

# 문서기술언어 SGML에 의한 전문 데이터베이스의 구축\*

## Construction of Full-text Database by SGML

김 창 봉 편역\*\*  
(Chang Bong Kim, Trans.)

### 초 록

문서기술언어 SGML(Standard Generalized Markup Language, 표준 범용 마크업 언어)과 그 전문 데이터베이스(도표나 화상 포함)에의 적용에 대하여 해설하였다. SGML을 적용한 전문 데이터베이스의 구조는 SGML로 작성한 DTD (document type definition)로 표현하고 텍스트 자체는 DTD에 따른 범용 마크업을 사용하여 기술한다. 본고에서는 장, 절, 단락 등의 계층구조와 각주, 도표, 화상 등의 비계층 구조로 되어 있는 문서 구조를 어떻게 표현하는가에 대하여 예를 들어 해설하였다. 그리고 SGML의 효과, 전자 출판, 검색 시스템, 하이퍼텍스트, SGML 관련 도구 등에 대해서도 해설하였다.

### 키 워 드

SGML, 전문 데이터베이스, DTD, 전자출판, 검색 시스템, 하이퍼텍스트

### ABSTRACT

SGML(Standard Generalized Markup Language) and its application to full-text database including a table, a figure and a picture are explained. A structure of SGML based full-text database is defined by DTD(document type definition) written in SGML, and full-text itself is described with generalized markup depending on DTD.

---

\* 본고는 ISHIZUKA Hidehiro의 「SGML and Full-text Database」를 번역·편집한 것임. 내용에 포함된 예문은 역자 나름대로 선정하여 다시 구성한 것임. ISHIZUKA Hidehiro, "SGML and Full-text Database", 情報管理 37(2), 1994, pp.148~159

\*\* 산업기술정보원 DB사업부 연구위원  
(DB Development & Management Dept., KINITI)

This article explains how to represent a document structure : a hierarchical structure like a chapter, a section, or a paragraph, or non-hierarchical(referencial) structure like a note, a table, a figure or a picture. Merits of SGML, electronic publishing, a retrieval system or hypertext and SGML tools are also described.

## KEYWORDS

SGML, Full-text database, DTD, Electronic publishing, Retrieval system, Hypertext

## I. 서 론

최근에 워드프로세서, 전자출판 및 컴퓨터 조판의 발달로 대량의 문서 정보가 디지털화되고 있다. 그러나 그 대부분은 정서(淨書)를 위하여 디지털 입력되고 있을 뿐이고 출판은 재래의 방식에 의존하고 있어 디지털화를 외면하고 있다. 이 때문에 출판된 문서중에는 귀중한 정보가 포함되어 있음에도 불구하고 그 재이용이나 정보 교환이 충분히 이루어지지 않고 있다.

1986년 국제표준화기구(ISO)는 문서 편집의 합리화와 검색 재이용의 활성화를 위하여 국제규격 ISO 8879로서 SGML(Standard Generalized Markup Language)을 제정하였다. SGML이란 공통부호화(Generic coding)를 기본으로 하는 문서기술 언어로서 문서의 입력, 교정, 편집에 이르는 문서 처리를 인간과 컴퓨터가 협력하여 수행함으로써, 합리적인 문서 처리와 검색 재이용을 가능케 하는 도구라고 할 수 있다.

SGML은 문자 정보 뿐만 아니라 도표, 수식, 화학식, 사진 등의 화상 정보 및 음성 정보까지도 취급함으로써 단일 매체를 이용한 전통적인 출판으로 부터 하이퍼텍스트 등의 전자화 문서나 전문 데이터베이스(Full-text Database)의 구축은 물론 복잡한 멀티미디어 데이터베이스의 구축까지를 포함하는 광의의 출판을 위하여 사용할 수 있다.

SGML은 인쇄 출판의 분야에서 도출된 기술로서 전문 데이터베이스의 구축을 의식하여 개발된 것이라고 볼 수 있다. 이와같이 SGML에 의하여 출판된

전문 데이터베이스는 디지털화되어 있어 인쇄 출판은 물론 CD-ROM화가 용이하고 Internet 등 컴퓨터 통신망에 탑재 및 검색이 가능하므로 정보의 활발한 재이용이나 교환에 크게 기여할 것으로 사료된다.

본고에서는 SGML의 개요, SGML 방식 전문 데이터베이스의 구조, 구축, 관리, 검색에 관하여 해설하고 SGML의 효과와 적용 사례를 살펴보기로 한다.

## II. SGML의 개요

SGML은 1986년 ISO 8879로 제정되고 1988년 ISO 8879/AMENDMENT 1로서 내용 수정을 거쳐 현재에 이르고 있다. 우리나라도 상기 두 국제 규격에 대응하는 규격으로서 1993년에 KS C 5913(문서 기술 언어 SGML)을 제정하였다.

SGML(Standard Generalized Markup Language)을 직역하면 「표준 범용 마크업 언어」가 된다. 이것은 인쇄 출판 분야의 용어를 사용한 표현이다. 한편 KS에서는 SGML의 개념과 기능을 고려하여 「문서 기술 언어 SGML」라고 하였다. 이 표현에는 전문 데이터베이스의 개념이 내포되어 있어 SGML이란 전문 데이터베이스의 구조를 명시적으로 표현하기 위한 언어라고 할 수 있다.

이제, SGML을 「표준 범용 마크업 언어」와 「문서 기술 언어」의 두가지 측면에서 설명하고 다음에 간단한 예를 들어 두 견해를 통합해 보기로 한다.

### 1. 표준 범용 마크업 언어

표준이란 말은 미국 IBM사가 사내 문서의 작성에 사용하였던 GML (Generalized Markup Language)을 개념적, 기능적으로 확장하여 ISO의 표준으로 하였기 때문에 붙인 말이다.

마크업(markup)이란 'mark up'으로 부터 만들어진 신조어로 편집자나 인

쇄 디자이너가 텍스트 또는 기타 부분에 서체나 사이즈 등 체제에 관한 지정을 원고상에 손으로 써넣는 행위를 지칭한다. 이와같은 지시를 손으로 써넣지 않고 전자적으로 부여하는 것을 'markup'이라고 할 수 있다.

'markup'은 고유와 범용의 두가지 방식이 있다. 고유의 'markup'은 문서의 체제를 기술하기 위한 명령으로서 사용하는 인쇄 기기나 그 소프트웨어에 전용으로 쓰인다. 그러나 범용 방식은 문서의 체제가 아니라 구조를 기술하는 'markup'으로서 제목, 저자명, 각주, 인용문 등 문서를 구성하는 요소의 시작점과 종말점을 마크업한다. 소프트웨어는 이 마크를 판정하여 요소마다 서체, 사이즈, 체제를 지정하여 인쇄한다. 다시 말해 마크업이 기기와는 상관이 없으므로 범용이라고 한다. 범용 마크업은 이를 판정하는 소프트웨어가 오류를 발생시키지 않도록 엄밀해야 한다. 컴퓨터 과학적인 측면에서 「언어」란 문법(syntax)을 가지고 있는 것을 말한다. SGML도 문법을 가지고 있기 때문에 컴퓨터학적으로 「언어」라고 한다.

## 2. 문서 기술 언어

SGML에서 문서(document)란 텍스트 뿐만 아니라 도표나 사진 등을 포함하여, 단위(unit)로 처리되는 정보의 집합으로 정의한다. 문서를 기술하기 위해서는 그 구조를 기술하는 부분(DTD)과 문서의 데이터 자체가 필요하다.

DTD(Document Type Definition, 문서형 정의)는 문서 구조를 SGML의 구문으로 정의하고 기술하는 것이다. 문서는 단행본, 논문, 보고서 등 문서의 종류에 따라 그 구조가 다르기 때문에 종류 마다 DTD가 다르다. DTD를 스스로 작성하기 위해서는 SGML의 구문을 알 필요가 있지만 문서 종류에 따라 DTD가 이미 준비되어 있는 경우도 있다. 이런 경우에는 스스로 DTD를 작성할 필요없이 준비된 DTD를 사용하면 된다.

문서의 데이터로서는 DTD에 따라 마크업된 텍스트(문서 텍스트)와 그림이나 사진을 bit pattern(비트 0과 1로 나타내는 코드)화한 파일(외부 엔티티)이 있다. 외부 엔티티도 DTD에 따르며 본문과의 참조관계는 문서 텍스트 중에 기입하면 된다. 표나 수식은 문자로 쓰여 있기 때문에 문서 텍스트중에 기입할 수 있다.

### III. SGML 문서 엔티티

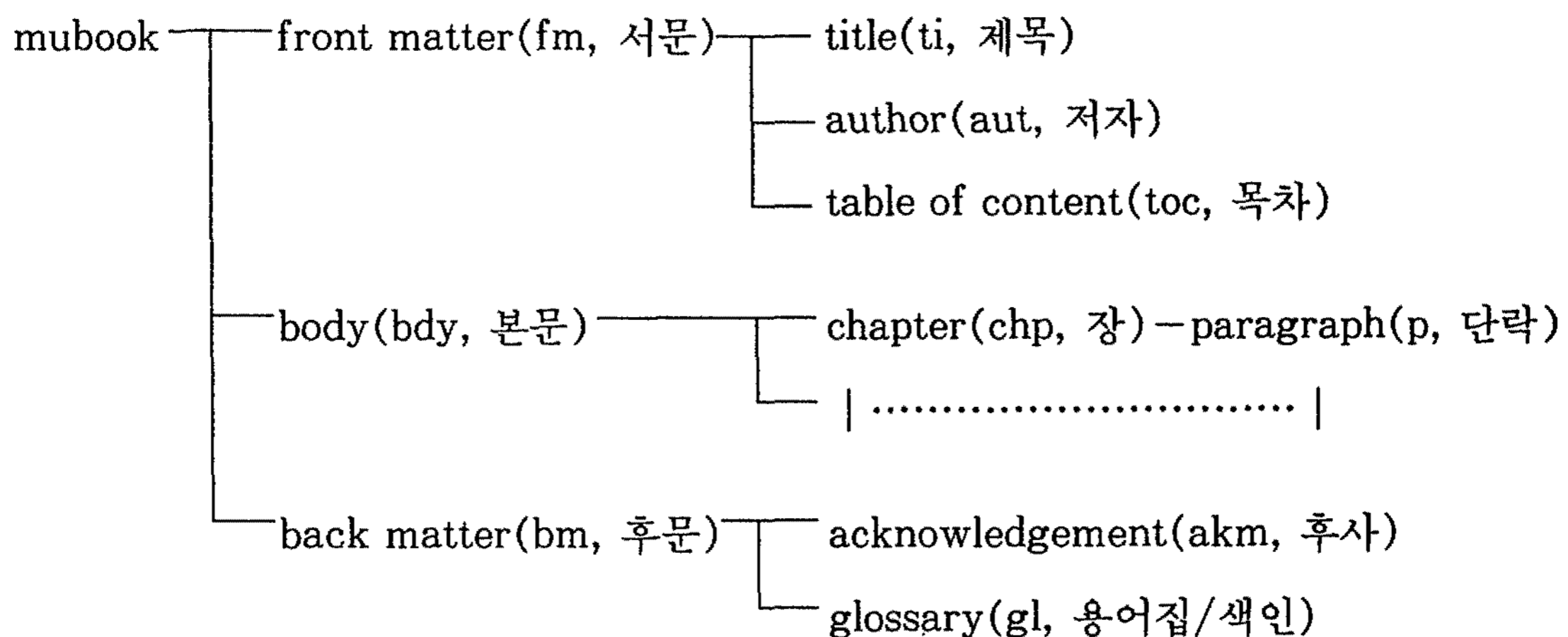
SGML 문서 엔티티란 SGML의 규칙에 따른 전문 데이터베이스로서 SGML 선언, DTD, 문서 텍스트의 세가지로 구성되어 있다. 그림이나 사진이 있을 경우에는 이를 비트 패턴화한 파일을 외부 엔티티로서 부여한다.

SGML 선언이란 사용할 문자집합, 처리에 필요한 기억용량, 구체적인 구문의 유효범위, 사용되고 있는 구체적인 구문, 사용되는 SGML 기능, 처리에 필요한 어플리케이션 고유기능 등을 선언하는 것이다. SGML 선언은 SGML 문서를 처리할 때 필요한 기능을 받는 측에 확실하게 전달하기 위한 것으로, 컴퓨터보다 오히려 받는 측의 사람이 읽어야 할 부분이다.

앞으로 「문서 기술 언어 SGML에 의한 전문 데이터베이스의 구축」이라는 본 해설의 내용을 예로 들어 DTD와 문서 텍스트를 설명하기로 한다. 본 해설 (mubook이라고 약칭)의 계층 구조를 나타내면 <그림 1>과 같다.

<그림 1>의 괄호 안에는 구성 요소에 대한 설명과 약호가 나타나 있다. 앞으로 이 약호를 사용하여 그 구조를 설명하기로 한다. 여기서 mubook은 fm, bdy, bm으로 구성되어 있고 fm은 ti, aut, toc로, bdy는 chp, p로, bm은 akm, gl로 구성되어 있는데 각각은 그 순서대로 출현된다. 특히 bdy는 chp

<그림 1> 'mubook'의 계층 구조



뿐만 아니라 p에도 직접 연결되어 있음을 나타내는데 이것은 bdy는 chp 또는 p로 구성되고 chp는 p로 구성된다는 것을 의미한다.

## 1. DTD(문서형 정의)

〈그림 1〉의 문서 구조를 SGML의 DTD로 정의하면 〈그림 2〉와 같다. 각 행은 ELEMENT문으로서 각 요소가 어느 요소로 구성되어 있는가를 정의하고 있다. 〈그림 1〉에서는 장의 타이틀(chpti로 표시)이 생략되어 있지만 DTD에서는 요소로서 추가되었다. 또한 akm이나 gl의 ELEMENT문은 생략하였다.

SGML의 문법에서는 대문자, 소문자를 동일한 것으로 취급한다. 여기서는 요소의 이름을 분명하게 알아볼 수 있도록 소문자만으로 표시하였다. 요소명을 약호로 표시하였기 때문에 이해하기 어렵다고 생각할 수 있다. 요소명은 DTD의 ELEMENT문으로 정의하기 때문에 이해하기 쉬운 이름으로 정의하여도 된다. 일반적으로 컴퓨터에서 이름의 길이는 8문자 이하로 정해져 있지만 DTD에서는 이름의 길이를 그 이상으로 할 수도 있다. 또한 요소명의 첫 문자는 A부터 Z사이의 문자를 사용해야 하지만 그 다음부터는 숫자, 마침표, 하이픈 등도 사용할 수 있다. 이 범위에서 적당한 요소명을 사용하면 된다.

〈그림 2〉 'mubook'의 DTD

---

〈!ELEMENT mubook	—	— (fm, bdy, bm?)〉
〈!ELEMENT fmk	—	0 (ti, aut, toc)〉
〈!ELEMENT bdy	—	0 (chp+   p+)〉
〈!ELEMENT bm	—	0 (akm?, gl?)〉
〈!ELEMENT chp	—	0 (chpti, p+)〉
〈!ELEMENT p	—	0 (#PCDATA)+)〉

---

ELEMENT문의 형식은 다음과 같다.

<!,ELEMENT, 시작 · 종료 태그 생략 여부, 시작 · 종료 태그 사이에 출현되는 구성 요소>

<그림 2>의 첫째 줄은 mubook이란 요소는 fm, bdy, bm이라는 세 요소로 구성되어 있음을 기술하고 있으며 또한 그 순서대로 출현될 것임을 나타낸다. #PCDATA는 해석 대상인 문자 데이터라는 의미로서 예약어이다. 데이터중에 실제 참조나 시작 태그 등의 마크를 포함하고 있는지의 여부를 SGML의 파서(parser)가 해석할 필요가 있다는 의미이다. #은 예약어를 나타내는 기호이다. SGML의 파서에 대해서는 다음에 설명한다.

<그림 3> 'mubook'의 문서 텍스트

```

<mubook>
<fm>
<ti> 문서기술언어 SGML에 의한 전문 데이터베이스의 구축
<aut> ISHIZUKA Hidehiro
<toc> I. 서론
      II. SGML의 개요
      III. SGML 문서 엔티티

      ((중간 생략))

<bdy> <chp> <chpti> I. 서론
<p> 최근에 워드프로세서, 전자출판 및 컴퓨터 조판의 ...
      ..... ((중간 생략)) ... 이루어지지 않고 있다.
<p> 1986년 국제표준화기구(ISO)는 .....
      ..... ((중간 생략)) ... 도구라고 할 수 있다.
      ((중간생략))

<chp> <chpti> II. SGML의 개요
<p> SGML은 1986년 ISO 8879로 ..... (중간 생략) ... 제정하였다.
      ((중간 생략))

</mubook>
    
```

DTD의 ELEMENT문에 의하여 문서의 구조를 다음과 같이 엄밀히 기술할 수 있다. 존재하지 않아도 되는 요소(optional, 출현빈도 0 또는 1) ‘?’로 나타낸다. bm, akm, gl 등은 optional이다. 장이나 단락 등은 1회 이상 반복하여 출현하는데 (1회부터 N회까지) 이러한 요소는 ‘+’로 나타낸다. 출현하지 않다가 반복하여 출현할 경우(0부터 N회까지)에는 ‘\*’로 나타낸다. 구성 요소에 따라서는 순서대로 나타나지 않고 어느 한 요소만 나타나는 경우도 있는데 이것은 | (OR 기호)로 나타낸다. 그밖에 & (AND 기호)는 &의 양측 요소가 반드시 나타나지만 순서에 관계없이 나타난다는 의미이다. 시작·종료 태그의 생략 여부에 대하여 설명하면 <그림 2>의 첫째줄은 ‘- -’, 둘째줄은 ‘- 0’로 표시되어 있는데 왼쪽은 시작 태그, 오른쪽은 종료 태그를 나타내고 ‘-’은 생략할 수 없음을, ‘0’은 생략할 수 있음을 나타낸다. 따라서 요소 mubook은 시작 태그, 종료 태그 모두를 생략할 수 없고, fm은 종료 태그를 생략할 수 있음을 의미한다. 시작 태그, 종료 태그는 다음 절의 문서 텍스트에서 설명한다.

## 2. 문서 텍스트

DTD에 따라 마크업된 텍스트를 문서 텍스트(Document text) 또는 문서 인스턴스(Document Instance)라고 한다. 문서 텍스트에는 SGML DTD의 마크가 삽입되어 있다. <그림 3>은 mubook의 문서 텍스트이다.

여기서 <mubook>, <ti>, <aut>, <p> 등 <요소명>의 형식으로 되어 있는 것은 각 요소의 시작점을 나타내는 마크로서 이것을 시작 태그라고 한다. 한편 요소의 종료점은 </mubook>과 같이 </요소명>으로 나타내는데 이것을 종료 태그라고 한다. <ti>나 <aut>의 종료 태그인 </ti>나 </aut>가 없는 것은 문서 텍스트용의 DTD에서 이 태그들을 생략할 수 있도록 허용하였기 때문이다. 다시 말하여 ti의 다음에는 aut, aut의 다음에는 toc가 나타난다고 정의되어 있기 때문에 <aut>가 나타나면 ti가 종료되고, <toc>가 나타나면 aut가 종료됨을 알 수 있으므로 해당 종료 태그를 생략할 수 있다. 한편 </mubook>은 전체의 종료를 나타내므로 이 종료 태그가 없으면 mubook의 종료를 판정할 수 없으므로 종료 태그 </mubook>은 생략할 수 없다.



이와같은 마크업 원고를 작성하는 것은 귀찮은 일이라고 생각할 수 있다. 그러나, 실제에 있어서는 <요소명>, </요소명> 없이 텍스트를 입력하는 방법이 있다. 예를들어 「SGML 대응 에디터」를 사용하여 입력자가 친숙한 기호를 DTD로 정의하고 이 기호를 시작 태그, 종료 태그로 하여 소프트웨어적으로 변환하는 방법 등이 있으며, 최근에는 워드프로세스 원고로부터 자동 변환시키는 시스템도 있다.

### 3. 비계층 구조의 표현

문서의 구조에는 장·절·단락 등의 계층 구조와 각주, 참고 문헌, 도표 등의 비계층 구조가 있다. 절은 장속에, 단락은 절속에 존재한다. 장이 절속에 존재하는 것은 있을 수 없는 일이다. 그러나 각주, 참고 문헌, 도표 등은 문서의 어느 곳이나 나타날 수 있다. 예를 들어 각주는 그 페이지의 하단에 기재하는 방식도 있으며 장의 끝에 기재하는 방식도 있다. 참고문헌도 마찬가지이다. 그림이나 표는 어느 곳에서도 존재한다. SGML에서는 이와같은 비계층 구조의 요소를 부동요소(floating element)라 한다.

또한 각주, 참고문헌, 도표 등은 존재 장소가 자유로울 뿐만 아니라 본문과의 사이에 참조 관계도 있다. 중요한 문헌이나 도표는 여러 번 참조되는 경우도 있다. 각주, 참고문헌, 도표 등은 존재장소의 면에서나 참조 관계면에서도 비계층적이다. SGML의 DTD에서는 이와 같은 비계층구조도 나타낼수 있도록 되어 있다.

#### (1) 부동 요소의 정의

먼저 그림이나 각주 등이 문서 데이터베이스 중의 임의의 장소에 존재할 수 있음을 DTD로 정의해 둘 필요가 있다. 예를 들어 다음과 같이 DTD를 정의한다.

```
<!ENTITY %floats "fig | fn">
```

```
<!ELEMENT mubook - - (fm,bdy,bm?) + (%floats;)>
```

여기서 %는 패러미터 엔티터를 나타내는 기호이다. 첫째줄은 %floats라는 이름의 패러미터값이 fig 또는 fn임을 정의하고 있으며, %floats의 내용은 그

림이나 각주이어야 함을 뜻한다. 둘째줄의 '+'는 포함(inclusions)을 나타내며 임의의 장소에 ( )내의 요소들이 존재하여도 됨을 의미한다. %floats 다음에 ;을 덧붙인 것은 % 등의 기호로 시작되는 엔티티를 인용할 때는 마지막에 ;를 덧붙이도록 규정되어 있기 때문이다.

## (2) 비계층 구조의 취급

각주, 참고문헌, 도표 등 비계층 구조를 표현하는 방법에는 ① 참조위치에 삽입하는 방법 ② 참조위치와 피참조 본체를 구분하여 ID로서 참조하는 방법의 두 가지 방법이 있다. 전자의 삽입 방법은 그 위치 이외에서는 참조하지 않을 때 편리하고 후자의 ID를 사용하는 참조링크법은 속성(attribute)을 사용하는 방법으로서 몇번이라도 참조할 수 있어 유리하다.

### 1) 삽입 방법

해당하는 요소를 DTD의 ELEMENT문으로서 정의 한 다음 본문중의 참조 위치에 삽입한다. 각주(fn)를 예로 들어 fn이 하나의 단락으로 구성되어 있다면 DTD는 다음과 같이 정의한다.

```
<!ELEMENT fm- - (p)>
```

본문중의 기술은 예를들어 다음과 같다.

「.....범용 마크업 언어 GML<fn> <p> 개발자 3인의 두문자를 나열한 것임. </fn>은 매뉴얼이나 명세서 .....」

이와 같이하여 작성된 SGML 문서를 인쇄하면 「.....범용 마크업 언어 GML」과 「은 매뉴얼이나 명세서.....」 사이에 각주 참조의 표시가 들어가고 「개발자 3인의 두문자를 나열한 것임.」이 각주로서 나오게 된다.

각주 이외도 마찬가지로이다. 그 요소를 ELEMENT문으로 정의하고 본문 중에서는 그 요소의 내용을 시작 태그와 종료 태그로 둘러싸서 삽입하면 된다. 여기서 ELEMENT문으로 정의한 구성 요소는 각각 다르다. 예를 들어 참고문헌의 경우 단락(p)이 없고 저자명과 제목 등의 서지 사항이 나열된다.

### 2) ID 참조 링크법

피참조 요소와 참조지시 요소를 ID를 사용하여 링크하는 방법이다. 삽입법과 달리 양자가 분리되어 있는 점이 특징이다. ID의 조작은 SGML의 속성 기능(ATTRLIST)을 사용한다. 각주를 예로 들어 설명한다.

피참조 요소에 대한 DTD는 다음과 같이 정의된다.

<!ELEMENT note - (p)>

<!ATTLIST note id ID#IMPLIED>

여기서 ATTLIST문은 속성 리스트를 선언하는 문으로서 그 형식은 다음과 같다.

<!, ATTLIST, 요소 명칭, 속성 명칭, 속성의 선언값, 속성의 디폴트 값>

속성 명칭으로 부터 속성의 디폴트 값까지는 묶어서 반복할 수 있다. note(각주)의 경우 요소 note가 id라고 하는 속성 명칭을 갖는데 그 선언된 값이 "ID" (식별자 값: Identifier value)인 경우 디폴트 값은 "#IMPLIED"나 "REQUIRED"이어야 한다. 디폴트 값이 #IMPLIED인 속성은 그 속성 값이 주어지지 않은 묵시 속성으로서 나중에 보충 가능성을 나타낸다. #REQUIRED인 속성은 속성 값이 주어진 필수 속성임을 나타낸다. 참조지시 데이터는 note(각주)의 refer(참조)라는 의미로서 그 요소 명칭을 ntref라고 하면 다음과 같은 DTD로 표시할 수 있다.

<!ELEMENT ntref - O EMPTY>

<!ATTLIST ntref refid IDREF #IMPLIED>

여기서, EMPTY란 구성 요소가 없음을 나타내고 IDREF(Identifier reference value, 식별자 참조 값)는 피참조 데이터의 식별자를 참조하라는 뜻이다.

따라서 문서 텍스트의 작성시에는 예를들어 다음과 같이 기술하면 된다. 피참조데이터는 「<note id = notel> 개발자 3명의 이름 첫 문자로 구성한 것이다. </note>」로 하면 본문에서는 「..... 범용 마크업 언어 GML<ntref id=notel>는 매뉴얼이나 명세서를 참고로 하여 .....」로 기술하고, 「..... GML<ntref id=notel> 이외에서도 시스템의 명명법으로는 .....」와 같이 본문의 다른 곳에도 동일한 각주를 참조할 수 있다.

### (3) 그림이나 화상의 취급

그림이나 화상은 그 번호, 제목 및 그림이나 화상 자체로 구성되어 있다. 번

〈그림 4〉                                  그림의 DTD

---

```

<!ELEMENT figure ..... (nf?, figcp?, figbdy?)>
<!ATTLIST figure      id      ID      #IMPLIED
                     file     CDATA   #IMPLIED
                     type     CDATA   #IMPLIED
                     sizex    CDATA   #IMPLIED
                     sizey    CDATA   #IMPLIED

<!ELEMENT figref - O EMPTY>
<!ATTLIST figref refid IDREF #IMPLIED>

```

---

호나 제목은 문자 데이터이다. 그러나 그림이나 화상 자체는 점의 조합으로 형성된 도트 데이터, 즉 비트 데이터인 것이 많으며 SGML에 의한 해석 대상에서 제외된다. 따라서 그림이나 화상의 기술 방법은 다음과 같이 조금 다르다.

- ① 그림이나 화상의 데이터는 별도 파일에 수록한다.
- ② 그림이나 화상의 표현 형식에는 PICT(사진), EPS(밀폐형 포스트스크립트), TIFF(컷 등의 간단한 그림) 등이 있는데 이들을 구별할 필요가 있다.
- ③ 그림이나 화상의 크기를 지정할 필요가 있다.

〈그림 4〉는 그림의 DTD를 나타낸 것인데 화상의 경우도 기본적으로는 동일하다.

여기서 nf는 그림의 번호, figcp는 제목, figbdy는 그림 자체를 의미한다. file은 그림이 수록된 컴퓨터 파일명으로서 그 값은 예를 들어 「b:\figure\figure1.dat」와 같이 할 수 있다. 즉, 디스크드라이브 B의 디렉토리 figure에 figure1.dat라고 하는 파일로서 보관되어 있음을 나타낸다. 또한, CDATA는 예약어로서 SGML의 해석 대상외의 문자 데이터임을 나타낸다. type는 PICT와 같은 그림의 표현 형식을 나타낸다. 그림의 크기는 가로, 세로의 sizex, sizey로 나타낸다. 요소 figref와 그 속성에 대해서는 이미 설명한 방법과 동일하다.

#### (4) 표의 취급

표는 DTD로서 정의하고 이에 따라 표자체의 데이터를 작성하고 속성을 이

용하여 참조하면 된다.

예를 들어 아래와 같은 표가 있다고 하자. 이 표의 DTD는 <그림 5>와 같이 정의 할 수 있다.

표는 표 번호, 표 제목, 칼럼 이름 및 표 자체로 구성된다. <그림 5>에서 - 과 -사이의 표시는 SGML 주석이다. #REQUIRED는 반드시 id값을 부여해야 한다는 뜻이다. 이를 #IMPLIED로 바꾸면 시스템에서 id의 값을 자동적으로 부여한다. <표 1>의 데이터는 <그림 5>의 DTD에 의하여 다음의 <그림 6>과 같이 입력한다. 한편 본문에서 표의 내용을 인용할 때는 「공작기계의 보유 현황은 <tblref refid=tbl 1>와 같다.」와 같이 입력하면 된다.

<표 1> 공작 기계 보유 현황

기 종 명	수 량(대)
밀 링 기	60
NC 선반	20
기어 연삭기	30
평면 연삭기	20
공구 연삭기	10
머시닝 센터	50

<그림 5> 표1의 DTD

```

<!ELEMENT table -- (nt?,ht?,hc?,bt+)>
<!ATTLIST table      id      ID              #REQUIRED
                    cols    NUMBER        #REQUIRED>
<!ELEMENT nt        - 0  (#PCDATA)          - number of table
<!ELEMENT ht        - 0  (#PCDATA)          - heading of table
<!ELEMENT hc        - 0  (r)+-(bt)          - headint of columns
<!ELEMENT bt        - 0  (r)+              - body of table
<!ELEMENT r         - 0  (c | hc | bt)+     - row of table
<!ELEMENT tblref    - 0  EMPTY>
<!ELEMENT tblref    refid  IDREF #REQUIRED>
    
```

〈그림 6〉

〈표 1〉의 데이터 입력

---

```

<table id="tbl 1" cols = 2>
<ht> 공작 기계의 보유 현황
<hc><c>기종명 <c>수량(대) </hc>
<bt>
<r><c> 밀 링 기 <c> 60
<r><c> NC 선반 <c> 20
<r><c> 기어 연삭기 <c> 30
<r><c> 평면 연삭기 <c> 20
<r><c> 공구 연삭기 <c> 10
<r><c> 머시닝 센터 <c> 50
</table>

```

---

#### IV. ISO의 전자 출판 모델

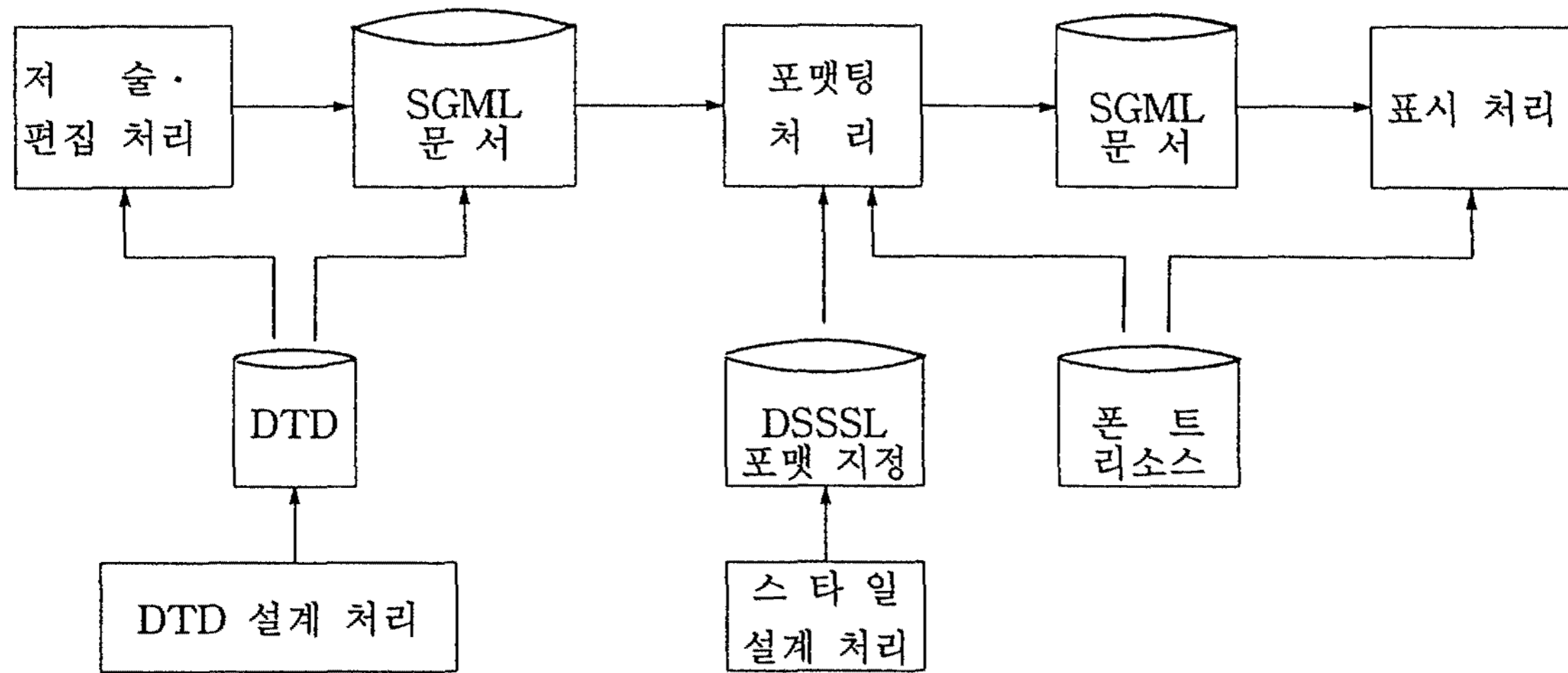
SGML은 출판 인쇄 분야에서 도출된 규격임에도 불구하고 지금까지의 설명은 전문 데이터베이스를 중심으로 한 것이고 인쇄 출판에 관한 것이 아니었다. SGML에서는 글자의 크기, 도표의 위치 등을 지정하는 편집 배정이나 페이지·레이아웃 등 인쇄에 관한 것이 별도의 독립된 규격으로 분리되어 있기 때문이다.

여기서, ISO의 전자출판 모델을 소개한다. 이에 의하면 출판·인쇄 과정에서 SGML이 어떤 역할을 담당하는지 확실해질 것이다. 전자출판이란 종이에 인쇄할 뿐만 아니라 검색 기능을 가진 전자 매체, 예를 들어 CD-ROM으로 출판하는 것도 포함한다.

보다 넓은 의미로서는 온라인·네트워크상의 전문 데이터베이스 검색 시스템도 포함한다.

ISO의 전자출판 모델은 〈그림 7〉과 같이 나타낼 수 있는데 ① SGML에 의한 마크업 ② DSSSL에 의한 편집지정 ③ SPDL에 의한 페이지 표시 등 3개로 구성되어 있다. 종래의 CTS(computerized typesetting, 전산 사식)에

〈그림 7〉 ISO의 전자 출판 모델



의한 출판은 원고에 인쇄용의 제어기호(명령어)를 수작업으로 삽입하는 것으로, 말하자면 인쇄를 위한 작업이었다.

이에 반하여 ISO의 전자 출판 모델에서는 문서 구조를 기본으로 하고 인쇄에 관한 것 (DSSSL, SPDL)은 분리하고 있다. 또한 CTS에서는 파일이 만들어 지지만 데이터 항목이 구별되어 있지 않아 데이터베이스라고 할 수 없다. 한편 SGML로서는 전문 데이터베이스가 만들어진다.

(1) SGML에 의한 마크업

SGML 마크업에 의하여 문서의 구조를 나타낸다. 문서를 구성하는 요소를 명시하고 요소 사이의 관계도 나타낸다. 이 방법에 의하여 문서는 단순한 텍스트 파일이 아니라 문서 데이터베이스가 된다. 이렇게 작성된 데이터베이스는 단순히 인쇄 측면에서 뿐만 아니라 CD-ROM 측면에서도 그리고 교환 유통에도 사용할 수 있다.

(2) DSSSL에 의한 편집 지정

DSSSL이란 Document Style Semantics and Specification Language ISO/DIS10179의 약어로서 문서의 요소마다 서체, 글자 크기, 한 줄의 글자 수, 다음 줄로의 물림 등을 지정한다. 여기서 ISO/DIS란 ISO의 규격으로 되기 직전의 DIS단계임을 나타낸다. DSSSL 규격에는 SGML파서와 포맷터가 포함되어 있다.

### (3) SPDL에 의한 페이지 기술

SPDL이란 Standard Page Description Language ISO/DIS10180의 약어로서 페이지 기술(記述) 언어이다. PostScript의 ISO판으로서 포맷을 끝낸 정보에 페이지·이미지를 만드는 출력 장치의 표준 인터페이스 언어이다. DSSSL은 각각의 요소마다 지정하지만 SPDL은 각 요소를 편성하여 페이지·이미지를 만든다.

### (4) 폰트 리소스

스타일 지시, 포매팅, 출력의 각 처리로부터 참조하여 각각의 통일된 폰트(font) 정보를 제공한다. DTP 등의 문서 정형에 대한 요구가 고도화됨에 따라 중요도가 점점 커지고 있다.

이상이 ISO 전자 출판의 모델이다. CD-ROM은 이상의 모든 단계를 거쳐 만들수도 있고 DSSSL에 의한 편집 지정이 종료된 단계나 SGML에 의한 마크업이 종료된 단계에서도 만들 수 있다.

그런데, SGML에 의하여 만들어진 전문 데이터베이스를 인쇄하는 경우에는 DSSSL과 SPDL의 기능을 가진 소프트웨어(포맷터)가 필요하다. SGML용의 포맷터가 없어도 인쇄 소프트웨어가 SGML에 대응하도록 하여 SGML 전문 데이터베이스로부터 그 인쇄 소프트웨어의 형식으로 변환하여도 된다.

DSSSL과 SPDL의 기능을 병합한 인쇄 포맷터로서는 1984년 Knuth가 개발한 TeX와 1986년 Lamport가 개발한 LaTeX가 있다. 이 소프트웨어를 이용하면 SGML의 문서 데이터로부터 TeX나 LaTeX의 커맨드(명령)가 포함된 파일을 자동 생성한다. 이것을 TeX나 LaTeX의 처리계에 걸면 인쇄 원고가 나온다. 특히 LaTeX는 문서 구조를 의식하고 있기 때문에 자동 생성이 용이하다.

## V. SGML의 효과

SGML을 적용하여 문서를 작성하면 전문 데이터베이스의 작성 유지, 교환



유통, 검색용 데이터베이스나 전자 도서의 작성 등이 용이하게 된다.

### (1) 작성 유지

전문 데이터베이스의 작성 유지로서는 예를 들어 메뉴얼의 작성과 내용관리  
가 있다. 문헌을 데이터베이스화 함으로써 빈번히 일어나는 개정 작업에 신속  
히 대응할 수 있다.

### (2) 교환 유통

SGML에 의하여 작성된 데이터베이스는 문헌의 교환 유통에도 적합하다. 각  
각의 시스템으로 부터 독립되어 있을 뿐만 아니라 구조가 엄밀하게 정의되어  
있기 때문에 각 시스템에 적합하도록 변환시킬 수도 있다. 교환 유통이라면 플  
로피나 전자메일로서도 가능하다는 주장이 제기될 수도 있다. 그러나 우리는  
무의식적으로 문서의 구조를 암시적으로 정하여 글을 쓴다. 예를 들어 제목의  
글자를 크게 한다든지, 인용을 표시할 경우에는 글자를 작게 한다든지, 절의 첫  
줄은 한 글자 들여 시작한다(indent). 이와 같은 암시적인 구조 표현은 기존의  
워드프로세서에서는 호환성이 없다. 글자의 크기를 바꾸는 방법은 다른 워드  
프로세서 소프트웨어에서는 잘 되지 않는다. 또한 인덴트는 한 줄당의 글자수  
가 바뀌면 파괴되어 버린다. 이러한 것들은 다른 워드프로세서 사이에서 플로  
피 디스크를 이용하여 문서를 교환해 본 사람들이 자주 경험하는 일이다.

### (3) 전자 도서

전자 도서의 하나의 형태로서 최근 주목을 받고 있는 하이퍼텍스트도 문서  
의 구조를 의식하고 있는데 각 구성요소 끼리의 링크(예를 들어 「장」에서  
「절」, 「본문」에서 「그림」이나 「표」 또는 「주」등)가 기본 구조이다. 이 때문  
에 SGML에 의한 전문 데이터베이스로 부터 자동 변환이 가능하다. 예를 들  
어 인터넷의 WWW는 SGML에 의한 하이퍼텍스트 시스템이다. 다만 DTD는  
전용의 DTD이므로 이에 맞는 데이터를 준비할 필요가 있다.

또한, 전자 도서를 더욱 진척시키면 SGML로 만들어진 전문 데이터베이스  
는 전자도서관 시스템의 자료로서 유용하게 된다. 예를 들어 워크스테이션을  
LAN으로 연결한 전자도서관 시스템으로서 미국 카네기멜론 대학의 Mercury  
나 코넬 대학의 CORE가 유명한데 모두 SGML을 채택하고 있다.

## VI. SGML의 적용 사례

먼저, SGML이 어느 분야에서 사용되는가를 살펴본 다음에 학회지에의 적용에 대하여 기술한다. SGML이 사용되는 분야는 다음과 같다.

### (1) 사내 문서의 작성

미국 IBM사나 Hewlett-Packard사에서는 컴퓨터 소프트웨어의 이용자 매뉴얼이나 명세서를 SGML의 방식으로 제작하고 있다. IBM사는 SGML의 기초가 된 GML을 개발한 회사이다.

### (2) 정부기관에서의 채택

미국에서는 국방부의 CALS(Computer aided Aquisition and Logistics Support : 미국 국방부가 조달하는 물자에 대한 방대한 정보를 전자문서화하는 프로젝트)를 비롯하여 표준국, 정부인쇄국, 특허청에서도 SGML을 채택하고 있다. EC에서는 의사록의 데이터베이스가 SGML 방식으로 작성되고 있다. ISO도 규격 서안 등의 작성에 SGML을 적용하고 있다.

### (3) 학술 도서의 작성

미국, 영국을 중심으로 학술 도서의 출판에 SGML 방식이 점차로 채택되고 있다. SGML을 채택하게 된 이유는 새로운 지식이 도출될 때마다 이에 신속히 대응하여 개정하기가 쉽기 때문이다.

### (4) 학술 논문의 작성

일본의 정보지식학회는 2년간의 검토와 준비를 거쳐 1990년말에 학회지를 SGML 방식으로 출판하였다. 투고자가 작성한 전자 원고로 부터 학회지용의 SGML방식 전문 데이터베이스를 구축하고 인쇄는 SGML에 링크된 LaTeX를 이용하였다. 복잡한 SGML의 태그를 사용하지 않고 보통의 워드프로세서를 사용하여 원고를 작성할 수 있는 간이 마크업법을 사용한 것이 특징이다.

미국 화학회는 Chemical Abstracts Service가 중심이 되어 출간되는 화학 관련 학회지들을 SGML을 적용하여 전문 데이터베이스로 제작할 것을 검토하고 있다.

일본 화학회도 1993년부터 그 영문 학회지를 전문 데이터베이스로 제작하

고, CD-ROM과 그 검색 시스템도 시험 제작하여 보급하고 있다.

## VII. 검색 시스템

SGML에 의한 전문 데이터베이스의 검색은 특정한 검색 시스템을 필요로 하지 않는다. 전문 데이터베이스의 검색 기능을 갖춘 시스템이라면 어느 것에도 탑재하여 사용할 수 있다. 예를 들어 일본화학회지(영문판) 데이터베이스는 학술정보센터의 전문 데이터베이스 검색 시스템에 탑재하고 있다. 이 시스템에서 도표는 FAX로 송신하고 있다. 또한, 일본 화학회는 학회지(영문판) 데이터베이스를 화학계 잡지의 전문 데이터베이스가 다량 탑재되어 있는 STN International의 검색 시스템에 탑재할 것을 검토하고 있다. STN International은 도표 제공 기능을 계획중에 있다.

데이터베이스 중에 도표나 사진이 있는 경우 도표도 표시할 수 있는 검색 시스템이 바람직하다. 또한, SGML에 의한 전문 데이터베이스는 장, 절, 단락 등의 계층 구조나 도표, 참조문헌 등 참조관계의 구조를 가지고 있으므로 이를 살리는 시스템, 예를 들어 하이퍼텍스트 시스템 또는 이와 같은 종류의 기능을 가진 시스템이 바람직하다. 일본화학회는 MS Windows 3.1의 MS Multimedia Viewer 상에서 검색되도록 학회지(영문판) 데이터베이스의 CD-ROM판을 시험 제작하였다. 또한, UNIX상의 하이퍼텍스트 시스템도 시험 제작하였다.

하이퍼텍스트 기능을 갖추고 SGML용 인터페이스 기능을 가진 검색 시스템 으로서는 Guide Professional Publisher 의 「Guide」나 Electronic Book Technology의 「DynaText」 및 CERN의 「WWW」가 있다.

## VIII. SGML 관련 도구

SGML 방식의 전문 데이터베이스를 제작하여 이용할 때 어떠한 과정을 거

치는가, 또한 어떤 SGML 관련 도구를 써야 하는가를 설명한다.

### (1) DTD의 작성

전문 데이터베이스용의 DTD는 기존의 다른 데이터베이스의 DTD를 개량하거나 나름대로의 사용에 편리하도록 DTD를 직접 작성한다. 작성한 DTD가 SGML의 문법에 합치되는가의 여부는 SGML 파서(parser)로 체크한다.

파서로서는 예를 들어 SEMA Group의 「Mark-It」, Exoterica의 「OmniMark」, 일본 후지사의 「SGML/Base」등이 있다.

### (2) 문서 텍스트의 작성

DTD에 따르는 문서 텍스트를 작성하는 방법은 SGML에디터를 사용하는 방법과 기존의 워드프로세서 소프트웨어를 사용하는 방법이 있다. SGML 에디터를 사용하면 프롬프트에 따라 해당 데이터를 입력함으로써 용이하게 문서 텍스트를 작성할 수 있다. 워드프로세서 소프트웨어를 사용하여 SGML 문서 텍스트를 작성하려면 ① 특정한 기호를 DTD에 설정하여 입력하고 SGML 파서를 써서 변환하거나 ② 워드프로세서 소프트웨어의 스타일 시이트의 기능을 이용하거나 ③ 워드프로세서 소프트웨어의 매크로 기능을 이용하거나 ④ SGML로의 변환 소프트웨어를 이용해야 한다.

그러면 시작 태그나 종료 태그 없이 문서 텍스트를 작성할 수 있다. ②와 ③의 경우는 SGML 파서를 거쳐 체크하는 것이 좋다.

SGML 에디터로서는 예를 들어 SoftQuad사의 「Author/Editor」, 일본 松下電送의 「SGML-Plus」등이 있다. 또한 워드프로세서 형식으로 부터 SGML 형식으로의 변환 소프트웨어도 최근에 많이 개발되고 있다.

### (3) 문서 텍스트의 인쇄

SGML 문서용의 포맷을 사용하거나 SGML 형식에서 LaTeX로 변환하여 인쇄하는데 SGML 대응 기능을 갖춘 인쇄 소프트웨어를 사용하는 방법도 있다. 예를 들어 LaTeX 출력이 가능한 것으로서는 「Mark-It」, 「SGML/Base」, 「SGML-Plus」 등이 있다. SGML 대응의 인쇄 소프트웨어로서는 일본 유니텍크의 「GRACE」, Interleaf의 「Interleaf5<SGML>」등이 있다.

#### (4) 전자 문서의 작성

SGML에 대응하는 전자 문서의 작성 프로그램으로서는 이미 언급한 「DynaText」, 「WWW」, 「Guide」 등이 있다. 이 이외에도 많은 변환 프로그램이 있다.

### IX. 앞으로의 전망

SGML에 의한 전문 데이터베이스의 구축은 이미 미국, 유럽 등에서는 널리 보급, 이용되고 있으며 일본에서도 최근 비상한 관심을 모으고 있다. 일본의 과학기술진흥사업단(JST; JICST와 신기술사업단의 통합기관)에서도 「정보 관리」등 정기 간행물과 매뉴얼, 사내 문서의 제작에 SGML을 적용할 계획을 세우고 JST-DTD의 개발에 착수하였다. 우리 나라도 정보화 시대에 부응하여 학술잡지의 전자 출판을 위한 SGML의 연구개발과 그 실천에 곧 착수해야 할 것으로 사료된다.

SGML에 동화상이나 음성을 도입한 HyTime도 ISO 규격화하려는 움직임이 있다. 이에 따라 SGML은 정보화 사회에 있어서 멀티미디어에 의한 정보 교환에 큰 역할을 할 것으로 기대된다.

#### 〈참 고 문 헌〉

- KS C 5913, “문서기술언어 SGML”, 한국표준협회, 1993.
- 田中洋一, “文書記述言語SGMLとその動向”, 情報處理, 32(10), 1991, pp. 1118-1125.
- 伊藤卓, “日本化學歐文誌の全文データベース化と電子出版化への移行について”, 化學と工業, 46(1), 1993, pp. 66-69.
- 石塚英弘, “SGML 形式による學會誌全文データベースの構策と印刷”, 情報知識學會誌, 2(1), 1991, pp. 23-48.
- ISO 8879-1986, “Information processing-Text and office systems-Standard

- Generalized Markup Language (SGML)”, 1986.
- ISO 8879/AMENDMENT 1, 1988.
  - ISO/IEC JTC 1/SC18/WG 8N 1078, “Operational Model for Text Description and Processing Language”, 1990.
  - ISO/IEC JTC 1/SC18/WG 8N 10179, “Document Style Semantics and Specification Language (DSSSL)”, 1991.
  - ISO/IEC DIS 10180, “Standard Page Description Language (SPDL)”, 1991.
  - ISO/IEC DIS 954-1.2, “Font information interchange—Part 1: Architecture”, 1990.
  - ISO/IEC DIS 954-1.2, “Font information interchange—Part 2: Interchange Format”, 1990.
  - ISO/IEC DIS 9541-3, “Font information interchange—Part 3: Glyph Shape Technologies”, 1991.
  - Bryan, M., “SGML: An Author’s Guide to the Standard Generalized Markup Language”, Addison—Wesley, 1988.
  - Goldfarb, C. F., “Document composition facility Generalized Markup Language: Concepts and design guide”, IBM Sh 20-9188, 1984.
  - Knuth, D. E., “The TeXbook”, Addison—Wesley, 1984.
  - Lamport, L., “LaTeX : A Document Preparation System”, Addison—Wesley, 1986.
  - Nielsen, J., “HYPERText & HYPERMedia”, Academic Press, 1990.