

SGML을 이용한 특허정보처리 연구

A Study of Patent Document Processing by SGML

권영숙*
(Young-Sook Kwon)

초 록

SGML(Standard Generalized Markup Language)의 디스크립션은 WIPO(World Intellectual Property Organization) Standard ST.32의 디스크립션과 더불어 자세히 기술되고있다. SGML의 이용에 대한 이점이 강조되고 있으며, 그것은 시스템 독립적이며 특허출판 및 전문 데이터베이스구축에 타당성이 있다는 것이다. WIPO Standard ST.32를 적용한 특허문헌의 내용구조는 ST.32에 따라 작성한 DTD로 표현하고 텍스트 자체는 DTD에 따른 범용 마크업을 사용하여 기술한다. 본고에서는 전체문헌, 특정 서브문서, 문단 등의 계층구조와 표, 도면, 화학구조식 등의 비계층구조로 되어 있는 문서구조를 어떻게 표현하는가에 대하여 예를들어 설명하였다. 그리고 특허문헌처리에서 SGML의 효과에 대하여 논의하였다.

키 워 드

SGML, WIPO Standard ST.32, DTD, 특허문헌처리, 특허출판, 특허전문데이터베이스

ABSTRACT

A description of SGML(Standard Generalized Markup Language) is given together with a detailed description of WIPO Standard ST.32. The benefits of the use of SGML are highlighted-its system independence and flexibility in building publication systems and full-text databases. A structure of WIPO Standard ST.32 based patent content is defined by DTD(document type definition) written in ST.32, and full-text itself is described with generalized markup depending on DTD. This article explains how to represent a document structure : a hierarchy structure like a entire document, a specific, sub-document, a paragraph, or non-hierarchy structure like a table drawings, or chemical structures. Merits of SGML in patent document processing are also discussed.

* 중앙대학교 대학원 박사과정
(University of Chung-Ang, Ph. D Student)

KEY WORDS

SGML, WIPO Standard ST.32, DTD, Patent document processing, Patent publication, Patent full-text database

1. 서 론

유럽 특허청에서는 지난 10여년 동안 SGML을 이용하여 특허출판을 하고 있다. SGML은 오늘날 특허출판 전략에서 핵심역할을 하고 있으며 모든 특허출원 및 명세서, 특허공보, 특허정보교환, 전문데이터베이스, CD-ROM제작에서 중요한 요소이다.

그러나 이러한 특허출판시스템이 아직은 많이 보급되지 않고 독일, 미국, 일본 등 소수 국가의 특허청에서 특허출판에 SGML을 이용하고 있다. 우리나라는 1999년 1월부터 SGML 문서처리기술에 의한 CD-ROM공보를 발간하고 있다.

본고에서는 SGML 디스크립션 및 특허문헌을 위하여 개발된 WIPO Standard ST.32에 대한 디스크립션을 자세히 기술하고, WIPO Standard ST.32가 특허청에서 어떠한 도구로 쓰여질 수 있는지를 살펴본다.

2. 특허문헌과 SGML

ISO 8879는 이용자의 문헌 식별을 위하여 선택사항을 기술하는 논리적이고 분명한 문법을 제공하지만 그것은 단순히 문법일 뿐이지 표준 태깅언어를 제공하지 못한다. 그러므로 사용자들은 자체 문헌구조에 대한 태

그생성을 위하여 자체의 장치를 마련해야한다. 예를 들어 문단시작 지시기호로 <PAR>을 사용할 수 있고 또는 <P>로 표시하기도 한다. 처음에는 이것이 혼란스럽게 보일 수 있지만 문헌이 서로 다른 구조를 갖게 되므로 상이한 태그셋이 필요하다.

특허문헌은 그 자체적으로 어휘를 가지고 있으므로 유럽 특허청에서는 SGML태그셋과 DTD 개발에 힘써왔다.

SGML은 실제문헌에 사용하는 특정 문자집합을 허용한다. PC상에서 쓰는 대부분의 문자코드는 ASCII이다. 그러나 그 대부분이 학술문헌, 특히 특허문헌에서 표현된 모든 문자들을 저장, 디스플레이, 인쇄하지 못한다. 예를 들어 모든 그리스문자 및 수학기호 등과 같은 것이다. 이러한 문자코드의 결점을 극복하기 위하여 SGML은 문자엔티티 참조를 사용하고, 그래픽문자는 선언된 기본 문자코드를 사용하여 표기법에 의하여 표현한다. ASCII의 예를 들어보면

- & permil; represents ‰ permille
- & Sigma; represents Σ capital sigma
- & nlt; represents < not less than

문자엔티티 참조의 사용은 문서이동이 가능하고 독립적이라는 점에서 커다란 이점이 있다. 예를 들어 워드프로세스에서는 활용가능한 모든 문자코드

와 폰트에 독립적인데, 그것은 다른 시스템으로 쉽게 변환되지 않는다.

다른 이점은 키보딩이 더 쉽고 그래픽문자가 스크린상에 디스플레이 되지 않는다는 점이다.

3. WIPO Standard ST.32

ISO 8879의 태그셋은 특허문헌에 대한 태그로 정의되지 못하기 때문에 유럽 특허청에서는 특허문헌 처리를 위한 태그를 개발하여 이 작업의 결과를 문서로서 WIPO에 제출했다. 이 문서를 유럽, 일본, 미국특허청의 협정 하에 WIPO표준으로 승인되어 출판하였다. 즉 그것은 1987년 10월에 특허문헌 텍스트의 공통부호화를 위하여 추천된 표준 포맷이다(ST.32).

1990년 12월에 WIPO에서는 이 표준 포맷을 개정하여 재출판하였다. 개정작업은 계속되어 1996년 말 버전3이 출판되었으며 현재 유럽 특허청은 이 버전3을 사용하고 있다. DTD 및 SGML태그의 개발은 문헌의 구조, 문헌의 내용, 데이터요소 등을 이해하여 문헌이 마크업될 수 있도록 문헌의 분석을 요구한다. 이로 인하여 최초로 유럽 특허문헌을 위한 ST.32의 개발이 이루어지게 되었다. 그후 세계도처의 특허문헌에까지 적용범위가 확대되었으며 또한 전자파일링에도 적용되었다. 다음에서 특허문헌을 위하여 개발된 마크업을 살펴본다.

3.1 특허문헌의 구조 및 마크업

ST.32에 사용된 SGML태그의 계층은 특허문헌의 일반구조를 따르며. 계층의 단계는 일반적이며 논리적인 문헌요소를 기술하는 SGML태그에 의하여 지시된다. 문헌요소는 전체문헌, 특정 서브문헌, 문단, 표 등과 같은 텍스트의 구성요소이다. 이들 문헌요소의 각각에 대한 시작태그와 종료태그는 아래와 같이 기술될 수 있다.

Level	SGML tag(example)
Document	<PATDOC>
Sub-document	<SDOBI>
Text Component(Paragraph)	<P>
Text Element(Subscript)	<SB>
Character	
End	</SB>
End	</P>
End	</SDOBI>
End	</PATDOC>

태그들이 특허문헌의 논리적 구조를 어떻게 표시하는지 주의해야 하며 어떤 태그는 다른 태그안에 포함되어야 한다.

서브문헌<SDOBI>은 문헌<PATDOC>의 외부에 위치할 수 없다. DTD는 이 구조를 정의하고 SGML파서는 이들 규칙의 오류를 검증한다. ISO 8879(SGML선언부 내의 ST.32)는 SGML태그에 대한 문법을 다음과 같이 정의한다.

시작태그	내용	종료태그
<P>	text	</p>

여기에서

< 시작태그에 대한 열림기호	<1문자>
</ 종료태그에 대한 열림기호	<2문자>
> 시작태그 및 종료태그에 대한 닫힘기호	<1문자>

P는 DTD에서 정의된 특정태그(문단)의 전체지시자이다.

기술정보로서 특허정보는 권리정보이며 특허정보와는 달리 6개 주요 부분으로 구성되어 있다.

- (1) 서지데이터 : 출원자, 출원자주소, 발명자, 발명의 명칭, 특허분류번호, 파일링날짜, 출원번호 등 이러한 데이터는 표제지에 수록되며 이들 데이터의 대부분은 출원자에 의하여 등록된다.
- (2) 초록 : 발명에 대한 간략한 기술 사항이며 이것은 일반적으로 표제지에 인쇄되며 그림초록이 포함되기도 한다. 초록의 분량은 보통 1페이지 이하이다.
- (3) 발명의 명세서 : 발명에 대한 상세한 설명으로 가장 긴 문서이다. 국내 특허출원의 경우 대략 10~20페이지 정도이나 100페이지 이상이 될 수도 있다.
- (4) 청구범위 : 법적보호를 요구하는 것으로 1이상의 페이지로 구성된다.
- (5) 도면 : 일반적으로 별도의 문서로서 기술도면, 흐름도, 다이어그램, 그래프 등을 포함한다.
- (6) 탐색보고서 : 탐색결과를 제시하는 것으로 보통 1페이지 정도이며 특허청에 의하여 제공된다.

이러한 부분은 하위 특허문서로서 각 문서마다 ST.32에 의하여 다음과 같이 표시된다.

<PATDOC>

Start of patent document

<SDOBI>

Bibliographic data

</SDOBI>

<SDOAB>

Abstract

</SDOAB>

<SDODE>

Description

</SDODE>

<SDOCL>

Claims

</SDOCL>

<SDODR>

Drawings

</SDODR>

<SDOSR>

Search report

</SDOSR>

End of patent document

</PATDOC>

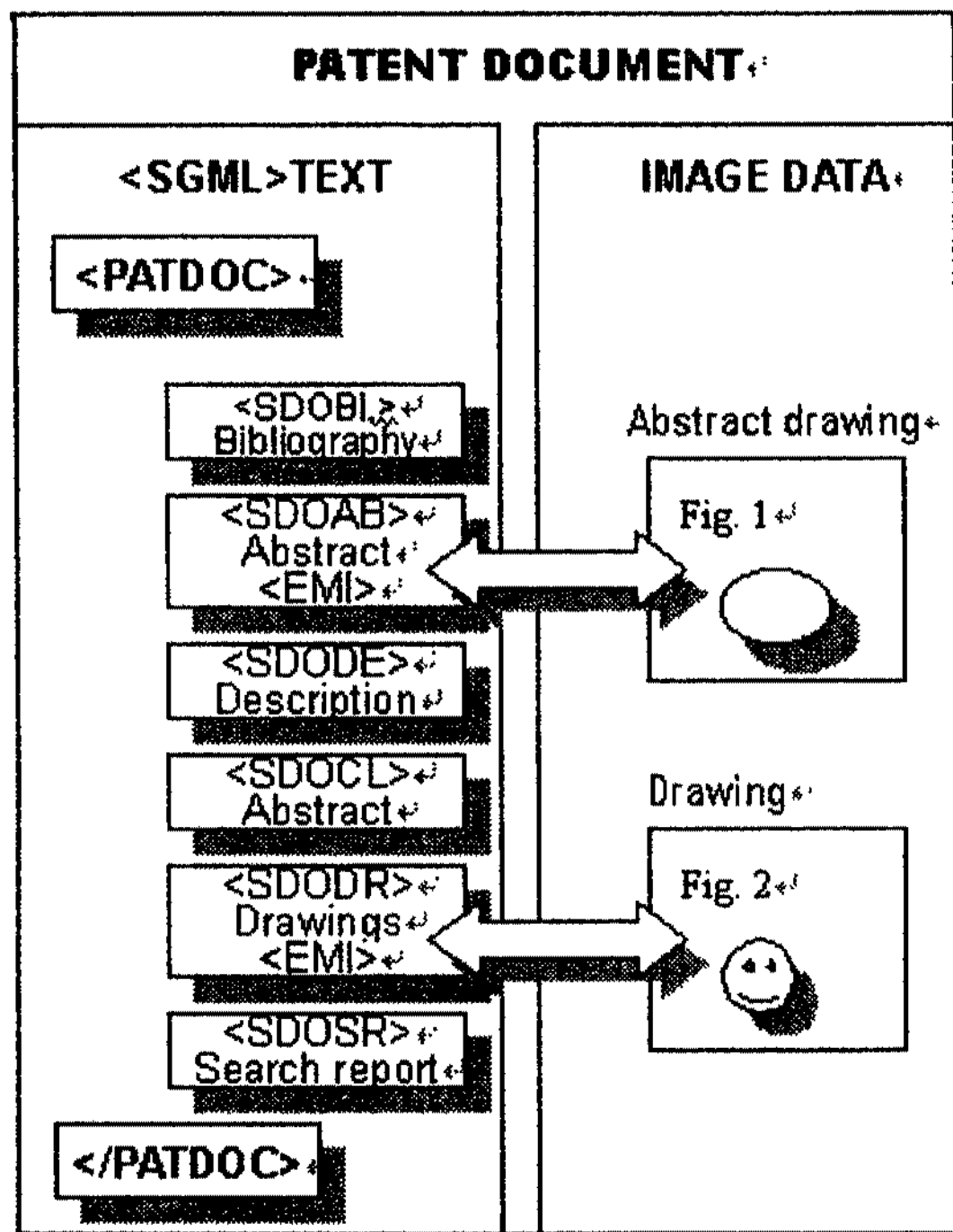
외부 엔티티인 이미지데이터와 연결되어 있는 이러한 구조는 [그림 1]에서 설명하고 있다.

[그림 1]에서 보면 삼입 이미지태그는 2개의 그림파일로 연결한 것으로 나타난다. 이들 그림파일은 특허문서를 인쇄하기 위한 사후처리과정이라고 할 수 있다. 특허서브문서는 각기 다른 처리가 필요하며 어떤 경우에는 태그가 아래에 기술되기도 한다.

3.1.1 서지 태그

서지 태그에 대한 구조는 WIPO Standard ST.9와 ST.30을 기초로 하고 있다. ST.9 표준에 있는 코드는 특허문헌의 표제지에 인쇄되는 주요 데이터요

[그림 1]



소를 위해서 제공되는데, 이것을 INID 코드라고 한다. 서지데이터에 대한 DTD설계에서는 SGML태그의 기본인 ST.9와 ST.30표준코드를 사용한다. 이 데이터의 계층구조는 다음과 같다.

- <SDOBI> Start of bibliographic data
- <B100> INID code 10-identification of the patent
- <B200> INID code 20-application data
- <B300> INID code 30-priority data
- <B400> INID code 40-public availability
- <B500> INID code 50-technical information
- <B600> INID code 60-related documents
- <B700> INID code 70-parties concerned with the patent
- <B800> INID code 80-data related international conventions
- </SDOBI> End of bibliographic data

이들 태그는 서지데이터를 더 세분

하여 태깅할 수 있도록 되어 있다. 예를 들면

- <B700> Parties concerned with patent
- <B710> Start of applicant data (INID code 71, ST.30 code 710)
- <B711> Start of application number one data
- <SNM> Name (may be further sub-divided)
- <ADR> Address (may be further sub-divided)
- </711> End of applicant number one data
- </B710> End of applicant data
- </B700> End of parties concerned with the patent

서지데이터 및 태그의 경우에 특허출원자가 SGML을 사용한다면 서지데이터 태그에 대하여 전혀 걱정할 필요가 없는데 그것은 데이터엔트리가 소프트웨어에 의하여 처리되기 때문이다. 특허출판을 위한 서지데이터 요소가 ST.32에 의하여 처리된다 할지라도 모든 국가의 특허청이나 특허분야의 요구에 부응하지 못하기 때문에 사무적이고 세부적인 데이터가 처리될 수 있도록 ST.32는 글자의 표준국가 코드에 약간의 새로운 태그번호를 DTD에 추가하여 다음과 같이 사용한다 <B 710> - <B 713 EP>.

3.1.2 초록, 명세서, 특허청구범위

초록 명세서 특허청구범위와 같은 서브문서는 일반적으로 출원자에 의하여 등록되는 데이터인데 학술문헌과 유사하다. 이런 점에서 데이터 마크업에

사용된 태그가 특허문헌에만 필요한 것이 아니므로 이 부분에 대한 태그를 다시 설계하지 않고 학술문헌에 사용하는 태그가 특허문헌에 적용될 수 있는지를 결정하여 사용한다. 그러나 일반적으로 일반문헌에 사용하는 태그는 특허문헌에 맞지 않는다. 예를 들어 내용 중에 있는 표, 색인, 장, 상호참조 등의 부분에서는 적용이 불가능하다.

3.1.3 텍스트

표목, 문단, 표 등과 같은 일반 텍스트를 위하여 ISO 8879는 ISO/IEC/TR9573과 마찬가지로 몇 가지 지침을 제시한다. <H> 표목; <P> 문단; 순차표; 표항목과 같은 태그는 ST.32와는 다른 SGML

응용이며 DTD이다. 특허문헌은 일반 문헌과는 달리 분모, 분자 같은 수학적 문자가 상당히 일반적으로 나타난다. 그러나 ISO 8879는 수학적 특수문자에 대한 마크업으로 맞지 않는다. [그림 2]는 간단한 수식에 대한 마크업을 보여준다.

3.1.4 표

특허문헌에서 표의 삽입은 흔히 나타나는데, 표는 마크업의 가장 중요한 부분을 차지한다. ST.32 표태그 개발에 대한 기초는 미국출판사 “Electronic Manuscript Series”와 연합하여 만들어졌다. 표는 특정 소프트웨어 “Wordperfect”를 사용하여 태깅한다. ST.32는 표 마크업에 맞지 않으므로

[그림 2]

Original(EP 0677937 AI):
 ...this average distance may be calculated to be approximately equal to

$$\sqrt{\frac{2}{\pi}}\alpha(1-e^{-\frac{\alpha^2}{2\sigma^2}})$$

α

For the above value of σ this may be calculated to be less than or equal to 0.15α .

SGML Markup:
 ... this average distance may be calculated to be approximately equal to
 <DF><SQRT><FRAC>2<OVER>π</FRAC><SQRT>&<sigma;(1-e^{-<FRAC>&alpha²<OVER>2&sigma²</FRAC>}</DF>
 For the above value of<CHF>α<CHFBR>σ</CHF> this may be calculated to be less than or equal to 0.15α.

Where:

<DF>=Display Formula <SQRT>=Square Root <FRAC>=FRACTION
 <SUP>=SUPerscript <CHF>=Character Fraction <CHFBR>=Character Fraction Break

Note also the use of character entity Reference, eg. σ for characters outside common character sets.

개선할 필요가 있으며 어떤 표는 이미 업은 [그림 3]에서 설명하고 있다. 지데이터로 처리되기도 한다. 표 마크

[그림 3]

COOLING SYSTEM COMPONENTS			
DESCRIPTION	RETAIL PRICE		QUANTITY
	ex VAT	inc VAT	in stock
Radiators	295.50	333.00	3.012
Hose Clips	5.25	6.25	27.435
Lower Pipes	23.66	26.50	12.445
Upper Pipes	21.35	23.00	13.752
Caps	15.50	17.00	4.049
Pumps	341.00	375.00	3.553

SGML Markup:

```
<TAB CO=4 RS='PS S AS' CS='PS FS AS'>
<ROW><TTI>COOLING SYSTEM COMPONENTS
<ROW><TCH AL=L>DESCRIPTION
<TCH CB=2 CE=3>RETAIL PRICE
<TCHAL=R>QUANTITY
<ROW><TSH>
<TSH AL=R>ex Vat
<TSH AL=R>inc Vat
<TSH AL=R>in stock
<ROW><TSB>Radiators<CEL
AL=D>295.50<CEL AL=D>333.00<CEL>3.012
<ROW><TSB>Hose Clips<CEL AL=D>5.25<CEL AL=D>6.25<CEL>27.435
<ROW><TSB>Lower pipes<CEL AL=D>23.66<CEL
AL=D>26.50<CEL>12.445
<ROW><TSB>Upper Pipes<CEL AL=D>21.35<CEL
AL=D>23.00<CEL>13.752
<ROW><TSB>Caps<CEL AL=D>15.50<CEL AL=D>17.00<CEL>4.049
<ROW><TSB>Pumps<CEL AL=D>341.00<CEL AL=D>375.00<CEL>3.553
</TAB>
```

Where:

```
<TAB CO=1 RS='PS FS AS' CS='PS FS AS'>=start table with 4
Columns. Row and Column Separators are all single line
<TTI>=Table Title <TCH AL=L>=Table Column Header, Aligned
left
<TSH>=Table Sub Header, and so on.
```

3.1.5 수 식

표와 마찬가지로 수식도 특허문헌에서 아주 일반적으로 나타나며 마크업도 복잡하다. 수식에 대한 ST.32 DTD는 ISO 기술보고서 9573의 DTD를 유사하게 따르며 다양한 수학적 구조를 코

드화할 수 있다. 수학적 마크업은 [그림 2]와 [그림 4]에 제시되어 있다.

3.1.6 도면과 이미지데이터

대부분의 특허문헌은 텍스트데이터에 첨부된 별도의 문서로서 도면을

[그림 4]

Original(EP 0677937 AI):
 For both the in-phase and (quadrature components this error probability is approximately equal to

$$P_{err} = 2 \int_{\frac{\alpha}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{t^2}{2\sigma^2}} dt = 2 \int_{\frac{\alpha}{\sigma}}^{\infty} \frac{1}{2\pi} e^{-\frac{u^2}{2}} du = \operatorname{erfc}\left(\frac{\alpha}{\sqrt{2}\sigma}\right)$$

Where α is half the length of a side of the square $\operatorname{erfc}(x)=1-\operatorname{erf}(x)$ is the complementary error functions.

SGML Markup:

For both the in-phraise and quadrature components this error probability is approximately equal to

```
<DF>
P<SUB>err</SUB>=2<INTEGRAL><FROM>&alpha;<TO>&INFIN;</INTEGRAL>
<FRAC>1<OVER><SQRT>2&pi;</SQRT>&sigma;</FRAC>E
<SUP><FRAC>t<SUP>2<SUP><OVER>-2&alpha<SUP>2</SUP></FRAC></SUP>
DT=2<INTEGRAL><FROM><FRAC>&alpha;<OVER>&sigma</FRAC>
<TO>&infin;</INTEGRAL><FRAC>1<OVER><SQRT>2&pi;</SQRT><FRAC>e
<SUP><FRAC>u<SUP>2</SUP><OVER>2<FRAC></SUP>
du=erfc<FENCE><FRAC>&alpha<OVER><SQRT>2</SQRT>&sigma</FRAC></FENCE>
</DF>
where &alpha is half(the length of(a side the sware  $\operatorname{erfc}(x)=1-\operatorname{erf}(x)$  is the complementary error function. ....
```

where:

<DF>=Display Formula <SUB>=SUBscript <FRAC>=FRAction
 <SQRT>=Square Root <SUP>=SUPerscript
 Other tags should be self explanatory

포함하고 있다. 그러나 이미지데이터(화학식, 특수문자, 복잡한 표 등)는 초록, 명세서, 청구범위와 같은 본문 텍스트 안에 나타날 수 있다. ST.32는 별도 페이지의 도면까지 포함하여 모든 이미지를 삽입이미지(tags<EMI>)로 분류한다. 이미지는 본문 텍스트외부의 이미지파일로서 외부엔티티 처리를 하는데[그림 1참조], 외부파일에 이미지가 저장되는 방법은 특허문헌의 특정요구에 따라 변할 수 있으므로 ST.32와 다른 표준인 WIPO표준 ST.33을 사용하여야 한다.

ST.33은 특히 이미지 색인방법을 제공하고 팩시밀리 Group4 이미지 압축을 이용하여 이미지를 저장한다. 유럽 특허청에서는 모든 이미지를 이러한 방법으로 저장한다.(BACON Standard)

도면 서브문서 -<SDODR>은 텍스트의 부분이 아닌 별도의 문서로 되어 있는 모든 도면을 대신하여 사용된다. 따라서 다음과 같은 마크업 예를 들 수 있다.

```
<SDODR>
<DP N=12 >
<EMI ID=12.1 HE=179 WI=157
    LX=217 LY=908 TI=DR>
<DP N=13 >
<EMI ID=13.1 HE=176 WI=156
    LX=207 LY=925 TI=DR>
</SDODR>
```

여기에서 DP N=12는 원문헌 페이지번호 12인데, 그것은 출판사항으로 기술될수있고 <EMI>태그는 다음의 속성을 갖는다.

ID=문헌의 식별 참조
HE=높이
WI=폭
LX=페이지의 왼쪽 끝까지 밀리미터로 표현한 X좌표
LY=페이지의 왼쪽 끝까지 밀리미터로 표현한 Y좌표
TI=이미지형태, DR=그림

특히 본문의 초록, 명세서, 청구범위 내에 이미지가 나타나는 경우가 있는데, 그것 또한 삽입이미지로 처리된다. LX와 LY의 이용은 후처리 상에서 페이지는 완전히 재구성되며 원래의 페이지와 관계가 없기 때문에 X와 Y의 사용이 도움이 되지 않는다.

3.2 탐색 보고서

탐색보고서의 데이터는 특허청에서 등록하는 서지데이터와 유사하고 태그는 특허문헌에 대한 세부사항이다.

ST.32는 유럽 특허청의 모든 특허문헌을 마크업하기 위하여 사용된다. 그러나 범위에 이르지 못하는 몇 가지 영역에 있어 개선되어야 하는데 화학구조식 및 생물학적 결과 데이터 등에 관한 마크업이 없기 때문이다. ST.32는 모든 표준에서나 마찬가지로 개정 및 갱신이 요구된다.

4. 특허문헌 처리를 위한 SGML의 이용

ST.32는 이론적인 개념이 아니고 실제 특허문헌처리에 사용하는 작업틀이다. 유럽특허청에서는 1997년 현재 50만건 이상의 전문과 70만건의

초록을 SGML포맷으로 저장하고 있다. 유럽 특허청의 계약자는 매주 5만 페이지의 출원 및 명세서를 처리한다. SGML을 사용하지 않는 특허청과 특허관련 부서에서는 이 정도 양의 데이터를 처리하지 못한다. 이것으로 보아 SGML과 ST.32는 많은 양의 데이터 엔트리에 적합하다는 것을 알 수 있다.

항공, 자동차, 석유화학, 방위, 컴퓨터, 전자, 출판 등에 관한 기술문서들이 SGML을 이용하여 보고서나 단행본처럼 단일문헌으로 처리된다. 이처럼 SGML로 마크업된 문헌의 다양성은 불과 몇년전 만해도 거의 생각을 못하였다. 그러나 오늘날 웹과 HTML은 모두 SGML로 바뀌어 가고 있다.

HTML은 현재 상당히 널리 이용되고 있으며 성공적인 마크업 언어이다. HTML문을 구성하는 Syntax는 매우 간단하여 많은 사람이 쉽게 배우고 사용한다. 그러나 HTML문은 복잡한 문서나 수학적식이 있는 문서, 정해지지 않은 표의 표현 등 여러측면에서 특허문헌을 표현하는 데는 적절하지 못하다. 특히 HTML은 SGML의 하나의 DTD이면서도 편의성을 위하여 문서의 논리구조를 어렵게 하는 몇가지 특징을 추가하여 논리적 구조를 바탕으로 하는 문서를 구현하는데 어려움이 있다. 이는 전문적이거나 학술적이거나 그리고 논리적 구조에 의한 문서의 표준에 있어서 SGML의 범위가 광범위함을 알 수 있다. HTML이 SGML의 하나의 DTD이듯이 SGML에 의한 DTD의 정의로 모든 문서에

대한 DTD를 생성함으로써 모든 문서를 표현하고 이를 유통시킬 수 있다는 것이다.

SGML을 이용하는 특허문헌 저작 개념은 SGML을 이용하여 특허출원이 가능하다는 것을 의미한다. 또한 당분간은 ST.32만이 완벽한 특허문헌 마크업을 수행할 수 있다. ST.32는 특정 하드웨어나 특정 소프트웨어 시스템을 이용하여 특허문헌을 저장, 교환, 생성 가능하게 한다.

특허문헌 데이터 입수에서 SGML의 사용은 많은 이점이 있다. 일반적인 이점을 본고의 시작부분에서 열거하였으나 다음의 사항은 더 심도있게 강조할 만하다.

- 특허문헌을 시스템/응용 독립적으로 읽어들이고 처리할 수 있다. 즉 표준코드를 사용하여 SGML규칙에 따라 문헌을 작성하고 태깅하므로 어떤 시스템에서도 문헌을 처리할 수 있다.
- 이기종간의 시스템에서 정보교환이 가능하다. SGML파일은 특정 인코딩없이 인터넷에 탑재할 수 있다.
- 다른 형태의 문헌제작을 위하여 동일한 SGML파일을 이용한다. 페이지문헌, CD-ROM 출판, 데이터베이스발간 등.
- CD-ROM 제작에서 다른 레이아웃을 생성할 경우 동일한 SGML파일을 이용한다.
- 문헌의 내용 및 구조를 관리한다.
- 전문탐색데이터베이스를 구축한다.

5. 결 론

SGML은 여러가지 응용분야에 맞는 완전한 표준이다. ST.32는 특허문헌을 처리하기 위하여 수년에 걸쳐 개발되었으며 특허문헌 마크업이 가능한 유일한 SGML응용이다. 변화하고 있는 문헌처리 시스템은 최소한의 과업이 아니다. 새로운 방법과 기술의 습득, 훈련, 투자 등이 따라야 한다. 그러나 급변하는 문헌처리 분야에서 SGML은 모든 유형의 문헌을 처리하기 위한 토대로써 음성처리를 가능하게 한다. SGML은 대부분의 산업분야에서 일반적으로 사용되며 특히 정보의 양이 많은 특허문헌처리 즉 특허출원, 특허문헌의 출판, 특허정보의 교환, 전문데이터베이스구축, CD-ROM공보의 발간 등에서 핵심역할을 하고 있다. 따라서 본고에서는 특허정보 전문데이터베이스의 구축을 위한 기초자료를 제공하고자 특허문헌을 구성하고 있는 주요 서브문서를 중심으로 각각의 서브문서에 대하여 SGML의 응용인 WIPO Standard ST.32를 사용한 특허문헌 구조화의 예를 설명하였다. 그러나 범위에 이르지 못하는 몇 가지영역 화학구조식, 생물학적 결과 데이터 등에 관한 마크업이 가능하도록 ST.32는 개정되고 보완되어야 할 것이다.

〈참 고 문 헌〉

- 根岸正光, 石塚英弘, 1994. SGML의活用, 東京: オーム社.
- 맹성현, “문서구조화와 정보검색”, 정보과학회지, 16(8), 1998, pp. 6~15.
- 박현주, 1997. “구조화된 문헌의 이미지 정보 검색시스템에 관한 연구”, 연세대학교 대학원 석사학위논문.
- 박홍서, 1997. 전자문서의 표준-SGML, 디지털도서관, pp. 72~82.
- 현규섭, 1995. “SGML의 서지정보 표식 방법에 관한 실험적 연구”, 도서관, 50(3), pp. 38~63.
- KSC 5913 1993. “문서기술언어 SGML”, 한국표준협회.
- Alexander, B. L., 1993. The US Patent and Trademark Office's Electronic Application System. World Patent Information, 15, 63~71.
- American National Standards Institute, 1991. Electronic Manuscript Preparation and Markup(z39.59). Transaction Publishers: New Brunswick, NJ and London, 1991.
- Goldfarb, C. F., 1991. The SGML Handbook: The Annotated Full Text of ISO 8879-Standard Generalized Markup Language. Oxford University Press: Oxford.
- Nakamura, A., 1991. Paperless Plan of the Japanese Patent Office and Its Dissemination Policy. World Patent Information, 13, 125~138.
- Spear, B. 1996. Patent Information on the Internet, World Patent Information, 18, pp.33~37.
- WIPO Standard ST.32. 1996. Recommendation for the Markup of Patent Documents using SGML. version 3. World Intellectual Property Organization: Geneva.
- WIPO Standard ST.33, 1990. Recommended Standard Format for Data Exchange of Facsimile Information of Patent Document. World Intellectual Property Organization: Geneva.