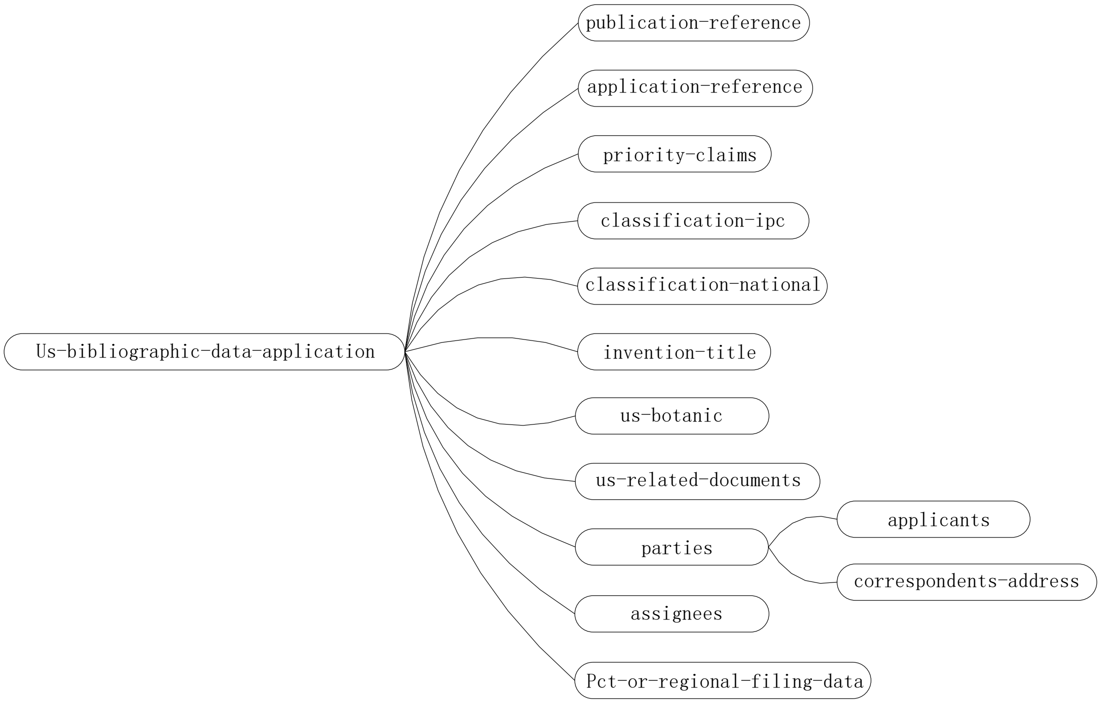

FIGURE 3. STRUCTURE OF APPLICATION.DTD_V4.0

②The descriptive information elements of grant.DTD

The storage format of descriptive information of grant.DTD in the US is not uniform until it achieves the changes from the SGML language to XML language in 2001 and the structure corresponds to the application.DTD at the beginning of 2005.

The descriptive information of grant.DTD not only includes all the elements in application.DTD, but also have the corresponding classification for designs: Locarno, cited patent information, cited non-patent information, patent search, and deceased or incapacitated the inventor information, the examiner information, the international patent disclosure information, the botanical information of the plant patent etc. Its structure is shown in fig.4.

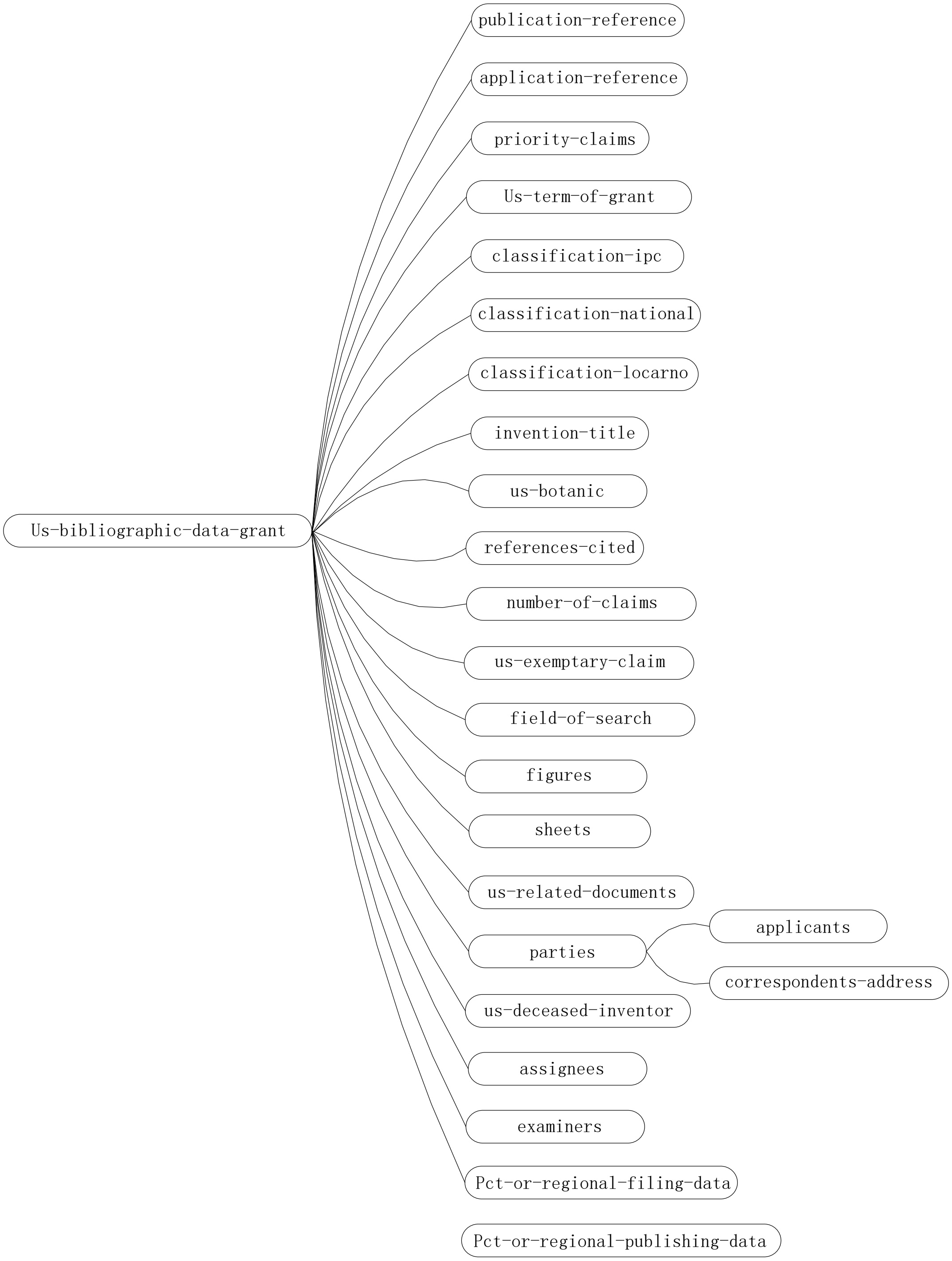

FIGURE 4. STRUCTURE OF GRANT.DTD_V4.0

(3) Korea

The data format of Korea's patent is XML, XML document structure mainly defined by schema, the patent document is stored and managed by the KIPO (Korea Intellectual

Property Office, South Korean Intellectual Property Office).

①The main elements

First of all, when the Korean invention and utility model are in the same state (as open state or the same as the authorized state), their structure is basically the same except names. Therefore, the following description will be given only by the way of inventions.

Fig.5 and Fig.6 are structural views of the elements of the disclosed invention and design. Authorized inventions and designs, whose structural drawings (not shown) are substantially the same as those of Figures 5 and 6, differ in that there is one additional element, authority-correction, that is, the change of information. This element is not required and often has no data in the actual sample.

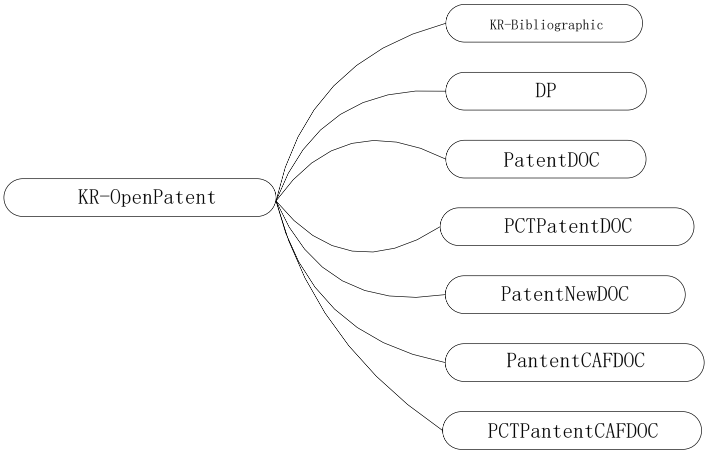

FIGURE 5. STRUCTURE OF KR-OPENPATENT

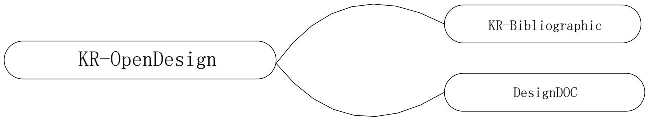

FIGURE 6. STRUCTURE OF KR-OPENDESIGN

②The descriptive information elements

1>Sub-elements of the KR-Bibliographic;

a.The unique elements KR-DomesticClassification and KR-BaseDesign of design. Korea designs use their own network classification number, the classification network is made up by of group, big category, medium category, small category and form classification, for example, C6-4490, D3-316A.

b.The unique elements KR-BaseDesign of design, meaning the basis of design, it is not common in the actual sample.

c.The unique elements LawTextCOntent and LawAmendContent of design, said the legal status. The public design does not have this element.

2>Sub-elements of the KR-Bibliographic

TABLE 2. THE ELEMENTS AND MEANING OF THE KR-BIBLIOGRAPHIC

| id | element | element meaning |
|---|---|---|
| 1 | KR—PublishingORG | publish organization information |
| 2 | KR—DocumentType | file type |
| 3 | KR—PublicationDate | publication date |
| 4 | KR—IPCInformation | IPC classification information |
| 5 | KR-OpenNumber | publication number |
| 6 | KR—RegisterNumber | registration Number |
| 7 | KR—RegisterDate | Registration Date |
| 8 | KR—ApplicationNumher | application number |
| 9 | R—ApplicationDate | application date |
| 10 | KR—ExaminationRequestDate | examination request date |
| 11 | KR—PublicationNumber | publication number |
| 12 | KR—OpenDate | open date |
| 13 | KR—0riginalApplicationKind | type of original application |
| 14 | KR 一 OriginalApplicationNumher | number of original application |
| 15 | KR 一 0riginalApplirationDate | date of original application |
| 16 | KR 一 0rignalExaminationRequest | original examination request date |
| 17 | KR-TranslationSubmitDate | translation submission date |
| 18 | KR—IntemationalApplicationNumher | international application number |
| 19 | KR—InternationalApplirationDate | international application date |
| 20 | KR—IntemationPublicationNumber | international publication number |
| 21 | KR—InternationPublicationDate | international publication date |
| 22 | KR—PriorityApplicationNumber | priority number |
| 23 | KR—PriorityApplicationDate | priority date |
| 24 | KR—ApplicationCountry | priority country |
| 25 | KR—ApplicantInformation | information of applicant |
| 26 | KR—RightHolder | patentee information |
| 27 | KR-Inventor | inventor information |
| 29 | KR—PriorArtdocuments | earlier technical documentation |
| 30 | KR—Examiner | name of examiner |
| 31 | KR—AppealInformation | appeal information |
| 32 | KR_TechnologyTransfer | technology transfer information |
| 33 | KR—NationResearchProjectAsistOflnvetation | National Development Plan |
| 34 | KR—LanguageCode | application language |
| 35 | KR—ClaimCount | number of claims |
| 36 | KR—InventionTille | title |

③full-text text elements

The structure and content of the full-text text elements of the same invention, utility model or design are fully consistent in both the public and the authorized states.

1>Sub-elements of the full-text text element of PatentDOC in the invention or utility model

TABLE 3. THE ELEMENTS OF THE PATENTDOC IN THE INVENTION OR UTILITY MODEL

| id | element | element meaning |
|---|---|---|
| 1 | Abstract | abstract |
| 2 | ApplicationBody | applicant body |
| 3 | Claims | claims |
| 4 | drawings | specification attached figure |
| 5 | SequenceList | biological sequence information (invented only) |

2>Sub-elements of the full-text text element of DesignDOC

TABLE 4. THE ELEMENTS AND MEANING OF THE DESIGNDOC

| id | element | element meaning |
|---|---|---|
| 1 | CubicDesignDrawings | Dimensional design information |
| 2 | PlaneDesignDrawings | Graphic design information |
| 3 | TypeFaceDrawings | Cartographic font design |
| 4 | DesignDrawings | Design drawings |

CubicDesignDrawings also contains a variety of view information, such as the left view, right view, front view, rear view, bottom view, top view and so on.[13]

### 3.2.2. **domestic.**

(1) SGML and XML

SGML is not a formatting language, even not a specific markup language, but rather a set of specifications that allow people to create their own markup languages.

The content identifier it specifies makes it easy to format text consistency, enabling the document management system to quickly locate information. SGML is well suited for projects involving a large number of structurally similar data, allowing developers to easily develop data structure specifications, create document type definitions (DTD), and then apply them to documents throughout the organization. Although it is a very powerful language, but both complex and extensive, while achieving and maintaining it is very expensive. And SGML DTD creation is also very complex. Only one of the simplest applications of SGML - HTML - has attracted widespread public attention, but HTML contains only a bit of the power of SGML. Some aspects of SGML have also become obsolete and already have become an obstacle to its widespread use.

XML is a subset of SGML, which allows information providers to define markup and attribute names on their own and structure the information content so that the structure of the XML file can be complex to any degree. XML gives Web-based applications the function and flexibility they need. Since XML is an open text-based format, it can be delivered using HTTP, without the need for changes to the existing network. Because XML separates the data from the display, the processor can nest procedural descriptions in structured data to show how to display the data. This is an incredibly powerful mechanism that minimizes client computer interaction with users while reducing server and the time that browser data exchange response, greatly enhancing the scalability of the server. In short, XML has the potential for development, suitable for the network.[14]

(2) DTD or Schema?

①DTD

1> DTD defect

a.DTD is not a well-formed XML, and use the different syntax with XML, so we need the different parser and API for DTD and XML.

b. DTD has very limited support for defining new data types and lacks the support for namespaces.

c. Using DTDs to represent certain types of document structures is very difficult.

2> DTD status quo

Although XML Schema can meet the demand of more and more areas, but in the short term, there are still some advantages of DTD:

a. At present, most of the XML-oriented applications has done a very good support for DTD, under normal circumstances, the ways of upgrade of these applications and tools will not use schema to replacement DTD, more options should be both supported. Of course, for those who require high data exchange or description capabilities, DTD has been unable to meet the functional requirements, the schema to replace the DTD has become an inevitable trend.

b. Most of the current algorithms related to XML schema are based on DTD. As a continuation of research, DTD research will not be abandoned easily. However, schema research will become a new hotspot.

c. In some relatively simple processing environment requires the DTD will still occupy its place.[15]

②Schema

A very important aspect that Schema is superior to DTD is: Schema is extensible; developers can according to their own needs to expand the data type. At the same time, Schema provides a wide range of types of inheritance support, allowing reuse and expansion of the defined structure. In this way, the same type can be reused for structurally coincident or mostly coincident data elements. For example, to define an element type DateType, the element ActionDate and the element Birthday can all use the same type. In this way, when the structure changes, you can only alter the type definition, do not have to modify all the relevant elements one by one.

In addition, the flexibility to use attributes can reduce element redundancy, such as: the original element <ChemicalTradeName> and <ChemicalTradeName_PinYin> were used to represent chemical substances in English and Chinese Pinyin, because the internal structure of the two elements are completely the same, so you can merge <ChemicalTradeName_PinYin> into <ChemicalTradeName>, just add a lang, and set the optional value as "PinYin". By such a process, the number of elements is reduced, and the patent service personnel can make less troublesome when dealing with patent data elements.

(3) XML representation of Chinese patent data elements

China's patent data elements can be divided into two major parts according to their functional divisions: basic class elements and business class elements. The basic class element is widely used in the XML representation of the patent data element. It can only express the exact data product information when it is called for other elements. Its meaning changes with the context, such as date, Country, paragraph, address and so on.

Business class element is the functional data element, which is corresponding to certain data item, expresses the certain information of IP data, so that it has certain function or purpose, such as application date, application number, public number, manual and so on.

The base class element and the business class element can be well distinguished by Schema-specific namespace mechanisms. The base class elements are prefaced by the namespace prefix base, while the business class elements are led by the namespace prefix business.

The representation of the Chinese patent data element is realized from the description of the following eleven attributes, and the description of the eleven attributes has its unique set of standard formats. The representation of Chinese patent data element shown in Figure 7.

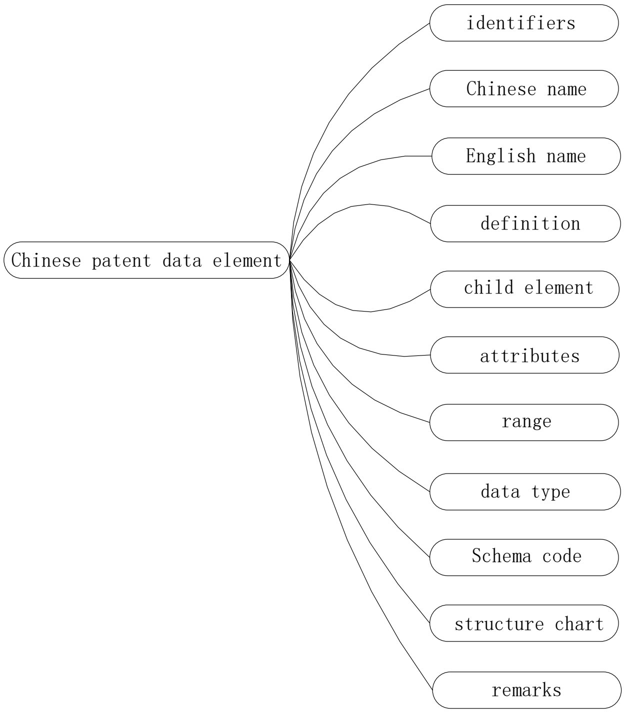

FIGURE 7. THE REPRESENTATION OF CHINESE PATENT DATA ELEMENT

Through the representation of identifiers, Chinese names, English names, definitions, sub-elements and attributes, the complete features of a Chinese patent data element can be abstractly abstracted from the logical level. The Schema code and structure chart is the XML materialization of these representations, that is, the XML representation of the patent data elements.

Through the above way, a patent data element can be fully expressed by the XML mechanism, the combination of these patent data elements can be the final complete expression of the relevant content of the patent document to facilitate the data transfer of intellectual property rights parties. At the same time, the "element" also makes the patent data retrieval and other aspects become more accurate and efficient.

## 4. **Examples of patent XML applications.**

### 4.1. **The science and technology literature push system based on XML.**

4.1.1. **System Architecture.** In view of the shortcomings of the traditional science and technology literature service system, XML technology is used as the data pushing medium in the science and technology literature push system, which makes it convenient for the client and obtain the scientific and technical documents according to their own needs, achieve a compatible platform for different systems, improve the service efficiency of the system. The main function modules of the system are as follows.

(1) Security certification: to obtain the resources of the user identity verification.

(2) Resource search: mainly provide the title, abstract, keyword, date of publication, document type, author and other aspects of scientific and technological literature retrieval, through resource retrieval users can obtain the information according to their own needs.

(3) Custom system: including information customization and customs management of a template, the user can customize their own information resources required according to their personal preferences and concerns of information, including journals, papers, patents, scientific and technological information. Through custom management, the information they have customized can be a query, delete and organize, at any time to adjust their own custom information.

(4) system: According to the needs of customization, retrieval of scientific and technical literature database, the search results will be converted into a batch of documents in line with the requirements, then pushed to the customer by e-mail.

(5) resource management and acquisition system: the push system push the scientific and technical literature into the resource management and acquisition system, according to the relevant data link to obtain full text of scientific and technical information and classify them.[16]

4.1.2. **The definition of XML document delivery format.** With DTD, you can exchange data and independently validate your own data with a standard that is used by independent groups all the time. The structure of scientific literature is neat, logical, regular and related, and it is easier to structure and semantic analysis than other styles. Based on the format which XML technology literature pushed the data, using the DTD specification definition, the format is as follows:

```
    <?XML version=”1.0”?>
    <DOCTYPE Article[
        <!ELEMENT Aticle(title,keyword,author,abstract,type,from,address,date)>
        <ELEMENT title(#PCDATA)>
        <ELEMENT keyword(#PCDATA)>
        <ELEMENT author(#PCDATA)>
        <ELEMENT abstract(#PCDATA)>
        <ELEMENT type(#PCDATA)>
        <ELEMENT from(#PCDATA)>
        <ELEMENT address(#PCDATA)>
        <ELEMENT date(#PCDATA)>
  ]>
    <Article>
    <title>title</title>
    <keyword>keyword</keyword>
    <author>author</author>
    <abstract>abstract</abstract>
    <type>data type</type>
    <from>data source</from>
    <address>download link</address>
    <date>release date</date>
    </Article>
```

### 4.2. **Chinese patent search system based on Tamino.**

4.2.1. **Tamino database.** Tamino (Transaction Architecture for the Management of INternet Objects) is a native XML document storage and management information server developed by Software AG, Germany. It is a complete database management system supporting Web, providing an integrated environment for digital resource exchange and application. Its main function modules include:

(1) X-Machine: Is the driving engine of Tamino, its basic function is to store XML objects and query them from different data sources. It is based on the schema that the administrator defines in the data map to handle storage and querying. XML objects are stored in Tamino, but Tmnino also provides an internal SQL store and supports other database types.

(2) Data mapping: Tamino is the knowledge base. These schemas are primarily governed by the administrator and contain rules for storing, querying and organizing XML objects. The development of these patterns can be supported by graphical tools and maintain the correctness of their pattern tree.

(3) XML parser: The XML objects stored in the X-Machine will be described by the schema stored in the Tamino data map. The XML parser within the X-Machine will check the syntax of the schema to ensure that the loaded XML object is "Well-formed".

(4) Object Processor: When Tamino stores data, the object processor stores the XML

data in the XML store using the relevant schema information provided in the data map.

(5) Query Interpreter: The query language of Tamino is X-Ouery.The query interpreter breaks down the query request and interacts with the object synthesizer by retrieving the object in a schema stored in the data map.

(6) object synthesizer: use the storage and query rules which is defined in the data mapping.

Tamino database's main technical advantages:

(1) Designed an extended XML structure for storing and processing XML data, supporting input and output of XML document format, providing the storage of native XML to improve the performance of XML document processing.

(2) Data query is based on XML-related standards such as Xquery, Xpath, XSLT, DOM and SAX established, providing a fast and efficient query response.

(3) The basic storage unit of the system is the XML document, which provides the completeness and consistency of XML data representation.

4.2.2. **Tamino patent search system's main function.**

(1) Data loading: According to the XML-based patent project standard DTD, the entry data and the patent literature full text data are converted into XML data. And the transformed XML data is loaded into the Tamino 3.11 database, providing a single record to add and bulk block loading. At the same time, through the full - text search engine based on TRS 4.5, an index library of full - text search of patent documents is set up for searching full - text patent data and complex units.

(2) Data management: Tamino 3.11, a pure XML database, is used as the back-end patent document base database, which stores and manages the huge XML format patent document base data (nearly one million patent documents). The system provides a simple and friendly data management module for the database administrator. The module can provide database administrators a management tool with the function including data validation, modification, deletion, backup, recovery, access control and log management. The data query provides a standard XML query language, Xquery or Xpath, which makes it easy for the administrator to flexibly query the data stored in the tree structure in the database. The data modification supports the query to the node-level, record-level data and batch modification. Data deletion Supports node-level, record-level data by article or batch delete. The data backup supports the automatic backup and manual backup of data, the backup mode includes full backup, incremental backup and phase backup. The data recovery supports the function of restoring the backup data into the database. The access control supports the user management function, and can be used for each user permission settings. Log management system log on the user and its operating content are automatically recorded.

(3) User management: This module provides a system management administrator, including access control and multi-level user management mechanism. Using hierarchical user management mechanism, each level of users have different administrative privileges. Including super administrators, administrators, and regular users.

(4) Data retrieval: The pure XML database Tarrino is combined with full-text search

engine to provide not only XML-based query mechanism, but also full-text search engine based on TRS as the index layer of the system. The basic search methods provided by the system to the public are classified into IPC (International Classification Standard for Patent Literature) Classification Search, Form Search, Expression Search, and Complex Elements of Chemical Documents (Chemical Expression and Mathematical Expression) Search and Search History. Each search supports the secondary retrieval, synonym retrieval, and retrieval results can be retained as a search history, in which the complex unit retrieval method to Applet embedded in the browser.[17]

### 4.2.3. **Effect evaluation.**

(1) To achieve the patent search system data loading and publishing functions, management of the huge number of documents at the same time, to ensure the consistency and integrity of the data being managed.

(2) Data on the management of the patent system include the data of the three parts (the full-text content is not included in the appearance patent record): patent full-text data in XML form; patent entry data and legal status data in TRS file form.

(3) Since the data has been proofread and archived, the integrity and consistency of the data is relatively easy to manage when the system is loaded in bulk, so the loader can keep the data in the Tamino database consistent with the TRS database data.

## 4.3. **A patent map based on XML Schema.**

### 4.3.1. **Production methods.**

(1) XML Schema patent file collection

Selecting a target technical area to be studied, and determining the patent subject and significance. Then, in the patent database, search the target area of the relevant patent documents. Finally, the retrieved patent documents are stored in the XML Schema file format, and the target technical field XML Schema patent file database is established.

(2) Patent unstructured item analysis

①Patent literature keyword extraction. First, extract the contents of the <Abstract> ... </ Abstract> tags in the XML Schema file. Then, the Chinese Academy of Sciences ICTCLS on the label of the content of the word, in order to obtain patent keywords. Then, the keywords are filtered, the initial keyword list is set up. Finally, the keywords with high frequency and technical significance and the results of expert visit are combined to sort the keywords and get the keyword list.

②Patent document clustering.

Different clustering algorithms, time complexity and clustering results are different. When the number of clustering documents is not too large and the clustering accuracy is high, the hierarchical clustering method can be used for clustering analysis. According to this, the key word matrix of the patent document is established, and the keyword clustering of the patent document is established.

③Key words semantic network construction

Semantic Networks (Semantic Networks) is a node and the relationship between the arc to represent the knowledge of the directed graph. According to the analysis of keywords and the clustering results of patent documents obtained by hierarchical clustering method, the keywords are represented by nodes, and the relations between keywords are expressed by directed arcs, and semantic networks can be built.

(3) Analysis of the structure of the patent

According to the acquired XML Schema document, the content of the <Application Date> ... </ ApplicationDate> tag in the document is regarded as the patent structured item, and the patent application date of the retrieved patent document is obtained.

(4) Reset keywords semantic network

The semantic network construction of keywords only takes into account the unstructured items of XML Schema files. Now, the semantic network is reset according to the unstructured items - date of application of XML Schema files.

(5) The formation of the patent map

According to the completed semantic network, the patent map is constructed. The application date is taken as the x-axis direction, and the frequency of the keyword is taken as the y-axis direction. Each node of the semantic network is rationally arranged to complete the patent map making.[18]

4.3.2. **Evaluation.**

(1) The method of making patent map based on XML Schema is put forward for the first time, and patent information of patent map is made by using XML Schema patent document structured item and unstructured item.

(2) The key word semantic network method is constructed by using unstructured item of XML Schema patent file, but this method is more complicated in practice. Therefore, how to optimize keyword semantic network construction method according to the characteristics of XML Schema, how to quickly form key words semantic network remains to be further studied.

5. **Conclusions.** Based on the above analysis, XML Schema has played an important role in the process of standardization, storage, transmission and analysis of patent documents by virtue of its consistency and unique advantages. Schema has more expressive power than DTD, can meet more needs, in most applications have a certain application. However, as the same with the development of other technologies, although the role of DTD will gradually weaken, but it will continue to exist and play a role. In the future, more systems and applications will support both DTD and schema, so that the XML language in the field of patent literature data to achieve a wider range of applications.

**Acknowledgment.** This work is supported by National Digital Composite Publication System Project (XWCB-ZDGC-FHCB/29). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

## REFERENCES

[1] QU Yuzhong, ZHANG Jianfeng, CHEN Zhen, WANG Conggang, A survey of XML and related technologies, *Computer Engineering*, vol.26, no.12, pp.4-6, 2000.

[2] LIU Zhi, SHI Rui, XIE Xinquan, D-Schema-based mapping from XML Schema to relational Schema, *Computer Applications*, vol.24, no.2, pp.128-131, 2004.

[3] FANG Xiang, LI Weisheng, Mapping from relational module to XML module, *Journal of Computer Applications*, vol.19, no.1, pp.130-132, 2002.

[4] LIU Zhengmin, NIU Yanfang, A survey of XML-related technologies, *Modern Information*, vol.08, pp.57-59, 2003.

[5] WANG Ru, SONG Hantao, XML document' s structure defining specification—XML Schema, *Journal of Computer Applications*, vol.19, no.1, pp.127-129, 2002.

[6] Philip Bohannon, Juliana Freire, Prasan Roy , Siméon J, From XML schema to relations: a cost-based approach to XML storage, *Proceedings of the 18t h International Conference on Data Engineering*, 2002.

[7] WU Xinyin, LIU Ping, Research on patent map, *Research and Development Management*, vol.15, no.5, pp.88-92, 2003.

[8] ZHANG Ping.What is the patent literature, *Locomotive & Rolling Stock Technology*, no.2, pp.18, 2003.

[9] CHEN Xuefang, ZHU Wei, Data collection, processing and maintenance of the European patent office, *Chinese Invention and Patent*, no.7, pp.71-73, 2006.

[10] ZHANG Fan, YIN Wenyuan, PENG Lei, CHEN XIaoyu, The application of DOCDB patent data in patent retrieval system, *China Standardization*, no.9, pp.42-49, 2012.

[11] LIU Huijing, ZHU Xinchao, QI Ping, JIANG Jun, Analysis on attributes and elements of EPO XML patent data, *Science and Technology Management Research*, no.17, pp.156-160, 2015.

[12] QI Ping, HUO Cuiting, LIU Huijing, Characteristics of US patent system and element analysis of DTD description project, *Digital Library Forum*, no.12, pp.44-50, 2013.

[13] LI Bingbing, GAO Lihua, ZHANG Fan, Research on Korea patent literature data, *Standard Science*, no.9, pp.35-58, 2012.

[14] DONG Tieying, Standardization and the development and utilization of Chinese patent information, *Modern Library and Information Technology*, no.s1, pp.110-112, 2002.

[15] DONG Quanling, HAO Chunhui, Comparison of XML DTD and XML Schema, *Modern Computer (Professional Edition)*, no.8, pp.64-66, 2006.

[16] LI Jiashen, Design and implementation of the science and technology literature push system based on XML, *Enterprise Science Technology and Development*, no.4, pp.54-56, 2014.

[17] Li Huajiag, Xing Chunxiao, Zhu Yi, Design and implementation of China patent search system based on J2EE and Tamino, *Computer Engineering and Applications*, vol.40, no.7, pp.168-171, 2004.

[18] ZHANG Ying, HUANG Weilai, ZHOU Quan, A new approach for patent information analysis: patent map based on XML Schema, *Journal of Intelligence*, vol.29, no.9, pp.59-63, 2010.