정보문서 표준화를 향한 SGML기술

송 윤 철/뉴미디어사업부 멀티미디어팀장

SGML 개요

현대는 "고도 정보화 사회"이다. 즉, 전산화된 정보의 사회적 비중이 날로 확대되고 있다. 그리고 이러한 정보들은 특정 시스템이나 응용소프트웨어에 의하여 작성되기 때문에, 컴퓨터 기종이 다를 경우나 응용소프트웨어의 버전이 올라갈 때 기존 정보가 그 대로 이용될 수 없어 재구성되거나 변환되어야 할 필요가 있었다.

이와같은 문제점을 해결하기 위하여 국제 표준화 기구(ISO : International Organization for Standardization)에서 문서처리에 대한 표준을 발표하였는데, 그것이 ODA(Office Document Architecture)와 SGML(Standard Generalized Markup Language)이다. 그 중에서 ODA는 페이지, 페이지 영역, 행 등의 레이아웃(layout)조직과 장, 절, 단락 등으로 구성되는 논리적인 조직으로 표현하며, SGML은 주로 문서의 논리구조와 속성을 이용하여 문서를 표현한다.

SGML은 문서내에서 어느 부분에 해당하는 지를 인지하는 기능과 그 부분을 처리하는 기능사이에 분명한 구분을 두는 서술적 마크업(descriptive markup) 개념에 기초한다. 이 서술적 마크업은 기존의 문서처리기에서 주로 사용되어 온 절차적 마크업(procedural markup)에서의 단점인 문서의 논리적 구조에 관한 정보를 기억하지 못하는 점과 문서의 스타일을 변형시킬 필요가 발생하였을 때의 번거러운 작업이 필요한 것 등을 개선한 마크업 방법이다.

SGML에 의한 마크업은 문서의 구조와 그에 따르는 속성들을 표현하여야 하는데 그것에 수행될 처리는 지정하지 않는다. 그리고 마크업은 규칙을 엄격하게 적용하여야 오차없이 정의된 처리에 유용하게 이용될 수 있을 것이다. 문서의 논리적 기본단위인 엘리먼트의 명칭을 공통식별자(GI:Generic Identifier)라 하고 GI는 SGML문서내의 각 시작 태그에서 선택적으로 속성을 가질 수 있으며 이 속성은 엘리먼트에 부가적인 정보를 제공한다.

SGML표준은 마크업 문법을 정의할 뿐이지 특정한 GI의 어휘들을 지정하지는 않는다. 그것은 SGML사용자의 뜻대로 임의로 정할 수 있으며 SGML문서의 한 부분인 문서형 정의에 기술된다. 이러한 SGML은 미국의 국방성에서의 군수 지원과정에 소요되는 시간과 비용을 절감하기 위한 "군수조달 및 관리정보시스템 계획"에 이용되었고, 현재는 민간부문의 활용에 까지 확산 되었다. 멀티미디어 문서와 HTML 등에 이용되고 있음은 물론이며, 특히 HTML은 인터넷의 World Wide Web(WWW)에서 상호 교환을 위한 문서형식으로서 인터넷에 기여하는 바가 크다.

SGML 구조

SGML문서는 문자로만 이루어지는 것이 아니고, 그림, 도형, 그리고 소리까지도 포함 될 수 있다. SGML문서는 ISO 8879에서 정한 엄격한 문법을 적용하여 문서의 구조 및 내용을 기술함으로써, 문서 구조의 기본단위인 엘리먼트들이 갖는 속성도 표현할 수 있다. 그리고 세계 여러나라에서 쓰이고 있는 각종의 언어를 사용한 문서도 SGML화 할 수 있는 방법을 제공하며, 한 문서에서 반복되는 문자열의 입력을 쉽게하기 위하여 엔터티를 선언하여 참조할 수 있게 한다.

SGML문서는 크게 세부분으로 분류되는데, 첫째는 SGML 선언부(SGML Declaration)로서 사용될 언어 및 글자들의 집합을 선언하고 엘리먼트의 수, 토큰의 수 등의 상한값을 정하게 된다. 둘째는 엘리먼트 선언에 의하여 SGML문서의 구조를 내포하며 각 엘리먼트가 가질 수 있는 속성선언, 그리고 단축 참조를 위한 엔터티 선언을 포함하는 문서형태 정의부(DTD : Document Type Definition)이며, 셋째는 실제의 문서를 DTD에 맞추어 작성한 SGML문서이다.

여기서 SGML선언부는 SGML문서가 사용할 문자의 집합과 코딩규칙 등의 주요 구체적 문법을 정의하며, SGML문서가 갖는 특수한 특징을 표현하기도 한다. 이것은 컴퓨터 내부처리시에 이용될 수도 있지만 인쇄된 형태로서 인간의 이해를 돕기도 한다. 즉, SGML문서를 수신할 때에 시스템이 문자의 어떤 번역을 하여야 하는지와 SGML의 기능 중 처리 가능한 것은 무엇무엇인지 등을 알 수 있게 한다.

문서 형태 정의부에서는 엘리먼트를 단위로 하는 트리 구조와 유사한 일종의 계층 구조를 이용하여 SGML문서의 구조를 정의하며 ISO 8879에서는 문서형태 정의부 자체를 작성하는 문법도 정의되어 있다. 문서형태 정의부에서는 어떤 특정부류의 문서에 적용하는 마크업 선언들의 집합이며, 시작부분에서 문서형태 고유의 명칭이 선언되는데, 크게 엘리먼트 선언, 속성정의 선언, 그리고 엔터티 선언의 세가지로 구분된다. 엘리먼트 선언은 각 엘리먼트의 내부에 존재할 수 있는 부엘리먼트들에 대한 공통 식별자들을 내용 모델로서 선언하며, 이때 부엘리먼트들의 출현순서, 반복출현 여부, 그리고 없을 수도 있는 선택성들을 seq(,), or(|), and(&)의 연결자와 opt(?), plus(+), rep(*)의 발생지시자를 이용하여 표현한다. 이때 연결자와 발생지시자를 잘못 사용하면 엘리먼트의 선언이 모호하게 되어 구문을 분석할 때 에러가 발생할 수 있으므로 엘리먼트의 내용 모델의 논리구조를 보다 명확하게 파악하여 주의 깊게 적용하여야 한다. 아래는 SGML DTD의 간단한 예이다.

```
〈!DOCTYPE sampldoc[
〈!ENTITY % paradt "#PCDATA | para "〉
〈!ELEMENT sampldoc - - (title, chapter+, appendx*)〉
〈!ELEMENT title - O (#PCDATA)〉
〈!ATTLIST title id ID #IMPLIED〉
〈!ELEMENT chapter - O (title?, section+ | #PCDATA)〉
〈!ELEMENT section - O (para*, %paradt;)〉
〈!ELEMENT (appendx | para) - O (%paradt;)〉
]〉
```

그림 1. SGML DTD의 예

위의 SGML DTD에서 〈!DOCTYPE는 그 다음에 주어진 sampldoc라는 이름의 문서 형태를 정의하기 시작한다는 의미이며, [다음부터 시작하여]이 올 때까지가 그 문서형태의 내용이 된다는 뜻이다. 그러

므로 정의한 문서형태 부류의 이름은 sampldoc인 것을 알 수 있으며 문서의 구조가 이 sampldoc DTD에 만족하는 실제 문서는 다수가 있을 수 있는 것이다. 즉, DTD는 어떤 문서들의 기본구조 형태를 정의한 것으로서 실제 문서의 내용은 다를 수 있다는 의미이다.

엔터티 선언을 위한 문법은 〈!ENTITY entity_name" 엔터티 내용"〉이며 위 예에서는 하나의 엔터티 선언이 있으며 엔터티 내용은 "#PCDATA:para"이고 그 이름은 paradt이다. 이 paradt엔터티를 참조하려면 %paradt; 라고 쓰며, 그렇게 하면 #PCDATA:para라고 한 것과 같은 효과를 갖는다. 여기에서의 %과 :도 엔터티 참조를 위한 SGML 어휘이다. 이 예제의 경우에는 엔터티의 내용이 비교적 짧으므로 큰 차이가 없지만 엔터티의 내용이 50여 글자 정도로 길고, 자주 참조된다면 효과가 좀더 크게 느껴질 것이다.

엘리먼트 선언을 위한 SGML 문법은 〈!ELEMENT element_name S E (content_model)〉인데 element_name은 엘리먼트의 식별자가 되는 엘리먼트의 명칭이고 S와 E는 각각 시작 태그와 끝 태그의 생략가능 정보를 표시하는 것으로서, - 는 생략이 불가능함을 의미하고 O는 생략이 가능함을 의미한다. 그리고 괄호 안의 content_model은 선언중인 엘리먼트의 내용이되는 부엘리먼트들과 파싱되는 SGML 문자열인 #PCDATA 등의 조합으로 구성되어 엘리먼트의 구조를 나타내며, 모든 엘리먼트 선언들의 구조가 모여서 문서형태를 표현한다. 이때 연결자와 발생지시자가 중요한 역할을 하는데 그 각각의 의미는 아래와 같다

연결자:

seq(,) --- 순서대로 , 연결자 전 후의 내용이 존재한다

or(|) --- | 연결자 앞의 내용과 뒤의 내용 중 하나만 존재한다

and(&) --- 순서에 관계없이 &연결자 전 후의 내용이 모두 존재한다

발생지시자:

opt(?) --- ? 발생지시자 앞의 내용이 있을 수도 있고 없을 수도 있다.

plus(+) --- + 발생지시자 앞의 내용은 하나 혹은 두 개 이상 있을 수 있다.

rep(*) -- * 발생지시자 앞의 내용은 없거나 하나 이상 있을 수 있다

다시 위의 sampldoc DTD로 돌아가서 제일 먼저 선언한 sampldoc 엘리먼트 선언을 보면

〈!ELEMENT sampldoc - - (title, chapter+, appendx*)〉이고 엘리먼트의 명칭은 sampldoc이며, 그 다음에 있는 두개의 -(마이너스)는 각각 시작 태그가 생략될 수 없음과 끝 태그가 생략될 수 없음을 의미한다. 그리고 괄호안의 title은 부엘리먼트로서 다음에 오는 부엘리먼트인 chapter보다 순서가 앞서야 한다. 그것은 seq(,)연결자가 있기 때문이다. chapter 뒤에 붙은 +는 발생지시자로서 chapter는 반드시 존재하여야 하며 두개 이상 올 수도 있음을 의미한다. 그 다음 부엘리먼트인 appendx는 앞에 seq(,) 연결자가 있으므로 chapter들 보다 뒤에 와야하고, 발생지시자 rep(*)에 의하여 없을 수도 있고 한개 또는 그 이상 있을 수 있음을 의미한다. 그 아래의 엘리먼트들도 같은 방법으로 해석할 수 있다. 맨 끝에 선언한 엘리먼트의 이름은 (appendx | para)인데 이것은 명칭 그룹이라고 하는데 엘리먼트의 내용이 같을 때 사용하며 동시에 두 엘리먼트를 선언하는 셈이다.

끝으로 속성 선언이 하나 있는데 〈!ATTLIST title

id ID #IMPLIED〉이며 title은 지금 선언중인 속성을 갖는 엘리먼트를 가리키고 id는 속성의 명칭, 그 다음의 ID는 속성이 갖는 값의 형태가 식별자라는 것을 나타낸다. 그림 1의 DTD가 내포하는 문서의 형태를 그림으로 표현하면 그림 2와 같다.

문서의 구조를 표현하는 방법으로 발생지시자와 연결자를 덧붙여 사용하여 의미를 부여 하였다. 다음에 위의 DTD에 만족하는 실제의 SGML 문서의 한 예를 들어서 설명해 본다.

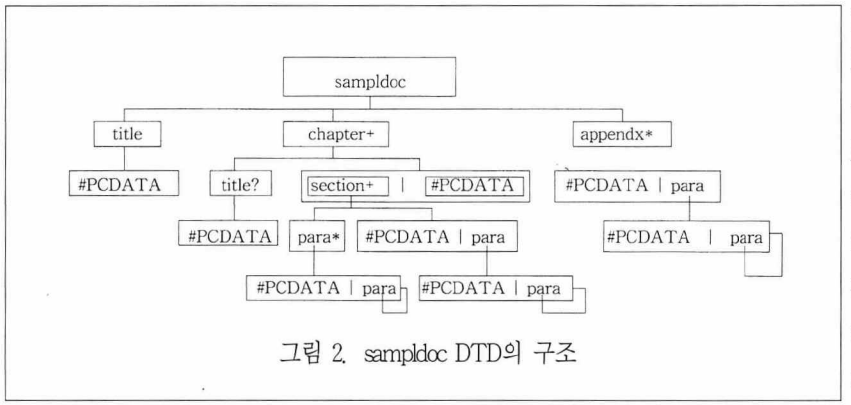

그림 2. sampldoc DTD의 구조

```
〈sampldoc〉
  〈title〉 UNIX의 이해 〈/tile〉
  〈chapter〉
    〈title〉 시스템의 개관 〈/title〉
    〈section〉 역사
      〈para〉 1965 벨 전화 연구소의      MAC 프로젝트와 공동으로_
      〈para〉 Multics 프로젝트가 끝남에 따라_
        .
        .
      〈/section〉
      〈section〉 시스템의 구조___
      〈section〉 사용자의 관점____
  〈chapter〉
      〈title〉 커널의 개관
      〈section〉 UNIX 운영 체제의 구조___
      〈section〉 시스템 개념의 소개 ____
〈/sampldoc〉
```

위에서 〈는 시작태그의 시작을 의미하고, 〈/는 종료태그의 시작을 의미하며, 〉는 시작태그와 종료태그의 끝을 의미한다. 그러므로 〈sampldoc〉은 sampldoc 엘리먼트의 시작태그 이고, 〈/sampldoc〉끝태그 이다. 그러므로 〈sampldoc〉과 〈/sampldoc〉의 사이가 sampldoc엘리먼트의 내용이다. 2행의 title 엘리먼트에서는 시작태그 〈title〉과 끝태그 〈/tit:e〉 사이의 " 시스템 개관"이 내용이다. DTD에서 title, chapter, section, para 등의 엘리먼트는 끝태그를 생략할 수 있도록 정의되었다. 그러므로 종료태그는 문맥상 혼동을 주지 않는 한 생략하여도 좋다. 다음에 chapter 엘리먼트에는 title 엘리먼트와 section 엘리먼트가 있는데, section은 복수개가 허용이 되어 있었고 그 하위에 para 엘리먼트들이 포함 된다. .....부분은 문서의 일부가 생략됨을 의미한다.

결언

이상으로 SGML 및 SGML문서에 대하여 간단한 설명을 하였다. 이러한 SGML은 ISO표준으로서 문법이나 규칙을 규정한 것으로서 워드프로세서나 Text 등과 같이 그 자체를 바로 사용할 수 있는 것은 아니다. 그리고 SGML이 특정한 문서형을 정의한 것도 아니다. SGML을 적용할 문서들에 맞는 문서형을 정의하는 규칙을 정했을 뿐이다.

따라서 SGML을 활용하기 위해 모든 사용자가 SGML을 완벽하게 이해하여야 한다면 누구나 가까

이 하기 싫어할 것이다. 일반 사용자들이 쉽고 편리하게 SGML을 이용할 수 있도록 전문가들이 SGML 환경을 개발하여야 하는 것이다. 개발이 필요한 SGML환경으로는 SGML 문서를 구문 분석하여 검사하고 오류를 찾으면서 문맥을 파악할 수 있는 파서가 있고, SGML 문서를 만들 수 있게 하는 입력 시스템인 SGML에디터, 그리고 구문 분석된 SGML 문서에 사용자가 요구하는 처리를 할 수 있는 처리기가 필요하다.

SGML 에디터는 파서를 내부에 포함하여 SGML DTD의 식별자 및 입력중인 식별자의 내용 모델도 파악하여 사용자가 입력화면에서 엘리먼트 태그들을 알아볼 수 있도록 표시하는 것이 좋으며, 사용자가 내용을 바르게 입력하는지를 확인하고 잘못 입력할 경우 적절한 메시지를 보여야 하겠다.

SGML 문서 처리기도 파서를 이용하여 문서의 문맥을 인식하여야 특정한 처리를 하는 것이 가능하다. 그 처리에는 문서의 표현이 있을 수 있는데, 보통의 텍스트 뿐아니라 그래픽이나 수식, 소리 등의 내용도 포함할 수 있다.

그리고 SGML 문서는 데이터베이스에도 이용할 수 있으며 시스템환경에 독립적이므로 생명주기가 길다는 장점이 있다. 이러한 SGML문서는 다목적으로 이용할 수 있는 특성을 갖는다. 예를 들면 같은 문서가 어떤 제품의 상표에도 들어갈 수가 있고 글자의 폰트, 크기, 색상, 그리고 배치를 다르게 하여 카탈로그 제작 등에도 이용 할 수 있다.

따라서 SGML은 컴퓨터를 이용한 출판, 전자출판, 문서변환, 그리고 HTML 등 다양한 분야에 응용이 가능하다.

한편, 이러한 SGML과 관련한 세미나가 한국정보통신협회 멀티미디어협의회 SGML분과위원회 주관으로 지난 6월11일 한국과학기술회관 대강당에서 산·학·연 관계자 약 300여명이 참석한 가운데 "SGML/XML의 기술동향과 전망"이라는 주제로 개최된 바 있으며, 여기서는 SGML개요 효과, 솔루션 및 적용시스템 소개와 사례발표를 통하여 문서 표준화 도구로서 SGML채택의 필요성이 강조됨과 함께 차세대 정보포멧표준인 XML기술이 소개되었다.

인터넷 인증시험 실시

협회는 지난 '98. 7. 5일 14시부터 16시까지 2시간동안 전국 9개 지역 9개고사장에서 정보검색사 1급 및 전문가(전문검색사, 시스템관리사, 정보설계사) 인증시험을 동시에 실시했다. 이날 인증시험은 1차 필기시험으로 총 2,256명이 지원하여 약 87%의 응시율을 보였으며 '98. 7. 14일 발표한 합격자를 대상으로 2차 실기시험을 '98. 8. 30일 실시, 최종합격자를 '98. 9. 15일발표할 예정이다.