8장. 표준화

부산대학교 문헌정보학과
이수상 교수
sslee@pusan.ac.kr

1. 표준화 대상과 영역

■ 디지털도서관 운영의 영역별 주요 표준 대상

적용 영역 및 구분		주요 표준	비고
디지털 객체	생산	HTML	8.1.1
		XML	8.1.2
	패키징	OAIS Information Model	3.4.1 참조
		METS	3.4.1 참조
메타데이터	콘텐트	DC	4.2.3 참조
		MODS	4.2.3 참조
		EAD	4.2.3 참조
		MPEG-7	4.2.3 참조
	장서	RSLP CD Schema	4.2.3 참조
		DC Collection AP	4.2.3 참조
정보서비스	웹 프로토콜	HTTP	8.2.1
		SOAP	8.2.2
	통합검색 프로토콜	Z39.50	6.6.1 참조
		SRU	6.6.3 참조
		OAI-PMH	6.6.4 참조
	콘텐트 배포	RSS	8.2.3
	이용자 인증	LDAP	6.5.1 참조
		Shibboleth	6.5.2 참조
	시맨틱 웹	RDF	8.2.4
		OWL	8.2.5
		Topic Maps	8.2.6

■ 디지털도서관 운영의 영역별 주요 표준 대상 (계속)

적용 영역 및 구분		주요 표준	비고
저작권	저작권 표현 언어	ODRL	9.2.3.2 참조
		XrML	9.2.3.2 참조
		CCL	9.4.3.1 참조
기타	식별체계	DOI	6.2.1 참조
		OpenURL	6.2.3 참조
	이용통계	COUNTER	6.7.2 참조
		SUSHI	8.4.1
	문자코드	ASCII	8.4.2
		ISO 8859	8.4.3
		UNICODE	8.4.4
		KSC 5601	8.4.5

2. 디지털 객체 관련 표준

HTML

- HTML(HyperText Markup Language)은 W3C의 표준으로 전 세계에서 볼 수 있는 정보를 작성하고, 모든 컴퓨터가 알 수 있는 웹 문서를 작성하기 위한 언어이다.
- HTML은 특히 하이퍼텍스트를 작성하기 위해 개발되었으며, 인터넷에서 웹을 통해 접근하는 대부분의 웹 페이지들은 HTML로 작성된다.
- HTML은 2.0 버전, 3.2 버전에 이어 4.0이 1997년 12월에 개발되었다. 현재 많이 사용하는 것은 1998년에 만든 HTML 4.01이며, 2012년 HTML 5가 개발되었다.
- HTML 4.01에서는 스크립팅(Scripting) 기능을 통해 동적인 페이지를 만들 수 있으며, 네트워크 응용 프로그램을 만드는 수단으로도 사용할 수 있다.
- HTML에서 사용하는 명령어를 태그(Tag)라고 하는데, 태그는 시작과 끝을 표시하는 2개의 쌍으로 이루어져 있고, 이를 통해 문서의 글자 모양, 글자 크기, 색 등을 표현할 수 있으며 관련 정보를 연결(hypertext link)시켜줄 수 있다.

NIS

XML

- XML(eXtensible Markup Language)은 인터넷 상에서 데이터 교환을 목적으로 모든 문서 및 응용에 대한 범용 마크업 정의 방법을 표준화한 메타언어(meta language)이다.
- HTML(HyperText Markup Language)이 데이터 구조를 기술하는 기능이 없다는 한계를 극복하고, SGML(Standard Generalized Markup Language)의 복잡함을 단순화함으로써, SGML과 HTML 양쪽 모두와의 상호운용이 가능하다.
- XML을 기반으로 하는 프로그래밍 기술이 발전함에 따라 인터넷, 전자상거래, 음악, 과학, 디지털도서관 등과 같은 매우 다양한 분야의 응용에 XML을 적용하고 있다.
- 현재는 XML을 기반으로 각 정보간의 관계를 명시함으로써 지식(knowledge)의 전달이 될 수 있게 하는 "시맨틱 웹"에 대한 개발이 진행중이다.

NIS

XML의 강점

- 첫째, 표준 규약을 따름으로써 응용 프로그램 호환성 문제를 극복하는데 커다란 구실을 한다.
- 둘째, 하드웨어 특성, 운영체제 특성, 프로그래밍 언어에 무관하게 중립적인 방식으로 정의된다.
- 셋째, XML 지원 소프트웨어가 풍부하다. 넷째, 응용 프로그램 설계와 개발을 편리하게 만든다.

XML 관련 표준

- XML Query: XML 문서에 대한 질의 언어
- XSL(eXtensible Stylesheet Language): XML을 위한 스타일시트(stylesheet) 언어
- DOM(Document Object Model): HTML과 XML 문서를 위한 API(Application Programming Interface)
- XML 스키마: XML 문서의 구조와 콘텐트를 정의하는 파일

NIS

3. 정보 서비스 관련 표준

HTTP

- HTTP는 인터넷에서 하이퍼텍스트(hypertext) 문서를 교환하기 위하여 사용되는 통신규약(프로토콜, protocol)이다.
- HTTP는 1989년 팀 버너스 리(Tim Berners Lee)에 의하여 처음 설계되어 인터넷을 통한 월드 와이드 웹(World-Wide Web) 기반에서 전 세계적인 정보공유를 이루는데 큰 역할을 하였다.
- HTTP 프로토콜은 요구/응답 (Request/Response) 방식을 이용하여 동작하고 있다. 즉, 원하는 프로토콜 기능(예: GET, HEAD, POST)에 대해 서비스 요구를 하면 데이타 송수신을 위한 TCP 연결이 만들어지고, 서버가 응답을 보내어 데이타 전송을 끝내면 자동적으로 연결이 끊어지게 되는 것이다.

SOAP

- SOAP(Simple Object Access Protocol)은 용어의 의미로 보면 단순객체 접근 프로토콜이며, 기능적으로 보면 XML과 HTTP 등을 기반으로 하여 다른 컴퓨터에 있는 데이터나 서비스를 호출하기 위한 통신규약이다.
- 기능이 분리되고 공간적으로 분산된 환경에서의 정보교환을 위해 단순하게 설계된 통신규약이다.
- SOAP은 가장 많이 사용되는 웹(World Wide Web: www)의 통신규약인 HTTP와 높은 유연성과 확장성을 갖춘 정보표현의 기술인 XML의 조합이다.
- 모든 SOAP 메시지는 필수적인 SOAP 패키지(envelope), 선택적인 SOAP 헤더(header), 그리고 필수적인 SOAP 바디(body)로 구성된 하나의 XML 문서이다.
- 패키지는 메시지를 표현하는 XML 문서의 최상위 요소이다. 헤더는 통신 주체들간에 사전 합의없이 분산화된 방법으로 특성들을 SOAP 메시지에 추가하기 위한 일반적인 메커니즘이다.

NIS

RSS

- RSS는 RDF Site Summary 또는 Really Simple Syndication의 줄임말
- 뉴스나 블로그와 같이 콘텐트가 자주 업데이트가 되는 사이트들이 업데이트된 정보를 쉽게 이용자들에게 제공하기 위해 만들어진 포맷이다.
- RSS는 가장 성공적인 XML서비스로서 웹사이트를 통해 콘텐트 정보를 교환하는 커뮤니티 표준으로 자리를 잡아 가고 있다.

RSS의 사용분야

사용분야	내용
뉴스 및 공지사항	매시간 새로운 정보가 추가, 변경되는 뉴스 또는 신규 소식 서비스
강좌	이용자가 매번 사이트를 방문하여 규칙적으로 확인하지 않는 콘텐트 서비스
일정	주요 행사, 마감일자, 휴일정보
검색결과	관심 키워드에 대한 변경 및 신규 정보 조회 서비스
메일링 리스트	주기적으로 이메일로 고객에게 서비스 한 내용 모음(입찰정보, 채용정보 등)

RDF

- RDF(Resource Description Framework)는 메타데이터의 기술과 교환을 위한 구조로서, 웹 상의 메타데이터를 지원하는데 필요한 구조를 정의하기 위해 W3C(World Wide Web Consortium)에서 제안한 표준이다.
- RDF는 상이한 메타데이터간의 어의, 구문 및 구조에 대한 공통적인 규칙을 지원하는 메커니즘을 통해 웹 상에서 기계적 이해가 가능 하도록 정보를 교환하는 즉, 구조화된 메타데이터간의 상호운용성(interoperability)을 지원하는 새로운 개념이라 할 수 있다.
- RDF는 인터넷 상에 존재하는 상이한 성격의 메타데이터간의 상호운용이 가능하도록 하는데 그 목적이 있다.

RDF의 특성

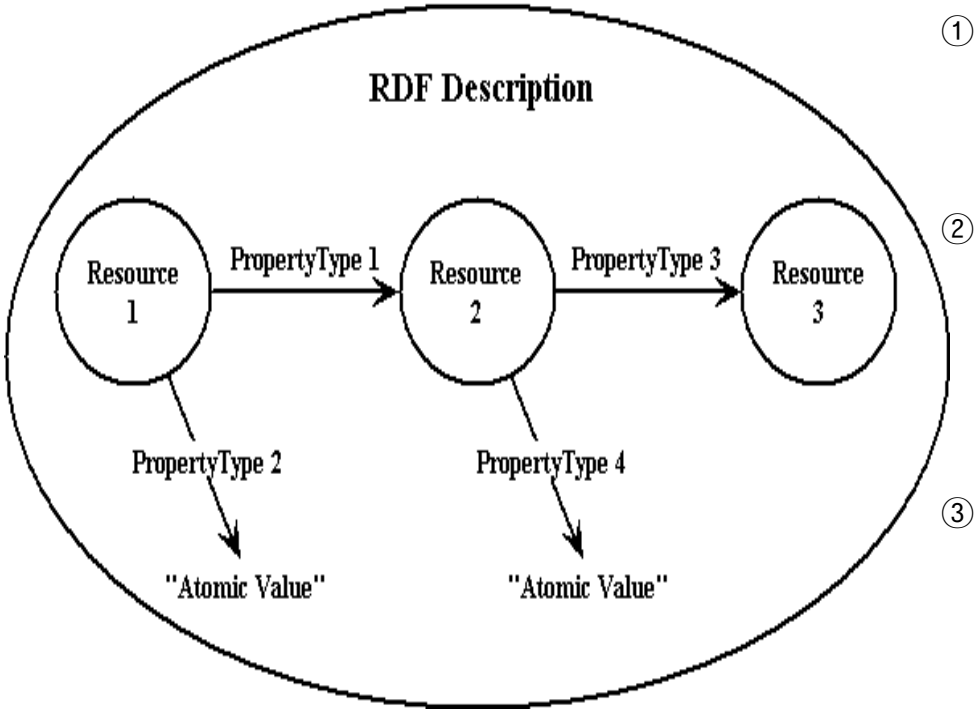
- 구문 독립성(syntax independence)
- 기술/검색의 용이성
- 확장성

RDF의 사용분야

- 자원 검색(resource discovery)
- 자원 편목(cataloging)
- 지능 소프트웨어 에이전트(intelligent software agent)
- 내용 순위부여 및 평가(content rating)
- 디지털 서명(digital signature)
- 지적 재산권 보호(intellectual property rights)

NIS

RDF의 구성요소: 자원, 속성, 속성값의 관계



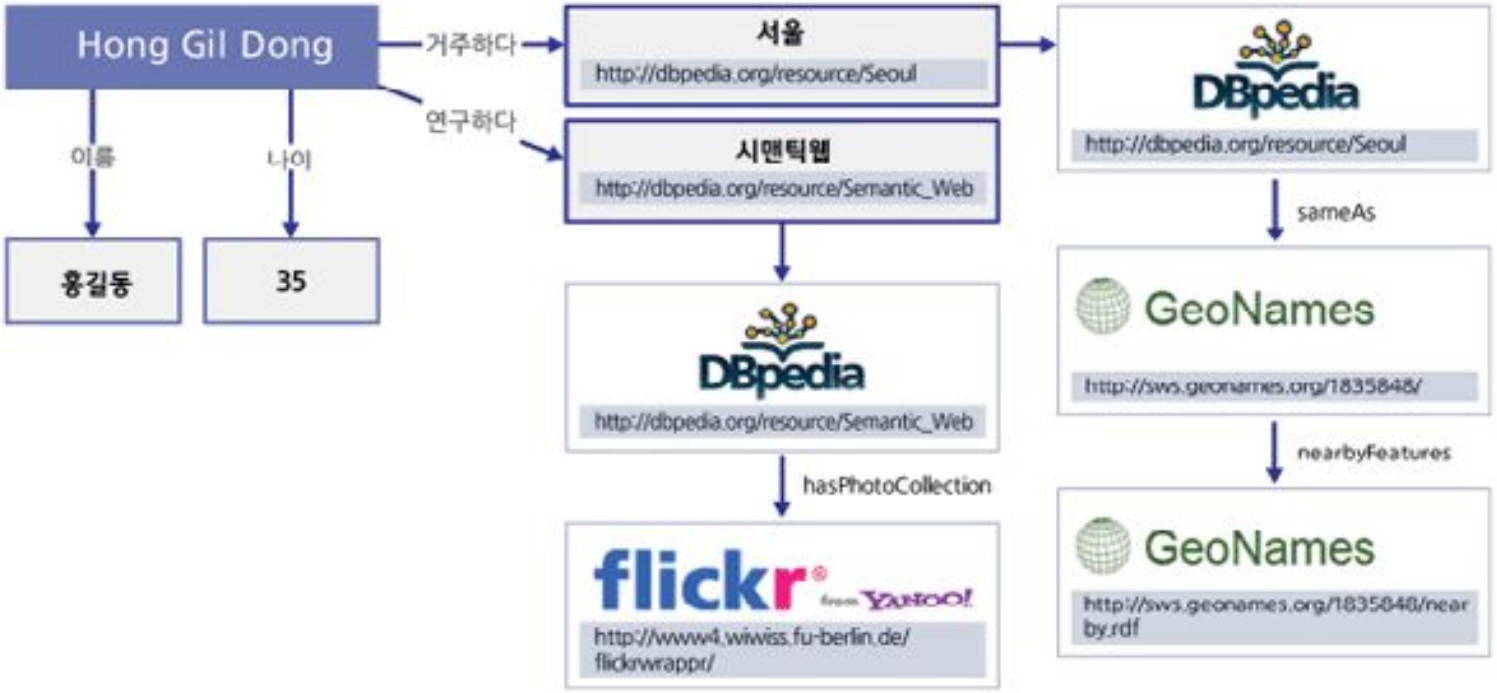
① 자원(Resource): 인터넷 상에서 존재하고 있는 모든 웹 페이지는 자원이 된다. 하나의 자원은 여러 속성 유형이나 여러 속성값을 가질 수 있다.

② 속성 유형(PropertyType): 자원을 표현하는 속성은 속성 유형으로 구분된다. 속성 유형은 '저자'나 '서명' 등과 같이 자원의 속성을 적절한 이름으로 표현한 것이다. 인터넷 상에서 속성 유형 자체가 자원이 되는 경우가 있다.

③ 속성값(Value): 속성 유형은 상응하는 값으로 표현되는데, 이를 속성값이라 하며, 문자열이나 숫자 등과 같이 자연어로 상세하게 기술된다. 속성값 역시 속성 유형처럼 자원이 될 수 있으며, 고유한 속성을 가지고 있다.

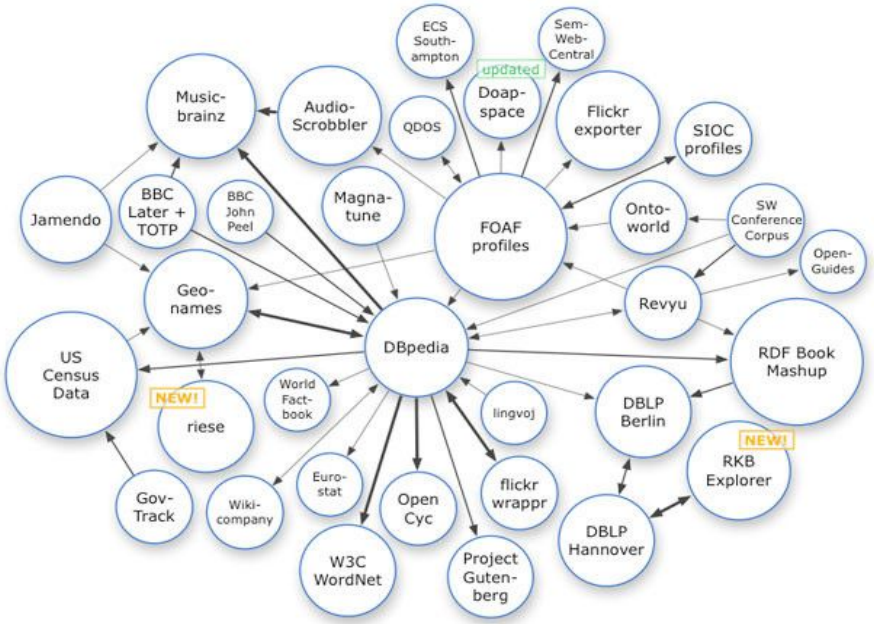
속성(Property)은 자원과 속성 유형, 그리고 속성값을 모두 조합한 것이다. 예를 들어 '<http://info.lib.pusan.ac.kr>' 이라는 홈페이지 제작자가 '홍길동' 이라면, 'URL' 은 자원이 되고, '제작자' 는 속성 유형이 되며, '홍길동' 은 속성값이 된다.

- RDF 적용사례: 시맨틱 웹의 Linked Data



Linked Data

- 링크드오픈데이터(Linked Open Data) 또는 링크드 데이터(Linked Data)
- 시맨틱 웹의 창시자인 팀버너스리에 의해 2006년 처음 주장: "시맨틱 웹은 단지 데이터를 웹으로 제공하는 것이 아니라, 데이터 간의 링크를 만듦으로써, 인간이나 기계 모두 데이터의 웹을 탐험할 수 있도록 해준다. Linked Data를 통해 유용한 데이터를 얻게 되면, 그 데이터에 관계된 데이터로 계속되는 항해가 가능하다."
- 시맨틱 웹이 표방하는 데이터 웹(Data Web)을 구체적으로 구현하는 방법으로, 웹으로 접근가능한 이름(URI)을 붙이고 이를 통해서 RDF 형태의 시맨틱 데이터를 서로 연결함으로써 데이터를 공개하고, 공유하고, 연결하기 위한 방법이다.



OWL

- OWL(Ontology Web Language)은 온톨로지 웹 언어 또는 웹 온톨로지 언어를 말하며 시맨틱 웹 영역에서 가장 중요한 언어이다.
- RDF와 RDF 스키마의 문제점을 보완하기 위해 W3C의 후원으로 개발이 시작되었으며, 2004년에 W3C의 권고안이 된 온톨로지 마크업 언어이다.
- 현재로서는 OWL이 다른 온톨로지 마크업 언어에 비해 표현력이나 추론 능력에 있어서 가장 뛰어난 언어라고 평가되고 있으며 또한 가장 널리 사용되고 있다.

OWL의 세 가지 유형

- OWL Lite: 클래스의 계층과 간단한 제약사항만을 제공하는 간략한 언어
- OWL DL: 기술논리(Description Logic)에 기반을 둔 언어이며, 계산적 완전성과 결정 가능성을 유지하면서 최대한 표현력을 제공
- OWL Full: RDF의 모든 문법을 사용할 수 있으며 최대의 표현력을 제공하는 언어

NIS

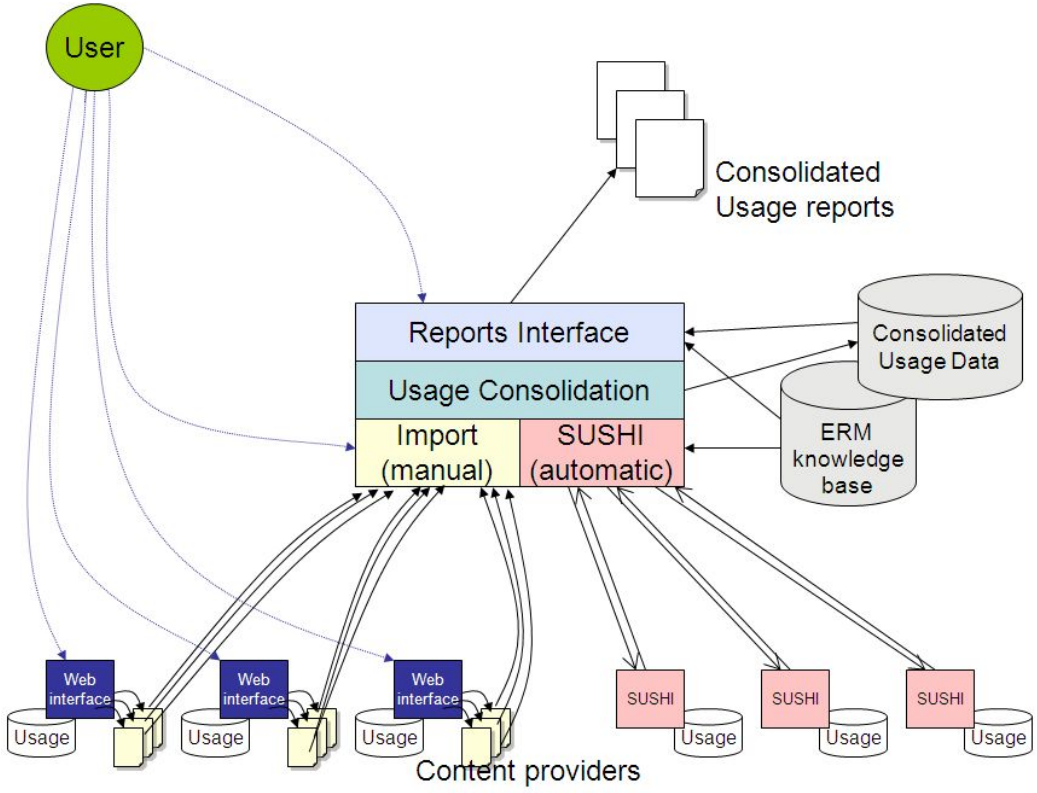
Topic Maps

- RDF와 RDF 스키마, 그리고 OWL이 W3C에서 연구하는 언어이지만, 토픽맵(Topic Maps)은 국제표준화기구(ISO)를 중심으로 연구하는 시맨틱 웹 온톨로지의 표현언어이다.
- 토픽맵의 표준은 SGML에 기반을 둔 HyTM(HyTime Topic Maps)과 XML에 기반을 둔 XTM(XML Topic Maps)으로 구분된다.
- 토픽맵은 용어집, 시소러스, 색인집 등 용어의 의미적 구조를 다루는 목적으로 개발되었지만, 현재는 정보자원을 의미적 관계를 표현하고 의미적 검색이 가능하도록 하는 시맨틱 웹의 핵심기술로 인정받고 있다.
- 토픽맵은 정보자원을 주제(토픽)별로 묶고 주제와 주제 간의 관계를 나타내어 정보자원에 대한 의미적 지식(온톨로지)을 표현하는 색인어 지도이다.

4. 기타 표준

SUSHI

- SUSHI(Standardized Usage Statistics Harvesting Initiative)는 COUNTER 프로젝트와 밀접한 관련이 있는 이용통계 관련 표준 프로토콜이며, SOAP의 요청/응답 웹서비스를 통해 전자자원의 이용통계 데이터를 수확하는데 사용한다.
- SUSHI 개념도



ASCII

- SCII(American Standard Code for Information Interchange) 표준은 1963년 미국표준협회(ASA)에 의해 결정된 미국 표준의 문자코드로 128개의 문자를 제공하는 7비트 코드이다.
- ASCII는 본래 전신(teletype) 터미널용으로 설계되었기 때문에 처음 32개의 코드는 프린트되지 않는 특별한 제어 문자로 사용된다.
- ASCII는 영어 문자를 표기하는데 필요한 문자와 소수의 특수문자로 구성되어 있어 세계 여러 나라에서 사용하는 모든 숫자, 국가언어, 기호 등을 충분히 표현할 수 없다.
- 그럼에도 불구하고 오늘날 사용되는 많은 문자세트가 ASCII를 시작점으로 하여 확장된 문자의 집합을 만들고 있다.

ISO 8859 표준

- 128자의 ASCII 문자들은 영어로 된 정보를 교환하는 데는 충분하지만, 스위스와 그 외 노르딕 언어들과 같은 로마자를 사용하는 대부분의 유럽 언어들은 ASCII 표준으로는 표현할 수 없는 추가적인 기호들이 필요하게 되었고, 그 결과 8비트 문자세트를 구현하게 되었다.
- 국제표준기구(ISO)는 여러 가지 8비트 문자세트를 15가지 범주로 구분하여 다양한 유형의 유럽 언어문자들을 정의하고 있다.
- ISO 8859 문자세트 범주 사례

문자세트
ISO/IEC 8859-1:1998 (또는 Latin-1)
ISO 8859-2:1999 (또는 Latin-2)
ISO 8859-3:1999 (또는 Latin-3)
ISO 8859-4:1998 (또는 Latin-4)
ISO 8859-5:1999

UNICODE 표준

- Apple, IBM, Microsoft 등이 컨소시엄을 통해 전 세계 문자 코드를 표현할 수 있는 문자세트를 제공하였고, ISO/IEC JTC1에서 1995년 9월 국제 표준으로 제정하였다.
- 공식 명칭은 ISO/IEC 10646-1(Universal Multiple-Octet Coded Character Set)이다. ISO/IEC 10646-1의 문자판에는 전 세계에서 사용하고 있는 26개 언어의 문자와 특수기호에 대해 일일이 코드값을 부여하고 있는데, 216 = 65,536개의 문자를 수용할 수 있다.
- 세계의 26개 언어의 문자와 특수기호에 2바이트(2byte=16bit=65,536)로 코드 체계로 만들어 세계 각국의 언어를 동시에 사용할 수 있도록 만든 것이다.
- 이 가운데 3만 8,885자는 주요 국가의 언어를 구현하는 용도로 이미 할당되어 있고 6400자는 사용자 정의 영역으로, 나머지는 2만여 자는 새로 추가될 언어 영역으로 각각 비워두고 있다. 코드 할당비율을 보면 한자가 39.89%로 가장 많고, 한글 17.04%, 아스키 및 기호문자 10.39% 등의 순이다.

NIS

UNICODE 인코딩(Encodings) 방식

- UTF-8, UTF-16, UTF-32의 세가지 방식이 있다.
- UTF는 UCS Transformation Format의 약자이며, 뒤에 붙은 숫자는 인코딩에 사용되는 단위의 비트수를 의미한다. 즉 UTF-8은 8비트 단위, UTF-16은 16비트 단위, UTF-32는 32비트 단위로 문자를 표현한다.
- 세가지 방식의 공통점이라면 16개의 보충언어판에 위치한 1,048,576개의 코드를 표현할 때는 4바이트를 사용한다는 점이다. 하지만 그 방식은 모두 다르다. UTF-8은 4개의 8비트로, UTF-16은 2개의 16비트로, UTF-32는 1개의 32비트 단위로 표현한다.

NIS

KSC 5601-87 표준

- 한국 정부가 제정한 한글코드표준
- KSC 5601 표준이 바로 ISO 2022 부호확장법에 따른 코드
- 2,350자로 한글을 제한한 완성형 코드이기 때문에 표기할 수 있는 문자에 대한 제약이 크고, 특히 문서 입력에 있어서 제대로 입력되지 못하는 글자가 속출하는 등의 문제가 발생하였다. 이러한 문제를 해결하고자 1992년에는 완성/조합의 이원적인 코드 체계인 KSC 5601-92를 표준으로 지정하기도 하였다.
- 1998년에 KSC 5601-87의 새로운 규격으로 KS X 1001로 개정하였다.

관련 표준

- KSC 5657: KSC 5601-87에 1930자의 한글을 추가한 한글코드 표준. 이 표준으로 KSC 5601-87이 한글을 2,350자로 제한하여 발생하는 문제를 해결하고자 했다.
- KS X 1005-1: 바로 국제표준기구의 문자세트 표준 규격인 ISO/IEC 10646(Unicode)을 국내 표준화 한 것