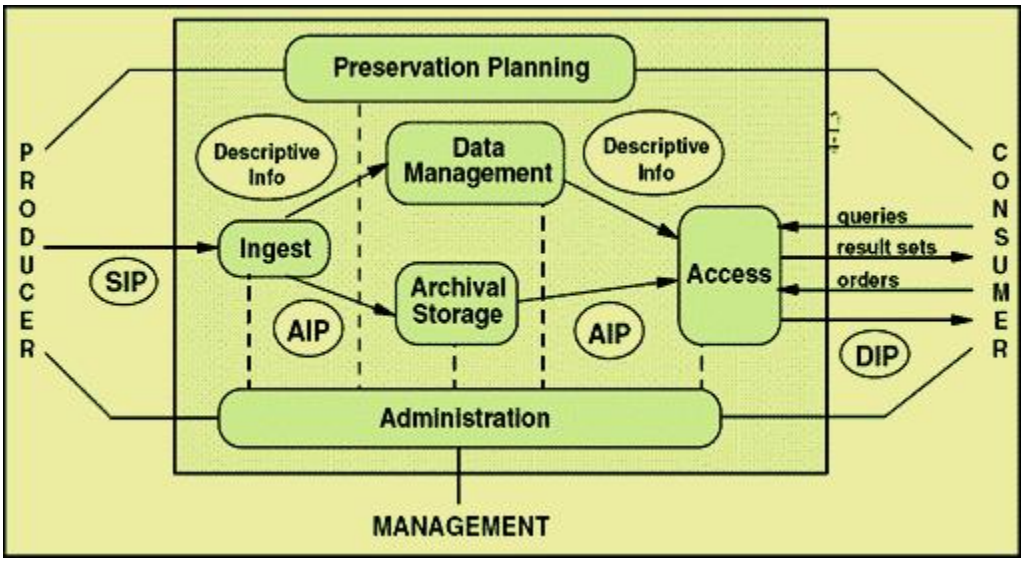
2014 정보검색 보조교재

Annotated by S.H. Kim



대구대학교 사회과학대학 문헌정보학과

p. lxvi

*** metadata: "data about data".**

이 용어의 개념이 모호한데 그 이유는 two fundamentally different concepts을 포함하고 있기 때문이다

Structural metadata; the design and specification of data structures에 대한 것이며, 보다 정확하게는 "data about the containers of data"라 부른다.

Descriptive metadata; individual instances of application data 즉, the data content에 관한 것이다. 따라서, "data about data content" or "content about content" 그러므로 metacontent라고 부르는 것이 더욱 좋을 것이다.

Metadata (metacontent)는 전통적으로 도서관 목록에서 찾을 수 있다. 정보가 점점 더 디지털화됨으로써, 메타데이터 역시 특별한 학문영역에 전문적으로 적용되는 메타데이터 표준을 사용하여 디지털 데이터를 기술하는데 이용되고 있다. 데이터 파일의 contents와 context를 기술함으로써, 오리지널 데이터/파일의 질이 크게 높아진다. 예를 들어, 이용자의 경험을 자동적으로 개선시킬 수 있도록 하는 브라우저를 통하여어떤 웹페이지에 사용된 언어, 만들 도구, 그 주제에 대해 더 알고 싶은 경우에 가야할 장소를 자세하게 나타내는 메타데이터를 포함할 수도 있다.

** Definition

Metadata (metacontent)란 데이터와 관련해서 다음과 같은 한 가지 이상의 aspect에 대한 정보를 제공하는 데이터이다;

1) Means of creation of the data
2) Purpose of the data
3) Time and date of creation
4) Creator or author of the data
5) Location on a computer network where the data were created
6) Standards used

예) *a digital image에 그 그림이 만들어졌을 때의 그 그림의 크기, 색도, 해상도 등과 같은 metadata가 포함될 수 있다. 텍스트 문서의 메타데이터에는 how long the document is, who the author is, when the document was written, and a short summary of the document 에 대한 정보가 포함될 수 있다.*

Metadata도 데이터이다. 따라서 메타데이터는 데이터베이스에 저장되어 관리될 수 있다. 종종 이 데이터베이스를 Metadata registry 또는 Metadata repository라 부른다. 그렇지만, the context 그리고 a point of reference가 없다면, 이 레포지토리는 메타데이터를 단지 살펴보기만 하고 메타데이터를 식별해내지 못할 수도 있다. 예를 들어, 모두 13 digits로 이루어진 여러 개의 번호들을 포함하고 있는 데이터베이스는 그 자체가 계산의 결과이거

나 방정식에 사용할 숫자의 리스트일 수도 있다 --만일 어떤 다른 context가 없다면, 번호들 그 자체가 데이터로 인식될 수 있다. 그러나 이 데이터베이스가 장서수집용 기록(log)이라는 context가 주어진다면, 이러한 13자리 번호들은 ISNSs - 그 책속에 들어 있는 정보 자체가 아니라 그 책에 대하여 말하는 정보- 으로 식별할 수 있다.

"메타데이터" 용어는 1968년 Philip Bagley가 자신의 저서 「Extension of programming language concepts」에서 처음으로 사용하였다. 그는 분명히 말해서, 데이터 콘텐트인 각각의 경우(instances)에 대한 콘텐트 또는 도서관 목록에서 대체로 발견되는 데이터의 종류인 메타콘텐트라는 대안적 의미보다는 구조적 데이터 즉, 데이터의 콘테이너에 대한 데이터라는 ISO 11179의 "전통적 의미"로 이 용어를 사용하였다. 그 이후로 정보관리, 정보학, 정보기술, 도서관학, 그리고 GIS에서 이 용어를 널리 채택하였다. 이들 분야에서 메타데이터란 단어는 "데이터에 대한 데이터"로 정의하고 있다. 이것이 일반적으로 받아들여지고 있는 정의인 반면에, 또다른 여러 분야에서는 스스로 보다 협의적인 정의를 채택하여 이 용어를 사용하고 있다.

** Metadata types

Bretheron & Singley (1994)의 two distinct classes: structural/control metadata and guide metadata:

Structural metadata: 구조적 메타데이터

tables, columns and indexes와 같은 computer systems의 structure를 기술하기 위하여 사용한다.

Guide metadata: 안내용 메타데이터

누구가에게 특별한 아이템을 찾는 것을 도와주기 위하여 사용되며, 대부분 자연어로 된 키워드의 세트로 표현된다. Ralph Kimball는 유사한 2 가지 categories: technical metadata와 business metadata를 주장하였다. 여기서 Technical metadata는 internal metadata이고, business metadata는 external metadata를 말한다. 또한 Kimball은 a third category로 process metadata를 추가하였다. 또 한편으로 NISO에서는 3 types of metadata: descriptive, structural and administrative로 구분하고 있다. Descriptive metadata란 title, author, subjects, keywords, publisher와 같은 객체(object)를 탐색하고 그 위치를 설정하기 위하여 사용되는 정보이고, structural metadata란 how the components of the object are organised에 대한 기술이며, administrative metadata란 file type을 포함하고 있는 technical information을 나타낸다. 그리고 Two sub-types of administrative metadata으로 rights management metadata와 preservation metadata를 제시하고 있다..

기술 메타데이터(Technical Metadata)

기술 메타데이터는 디지털 캡쳐과정에서 생산 및 제작 정보, 다시 말해서, 미래에 디지털 자원들의 재생을 위하여 디지털 객체, 주파일과 보조 파일의 포맷, 해상도, 칼러 기록파일 등을 수십하개 위하여 사용된 하드웨어와 소프트웨어에 대한 정보를 포함하고 있는 디지

털 객체의 기술적 속성을 기록하기 위하여 사용된다. 이런 이유로 기술적 메타데이터는 행정 그리고 보존 메타데이터의 범주에 속한다.

기술 메타데이터는 보존 메타데이터(PREMIS) 또는 구조 메타데이터(METS)에 포함될 수 있다. 기술 메타데이터의 추가와 관련된 결정은 METs와 PREMIS가 자체의 스키마에서 기술 메타데이터를 포함시키는 서로 다른 방법을 가지고 있으므로, 보존되어야만 하는 기술 메타데이터의 유형에 의해 결정되기도 한다.

보존용 메타데이터(PREMIS); *See Also p.72; PREMIS schema 와 METS*

In June 2003, OCLC and RLG jointly sponsored the formation of the PREMIS (***Pre****servation* ***M****etadata:* ***I****mplementation* ***S****trategies*) working group, comprised of international experts in the use of metadata to support digital preservation activities.

** Metadata structures

메타데이터(메타콘텐트) 또는 보다 정확하게 말해서 메타데이터(메타콘텐트) statements를 모으기 위하여 사용된 어휘들은 잘 정의된 메타데이터 스킴(요강) -메타데이터 표준과 메타데이터 모델을 포함하는 - 을 사용하는 표준화된 개념에 따라 전형적으로 조직화 된다. 통제어휘집, 택소노미, 시소러스, 데이터 사전, 그리고 메타데이터 등록부과 같은 도구들은 추가적인 표준을 메타데이터에 적용하기 위하여 이용될 수 있다. 구조 메타데이터 공통적 속성은 데이터 모델 개발과 데이터베이스 디자인에 있어서 매우 중요하다.

** Metadata syntax

메타데이터 구문은 메타데이터의 필드 또는 요소를 조직하기 위하여 만들어진 규칙을 말한다. 단일 메타데이터 스킴은 서로 다른 구문을 필요로 하는 수많은 다양한 markup 또는 프로그래밍 언어로 표현될 수도 있다. 예를 들어, Dublin Core는 plain text, HTML, XML 그리고 RDF로 표현 가능하다.

(가이드) 메타콘텐트의 한가지 일반적인 예는 서지 분류, 주제, DDC 분류번호이다. 이것은 항상 어떤 객체의 “분류”에 있는 하나의 암시적 표현이다. 예를 들어, 듀이분류번호 514(형태학)처럼 어떤 객체(다시 말해서 자신의 등에 514라는 숫자를 가지고 있는 책)를 분류하기 위하여 그것의 암시적 표현은 “<book><subject heading><514>” 이다.

이것은 주체-술어-객체 트리플 이거나 더욱 중요하게 말해서, 부류-속성-값 트리플이다. 이 트리플의 첫 번째 2가지 요소인 부류와 속성은 분명하게 정의된 어의를 갖고 있는 어떤 구조 메타데이터의 단편들이다. 3번째 요소는 어떤 참고(마스터) 데이터, 즉 어떤 통제어휘에서 나온 값이다. 메타데이터와 마스터 데이터 요소의 결합은 메타콘텐트인 하나의 문장, 다시 말해서 "metacontent = metadata + master data"을 만들어냈다. 이러한 요소 모두는 “어휘”로 여겨질 수 있다. 메타데이터와 마스터 데이터 둘 다 메타콘텐트 문자에서 모여질 수 있는 어휘들이다. 메타 와 마스터 데이터 둘 다를 가지고 있는 어휘집에 대한 많은 정보원이 있다.

1) UML(Unified Modeling Language (UML); UML은 소프트웨어 공학 분야에서 사용하는 표준화된고 범용의 모델링 언어이다. 이 언어에는 객체지향적 소프트웨어-강화형 시스템의 시각적 모델을 만들기 위한 한 세트의 그래픽 표기 기법이 포함되어 있다.

2) **EDIFACT**(United Nations/Electronic Data Interchange For Administration, International Organization for Standardization (ISO) as the ISO standard ISO 9735).; 유엔에서 개발한 국제 EDI 표준이다. 이 표준은 ISO standard ISO 9735로 채택되었다.

3) **XSD**(XML Schema); **XSD**는 2001년에 W3C 권고문으로 출판된 XML 스키마이며 여러 가지 XML 스키 마 언어들 중의 하나이다. 이것은 W3C에서 만든 권고사항을 충족시키기 위하여 XML 용으로 만든 첫 번째 독립된 스키마 언어이다. 특별한 W3C의 세부규정으로서의 XML 스키마와 일반적으로 스키마 언어를 기술하기 위하여 동일한 용어를 사용하는 것 사이에서의 혼란으로 인하여, 몇몇 이용자 집단에서는 이 언어를 WXS(W3C XML Schema의 두문자어)로 언급하고 있는 반면에 다른 집단에서는 XSD(XML Schema Definition의 두문자어)라고 부르고 있다. 버전 1.1에서 W3C는 XSD라는 용어를 채택하였다.

4) **Dewey/UDC/LoC**,

5) **SKOS**(Simple Knowledge Organization System (SKOS); 시소러스, 분류표, 택소노미, 주제표목 시스템 또는 어떤 다른 정형화된 통제어휘의 표현용으로 디자인된 W3C의 권고이다. SKOS는 RDF와 RDFS를 근거로 구축된 시멘틱 웹 부류의 한 부분이며, 이것의 주요 목적은 링크된 데이터처럼 어휘들을 사용하여 쉽게 출판할 수 있도록 하는 것이다.

6) **ISO-25964**(ISO 25964 is the international standard for thesauri); 아래처럼 2 부분으로 된 시소러스의 국제 표준이다.

Part 1: 정보검색용 시소러스
Part 2: 기타 어휘와의 상호운영성

메타콘텐트 문장의 구성요소용으로 통제어휘를 사용하는 것은, 색인용이든 또는 찾기용이든, ISO-25964에 의해 정해져 있다: " 만일 색인자와 탐색자 둘 다 똑같은 개념용으로 똑같은 용어를 선택하도록 유도된다면, 적절한 문헌이 검색될 것이다." 이것은 인터넷의 거물인 Google이 텍스트 스트링을 간단하게 색인한 다음에 매칭시키는 것을 고려할 때 특히 적절하다. 따라서 어떠한 지능이나 "간섭"도 필요하거나 발생하지 않는다

7) **Pantone:** Pantone Inc.; Pantone Matching Systme으로 가장 잘 알려져 있으며, 비록 때때로 유색의 페인트, 직물, 플라스틱의 제조에 사용되더라도, 이 시스템은 여러 산업에서 기본적으로 printing에서 사용하는 독점적인 칼러 스페이스이다. Pantone Color Matching System은 거의가 표준화된 칼러 재생 시스템이다. 칼러를 표준화함으로써, 서로 다른 위치에서 서로다른 제조업자들 모두 서로 간의 직접적인 접촉없이 칼러의 매치를 확실하게 하기 위하여 Pantone system으로 말할 수 있다.

8) **Linnaean Binomial Nomenclature**(Binomial nomenclature (also called binominal nomenclature or binary nomenclature) ; 이항식 명명법은 비록 다른 언어로 된 단어를 근거로 하더라도 각각의 생물 종에게 두 부분으로 구성된 이름을 둘 다 라틴어의 문법을 사용하여 줌으로써 그것의 이름을 붙이는 공식적 시스템이다. 그러한 이름은 binomial name(줄여서 "binomial"), binomen 또는 scientific name이라 하며, 보다 비공식적으로는 그것을 라틴이름으로 부르는 것이다. 이 이름의 첫 번째 부분은 그 종이 속한 類를 구분하는 것이며, 두 번째는 類 안에 있는 種을 구분하는 것이다. 예) **인간은 Homo 류에 속하며 그 류속에 있는 Homo sapiens에 속한다**.

** Metadata standards

핵심 표준은 ISO/IEC 11179-1:2004와 후속 표준(see ISO/IEC 11179)이다. 이 표준에 따라 모두 이미 출판된 등록들은 단지 메타데이터의 정의만을 취급하고 있으며, 어떤 행정적 표준과 마찬가지로 메타데이터의 저장 또는 검색의 구조를 지원하지 못하고 있다. 중요한 것은 이 표준이 메타데이터를 데이터의 콘테이너에 대한 데이터로 언급하고 있지 데이

터의 콘텐트에 대한 데이터로 메타데이터를 언급하지는 않고 있다는 것이다.

Dublin Core 메타데이터 용어들은 한 세트의 어휘 용어들이며 찾기용의 자원을 기술하는데 사용될 수 있다. Dublin Core Metadata Element로 알려진 15개의 전통적인 메타데이터 용어들의 초기 세트는 다음과 같은 표준 문서들에 의해 보증을 받았다:

IETF RFC 5013
ISO Standard 15836-2009
NISO Standard Z39.85

비록 표준은 아니지만, Microformat는 시멘틱 마크업을 위한 웹 의존형 방식이매 메타데이터를 전달하기 위하여 기존의 HTML/XHTML 태그를 재사용하려고 한다. microformat은 XHTML과 HTML 기준을 따르지만 그 자체가 표준은 아니다.

** Library and information science and metadata

도서관은 가장 일반적으로 통합도서관관리시스템의 한 부분으로 도서관 목록에서 메타데이터를 채택하고 있다. 메타데이터는 책, 정간물, DVD, 웹 페이지 또는 디지털 이미지와 같은 자원을 편목하는데서 얻어진다. 이러한 데이터는 MARC 메타데이터 표준을 사용하는 통합도서관관리시스템인 ILMS에 저장된다. 이것의 목적은 이용자에게 그들이 요구하는 아이템이나 지역의 물리적 또는 전자적 위치로 직접 유도해 줄 뿐만 아니라 의문시하는 아이템에 대한 설명을 제공하는 것이다.

도서관 메타데이터에 대한 보다 최신의 그리고 전문화된 경우에는 e-print repositories와 digital image libraries를 포함하여 디지털 도서관의 설립이 포함된다. 종종 도서관 원칙에 의존하는 동안, 특히 메타데이터를 제공하는데 있어서 비-도서관적 사용에 초점을 맞춘다는 것은 그것들이 전통적이거나 일반적인 편목방법을 따르지 않는다는 것을 의미한다. 관련된 자료에 관습적 성질이 주어진다면, 예를 들어 분류 분야, 위치 분야, 키워드 또는 저작권 문장에 대한 메타데이터 분야는 가끔 특별하게 만들어진다. 파일 크기와 포맷과 같은 표준 파일 정보는 일반적으로 자동적으로 포함된다.

도서관 운영을 위한 표준은 수십년 동안 ISO의 중요한 논제이다. 디지털 도서관의 메타데이터 표준에는 Dublin Core, METS, MODS, DDI, ISO standard Digital Object Identifier (DOI), ISO standard Uniform Resource Name (URN), PREMIS schema, Ecological Metadata Language, 그리고 OAI-PMH가 포함된다. 세상을 선도하는 도서관들은 자신들의 메타데이터 표준 전략에 대하여 힌트를 제공하고 있다.

* Dublin Core

See p.72

* METS

METS는 W3C의 XML 스키마를 사용하여 표현된 디지털 도서관의 객체에 대한 기술, 행정 및 구조 메타데이터를 암호화하기 위한 메타데이터 표준이다. 이 표준은 the Network

Development and MARC Standards Office of the Library of Congress에 의해 유지관리되며, the Digital Library Federation의 주도하에 개발 중이다.

METS는 다음과 같은 목적을 위하여 디자인된 XML 스키마이다:

(1) Creating XML document instances that express the hierarchical structure of digital library objects.
디지털 도서관 객체의 계층적 구조를 표현하는 XML 도큐멘트 경우의 창조.
(2) Recording the names and locations of the files that comprise those objects.
그 같은 객체를 구성하는 필드들의 이름과 위치의 기록.
(3) Recording associated metadata. METS can, therefore, be used as a tool for modeling real world objects, such as particular document types.
관련된 메타데이터의 기록. 그러므로 METS는 특별한 도큐멘트 유형과 같은 실세계의 객체를 모델화하는 도구로 사용될 수 있다.

이러한 용도에 따라서, METS 도큐멘트는 Open Archival Information System Reference Model의 Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP)의 역할을 담당하도록 할 수 있다.

* OAIS: An Open Archival Information System

OAIS는 정보를 보존하여 어떤 특정한 커뮤니티용으로 이용할 수 있도록 해야 할 책임을 갖고 있는 사람과 시스템의 조직으로 구성된 아카이브 이다. 디지털 정보를 장기간 보존하는 데 필요한 시스템, 즉 아카이브를 위한 개념적 구조틀을 마련한 ISO 표준(ISO 14721)이다. ISO의 요청으로 미국 항공 우주국(NASA)의 CCSDS(Consultative Committee for Space Data Systems)가 주체가 되어 개발했다.

1999년 초안이 발표된 후 국제적 · 다학문적 의견 수렴 과정을 거쳐 2002년 국제 표준으로 확정되었다. 정보 모형, 정보 패키지 모형, 아카이브 기능 모형 등 디지털 아카이빙과 관련된 기본적인 개념틀을 정의하였다. 현재 디지털 아카이빙과 관련된 거의 모든 실험과 프로젝트가 이 표준을 기반으로 진행되고 있을 정도로 커다란 의미를 갖는 성과이다. OAIS 참조 모형의 서문은 이 문건을 “한 아카이브가 디지털 정보를 영구 혹은 무기한 장기간 보존하는 데 있어서 광범위한 의견 일치에 도달하기 위해 개발된 기술적 권고안”이라고 소개하고 있다. 광범위한 의견 일치란 디지털 정보를 장기간 보존하는 활동을 수행하는 모든 기관들 사이의 의사소통 기반을 마련해 앞으로의 협력을 활성화시킨다는 의도를 요약한 것이다. 따라서 이 참조 모형의 가장 즉각적인 의미는 수년에 걸친 개발과 의견 수렴 과정을 통해 디지털 정보 보존에 관한 기본 개념과 용어에 대한 의견 일치를 도출해낸 데 있다.

OAIS 참조 모형은 정부 기관, 도서관, 아카이브즈, 그리고 기업체나 대학 등 디지털 정보를 보존하여 이용할 수 있게 하는 모든 기관, 심지어는 현재로서는 스스로가 아카이빙 기능을 수행하고 있다고 믿지 않는 기관들까지 그 적용 대상으로 상정하고 있다.

** The reference model:

(1) 장기간의 디지털 정보 보존과 접근에 필요한 문서 개념에 대한 이해력과 자각심의 기본틀을 제공한다.

(2) 보존과정에서 효과적인 참여자가 되기 위하여 비-문서적 기관에 의해 요구되는 개념을 제공한다.

(3) 현재와 과거의 문서관의 운영과 구조를 비교하고 설명하기 위한 용어와 개념을 포함하는 기본틀을 제공한다.

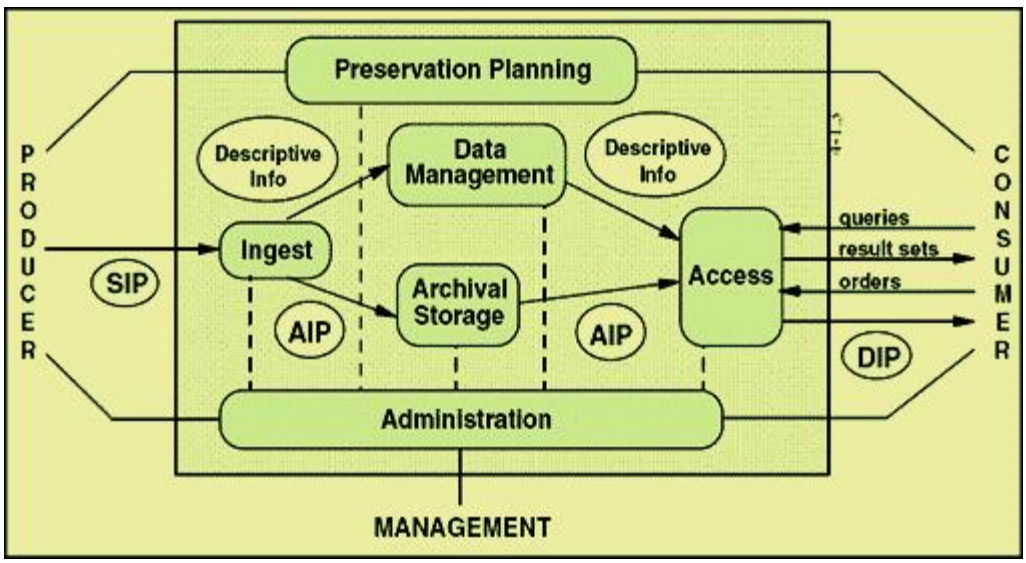
(4) 서로 다른 장기 보존 전략과 기술을 비교하고 설명하는 위한 기본틀을 제공한다.

(5) 문서관에 이해 보존된 디지털 정보의 데이터 모델을 비교하고 데이터 모델과 그것의 주요 정보가 시간이 지나면서 어떻게 변하는지를 논의하기 위한 근거를 제공한다.

(6) 비 디지털 형태(예, 물리적 매체와 물리적 표본)의 정보의 장기보존을 다루기 위하여 다른 연구에 의해 확장될 수도 있는 기초를 제공한다.

(7) 장기 디지털 정보의 보존과 접근을 위한 요소와 과정에 대한 일치된 의견을 확장시키고, 기업이 지원할 수 있는 보다 큰 시장을 육성시킨다.

(8) OAIS-관련 표준의 확인과 제작을 안내한다.



OAIS Functional Entities

*** Submission Information Package (SIP)

*** Archival Information Package (AIP)

*** Dissemination Information Package (DIP)

*** BagIt**

BagIt는 임의적인 디지털 콘텐트의 네트워크 전달과 디스크형 저장을 지원하기 위한 계층형 파일 패키징 포맷이다. 하나의 “bag(임의적 콘텐트)”은 bag의 저장과 전송을 도큐멘

트하려는 목적을 가진 메타데이터 파일들인 “payload”와 “tags”로 구성된다. 필요로 하는 태그 파일에는 그것에 상응하는 checksum을 갖고 있는 payload에 있는 모든 파일을 리스트하고 있는 manifest(적하목록, 승객명단)를 포함하고 있다. BagIt이란 이름은 때때로 “bat it and tag it”으로 말하기도 하는 “enclose and deposit” 방법에 의해 유래하였다.

Bags은 하나의 파일집단처럼 정상적으로 보관된 디지털 콘텐트용으로는 이상적이다. 이것들은 또한 문서보관 목적으로, 접수 부분의 지원이 쉽지 않은 데이터베이스 구조에 정상적으로 보관된 콘텐트를 export하는데 매우 적합하다. cross-platform (Windows and Unix) file system naming conventions에 따라, bag의 payload에는 어느 정도의 디렉토리들과 하위 디렉토리들(폴더와 하위 폴더)이 포함될 수 있다. bag은 그 bag을 완성하기 위하여 그 네트워크의 밖에서 가져올 수도 있는 콘텐트용의 URL들을 리스트하고 있는 “fetch.txt”파일을 통하여 간접적으로 payload content를 검사할 수 있다: 간단한 paralleization(평행, 비교) - 예를 들어, Wget의 연속적인 10가지 경우들 - 은 커다란 bags를 매우 빠르게 전송하기 위하여 이러한 기능을 활용하고 있다. bags의 장점은 다음과 같다:

디지털 도서관에서 널리 채택하고 있다(예, 미 의회 도서관)
일반적으로 잘 알려진 파일 시스템 도구를 사용하여 설치하기가 용이하다.
파일처럼 취급받는 콘텐트는 단지 payload 디렉토리에서만 복사될 수 있다.
ML wrapping과 비교해서, 콘텐트는 시간과 저장 공간을 절약하기 위하여 암화화될 필요가 없다.
접수된 콘텐트는 친숙한 파일시스템 트리로 쉽게 갈 수 있다.
병행적으로 일반적인 전이 도구를 기동시킴으로써 신속한 네트워크 전이가 용이하다.

1) In computing, **cross-platform**, or multi-platform, is an attribute conferred to computer software or computing methods and concepts that are implemented and inter-operate on multiple computer platforms.

2) **GNU Wget** (or just Wget, formerly Geturl) is a computer program that retrieves content from web servers, and is part of the GNU Project. Its name is derived from World Wide Web and get. It supports downloading via HTTP, HTTPS, and FTP protocols.

*MODS

MODS는 XML 기반의 서지 기술 구조형식이며 미 의회에서 개발하였다. MODS는 도서관에서 사용하는 MARC 포맷의 복잡성과 Dublin Core 메타데이터의 지나친 단순성 간의 타협으로 디자인되었다. MODS 레코드는 MARC 레코드로부터 key data elements를 전달하도록 디자인되었지만 모든 MARC 필드들을 정의하지는 않으며 MARC 표준으로부터 tag하고 있는 필드나 하위필드를 사용하지도 않는다. MODS에는 MARC 레코드와 호환할 수 없는 데이터 요소들이 존재하기 때문에, MARC에서부터 MODS로 변환시킬 때, 또는 그 반대의 경우에 어느 정도의 손실이 발생한다. MODS 이용자 집단에게 아무리 편리하다고 하더라도 이 두 가지 메타데이터 포맷의 호환성을 유지하는데 미 의회도서관은 어떠한 책임도 지지 않는다. MODS의 사용은 다른 메타데이터 구조식과 비교하여 여러 가지 장점을 제공한다:

기존의 자원 기술들과 높은 호환성을 갖는다.
다양한 내적 레코드 요소 세트들을 MODS용으로 작성할 수 있도록 하기 위하여 MARC보다 세부내용이 덜 까다롭다.
DC에서 outside로부터의 아이템 기술과 기타 보다 간단한 포맷은 mapped되어 개선될 수 있다.

*** DDI:** The Data Documentation Initiative (DDI)

DDI는 통계 및 사회과학 데이터를 기술하기 위한 정보 표준을 작성하기 위한 국제적 프로젝트이다. 1995년에 시작하였으며, 전세계의 데이터 전문가들이 참여하였다. XML로 쓰여진 DDI 규칙은 정보의 콘텐트, 교환, 보존용의 포맷을 제공하고 있다.

*** ISO standard Digital Object Identifier (DOI)**

See p.85

*** ISO standard Uniform Resource Name (URN)**

URN이란 the urn:scheme을 사용하는 URI의 오래된 역사적 이름이다. RFC 2141에서 1997년에 정의된, URNs는 단일 URN namespace에 간단하게 namespace의 mapping이 가능하도록 함으로써 자원들에 대한 항구적이고 위치-독립적인 식별자로서 활동하도록 만들어졌다. 이 같은 URI가 있다는 것은 식별된 자원의 이용가능성을 의미하는 것이 아니라, 자원들이 사라지거나 이용불가능하게 될 때조차도 그 URIs는 전 세계적으로 유일하고도 항구적으로 남아있어야 한다는 것이다.

1) A **Request for Comments (RFC)** is a publication of the Internet Engineering Task Force (IETF) and the Internet Society, the principal technical development and standards-setting bodies for the Internet.

2005년에 RFC 3986 이래로, 이 용어의 사용은 W3C와 IETF의 공동실무진에 의해 제안된 개념이면서 엄격성이 다소 떨어진 "URI"에 대한 선호로 인하여 줄어들었다. URNs(이름)과 URLs(위치명) 둘 다는 URIs이며 하나의 특별한 URI는 동시에 이름이면서 위치이름일 수 있다. URNs은 원리 1990년대에 하나의 메타데이터 기본틀로 URLs과 URCs와 함께 인터넷의 3-부분 정보구조의 일부로 만들어졌다. 그렇지만, URCs는 결코 과거의 개념적 단계를 벗어나지 못했으며 RDF와 같은 다른 기술이 후에 그 자리를 차지하였다.

**** URC: a uniform resource characteristic (URC)**

URC는 일련의 문자열이며 URI(웹 자원을 식별할 수 있는 문자열)의 메타데이터를 표현한다. URC는 URI의 조합적 URN(웹 자원의 유일한 이름)을 그것의 URL(웹 자원을 찾을 수 있는 위치)에 결합시킨다.

* PREMIS schema:

PREMIS (PREservation Metadata: Implementation Strategies)는 디지털 보존을 위해 사용할 수 있는 메타데이터의 개발을 위한 국제실무집단이다. 2003년에 OCLC와 RLG는 PREMIS 실무그룹을 만들었으며, 관리와 사용을 위한 가이드라인과 권고문과 더불어 실행 가능하고 핵심적인 보존용 메타데이터를 정의하기 위하여 이들은 문화, 정부, 사설 기관에서 온 30명 이상의 대표들로 구성된 다국적 멤버들이다. PREMIS는 설치에 있어 독립적이고, 실무 지향적이며 대부분의 보존기관에서 필요할 수 있는 어의적 요소의 집합을 정의하는데 책임이 있다.

2005년 5월에, PREMIS는 Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group를 해제하였다. 이 237 페이지짜리 보고서에는 다음과 같은 것들이 포함되어 있다:

PREMIS Data Dictionary 1.0- 디지털 아카이빙 시스템에서 보존 메타데이터를 실행하기 위한 포괄적이고 실재적인 자원;
딸림 보고서 - context, data model, assumptions을 제공;
특별한 논제, 용어, 사용의 예제들;
Data Dictionary의 사용을 지원하기 위하여 개발된 XML 구조식의 세트.

PREMIS 2.0 번전은 2008년 3월에 해제되었다.

** Entities

PREMIS 데이터 모델은 5개의 상호연관된 객체들로 이루어져 있다:

Intellectual,
Object,
Event,
Agent,
Rights - 위의 4가지 분야 중 하나에 mapped된 각각의 어의적 개체에 대한 권리.

지적 객체는 책이나 데이터베이스처럼 독립적이고 서로 연결되어 있는 지적 개체로 구성되어 있는 한 세트의 콘텐트이다. 이것들은 다른 지적 실체를 포함하고 있는 복합적인 객체일 수 있으며 복수적으로 디지털 표현이 가능할 수 있다. 기술 메타데이터는 일반적으로 이 단계에서 적용 된다; 경쟁 대상의 스킴들이 늘어남으로써, 이 실무 그룹은 어떤 추가적인 기술적 어의적 개체를 정의하지 않고 외부의 스킴용을 사용 가능한 “extension containers(콘테이너들은 관련된 그룹의 어의적 개체를 소장하고 있다)”를 통해 상호 호환하도록 하였다.

데이터 사전에 리스트되어 있는 대부분의 어의적 개체는 객체와 관련 실체와 연관되어 있다. 전자는 추가적으로 file, bitstream, representation인 3가지의 하위 유형으로 세분된다. 파일은 대부분의 최종 이용자가 “운영체계에 의해 알려진 바이트들의 이름이 붙고 주문을 한 순서” 작업을 하기 위하여 사용하는 수준이다. 여기에는 보존 목적을 위하여 의미있

는 공동의 성질을 가지고 있는 파일 속에 있는 연속 또는 비연속 데이터인 bitstreams을 품고 있는 운영체계에 의해 그것을 이해할 수 있도록 제공되는 다양한 파일 시스템 속성들이 포함된다. 대표는 어떤 의미에서 이 모델의 가장 높은 수준이다. 왜냐하면 이것은 지적 실체의 구조와 콘텐트를 적합하게 제공하기 위하여 여러 가지 파일을 포함할 수 있기 때문이다. 모든 레코지토리가 실체의 디지털적인 "본질적 가치"라고 여겨지는 것을 보존하기 위한 관리기관의 필요와 자신들의 목적에 따라 대표를 보존하는 것과 관련이 있는 것은 아니다. 더구나 지적 실체는 한 레포지토리에서 다수의 대표일 수도 있다. 이벤트는 객체와 결합된 이벤트나 에이전트("events와 결합되어 있는 사람, 기관, 또는 소프트웨어 또는 한 객체에 붙어있는 Rights)에 영향을 끼치고 있는 행동과 관련이 있는 한 객체들과 상호 관련되어 있다.

최종적으로 권리 실체의 내포는 저작권과 계약권의 법률적 필요조건에 대한 관심과 자성을 높이게 되었다. 또한 여기에는 허가된 특별한 행동들에 대한 정보가 포함되어 있다. 예를 들어, 어의 개체 4.1.6.1 조항: "보존 레포지토리가 취할 수 있도록 허용된 행동"에는 복제, 전달, 삭제에 대하여 제안된 값들을 포함하고 있다.

** Data dictionary

PREMIS 데이터 사전 항목들에는 12가지의 속성분야가 포함되어 있다. 이것들 모두가 모든 어의적 개체(다른 메타데이터에 있는 한 "요소"와 비슷한)에 응용되는 것은 아니다. 그 개체의 이름과 정의와 더불어, 그 필드들은 그 개체용으로 포함되어져야 하는 이론적 근거, 용도 설명서, 그리고 그 값이 채워지는 방법과 같은 것들을 기록한다. 이 속성들 중에서 4가지 – 객체범주, 적용성, 반복성, 그리고 책임성 –은 서로 연결되어 있으며, 마지막 3가지는 각각의 file, bitstream, representation인 객체 실체 수준용으로 정의된다. 이 사전은 계층적이다: 어떤 어의적 개체들은 다른 것들 속에 포함되어진다.

*** Ecological Metadata Language**

EML은 생태학 분야에서 개발된 메타데이터 표준이다. 이것은 the Knowledge Network for Biocomplexity를 포함하여 the Ecological Society of America 등에 의해 수행된 이전의 연구를 근거로 삼고 있다. EML은 메타데이터의 구조적 표현용으로 사용 가능한 한 세트의 XML 구조식 도큐멘트이다. 이것은 생태학에서 연구자들로 하여금 전형적인 데이터 세트를 도큐멘테이션 할 수 있도록 특별히 개발되었다. EML은 주로 디지털 자원을 기술하도록 설계되었지만, 그것은 또한 종이 지도와 기타 비 디지털 매체와 같은 비 디지털 자원을 기술하는데도 사용할 수 있다.

*** OAI-PMH(Open Archives Initiative Protocol for Metadata Harvesting)**

the Open Archives Initiative에 의해 개발된 프로토콜이다. 많은 아카이브즈들로부터 메타데이터를 이용하여 서비스가 이루어지도록 하기 위하여, 이것은 아카이브에 있는 레코드들의 메타데이터 기술을 수집하기 위하여 사용된다. OAI-PMH의 실행은 Dublin Core에 있는 메타데이터의 표현을 지원해야만 하지만, 또한 추가적인 표현도 지원하기도 한다. 이 프로토콜은 대체로 OAI Protocol로 언급되기도 하며, HTTP보다는 XML을 사용한다.

* Z39.50

Z39.50은 원거리 컴퓨터 데이터베이스로부터 정보를 찾거나 검색하기 위하여 사용되는 클라이언트-서버 프로토콜이다. 이것은 ANSI/NISO standard Z39.50 이면서 ISO standard 23950 이다. 이 표준의 관리기관은 미 의회도서관이다. Z39.50은 도서관 환경에서 널리 사용되고 있으며 종종 통합도서관시스템과 인적 서지 참고 소프트웨어에 사용되고 있다. 도서관 상호대차를 위한 상호대차목록 탐색은 종종 Z39.50 쿼리로 이루어진다.

Contextual Query Language (formerly called the Common Query Language)는 Z39.50 어의론에 근거하고 있다. Z39.50은 웹 이전의 기술이며 다양한 실무 그룹이 오늘날의 환경에 보다 잘 적응하도록 하기 위하여 갱신을 시도하고 있다. 이러한 시도는 the designation ZING(Z39.50 국제판)에서 나타나고 있으며 다양한 전략이 추구되고 있다.

1) **Contextual Query Language (CQL)**, previously known as Common Query Language, is a formal language for representing queries to information retrieval systems such as search engines, bibliographic catalogs and museum collection information. Based on the semantics of Z39.50, its design objective is that queries be human readable and writable, and that the language be intuitive while maintaining the expressiveness of more complex query languages.

가장 중요한 것은 쌍둥이 프로토콜들인 SRU/SRW이다. 이것들은 Z39.50 통신 프로토콜(HTTP로 그것을 대체한)에 속하지만 쿼리 구문의 장점을 보존하려고 시도하고 있다. SRU는 REST을 기반으로 하고 있으며, 쿼리를 URL 쿼리 스트링에서 표현할 수 있도록 한다: SRW는 SOAP를 사용하고 있다. 둘 다 탐색결가를 XML로 바꿀 수 있다.

이러한 프로젝트들은 상대적으로 시장이 작은 도서관 소프트웨어를 가지고 훨씬 커다란 시장용 개발된 웹 서비스 도구로부터 이익을 얻을 수 있도록 함으로써, 최초의 Z39.50 프로토콜보다 개발자가 참여하기 훨씬 좋아졌다.

이것의 대안들은 다음과 같다:

Search/Retrieve Web Service, successor to Z39.50
Open Archives Initiative Protocol for Metadata Harvesting
SPARQL

1) **SPARQL** (pronounced "sparkle", a recursive acronym for SPARQL Protocol and RDF Query Language) is an RDF query language, that is, a query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format.

* UDDI: Universal Description Discovery and Integration

UDDI는 전세계의 기업들이 인터넷 상에 스스로 리스트를 올릴 수 있는 프랫폼-독립적이며 XML-의존형 레지스트리이며, 웹 서비스 어플을 등록하고 위치를 설정할 수 있는 메카니즘이다. UDDI는 the Organization for the Advancement of Structured

Information Standards (OASIS)에 의해 후원을 받는 open industry initiative이며, 그 목적은 기업들로 하여금 서비스 리스팅즈를 출판할 수 있고, 서로서로를 발견할 수 있고, 그리고 그 서비스와 소프트웨어 어플이 인터넷으로 상호작용하는 방법을 정의할 수 있도록 하기 위한 것이다.

UDDI는 원래는 하나의 핵심적인 웹 서비스 표준으로 제안되었다. 이것은 SOAP 메시지에 의해 질문하도록 그리고 그 프로토콜 바이딩즈를 기술하고 있는 Web Services Description Language (WSDL) document로의 접근을 제공하도록 디자인되었다. 그리고 웹 서비스와 상호작용하는데 요구되는 메시지 포맷은 그것의 디렉토리에 열거되어 있다.

* SRU: Search/Retrieval via URl

Search/Retrieve via URL (SRU)는 인터넷 탐색 쿼리용으로 만든 표준 탐색 프로토콜이며 쿼리를 표현하기 위한 표준 쿼리 구문으로 Contextual Query Language (CQL)를 사용하고 있다.

Applications: Image search, Video search engine, Enterprise search, Semantic search;

* Search/Retrieve Web service (SRW)

이것은 탐색 및 검색용 웹 서비스이다. SRW는 쿼리용으로, 그리고 그것과 동반자적인 프로토콜 SRU에 의해 제공된 URL 인터페이스를 증대시키기 위하여 SOAP 인터페이스를 제공한다. SRU와 SRW에서의 쿼리들은 CQL을 사용하여 표현된다. SRW, SRU, and CQL의 표준은 미 의회도서관에 의해 공표된다.

* Representational State Transfer (REST)

웹과 같은 분산형 시스템을 위한 소프트웨어 구조의 스타일이다. REST는 탁월한 웹 API(응용 어플 프로그램 – 서로 통신하기 위하여 소프트웨어 구성요소들에 의해 하나의 인터페이스퍼엄 사용될 목적으로 만든 프로토콜)처럼 등장하였다. API는 routines, data structures, object classes, and variables) design model용의 스펙을 포함하기도 하는 하나의 도서관이다.

REST는 서로 다른 서비스간의 헐거운 이중성을 허용함으로써 웹 서버들 간의 교신을 용이하게 한다. REST는 그것의 상대방인 SOAP보다는 형식에 있어서 다소 강하지 못하다. REST 언어는 명사와 동사가 사용되며 가독성을 강조하고 있다. SOAP와 달리, REST는 XML 검사가 필요하지 않으며 서비스 제공자로부터 또는 그것에게로 가는 메시지 헤더를 필요로 하지도 않는다. 이것은 궁극적으로 적은 bandwidth를 사용한다. REST의 에러처리 또한 SOAP에서 사용되는 것과는 차이가 난다.

1) In computer networking and computer science, **bandwidth, network bandwidth, data bandwidth, or digital bandwidth** is a measurement of bit-rate of available or consumed data communication resources expressed in bits per second or multiples of it (bit/s, kbit/s, Mbit/s, Gbit/s, etc.).

* SOAP: Simple Object Access Protocol

컴퓨터 네트워크에서 웹 서비스를 실행하는데 있어서 정형화된 정보를 교환하기 위한 프로토콜 스펙이다. 이것은 그것의 메시지 포맷용으로 XML을 사용하고 있으며, 보통 다른 Application Layer 프로토콜 - 가장 유명한 것은 HTTP나 SMTP - 에 의존하고 있다.

1) In the Internet model, the **application layer** is an abstraction layer reserved for communications protocols and methods designed for process-to-process communications across an Internet Protocol (IP) computer network. Application layer protocols use the underlying transport layer protocols to establish process-to-process connections via ports.

In the OSI model, the definition of its application layer is narrower in scope. The OSI model defines the application layer as being the user interface. The OSI application layer is responsible for displaying data and images to the user in a human-recognizable format and to interface with the presentation layer below it.

* SPARQL (pronounced "sparkle", a recursive acronym for SPARQL Protocol and RDF Query Language)

SPARQL은 RDF 쿼리 언어, 즉 데이터베이스 언어로서 RDF 포맷에 저장된 데이터를 검색하고 처리하도록 한다. 이것은 the RDF Data Access Working Group (DAWG) of the World Wide Web Consortium의 표준이며 시멘틱 웹의 주요 기술 중의 하나로 인식되고 있다. SPARQL은 triple patterns, conjunctions, disjunctions, 그리고 optional patterns으로 구성된 쿼리를 사용한다. 복수의 프로그래밍 언어용으로 실행이 가능하다.

1) A **triplestore** is a purpose-built database for the storage and retrieval of triples, a triple being a data entity composed of subject-predicate-object, like "Bob is 35" or "Bob knows Fred".

Much like a relational database, one stores information in a triplestore and retrieves it via a query language. Unlike a relational database, a triplestore is optimized for the storage and retrieval of triples. In addition to queries, triples can usually be imported/exported using Resource Description Framework (RDF) and other formats.

2) In logic and mathematics, a two-place logical operator **and**, also known as logical **conjunction**, results in true if both of its operands are true, otherwise the value of false.

3) In logic and mathematics, **or** is a truth-functional operator also known as (inclusive) **disjunction** and alternation. The logical connective that represents this operator is also known as "or", and typically written as or . The "or" operator produces a result of true whenever one or more of its operands are true. For example, in this context, "A or B" is true if A is true, or if B is true, or if both A and B are true.

* WSDL: Web Service Description Language

WSDL은 XML-기반 인터페이스 기술 언어이며, 웹 서비스에서 제공된 기능성을 기술하기 위하여 사용된다. 웹 서비스의 WSDL 기술(WSDL 파일이라고도 함)에서는 서비스를 어떻게 부를 수 있는 방법, 예상되는 그것의 매개변수, 반환되는 데이터의 구조 형태에 대한 기계가독형 기술을 제공한다. 그러므로 프로그래밍 언어에 있는 a method signature의

that에 억지로 일치시키려는 목적에 도움이 된다. 2.0 버전에서는 두문자 D가 Definition으로 바뀌었다.

* CGI: Common Gateway Interface

CGI란 웹 페이지와 웹 어플에서 역동적인 콘텐트를 생산하는데 사용되는 표준 방법이다. 웹 서버를 실행할 때, CGI는 웹 서버와 그것의 웹 콘텐트를 생산하는 프로그램 사이에 인터페이스를 제공한다. 이러한 프로그램들은 CGI scripts 또는 단순히 CGIs로 알려져 있다; 이것들은 보통 scripting language로 작성된다.

1) A **scripting language** or script language is a programming language that supports scripts, programs written for a special run-time environment that can interpret (rather than compile) and automate the execution of tasks which could alternatively be executed one-by-one by a human operator.

p. lxvii

* Open URL

OpenURL은 URL의 표준 포맷이며, 인터넷 이용자로 하여금 접근할 수 있는 자원의 사본을 보다 쉽게 발견할 수 있도록 하기 위한 것이다. 비록 OpenURL이 인터넷에서 다양한 종류의 자원에 사용될 수 있다하더라도, 이용자를 구독예약 콘텐트에 연결시켜주는 것을 도와주는 도서관에서 가장 많이 사용하고 있다.

OpenURL 표준은 초록 및 색인 데이터베이스(정보원)와 같은 정보자원으로부터 온라인으로나 인쇄물로나 또는 다른 포맷으로 학술지와 같은 도서관 서비스(목표물)로의 링크가 가능하도록 디자인되었다. 그 링킹은 OpenURL의 요소들을 조사하여, OpenURL 지식 베이스를 이용함으로써 도서관을 통해 이용 가능한 올바른 목표물로 링크를 제공하는 “link resolvers” 또는 “link-servers”에 의해 이루어진다.

OpenURL을 생산하는 정보원은 전형적으로 학술 기사, 책, 특허 등과 같은 도서관에서 종종 발견되는 정보자원을 색인하고 있는 데이터베이스에 있는 서지 인용이나 서지 레코드이다. 이러한 데이터베이스의 예로는 Ovid, Web of Science, SciFinder, Modern Languages Association Bibliography , Google Scholar가 있다.

1) Ovid Technologies, Inc. (or just **Ovid** for short), part of the Wolters Kluwer group of companies, provides access to online bibliographic databases, academic journals, and other products, chiefly in the area of health sciences. The National Library of Medicine's MEDLINE database was once its chief product but, as this is now freely available through PubMed.

2) **Web of Science (WoS)** is an online subscription-based scientific citation indexing service maintained by Thomson Reuters that provides a comprehensive citation search. It gives access to multiple databases that reference cross-disciplinary research, which allows for in-depth exploration of specialized sub-fields within an academic or scientific discipline.

3) CAS databases are available via two principal database systems, **STN**, and **SciFinder**.

STN

STN (Scientific & Technical Information Network) International is operated jointly by CAS and FIZ Karlsruhe, and is intended primarily for information professionals, using a command language interface. In addition to CAS databases, STN also provides access to many other databases, similar to Dialog.

SciFinder

SciFinder is a database of chemical and bibliographic information. Originally a client application, a web version was released in 2008. It has a graphics interface, and can be searched for chemical structures.

CASSI

CASSI stands for Chemical Abstracts Service Source Index. This formerly print-only database is now a free online resource to look up and confirm publication information. CASSI provides titles and abbreviations, CODEN, ISSN, publisher, and date of first issue (history) for a selected journal. Also included is its language of text and language of summaries. The range is from 1907 to the present, including both serial and non-serial scientific and technical publications.

4) The **Modern Language Association of America** (referred to as the Modern Language Association or MLA) is the principal professional association in the United States for scholars of language and literature. The MLA aims to "strengthen the study and teaching of language and literature." The organization includes 30,000 members in 100 countries, primarily academic scholars, professors, and graduate students who study or teach language and literature, including English, other modern languages, and comparative literature.

5) **Google Scholar** is a freely accessible web search engine that indexes the full text of scholarly literature across an array of publishing formats and disciplines. Released in beta in November 2004, the Google Scholar index includes most peer-reviewed online journals of Europe and America's largest scholarly publishers, plus scholarly books and other non-peer reviewed journals. It is similar in function to the freely available Scirus from Elsevier, CiteSeerX, and getCITED. It is also similar to the subscription-based tools, Elsevier's Scopus and Thomson ISI's Web of Science. Its advertising slogan – "Stand on the shoulders of giants" – is taken from a quote by Isaac Newton and is a nod to the scholars who have contributed to their fields over the centuries, providing the foundation for new intellectual achievements

목표물이란 자원이나 서비스이며 이용자의 정보 요구를 만족시키는 것을 돕는다. 목표물의 예로는 full-text repositories, online journals, online library catalogs 그리고 기타 Web resources와 services가 있다. NISO는 ANSI 표준 Z39.88로서 OpenURL과 그것의 데이터 콘테이너(the Context Object)를 개발하였으며, 2006년 6월 22일에 OCLC가 이 표준의 유지관리기관으로 지명되었다.

** Use

OpenURL의 가장 일반적인 응용은 웹 자원(온라인 학술기사와 같은)의 요청에 대한 해결책을 제시하는 것이다. 하나의 OpenURL에는 참고한 자원 그 자체, 그리고 콘텍스트 정보 – OpenURL이 발생한 콘텍스트(예를 들어, 도서관 목록으로 얻은 탐색 결과의 페이지)와 리퀘스트의 콘텍스트(예를 들어, 리퀘스트를 한 특별한 이용자)에 대한 정보 둘 다를 포함하고 있다. 만일 다른 콘텍스트가 그 URL에 표현되어 있다면, 다른 카피가 제시되어 해결하게 된다. 콘텍스트에서 변화는 예측 가능하며 다른 콘텍스트용으로 다른 URLs를 처리하기 위하여 하이퍼링크(예를 들어, 학술지 출판사)의 최초 제작자를 필요로 하지 않는다.

예를 들어, 쿼리 스트링에서 기본 URL 이나 매개변수의 변화는 그 OpenURL이 어떤

다른 도서관에 있는 자원의 사본으로 해결한다는 것을 의미할 수 있다. 그래서 전자저널에서 어떤 경우를 포함하고 있는 동일한 OpenURL은 자원에 대한 자신들의 사본에 접근을 제공하는 어떠한 도서관에 의해서도 그 저널의 하이퍼링크를 완전히 덮어쓰지 않고 조정될 수 있다. COinS 도 봐라.

1) ContextObjects in Spans, commonly abbreviated **COinS**, is a method to embed bibliographic metadata in the HTML code of web pages. This allows bibliographic software to publish machine-readable bibliographic items and client reference management software to retrieve bibliographic metadata. The metadata can also be sent to an OpenURL resolver. This allows, for instance, searching for a copy of a book in one's own library.

** Format

OpenURL은 이용자의 기관 링크-서버의 주소를 포함하고 있는 하나의 기본적 URL과 그 뒤에 전형적으로 key-value pairs의 형태로 되어 있으면서 contextual data를 포함하고 있는 쿼리 스트링으로 구성되어 있다. contextual data는 대부분이 서지 데이터이지만, 1.0 버전에서처럼, OpenURL 역시 요구자, 하이퍼링크를 포함하고 있는 자원, 요청된 서비스의 유형 등에 대한 정보를 포함할 수 있다.

For example:

```
 http://resolver.example.edu/cgi
?genre=book
&isbn=0836218310
&title=The+ Far+ Side+ Gallery+ 3
```

is a version 0.1 OpenURL describing a book.

http://resolver.example.edu/cgi is the base URL of an example link-server.
In version 1.0, this same link becomes somewhat longer:

```
http://resolver.example.edu/cgi
?url_ver=Z39.88-2004
&rft_val_fmt=info:ofi/fmt:kev:mtx:book
&rft.isbn=0836218310&rft.btitle=The+ Far+ Side+ Gallery+ 3
```

A breakdown of the query string above shows that the following values are set:

1) The URL version **url_ver = Z39.88-2004**
2) Custom metadata **rft_val_fmt = info:ofi/fmt:kev:mtx:book**
3) And an object named "rft" is set which could be represented as
   rft = { isbn:"0836218310", btitle:"The Far Side Gallery 3" }

OpenURL was created by Herbert Van de Sompel, a librarian at the University of Ghent, in the late 1990s.

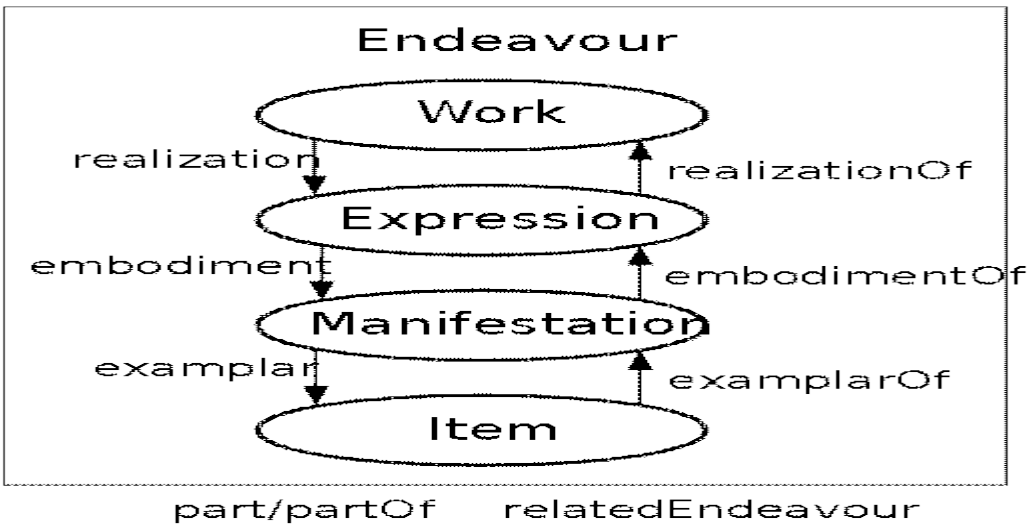
* FRBR: Functional Requirements for Bibliographic Records; (/ˈfɜrbər/)

FRBR는 IFLA에서 개발한 개념적 객체-관계 모델이며 이것은 이용자의 관점에서부터 온라인 도서관 목록과 서지 데이터베이스의 접근과 검색에 대한 이용자 임무와 관련되어 있다. 객체간의 관계는 관계의 계층을 항해하도록 링크를 제공하기 때문에 이것은 검색과 접근에 대한 보다 전체적인 접근방법을 나타낸다. 이 모델은 중요한데 그 이유는 AACR2나 ISBD와 같은 특수한 편목 기준과는 별개이기 때문이다.

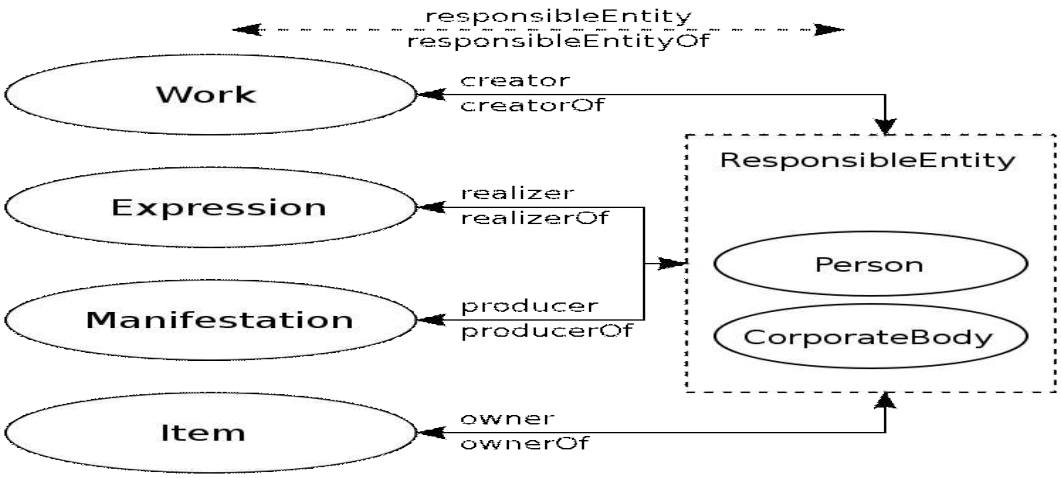
** FRBR entities

Group 1 entities and basic relations (RDF version)

Group 2 entities and relations



Group 1 entities and basic relation



Group 2 entities and relation

FRBR comprises groups of entities:

그룹 1 엔티티는 WEMI(work, expression, manifestation, and item) 이다. 이것들은 지능적 또는 예술적 노력의 결과물을 표현한다.

그룹 2 엔티티는 사람, 가족, 그리고 공동체이며, 그룹 1의 지능적 또는 예술적 노력의 관리에 책임을 진다.

그룹 3 엔티티는 그룹 1과 그룹1 2의 지능적 노력의 주제이며 개념, 대상, 사건, 장소가 포함된다.

그룹 1 엔티티는 FRBR 모델의 기초이다.

Work는 “명확한 지능적 또는 예술적 창조물”이다. 예를 들어, 베토벤의 교향곡 9은 그것을 표현하는 모든 방법과는 달리 하나의 work 이다. 우리가 “베토벤 9은 훌륭하다”고 말할 때, 일반적으로 그 작품을 말하는 것이다.

Expression은 “특별한 지적 또는 예술적 형태이며 한 작품이 ‘현실화’될 때 발생한다.” 베토벤의 9의 표현은 그가 쓴 음악악보의 초안일 수도 있다. 종이 그 자체가 아니라 그것에 의해 음악이 표현된다.)

Manifestation은 “어떤 작품의 표현의 물리적 구체화”이다. 하나의 엔티티로서 mainfestation은 지적인 콘텐트와 물리적 형태 둘 다와 관련해서 동일한 특성을 갖고 있는 모든 물리적 객체를 manifestation은 대표한다.“ 1996년에 베토벤 9번 교향곡의 런던 필하모니 연구는 manifestation이다. 비록 기록되어지지는 않는다 하더라도, 물론 manifestation이 레코딩이나 프린팅과 같은 항구적인 형태로 표현되어질 때 많은 관심을 받는다 하더라도, 이것은 물리적 구체화이다. ”런던 필하모니의 1996년 연주에 대한 레코딩은 베토벤 교향곡 9번의 본질을 포함하고 있다. 우리는 일반직으로 이것을 manifestation이라고 표현하고 있다.

Item은 “a manifestation의 단일 예이다.” 아이템으로 정의된 엔티티는 확실한 엔티티이다.“ 1996년 레코딩의 1996년 레코드의 각 사본은 하나의 아이템이다. 우리가 ”베토벤 교향곡 9번의 런던 필하모니의 1996년 연주에 대한 양 쪽의 사본들이 나의 지역 도서관에서 대출되었다“고 말할 때, 우리는 일반적으로 아이템들에 대해 말하는 것이다.

** Relationships

FRBR은 엔티티 간에 그리고 중에 있는 관계를 구축한다. “관계는 엔티티간의 링크를 설명하기 위한 근본적 도움을 제공하며, 그러므로 서지, 목록, 또는 서지 데이터베이스에서 표현하고 있는 우주를 항해하는 이용자를 도와주는 수단으로도 도움을 준다. 관계 유형의 예에는 포함되지만 다음과 같은 것을 제한하지는 않는다:

Equivalence relationships: 지적 콘텐트와 저작권이 보존되는 한, 한 작품의 동일한 manifestation의 확실한 사본들 간에 존재하거나 또는 원작과 그것의 재생품 사이에 존재한

다. 예로는 copies, issues, facsimiles and reprints, photocopies, and microfilms과 같은 복사물이 포함된다.

Derivative relationships: 저작을 근거로 한 서지 저작과 수정판 사이에 존재한다. 예를 들어, Editions, versions, translations, summaries, abstracts, and digests 가 있다. 새로운 저작이지만 옛 작품을 근거로 한 개작물(Adaptation), 장르의 변경, 작품의 주제적 콘텐트와 스타일은 유지하는 새로운 저작,

Descriptive relationships: 저작과 그것을 기술하고 있는 서평간에서처럼 서지 엔티티와 그 엔티티의 a description, criticism, evaluation, or review 사이에 존재한다. 기술관계는 또한 기존 저작의 annotated editions, casebooks, commentaries, critiques가 포함된다.

p. lxxii <제 3장 메타데이터>

* Dublin Core

DC 메타데이터 용어들은 한 세트의 어휘 용어들이며 찾기를 목적으로 자원을 기술하는데 사용할 수 있다. 이 용어들은 모든 웹자원(비디오, 이미지, 웹 페이지 등), 물리적 자원(책과 예술작품)과 같은 모든 것을 기술하는데 사용될 수 있다. DC 메타데이터 용어들의 풀 세트는 Dublin Core Metadata Initiative (DCMI) website에서 찾을 수 있으며, Dublin Core Metadata Element Set로 알려진 전통적인 15개의 메타데이터 용어 세트는 다음과 같은 표준에서 인증하고 있다.

IETF RFC 5013
ISO Standard 15836-2009
NISO Standard Z39.85

더불린 코어 메타데이터는 간단한 자원 기술에서부터 서로 다른 메타데이터 기준의 메타데이터 어휘를 결합하는데, 링크된 데이터 클라우드와 시멘틱 웹 실행에 있어서 메타데이터 어휘들에 대한 상호호환성을 제공하는데 까지 사용할 수 있다.

** Levels of the standard

The Dublin Core standard includes two levels: Simple and Qualified. Simple Dublin Core comprises 15 elements; Qualified Dublin Core includes three additional elements (Audience, Provenance and RightsHolder), as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in ways that may be useful in resource discovery.

Simple Dublin CoreThe Simple Dublin Core Metadata Element Set (DCMES) consists of 15 metadata elements:

1. Title 2. Creator 3. Subject 4. Description 5. Publisher

6. Contributor 7. Date 8. Type 9. Format 10. Identifier
11. Source 12. Language 13. Relation 14. Coverage 15. Rights

** Example of code; <meta name="DC.Publisher" content="publisher-name" >

The Dublin Core Metadata Initiative ; an open organization engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include work on architecture and modeling, discussions and collaborative work in DCMI Communities and DCMI Task Groups, annual conferences and workshops, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

DC-dot: Dublin Core metadata editor); This service will retrieve a Web page and automatically generate Dublin Core metadata, either as HTML tags or as RDF/XML, suitable for embedding in the …section of the page. The generated metadata can be edited using the form provided and converted to various other formats (USMARC, SOIF, IAFA/ROADS, TEI headers, GILS, IMS or RDF) if required. Optional, context sensitive, help is available while editing.

Editor-Converter Dublin Core metadata; This online program can be used for two purposes: as a Dublin Core metadata editor, and as a converter to UNIMARC. After conversion to UNIMARC format, metadata can be saved to your local hard drive as an ISO-2709 file. Viewable in Ukrainian, Russian and English.

DC-assist; A small, flexible help utility for metadata applications and is intended to complement the help pages embedded within existing software.

* SOIF(Summary Object Interchange Format)

객체개념을 도입한 SOIF는 각각의 자원에 대한 기술을 객체로 저장하고 하나의 SOIF 스트림에 여러 개의 객체를 동시에 전송하게 된다. SOIF는 인터넷상의 각종 자원을 표현하고 이를 네트워크상에서 운용할 수 있는 효과적인 메타데이터이다. SOIF는 미국 콜로라도 대학에서 개발한 Harvest Architecture의 일부로 디자인되었다.

Harvest:

하베스트는 인터넷 정보를 접근하고, 복사하고, 캐슁하고, 색인하고, 수집하기 위하여 측정하고 맞춤화할 수 있는 구조를 제공하는 시스템이다. 수집자들과 브로커들은 일련의 객체 요약들로 기술되어 있는 SOIF라고 부르는 속성-값 스트림 프로토콜을 사용하여 통신한다.

Constituency of use:

Records in SOIF are designed to be generated by Harvest gatherers and then used for user searches by Harvest brokers. They provide a summary of the resources that a Harvest gatherer has found. The Harvest distribution contains a number of stock gatherer programs that can generate SOIF summaries from plain text, SGML (including HTML), PostScript, MIF and RTF formats.

* IAFA/ROADS

ROAD templates는 ROADS 소프트웨어를 사용하는 the subjects services에 의해 사용된다. 이 템플릿들은 1994년의 the Internet Anonymous FTP Archive(IAFA) templates를 발전시킨 것이다. Dublin Core를 이러한 템플릿으로 mapping하는 것은 여러 메타데이터 종류들 간에 교환이 가능한 포맷으로 사용하기 위해서이다. 이 템플릿들은 1994년의 the Internet Anonymous FTP Archive(IAFA) templates를 발전시킨 것이다. Dublin Core를 이들 템플릿으로 mapping하는 것은 메타데이터 타입들 사이에 교환가능한 포맷으로 사용하려는 Dublin Core의 잠재적 역할에 대하여 관심을 갖고 조사하도록 한다.

** sample

Dublin Core record:

Title: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web.
Title: (Subtitle) Universal Resource Identifiers in WWW
Creator: Berners-Lee, T.
Subject: IETF, URI, Uniform Resource Identifiers
Publisher: CERN
Date: 1994
Type: Internet RFC
Format (scheme=IMT): text/plain
Identifier(scheme=URL): gopher://gopher.es.net:70/0R0-57601-/pub/rfcs/rfc1630.txt
Relation (type=child)(identifier=URL): http://ds.internic.net/ds/dspg1intdoc.html
Relation (type=sibling)(identifier=URL): http://ds.internic.net/rfc/rfc1738.txt

IAFA / ROADS template record:

Author-Name: Berners-Lee, T.
Category: Internet RFC
Creation-Date: 1994
Format: text/plain
Keyword: IETF, URI, Uniform Resource Identifiers
Publisher-Name: CERN
Title: A Unifying Syntax for the Expression of Names and Addresses of Objects on the Network as used in the World-Wide Web.
Title: Universal Resource Identifiers in WWW
Template-Type: DOCUMENT
URI-v1: gopher://gopher.es.net:70/0R0-57601-/pub/rfcs/rfc1630.txt

* TEI headers: Text Encoding Initiative

Text Encoding Initiative는 디지털 형태로 된 텍스트의 표현을 위한 표준을 공동으로 게 개발하고 유지관리하기 위한 a consortium. 주요 목적은 인문학, 사회과학, 그리고 언어학에 초점을 맞추어 기계가독형 텍스트의 encoding 방법을 규정하는 Electronic Text Encoding and Interchange를 위한 가이드라인을 제공하는 것이다. 이것은 모든 TEI-conformant text에 첫머리(prefixed)의 전자 타이틀 페이지를 만드는 데 필요한 기술적이고 선언적인 정보(descriptive and declarative information)을 제공한다.

TEI는 1980년대 이래로 지속적으로 활동하고 있는 디지털 인문학분야에서 실질적인 텍스트 지향적 코뮤니티이다. 이 코뮤니티는 현재 a mailing list, meetings and conference series를 발간하고 있으며 a wiki, a SourceForge repository 그리고 a toolchain과 같은 이름의 시조가 되는 기술 표준을 관리하고 있다.

1) A **wiki** (i/ˈwɪki/ WIK-ee) is usually a web application which allows people to add, modify, or delete content in collaboration with others. Text is usually written using a simplified markup language or a rich-text editor. While a wiki is a type of content management system, it differs from a blog or most other such systems in that the content is created without any defined owner or leader, and wikis have little implicit structure, allowing structure to emerge according to the needs of the users.

The encyclopedia project Wikipedia is the most popular wiki on the public web in terms of page views, but there are many sites running many different kinds of wiki software.

2) **SourceForge** is a web-based source code repository. It acts as a centralized location for software developers to control and manage free and open source software development. It was the first to offer this service for free to open source projects.

3) In software, a **toolchain** is the set of programming tools that are used to create a product (typically another computer program or system of programs). The tools may be used in a chain, so that the output of each tool becomes the input for the next, but the term is used widely to refer to any set of linked development tools.[

총괄적으로 XML 포맷을 정의한 TEI Guidelines는 실재적으로 그 커뮤니티의 뚜렷한 결과이다. 이 포맷은 텍스트용으로 다른 잘 알려진 개방형 포맷(프레젠테이션보다는 기본적으로 시멘틱적인 포맷인 HTML과 OpenDocument)과는 다르다; 모든 태그와 속성의 시멘틱스와 해석은 규정되어 있다. 약 500개의 서로다른 텍스트 구성요소왕 개념들(단어, 문장, 인물, 상형문자, 사람, 등); 각각은 하나 이상의 학술분야에 그 뿌리를 두고 있으며 예들을 제공하고 있다.

1) The Open Document Format for Office Applications (ODF), also known as **OpenDocument**, is an XML-based file format for spreadsheets, charts, presentations and word processing documents. It was developed with the aim of providing an open, XML-based file format specification for office applications.

The standard was developed by a technical committee in the Organization for the Advancement of Structured Information Standards (OASIS) consortium. It was based on the Sun Microsystems specification for OpenOffice.org XML, the default format for OpenOffice.org, which had been specifically intended "to provide an open standard for office documents."

In addition to being an OASIS standard, version 1.1 is published as an ISO/IEC international standard, ISO/IEC 26300:2006/Amd 1:2012 — Open Document Format for Office Applications (OpenDocument) v1.1.

이 표준은 두 부분으로 나눈다: 확장된 예와 토론과 함께 추론적인 텍스트 기술 그리고 tag-by-tag 세트의 정의. 대부분의 현대적 포맷(DTD, RELAX NG, W3C Schema)에서의 구조식은 자동적으로 tag-by-tag 정의로부터 생산된다. 수많은 도구가 이 가이드라인의 생산과 특별한 프로젝트에 적용하는데 도움을 주었다. 수많은 특별한 태그들이 기초가 되는 Unicode(요구되는 엄격한 선형성의 극복을 위하여 Unicode의 내포와 선택에 대한 자격을 요구하지 않는 문자들의 표현을 허용하는 그림문자)에 의해 부여된 제한을 완화시키는데 사용되고 있다.

1) A **document type definition (DTD)** is a set of markup declarations that define a document type for an SGML-family markup language (SGML, XML, HTML). A DTD uses a terse formal syntax that declares precisely which elements and references may appear where in the document of the particular type, and what the elements' contents and attributes are. A DTD can also declare entities that may be used in the instance document. XML uses a subset of SGML DTD.

2) In computing, **RELAX NG** (REgular LAnguage for XML Next Generation) is a schema language for XML – a RELAX NG schema specifies a pattern for the structure and content of an XML document. A RELAX NG schema is itself an XML document but RELAX NG also offers a popular compact, non-XML syntax. Compared to other XML schema languages RELAX NG is considered relatively simple.

3) The **World Wide Web Consortium (W3C)** is the main international standards organization for the World Wide Web (abbreviated WWW or W3).

<History> The World Wide Web Consortium (W3C) was founded by Tim Berners-Lee after he left the European Organization for Nuclear Research (CERN) in October, 1994. It was founded at the Massachusetts Institute of Technology Laboratory for Computer Science (MIT/LCS) with support from the European Commission and the Defense Advanced Research Projects Agency (DARPA), which had pioneered the Internet and its predecessor ARPANET.

W3C tries to enforce compatibility and agreement among industry members in the adoption of new standards defined by the W3C. Incompatible versions of HTML are offered by different vendors, causing inconsistency in how Web pages are displayed. The consortium tries to get all those vendors to implement a set of core principles and components which are chosen by the consortium.

It was originally intended that CERN host the European branch of W3C; however, CERN wished to focus on particle physics, not information technology. In April 1995 the Institut national de recherche en informatique et en automatique (INRIA) became the European host of W3C, with Keio University becoming the Japanese branch in September 1996. Starting in 1997, W3C created regional offices around the world; as of September 2009, it has eighteen World Offices covering Australia, the Benelux countries (Netherlands, Luxembourg, and Belgium), Brazil, China, Finland, Germany, Austria, Greece, Hong Kong, Hungary, India, Israel, Italy, South Korea, Morocco, South Africa, Spain, Sweden, and the United Kingdom and Ireland.

In January 2003, the European host was transferred from INRIA to the European Research Consortium for Informatics and Mathematics (ERCIM), an organization that represents European national computer science laboratories. In October 2012, W3C convened a community of large Web players and publishers to establish a MediaWiki wiki that seeks to documents open Web standards called WebPlatform and WebPlatform Docs.

이 포맷의 많은 이용자들은 모든 범위의 태그를 사용하는 것이 아니라 하위세트를 가지고 맞춤화를 이룬다. 이 포맷은 TEI 가이드라인과 관련된 학술분야의 집단 속에 있는 한 챕터에 상응하는 각각 상응하는 세트에서 태그들을 그룹화시킴으로써 이것을 지원한다. 이

포맷의 몇몇 이용자들은 더욱 나아가서 출판을 보다 쉽게 하기 위하여 자신들의 local house style을 확고히 하기 위하여 schematron stylesheet를 설명하고 있다.

1) In markup languages, **Schematron** is a rule-based validation language for making assertions about the presence or absence of patterns in XML trees. It is a structural schema language expressed in XML using a small number of elements and XPath.

TEI Lite is an XML-based file format for exchanging texts. It is a manageable selection from the extensive set of elements available in the full TEI Guidelines. TEI Lite는 텍스트를 교환하기 위한 XML 의존형 파일 포맷이다. 이것은 TEI Guidelines의 완전판에서 이용할 수 있는 요소들의 광대한 세트들로부터 다루기 쉬운 것만을 선택한 것이다.

*** GILS:Global Information Locator Service**

사람들이 자신이 필요로 하는 모든 정보를 보다 쉽게 찾도록 하는 서비스; 도서관 이용 경험이 있는 사람이면 누구나 이 서비스를 이용할 수 있으며, ISO 23950 search standard를 근거로, title, author, publish, date and place와 같은 도서관의 개념을 도입하여 전세계 누구나 정보자원을 찾는데 가장 쉽게 이해할 수 있는 개념을 포함하고 있다. GILS record는 도서관의 목록 레코드의 일종의 souped-up version 이다. 누구나 자신이 필요로 하는 모든 정보를 보다 쉽게 찾도록 하는 서비스이다. 도서관 이용 경험이 있는 사람이면 누구나 이 서비스를 이용할 수 있으며, ISO 23950 search standard를 근거로, title, author, publish, date and place와 같은 도서관의 개념을 도입하여 전 세계 누구나 정보자원을 찾는데 있어서 가장 쉽게 이해할 수 있는 개념을 포함하고 있다. GILS record는 도서관 목록 레코드의 일종의 매력적 버전(souped-up version) 이다.

*** IMS: IP Multimedia Subsystem or IP Multimedia Core Network subsystem**

Internet Protocal(IP) multimedia services를 전달하기 위한 구조적 기본 틀이며, 제3세대 이동통신망의 IP 멀티미디어 서비스를 위하여 개발된 기술이다. IP Multimedia Subsystem 또는 IP Multimedia Core Network Subsystem (IMS)은 IP 멀티미디어 서비스를 전달하기 위한 구조적 기본틀이다. 본래 이것은 GSM 이상으로 모바일 네트워크를 발전시키려는 견해의 일부분으로 the wireless standards body 3rd Generation Partnership Project (3GPP)에 의해 디자인되었다. 이것의 원래의 공식(3GPP Rel 5)은 GPRS를 넘어서 “인터넷 서비스”를 전달하려는 방법을 나타내었다. 이러한 비젼은 나중에 Wireless LAN, CDMA2000 그리고 fixed lines과 같은 GPRS보다 다른 네트워크를 지원해야 한다는 요구에 따라 3GPP, 3GPP2 그리고 ETSI TISPAN에 의해 갱신되었다.

1) **GSM** (Global System for Mobile Communications, originally Group Spécial Mobile), is a standard developed by the European Telecommunications Standards Institute (ETSI) to describe protocols for second generation (2G) digital cellular networks used by mobile phones. It became the de facto global standard for mobile communications with over 80% market share.

The GSM standard was developed as a replacement for first generation (1G) analog cellular

networks, and originally described a digital, circuit-switched network optimized for full duplex voice telephony. This was expanded over time to include data communications, first by circuit-switched transport, then packet data transport via GPRS (General Packet Radio Services) and EDGE (Enhanced Data rates for GSM Evolution or EGPRS).

Subsequently, the 3GPP developed third generation (3G) UMTS standards followed by fourth generation (4G) LTE Advanced standards, which are not part of the ETSI GSM standard.

"GSM" is a trademark owned by the GSM Association. It may also refer to the initially most common voice codec used, Full Rate.

2) **General packet radio service (GPRS)** is a packet oriented mobile data service on the 2G and 3G cellular communication system's global system for mobile communications (GSM). GPRS was originally standardized by European Telecommunications Standards Institute (ETSI) in response to the earlier CDPD and i-mode packet-switched cellular technologies. It is now maintained by the 3rd Generation Partnership Project (3GPP).

GPRS usage is typically charged based on volume of data transferred, contrasting with circuit switched data, which is usually billed per minute of connection time. Usage above the bundle cap is either charged per megabyte or disallowed.

GPRS is a best-effort service, implying variable throughput and latency that depend on the number of other users sharing the service concurrently, as opposed to circuit switching, where a certain quality of service (QoS) is guaranteed during the connection. In 2G systems, GPRS provides data rates of 56–114 kbit/second. 2G cellular technology combined with GPRS is sometimes described as 2.5G, that is, a technology between the second (2G) and third (3G) generations of mobile telephony. It provides moderate-speed data transfer, by using unused time division multiple access (TDMA) channels in, for example, the GSM system. GPRS is integrated into GSM Release 97 and newer releases.

인터넷과 함께 통합을 쉽게하기 위하여, IMS는 예를 들어, SIP처럼 가능하다면 어디에서나 IETRF 프로토콜을 사용한다. 3GPP에 따라, IMS는 어플을 표준화하려는 것이 아니라 그 보다는 무선과 유선 터미널로부터 멀티미디어와 음성 어플의 접근, 다시 말해서 FMC의 형태를 만들기 위하여 도움을 제공하는 것이다. 이것은 서비스 레이어로부터 접근 네트워크를 고립시키는 a horizontal control layer를 가짐으로써 이루어진다. 논리적인 구조 측면에서 보면, 제어 레이어가 일반적인 수평적 레이어이므로 서비스들은 그것들 자신만의 제어 기능을 가질 필요가 없다. 그렇지만, 실행하는데 있어서 이것이 반드시 비용과 복잡성을 크게 줄인다고 말할 수는 없다.

1) The **Session Initiation Protocol (SIP)** is a signaling communications protocol, widely used for controlling multimedia communication sessions such as voice and video calls over Internet Protocol (IP) networks.

The protocol defines the messages that are sent between peers which govern establishment, termination and other essential elements of a call. SIP can be used for creating, modifying and terminating sessions consisting of one or several media streams. SIP can be used for two-party (unicast) or multiparty (multicast) sessions. Other SIP applications include video conferencing, streaming multimedia distribution, instant messaging, presence information, file transfer, fax over IP and online games.

2) **Fixed-mobile convergence (FMC)** is a change in telecommunications that removes differences between fixed and mobile networks. In the 2004 press release announcing its formation, the Fixed Mobile Convergence Alliance said: Fixed Mobile Convergence is a transition point in the telecommunications industry that will finally remove the distinctions between fixed and mobile networks, providing a superior experience to customers by creating seamless services using a combination of fixed broadband and local

access wireless technologies to meet their needs in homes, offices, other buildings and on the go.

In this definition “fixed broadband” means a connection to the Internet, such as DSL, cable or T1. “Local access wireless” means Wi-Fi or something like it. BT’s initial FMC service used Bluetooth rather than Wi-Fi for the local access wireless. The advent of picocells and femtocells means that local access wireless can be cellular radio technology.

The term “seamless services” in the quotation above is ambiguous. When talking about FMC, the word “seamless” usually refers to “seamless handover,” which means that a call in progress can move from the mobile (cellular) network to the fixed network on the same phone without interruption, as described in one of the FMCA specification documents: Seamless is defined as there being no perceptible break in voice or data transmission due to handover (from the calling party or the called party”s perspective).

The term “seamless services” sometimes means service equivalence across any termination point, fixed or mobile, so for example, dialing plans are identical and no change in dialed digits is needed on a desk phone versus a mobile. A less ambiguous term for this might be “network agnostic services.”

* RDF: Resource Description Framework; see p. 186

The Resource Description Framework (RDF)는 웹 페이지의 the title, author, modification date, content, and copyright information 와 같은 웹 자원을 기술하기 위한 W3C 표준이다.

**RDF 정의

- RFD는 웹에서 데이터를 교환하기 위한 표준 모델이다.
- RDF는 비록 근간이 되는 구조식이 서로 다르다 하더라도 데이터 합병을 원활하게 하는 기능을 가지고 있으며, 특별히 모든 데이터 소비자의 변화를 요구하지 않고 시간이 지남에 따라 구조식의 발달을 지원한다.
- RDF는 링크의 양쪽 끝뿐만 아니라 사물간의 관계를 명명하기 위하여(대체로 이것을 “‘triple”이라고 말함) URIs를 사용할 수 있도록 웹의 링크 구조를 확장시킨다.
- 이 간단한 모델을 사용함으로써, 이것은 정형화 및 반-정형화된 데이터를 가지고도 서로 다른 어플 간에서 섞이고, 노출되고, 공유되도록 한다.

* PDF: Portable Document Format

PDF는 응용 소프트웨어, 하드웨어, 그리고 운영체계와는 독립적인 방식으로 도큐멘트를 표현하기 위하여 사용되는 파일 포맷이다. 각각의 PDF 파일은 그것을 디스플레이하는데 필요한 text, fonts, graphics, 그리고 기타 정보를 포함하고 있는 a fixed-layout flat document의 완전한 기술로 싸여 있다. 1991년에 Adobe Systems의 공동설립자인 John Warnock는 나중에 PDF로 진화한 “Camelot”이라는 시스템을 발표하였다.

사용자가 보거나 탐색하거나 프린트하거나 또는 다른 사람에게 전달할 수 있도록 만들어진 이미지로서 프린트 출력된 문서의 모든 요소를 갖추고 있는 파일 형식이다(Adobe Reader)

* TIFF(Tag Image File Format)

TIFF는 graphic artists, the publishing industry, and both amateur and professional photographers in general 간에서 인기 있는 raster graphics images를 저장하기 위한 컴퓨터 파일 포맷이다. 이 포맷은 원래 Aldus 사에서 데스크 탑 출판용으로 만들었지만, 2009년부터 에 Adobe Systems에서 관리하고 있다.

이미지 저장 포맷으로 사용자가 고쳐서 쓸 수 있는 유연성이 특징이며, 1980년대 스캐너 제조사들이 일반적인 파일 형식을 사용하기 위하여 개발되었으며, 확장자는 .tiff 나 .tif 이다; 그리고 파일 포맷의 유형은 raster image(bit map style)이다.

* library 2.0

Library 2.0은 서비스가 이용자에게 전달되어지는 방식에 있어서 도서관계의 전환을 반영하고 있는 현대화된 도서관 서비스의 형태에 대하여 완만하게 정의한 모델이다. 이것의 중점 사항은 콘텐트와 커뮤니티의 참여와 변화가 이용자 중심으로 이루어져야 한다는 것이다. Library 2.0의 개념은 Business 2.0과 Web 2.0의 개념에서부터 온 것이며 몇 가지 동일한 기본적 철학을 가지고 있다. 이것은 OPAC 시스템의 사용과 이용자로부터 도서관으로 되돌아오는 정보량의 증대와 같은 온라인 서비스를 포함,하고 있다.

** Overview

The term "Library 2.0" was coined by **Michael Casey** on his blog LibraryCrunch as a direct spin-off of the terms Business 2.0 and Web 2.0. Casey suggested that libraries, especially public libraries, are at a crossroads where many of the elements of Web 2.0 have applicable value within the library community, both in technology-driven services and in non-technology based services. In particular, he described the need for libraries to adopt a strategy for constant change while promoting a participatory role for library users.

With Library 2.0, library services are frequently evaluated and updated to meet the changing needs of library users. Library 2.0 also calls for libraries **to encourage user participation and feedback in the development and maintenance of library services.** The active and empowered library user is a significant component of Library 2.0. With information and ideas flowing in both directions – from the library to the user and from the user to the library – library services have the ability to evolve and improve on a constant and rapid basis. The user is participant, co-creator, builder and consultant – whether the product is virtual or physical.

** Key principles

Browser + Web 2.0 Applications + Connectivity = Full-featured OPAC
Harness the library user in both design and implementation of services
Library users should be able to craft and modify library provided services
Harvest and integrate ideas and products from peripheral fields into library

service models
Continue to examine and improve services and be willing to replace them at any time with newer and better services.

라이브러리 2.0은 도서관이 오랫동안 추구해 온 이용자 중심의 서비스 제공을 위해서 참여, 공유, 개방, 소통을 모토로 하는 웹 2.0의 개념과 기술을 도서관에 접목한 개념이다. 2005년 Micheal Casey에 의해 처음으로 언급되었으며 이용자 위주의 서비스로 변화하고 있는 도서관의 트렌드를 반영하고 있다고 할 수 있다.

장점:
1. 정보 공유의 플랫폼으로서의 도서관
2. 학술,연구 커뮤니티의 중심으로서의 도서관
3. 개인별 맞춤 다이어리로서의 도서관: SDI서비스, AJAX(Asynchronous JavaScript and XML, 아이아스: 대화식 웹 애플리케이션의 제작을 위해 아래와 같은 조합을 이용하는 웹 개발 기법이다; 표현정보를 위한 HTML or XHTML과 CSS, 동적 화면 출력 및 표시 정보와의 상호작용을 위한 DOM, JavaScript, 웹서버와 비동기적으로 데이터를 교환하고 조작하기 위한 XML, XSLT, XMLHttpRequest)나 RSS feed.
4. 이용자와 소통하는 도서관: 온라인 참고서비스(실시간 메신저, 게시판, 이메일, 전화, SMS, PDA 등 다양한 커뮤니케이션 매체를 통해 가능)
5. 정보 활용 능력을 교육하는 도서관: 온라인 튜토리얼을 통해 이용자의 정보활용능력을 향상.

p. lxxiv

*** MODS(Metadata Object Description Schema, Library of Congress)**

The Library of Congress' Network Development 그리고 MARC Standards Office, 에서 2002년에 다양한 목적 그리고 특히 도서관 어플용으로 사용할 수도 있는 하나의 서지 요소 세트인 MODS를 개발하였다. 하나의 XML 구조식으로 이것은 기존의 MARC 21 레코드로부터 선택된 데이터의 전달 뿐만 아니라 본래의 자원 기술 레코드를 만들 수 있도록 하기 위한 것이다. 이것은 어떤 경우에는 MARC 21 서지 포맷에서 나온 요소들을 재집단화기 위하여MARC 필드의 하위 세트를 포함하고 있으며 수자적인 것보다는 언어-의존형 태그들을 사용하고 있다.

2009년 6월 현재 이 구조식은 버전 3.3이며, W3C의 XML 스키마 언어를 사용하여 표현하고 있다. 이 기준은 the Network Development and MARC Standards Office of the Library of Congress의 지원을 받아 the MODS Editorial Committee에서 관리하고 있다.

*** MOA2 DTD;** METS로 대체됨.

The Making of America II Testbed Project는 디지털 객체 관리에 필요한 메타데이

터 요소들의 잘 정리된 세트의 개발을 통해 디지털 도서관 객체용 메타데이타와 관련된 문제들을 해결하려고 시도하였다. 이 메타데이터 세트는 XML 도큐멘트 유형의 정의, MOA2 DTD를 통해 그것의 기술적 표현을 완성하였다. 디지털 객체 메타데이터의 토론에 관한 훌륭한 출발점을 제공하는 동안, MOA2 DTD는 단지 제한된 범위의 디지털 객체(diaries, still images, ledgers, and letterpress books)만을 코드화하도록 디자인되었다. DTD가 보다 널리 응용됨으로써, 그것의 본래의 디자인이 가지고 있는 문제가 더욱 분명하게 나타났다. 말 그대로 DTD는 기술 메타데이터의 코딩을 위한 올바른 준비가 부족하다. 단지 협의적으로 text-와 imaged-의존형 자원을 위한 기술적 메타데이터를 지원한다. 그러므로 audio, video, and other time dependent media에 대한 어떠한 지원도 제공하지 않으며 단지 최소한의 내부적 그리고 외부적 링크 기능만을 제공한다.

이러한 단점에도 불구하고, MOA2 DTD는 디지털 도서관 객체의 기술과 관리를 위한 표준화된 데이터 요소 세트와 그 정보를 표현하기 위한 기술적 메카니즘 양쪽 다를 발전시키는데 커다란 역할을 하였다. 이것의 워크숍에서 MOA2 DTD를 구축하고 확대시킬 기회를 제공하게 될 것이며, 보다 다양한 디지털 도서관 객체와 운영을 지원하고, 미래에 그 DTD를 발전시키고 유지하기 위하여 추가로 취해야할 조치에 대하여 논의하게 될 것이다.

MOA2 DTD 개정판에 영향을 끼친 관련 표준은 다음과 같다: Dublin Core, SMARC, Encoded Archival Description, Indecs Metadata Framework, VRA Core, NISO Technical Metadata for Digital Still Images, Library of Congress audio/visual technical metadata, National Library of Australia Preservation Metadata for Digital Collections, Resource Description Framework, Synchronized Multimedia Integration Language, MPEG-7

1) **Encoded Archival Description** (EAD) is an XML standard for encoding archival finding aids, maintained by the Technical Subcommittee for Encoded Archival Description of the Society of American Archivists, in partnership with the Library of Congress. *See p.78*

2) **indecs** (an acronym of "interoperability of data in e-commerce systems"; written in lower case) was a project part funded by the European Community Info 2000 initiative and by several organisations representing the music, rights, text publishing, authors, library and other sectors in 1998-2000, which has since been used in a number of metadata activities. A final report and related documents were published; the **indecs Metadata Framework** document "Principles, model and data dictionary" is a concise summary.

indecs provided an analysis of the requirements for metadata for e-commerce of content (intellectual property) in the network environment, focussing on semantic interoperability. Semantic interoperability deals with the question of how one computer system knows what the terms from another computer system mean (e.g. if A says "owner" and B says "owner", are they referring to the same thing? If A says "released" and B says "disseminated", do they mean different things?).

Use of indecs[edit]The indecs Framework does not presuppose any specific business model or legal framework; it can be used to describe transactions of copyrighted, open source, or freely available material.

3) **Synchronized Multimedia Integration Language** (SMIL (/'smaɪl/)) is a World Wide Web Consortium recommended Extensible Markup Language (XML) markup language to describe multimedia presentations. It defines markup for timing, layout, animations, visual transitions, and media embedding, among other things. SMIL allows presenting media items such as text, images, video, audio, links to other SMIL presentations, and files from multiple web servers. SMIL markup is written in XML, and has similarities

to HTML.

4) **MPEG-7** is a multimedia content description standard. It was standardized in ISO/IEC 15938 (Multimedia content description interface). This description will be associated with the content itself, to allow fast and efficient searching for material that is of interest to the user. MPEG-7 is formally called Multimedia Content Description Interface. Thus, it is not a standard which deals with the actual encoding of moving pictures and audio, like MPEG-1, MPEG-2 and MPEG-4. It uses XML to store metadata, and can be attached to timecode in order to tag particular events, or synchronise lyrics to a song, for example.

p. lxxviii

*** EAD(Encoded Archival Description)**

EAD DTD의 개발은 1993년 the University of California, Berkeley, Library에서 프로젝트를 시작하였다. 이 버클리 프로젝트의 목표는 자신들의 소장 자료의 이용을 지원하기 위하여 archives, libraries, museums, and manuscript repositories에서 만든 inventories, registers, indexes, and other documents와 같은 기계가독형 찾기 도구용으로 비독점권적인 코딩 표준을 개발하기 위한 열망과 가능성을 조사하는 것이었다.

이 프로젝트의 감독은 소장자료의 정보에 접근하는데 있어서 네트워크의 역할이 늘어났다는 것을 인정하고 전통적인 MARC에서 제공하는 것 이상의 정보를 포함하려고 노력하였다. EAD DTD의 개발은 여러 명의 전문가가 초기부터 공동로 참여한 벤처였다. 이 버클리 프로젝츠의 주요 조사자인 Daniel Pitti는 다음과 같은 기준을 가지고 코딩 표준의 필요조건을 개발하였다:

1) 고문서찾기도구로 발견된 포괄적이고 상호연관된 기술정보의 표현 능력;
2) 기술 수준간에 존재하는 계층적 관계의 보존 능력;
3) 하나의 계층적 단계에서부터 다른 단계로 유전되는 기술정보의 표현 능력;
4) 계층적 정보구조내에서 이동할 수 있는 능력;
5) 요소-맞춤형 색인과 검색의 지원.

p. lxxxv

*** DOI:** A digital object identifier

DOI는 전자 도큐멘트와 같은 객체를 유일하게 식별하기 위하여 사용되는 문자열(디지털 식별자)이다. 객체에 대한 메타데이터는 DOI 이름과 결합되어 저장되며 이 메타데이터에는 그 객체를 찾을 수 있는 URL과 같은 위치가 포함될 수 있다. 도큐멘트용 DOI는 항구적인 반면에 그것의 위치와 기타 메타데이터는 변할 수 있다. 그것의 DOI로 온라인 도큐멘트를 말하는 것은 단지 그것의 URL로 말하는 것보다 더욱 안정된 링킹을 제공하는데, 왜냐하

면 만일 그것의 URL이 변한다면, 출판사는 새로운 URL로 링크하도록 그 DOI용 메타데이터만을 갱신하면 되기 때문이다.

** DOI names

DOI name은 문자열 형태를 취하며, 슬래쉬로 구분되는 접두사와 접미사, 두 부분으로 나뉘어져 있다. 접두사는 그 이름의 등록자를 나타내며, 접미사는 그 등록자에 의해 선택되어 그 DOI와 결합된 특별한 객체를 나타낸다. 대부분의 합법적인 Unicode 문자들은 이 문자열에 포함될 수 있으며 대소문자를 구별하여 해석된다.

예를 들어, DOI name 10.1000/182에서 접두어는 10.1000이고, 접미사는 182이다. 접두사의 "10."은 그 DOI의 등록지를 나타내며, 접두사에 있는 문자 1000은 등록자를 나타낸다; 이러한 경우에 있어서 그 등록자는 the International DOI Foundation 그 자체이다. 182는 접미사이거나 아이템 ID이며 단일 객체를 나타낸다(이 경우에서는 the DOI Handbook의 최신 버전이다). DOI name을 이용하는 인용에서는 doi:10.1000/182처럼 인쇄되어야 한다. 인용이 하이퍼링크일 경우에는 그것의 "doi:" 접두사를 생략하고, 그 DOI name에 "http://dx.doi.org/"를 대체하여 하나의 URL처럼 링크로 사용하길 권장하고 있다. 예를 들어, DOI name doi:10.1000/182는 http://dx.doi.org/10.1000/182처럼 링크된다. 이 URL에서는 링크된 아이템의 정확한 온라인 위치로 웹 접근의 방향을 조정하는 HTTP proxy server의 위치를 제공한다.

DOI name은 전자 또는 물리적 형태 둘 모두에 포함되는 창조적 작품(such as texts, images, audio or video items, and software) 그리고 퍼포먼스, 교신 대상 등의 추상적 작품을 구분할 수 있다. 이 name은 세부적인 내용이 변화하는 단계에서 객체들을 표현할 수 있다: 그러므로 DOI name은 a journal, an individual issue of a journal, an individual article in the journal, 또는 a single table in that article을 구별할 수 있다. 세부내용의 수준에 대한 선택은 배정자에게 달려있지만, DOI 시스템에서, 이것은 the indecs Content Model에 근거한 데이터 사전을 이용하므로 하나의 DOI name에 결합되어 있는 메타데이터의 일부로 분명히 선언되어야 한다.

** Comparison with other identifier schemes

DOI name은 URL과 같이 자료에 대한 일반적인 인터넷 포인터와는 다르다. URL은 첫 번째 부류의 실체처럼 객체를 나타내지만 그 객체가 위치하고 있는 장소는 간단하지가 않다. DOI name은 URN의 개념을 실행하고 있으며 그것에 데이터 모델과 사회적 하부조직을 추가하고 있다.

DOI name 또한 ISBN, ISRC, 등과 같은 표준 식별자 레지트리와도 다르다. 식별자 레지스트리의 목적은 특정한 집단의 식별자를 관리하는 것이지만, DOI 시스템의 기본 목적은 식별자 집단을 가지고 활동하게 하고 상호작용하도록 만드는 것이다. 그러므로 그러한 집단이 많은 서로 다른 통제된 집단에 발생하는 식별자를 포함할 수 있다.

1) The International Standard Recording Code (**ISRC**) is an international standard code for uniquely identifying sound recordings and music video recordings.

DOI 시스템은 관련된 현재의 데이터에 대하여 항구적이고 어의적으로 상호 통할 수 있

는 해답을 제공하며 예를 들어 public citation 또는 managing content of value인 issuing assigner의 직접적인 통제를 벗어나 있는 서비스에서 사용할 자료에 가장 적합하다. 이것은 사회적 및 기술적 하부구조를 제공하는 관리되는 레지스트리를 사용한다. 이것은 식별자 또는 서비스의 제공을 위한 어떤 특수한 사업 모델을 가정하지 않으며, 다른 기존의 서비스를 지정된 방식으로 그것에 링크시킬 수 있다. 식별자를 항구적으로 만드는 여러 가지 방법이 제안되고 있다.

항구적 식별자 방법의 비교는 어려운데 그 이유는 그것들이 모두 동일 일을 하지 않기 때문이다. 부정확하게 "identifiers"로 한 세트의 스킴을 말하는 것이 그것들을 쉽게 비교할 수 있다는 의미는 아니다. 다른 "identifier systems"는 누구에게나 새로운 경우의 설치를 허용하는 사용하기 편한 labeling mechanism을 제공함으로써 기입하는데 작은 방해만을 받는 기술일 수도 있지만(예, Persistent Uniform Resource Locator (PURL), URLs, Globally Unique Identifiers (GUIDs), etc.), 레지스트리-통제 스킴의 기능성 중에서 어떤 것이 부족할 수도 있으며, 대체로 통제 스킴에 메타데이터를 동반하는데 문제점이 있다.

DOI 시스템은 이러한 방식을 취하지 않으며 그러한 식별자 스킴과 직접적으로 비교해서도 안된다. 추가된 기능을 갖춘 기술을 사용하는 다양한 어플들은 특수한 분야(예, ARK)를 위해 DOI 시스템에 의해 제공되는 몇 가지의 특징을 충족하도록 마련되어야 한다.

1) An **Archival Resource Key (ARK)** is a Uniform Resource Locator (URL) that is a multi-purpose identifier for information objects of any type. An ARK contains the label **ark:** after the URL's hostname, which sets the expectation that, when submitted to a web browser, the URL terminated by '?' returns a brief metadata record, and the URL terminated by '??' returns metadata that includes a commitment statement from the current service provider. The ARK and its inflections ('?' and '??') gain access to three facets of a provider's ability to provide persistence.

DOI name은 객체의 위치에 의존하지 않으며 이 때문에 URN이나 PURL과 비슷하지만 일반적인 URL과는 다르다. URLs는 비록 동일한 도큐멘트가 두 개의 다른 장소에 있어 두 개의 URLs을 가지고 있다하더라도 인터넷의 도큐먼트를 위한 대체 식별자(URIs처럼 보다 나은 특징을 갖는)로 종종 사용되기도 한다. 대조적으로 DOI name가 같은 항구적 식별자들은 일급 엔티티로 객체를 나타낸다: 동일한 객체의 두 가지 경우는 동일한 DOI name을 가질 수 있다.

** Resolution

DOI name resolution은 Handle System을 통해 제공되며, DOI name을 만나는 어떠한 이용자도 무료로 이용할 수 있다. 리솔루션은 하나의 DOI name에서부터, URLs 객체의 경우들을 표현하는 URLs, 이-메일과 같은 서비스, 또는 하나 이상의 메타데이터 아이템들 같이 하나이상의 타이프된 데이터 조각으로 이용자의 방향을 수정해 준다. Handle System에서 DOI name은 하나의 핸들이며 그렇게 때문에 그것에 할당된 한 세트의 값을 가지고 있고 한 그룹의 필드로 구성된 하나의 레코드처럼 생각할 수도 있다. 각 핸들의 값은 데이터의 구문과 어의를 정의하고 있는 그것의 "<type>" 필드에 지정된 데이터 유형만을 가져야만 한다.

1) The **Handle System** is a technology specification for assigning, managing, and resolving persistent identifiers for digital objects and other resources on the Internet. The protocols specified enable a

distributed computer system to store identifiers (names, or handles), of digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources. That information can be changed as needed to reflect the current state and/or location of the identified resource without changing the handle.

DOI name을 리솔브하기 위하여, DOI resolver(예: www.doi.org)에 그것을 입력할 수도 있고 다음과 같은 문자열 http://dx.doi.org/를 사용하여 DOI name을 표현할 수도 있다. 예를 들어, DOI name 10.1000/182는 어드레스 "http://dx.doi.org/10.1000/182"에서 해결될 수 있다. 웹 페이지나 기타 하이퍼텍스트 도큐멘트들은 이러한 형태 속에 하이퍼텍스트 링크들을 포함할 수 있다. 어떤 브라우저들은 추가기능(add-on)을 갖추고 DOI(또는 다른 핸들)의 직접적인 해결을 허용하고 있다(예: CNRI Handle Extension for Firefox). The CNRI Handle Extension for Firefox는 그 브라우저가 고유한 Handle System 프로토콜을 사용하는 hdl:4263537/4000 or doi:10.1000/1와 같은 핸들이나 DOI RURIs에 접근할 수 있도록 하고 있다. 이것은 web-to-handle 프록시 서버에 관한 레퍼런스를 native resolution으로 심지어 대체할 것이다.

1) The **Corporation for National Research Initiatives (CNRI)**, based in Reston, Virginia, is a non-profit organization founded in 1986 by Robert E. Kahn as an "activities center around strategic development of network-based information technologies", including the National Information Infrastructure in the United States. CNRI publishes D-Lib Magazine, a journal of digital library research and development. It also develops the Handle System for managing and locating digital information[citation needed]. CNRI formerly operated the Secretariat of the Internet Engineering Task Force

** Organizational structure

IDF는 비영리기관으로 1998년에 수립되었으며 DOI 시스템의 총괄기관이다. 이곳에서 DOI 시스템과 관련된 모든 지적 재산권을 보호하고 있으며, 일반적운 운영상의 기능을 관리하고 있고 DOI 시스템의 개발과 발전을 지원하고 있다. IDF에 의해 지명된 등록기관은 DOI 등록자에게 서비스를 제공한다: 그것들은 DOI 접두어를 할당하고, DOI names을 등록하며 등록자로 하여금 메타데이터와 상태 데이터를 선언하고 유지하는데 필요한 하부구조를 제공한다. 등록기관들은 또한 IDF와 협력하여 전체적인 DOI 시스템를 개발하는데 있어서 활발하게 DOI 시스템의 확산을 촉진시키고 그것들의 특수한 이용자 집단을 위하여 서비스를 제공할 것으로 기대되고 있다. 최신의 RAs의 리스트는 the International DOI Foundation에 의해 관리되고 있다.

등록기관은 일반적으로 새로운 DOI name의 할당은 무료로 이루어진다; 이러한 비용의 일부는 IDF를 지원하는데 사용된다. 전반적으로 DOI 시스템은 IDF를 통하여 비영리적인 비용보존방식으로 운영된다.

** Standardization

DOI 시스템은 IOS에서 개발한 국제적 표준이며, 최종 표준은 2012년 4월 23일 발표되었다. DOI는 infoURI 스펙(IETE RFC 4452)인 "Public Namespaces에 식별자를 갖춘 정보자산용 "info" URI 스킴"에 따라 등록된 URI이다; info:doi/ 는 DOI의 infoURI Namespace이다. DOI 구문법은 먼저 2000년에 표준화된 NISO 표준이다; ANSI/NISO Z39.84-2005 Syntax for the Digital Object Identifier.

1) In computer science, **info:** is a Uniform Resource Identifier (URI) scheme for information assets with identifiers in public namespaces that allows legacy namespaces such as Library of Congress Identifiers and Digital object identifiers to be represented as URIs. It acts as a bridging mechanism for older information identifiers to be used in the more generalised and standard URI allocation.

2) In general, a **namespace** is a container for a set of identifiers (also known as symbols, names). Namespaces provide a level of indirection to specific identifiers, thus making it possible to distinguish between identifiers with the same exact name. For example, a surname could be thought of as a namespace that makes it possible to distinguish people who have the same first name. In computer programming, namespaces are typically employed for the purpose of grouping symbols and identifiers around a particular functionality.

p. civ

*n-gram 방식

전산언어학 및 확률학에서, n-gram이란 텍스트나 언어의 지정된 순서로부터 이루어진 n개의 아이템의 연속적인 순서를 말한다. 의문속의 아이템은 음소, 음절, 글자, 단어, 또는 어플에 따른 base pairs(염기쌍, 이중사슬)일 수 있다. n-grams는 한나의 텍스트나 언어의 집성으로부터 수집된다. size 1의 n- gram 은 “unigram”; size 2는 “bigram”(또는 거의 사용되지는 않지만 “digram”; size 3는 “trigram”이라고 부른다. 보다 큰 사이즈는 때때로 그 n의 값에 의해 이름이 결정 된다; 예를 들어 “four-gram”, “five-gram”, 등등.

n-grams for approximate matching

n-grams can also be used for efficient approximate matching. By converting a sequence of items to a set of n-grams, it can be embedded in a vector space, thus allowing the sequence to be compared to other sequences in an efficient manner. For example, if we convert strings with only letters in the English alphabet into 3-grams, we get a -dimensional space (the first dimension measures the number of occurrences of "aaa", the second "aab", and so forth for all possible combinations of three letters). Using this representation, we lose information about the string. For example, both the strings "abc" and "bca" give rise to exactly the same 2-gram "bc" (although {"ab", "bc"} is clearly not the same as {"bc", "ca"}). However, we know empirically that if two strings of real text have a similar vector representation (as measured by cosine distance) then they are likely to be similar. Other metrics have also been applied to vectors of n-grams with varying, sometimes better, results. For example z-scores have been used to compare documents by examining how many standard deviations each n-gram differs from its mean occurrence in a large collection, or text corpus, of documents (which form the "background" vector). In the event of small counts, the g-score may give better results for comparing alternative models.

It is also possible to take a more principled approach to the statistics of n-grams, modeling similarity as the likelihood that two strings came from the same source directly in terms of a problem in Bayesian inference.

n-gram-based searching can also be used for *plagiarism detection*.

p. cxxi

* Word lists by frequency

빈도에 의한 단어 리스트는 어휘 수집을 목적으로 어떤 차원이나 서열 리스트처럼 특정한 텍스트 집단에서 발생한 빈도를 가지고 언어의 단어들을 집단화한 리스트이다. 빈도에 의한 단어 리스트는 학습자들이 자신들의 어휘 학습 노력으로 가장 좋은 결과를 확실하게 얻을 수 있다는 합리적 근거를 제공한다. 그러나 이것은 주로 course writer를 목적으로 한 것이며 학습자를 직접적으로 대상으로 한 것은 아니다. 이것의 몇 가지 주요한 단점은 the corpus content, the corpus register, and the definition of "word"이다. 단어계산이 수천년된 것이지만 아직 20세기 중반까지 수작업으로 이루어지는 대규모의 분석에서 사용하고 있으며, 영화 부제목(SUBTLEX megastudy)과 같은 대규모의 집단의 자연어 전자처리에서는 그 연구가 활성화되고 있다.

전산언어학에서, frequency list는 빈도를 가지고 단어들을 분류한 리스트이다. 여기서 빈도는 대체로 특정한 집단에서 발생한 수를 의미하며, 그것으로부터 의미가 부족하지만 서열이 발생한다.

* Ziph의 법칙:

지프의 법칙은 자연어로된 어떤 집단이 주어졌을 때, 그 속에 있는 어떤 단어의 빈도는 빈도 테이블에 있는 그것의 서열과는 반비례한다는 것이다. 그러므로 가장 높은 빈도의 단어는 두 번째로 가장 높은 단어의 빈도보다 약 2배 정도 많고, 세 번째 빈도높은 단어보다는 3개가 높다는 것이다. 예를 들어, “the Brown Corpus of American English” 텍스트에서, 단어 “the”는 가장 빈도수가 높은 단어이며 그 자체가 모든 단어 빈도의 약 7%를 차지하고 있다. 지프의 법칙에 따라, 두 번째 빈도의 단어 “of”는 약 3.5%이다.

** Ziph의 제 1 법칙

텍스트의 단어들을 출현빈도순으로 배열하여 순위를 매기면 출현빈도와 순위를 곱한 값이 일정하다는 법칙이며, 고빈도 단어에만 적용되며, 저빈도 단어에는 적용되지 않는다는 단점이 있다.

예

단어	순위(r)	출현빈도(f)	값(r x f)
the	1	301	301
of	2	152	304
for	3	108	324
to	4	81	324
and	5	68	340

** Ziph의 제 2 법칙: Booth(A.D. Booth)가 수정한 법칙이다.

텍스트에 한번만 출현한 단어의 수와 n번 출현한 단어의 수의 비율은 텍스트와 상관없이 일정하다는 법칙이며, 저빈도 단어에 적용된다.

공식

$I_n/I_1 = 3/4n-1$

I_n: 텍스트에 n번 나타난 단어의 수

I_1 : 텍스트에 1번 나타난 단어의 수

** 최소노력의 법칙(Principle of Least Effort)

최소노력의 원칙은 진화생물학에서부터 웹페이지 디자인까지 다양한 분야에서 다루어지는 포괄적 이론이다. 이것은 동물, 사람, 심지어 잘 디자인된 기계조차도 최소한의 저항이나 노력의 통로를 자연스럽게 선택한다는 것을 가정한다. 문헌정보학과 관련해서, 이 원칙은 정보입수 이용자는 이용 가능하며 최소한의 정확성을 사용하는 방법으로 가장 편리한 탐색 방법을 사용하려는 경향이 있다는 것이다. 정보입수행위는 최소한의 수용 가능한 결과를 얻자마자 중지된다.

* Bradford's law

1934년 Samuel C. Bradford에 의해 처음으로 주장된 하나의 패턴이며, 이것은 과학지에서 참고문헌의 수가 탐색을 확대해가면 그 결과가 기하급수적으로 줄어든다는 것이다. 한 가지 공식은 만일 어떤 분야의 학술지가 그것의 기사 수를 가지고 각각이 모든 기사의 약 1/3을 가지고 있는 3집단으로 분류해 보면, 각 그룹에 있는 학술지의 수는 그 비율이 $1:n:n^2$ 이 된다.

많은 학분 분야에서 이러한 패턴을 Pareto 분산이라 부른다. 예를 들어, 한 연구자가 자신의 연구주제를 위하여 5가지의 핵심 저널을 갖고 있다고 가정해 보자. 한단에 이들 저널에 12개의 기사가 관심대상이라고 가정해 보자. 추가로 이 연구자가 또다른 12개의 관심 기사를 찾기 위해서 그는 추가로 10개의 저널을 봐야할 것이다. 그러므로 이 연구자의 브래포드 승수 bm(Bradford multiplier) 은 2이다: 다시 말해서, 10/5이다. 각각의 새로운 12개의 기사용으로 이 연구자는 bm 배만큼 많은 저널을 봐야할 것이다. 5, 10, 20, 40, 등의 저널을 본 다음에, 대부분의 연구자들은 신속하게 깨닫는다: “there is little point in looking further.”

The result of this is pressure on scientists to publish in the best journals, and pressure on universities to ensure access to that core set of journals. On the other hand, the set of "core journals" may vary more or less strongly with the individual researchers, and even more strongly along schools-of-thought divides. There is also a danger of over-representing majority views if journals are selected in this fashion. Bradford's law is also known as Bradford's law of scattering and as the Bradford distribution. This law or distribution in bibliometrics can be applied to the World Wide Web.

* Lotka's law

지프 법칙을 특별하게 응용한 여러 가지 중의 하나이며, 이것은 특정분야 저자의 출판 빈도를 묘사한다. 이것이 주장하는 것은 n 편을 투고한 저자의 수는 한 편을 투고한 저자의 $1/n^a$이다. 이 공식에서 a는 거의 항상 2를 나타낸다. 보다 평범하게 말해서, 어떤 수의 기사를 출판하는 저자의 수는 단일 기사를 출판하는 저자의 수에 대한 고정비율로 나타난다. 출판된 기사의 수가 늘어남으로서, 많이 출판하는 저자는 그 빈도가 점점 떨어진다. 특정한 기간내에 2개의 기사를 출한하는 저자의 수는 단일 출판 저자의 수의 1/4는 이다. 3개의 기사를 쓴 저자는 1/9이고, 4개를 쓴 저자는 1/16이다.

*** The Pareto principle**

이 원칙은 80-20 규칙으로도 알려져 있으며, 많은 사건에 있어서 결과의 약 80%가 약 20%의 원인들로부터 발생한다는 것이다. 1906년 이태리 경제학자 Vilfredo Pareto는 이태리 국토의 80%를 인구의 약 20%가 소유하고 있다는 것을 관찰하였으며, 또한 자신의 정원에서 20%의 콩깍지가 80%의 콩알을 포함하고 있다는 것을 관찰하여 이 원칙을 발전시켰다.

The Pareto principle (also known as the 80–20 rule, the law of the vital few, and the principle of factor sparsity) states that, for many events, roughly 80% of the effects come from 20% of the causes.

*** Luhn의 가설**

KWIC 색인을 1959년에 제안; 단어의 출현빈도는 텍스트의 내용을 나타내는 주제어로 중요성을 측정하는 기준이 된다.; 통계적 기법의 자동색인에 있어 가장 중요한 가정이다:

" 고빈도의 단어는 너무 일반적인 단어로서 주제어로 가치가 없어 정확률이 떨어지고, 저빈도의 단어는 주제어로서의 의미가 없으며, 재현율을 낮추므로, 중간빈도의 단어를 색인어로 선정하여야 한다."

Hans Peter Luhn (July 1, 1896 – August 19, 1964) was a computer scientist for IBM, and creator of the Luhn algorithm and KWIC (Key Words In Context) indexing.

Two of Luhn's greatest achievements are the idea for an SDI system and the KWIC method of indexing.

KWIC

KWIC is an acronym for Key Word In Context, the most common format for concordance(용어색인) lines. The term KWIC was first coined by Hans Peter Luhn. The system was based on a concept called keyword in titles which was first proposed for Manchester libraries in 1864 by Andrea Crestadoro.

A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index. It was a useful indexing method for technical manuals

before computerized full text search became common.

For example, the title statement of this article and the Wikipedia slogan would appear as follows in a KWIC index. A KWIC index usually uses a wide layout to allow the display of maximum 'in context' information (not shown in the following example).

KWIC is an **acronym** for Key Word In Context, ...	page 1
... Key Word In Context, the most **common** format for concordance lines.	page 1
... the most common format for **concordance** lines.	page 1
... is an acronym for Key Word In **Context**, the most common format ...	page 1
Wikipedia, The Free **Encyclopedia**	page 0
... In Context, the most common **format** for concordance lines.	page 1
Wikipedia, The **Free** Encyclopedia	page 0
KWIC is an acronym for **Key** Word In Context, the most ...	page 1
KWIC is an acronym for Key Word ...	page 1
common format for concordance **lines**.	page 1
... for Key Word In Context, the **most** common format for concordance ...	page 1
Wikipedia, The Free Encyclopedia	page 0
KWIC is an acronym for Key **Word** In Context, the most common ...	page 1

The term permuted index(순열색인) is another name for a KWIC index, referring to the fact that it indexes all cyclic permutations of the headings. Books composed of many short sections with their own descriptive headings, most notably collections of manual pages, often ended with a permuted index section, allowing the reader to easily find a section by any word from its heading. This practice is no longer common.

1) A **concordance** is an alphabetical list of the principal words used in a book or body of work, with their immediate contexts. Because of the time, difficulty, and expense involved in creating a concordance in the pre-computer era, only works of special importance, such as the Vedas, Bible, Qur'an or the works of Shakespeare and other classical Latin and Greek authors, had concordances prepared for them.

2) **Text mining,** also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

A typical application is to scan a set of documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the

information extracted.

* Moore's law

무어의 법칙은 관찰법이며, 컴퓨터 하드웨어의 역사에서 집적회로에 있는 트랜지스터의 수가 매 2년마다 약 2배가 된다는 것이다.

Although this trend has continued for more than half a century, Moore's law should be considered an observation or conjecture and not a physical or natural law. Sources in 2005 expected it to continue until at least 2015 or 2020. However, the 2010 update to the International Technology Roadmap for Semiconductors predicts that growth will slow at the end of 2013, when transistor counts and densities are to double only every three years.

* 1% rule

인터넷 문화에서, 1%의 규칙은 인터넷 코뮤니티에 참가하는 것과 관련된 경험의 법칙이며, 웹사이트의 단지 1%의 이용자만이 활발하게 새로운 콘텐트를 만드는 반면에 참가자의 99%는 단지 이용만 한다(lurk)는 것이다.

A variant is the "90–9–1 principle" (sometimes also presented as the 89:10:1 ratio), which states that in a collaborative website such as a wiki, 90% of the participants of a community only view content, 9% of the participants edit content, and 1% of the participants actively create new content.

Both can be compared with the similar rules known to information science, such as the 80/20 rule known as the Pareto principle, that 20 percent of a group will produce 80 percent of the activity, however the activity may be defined.

p. cxxvi

(이미지자료의 내용기반색인을 위한 대표적인 기법)

* tree(data structure)

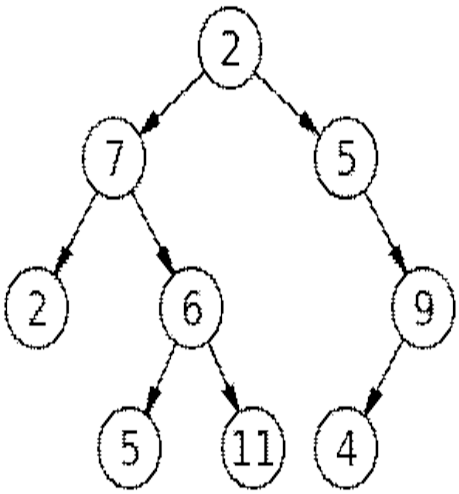
컴퓨터학에서, 트리는 abstract data type (ADT) 또는 한 세트의 링크된 노드들로 표현되면서 a root value 와 subtrees of children을 갖고 있는 계층적 트리 구조를 모방한 이 ADT를 실행하는 data structure에 폭넓게 사용된다.

트리 데이터 구조는 부분적으로 보면 하나의 노드 집단(루트 노드에서 출발하는)처럼 귀납적으로 정의될 수 있으며, 각 노드는 값과 더불어 노드들(the children)에 대한 레퍼런스의 리스트로 이루어진 데이터 구조이다. 그리고 이 노드는 레퍼런스의 중복이 없으며 어떤 것도 루트로 연결되지 않는다는 한계를 가지고 있다.

대안적으로 트리는 하나의 질서있는 트리처럼 전체적으로 봐서 추상적으로 정의될 수 있다. 이 경우에 각각의 노드에는 할당된 값을 갖는다. 이 같은 견해 둘 다 유용하다: 트리

는 수학적으로 전체를 분석할 수 있는 반면에, 실재적으로 데이터의 구조처럼 표현될 때, 그것은 누구나 diagraph를 표현할 수 있는 것처럼 노드의 리스트와 노드간의 경계에 있는 인접 리스트처럼이라기 보다는 노드에 의해 대체로 독립적으로 표현되거나 작업한다. 예를 들어, 전체적으로 트리를 살펴본 다음에, 누구나 특정한 노드의 “parent node”를 말한 수 있지만 일반적으로 하나의 데이터 구조처럼 특정한 노드는 단지 그것의 children의 리스트만을 포함하고 있지 그것의 parent에 대한 레퍼런스는 포함하고 있지 않다.

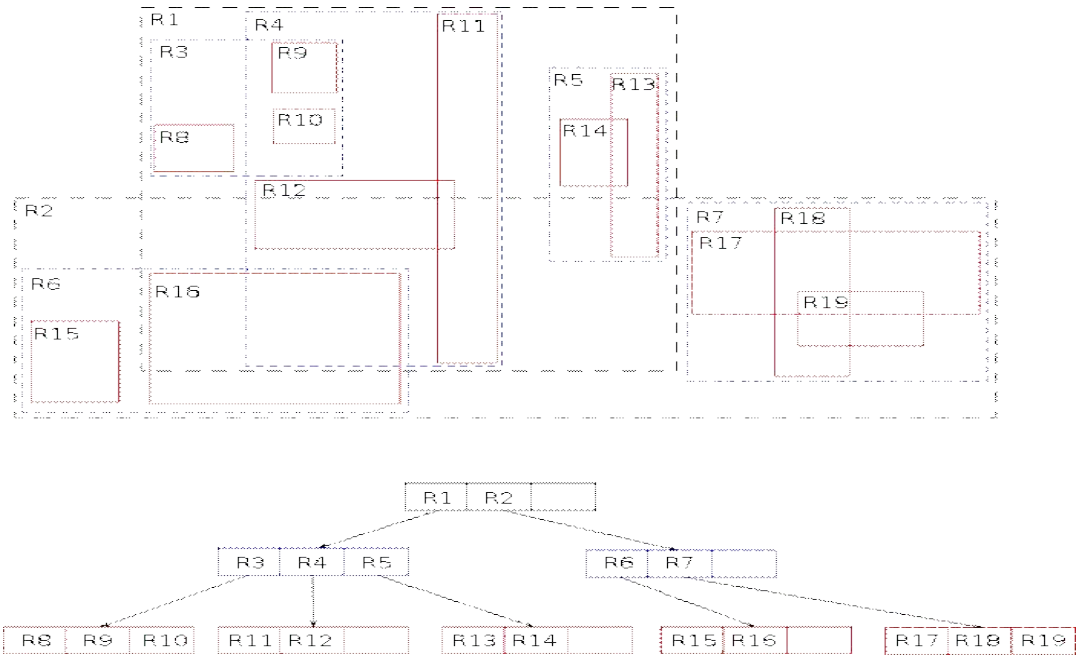
Binary tree



*** R-tree 기법:** 공간 데이터나 다차원 데이터의 효율적인 질의 처리를 위한 인덱스 구조.

R-trees는 공간적 접근방법용인 트리 데이터 구조이다. 다시 말해서, geographical coordinates, rectangles or polygons과 같은 다차원 정보를 색인하기 위하여 사용된다. R-tree의 일반적인 실생활 용도는 레스토랑 위치나 전형적인 지도로 이루어진 streets, buildings, outlines of lakes, coastlines, etc.의 다각형과 같은 공간적 객체를 저장하여, 예를 들어, “현 위치에서 2km내에 있는 모든 박물관을 찾아라”라는 질문에 신속하게 응답하는 것이다.

R-tree



*** B-tree**

컴퓨터학에서, B-tree는 분류된 데이터를 보관하고, logarithmic time으로 searches, sequential access, insertions, and deletions을 가능하는 트리 데이터 구조이다. B-tree는 하나의 노드가 2개이상의 children을 가질 수 있는 이진탐색 트리를 일반화한 것이다. 스스로 균형을 유지하는 이진 탐색 트리와 달리, B-tree는 커다란 블록의 데이터를 읽고 쓰는 시스템을 최적화시킨다. 일반적으로 이것은 데이터베이스 및 파일 시스템에서 사용한다.

For example, in a 2-3 B-tree (often simply referred to as a **2-3 tree**), each internal node may have only 2 or 3 child nodes. Each internal node of a B-tree will contain a number of keys. The keys act as separation values which divide its subtrees. For example, if an internal node has 3 child nodes (or subtrees) then it must have 2 keys: a_1 and a_2. All values in the leftmost subtree will be less than a_1, all values in the middle subtree will be between a_1 and a_2, and all values in the rightmost subtree will be greater than a_2.

1) In computer science, a **binary search tree (BST)**, sometimes also called an ordered or sorted binary tree, is a node-based binary tree data structure which has the following properties:

The left subtree of a node contains only nodes with keys less than the node's key.
The right subtree of a node contains only nodes with keys greater than the node's key.
The left and right subtree each must also be a binary search tree.
There must be no duplicate nodes.

Generally, the information represented by each node is a record rather than a single data element. However, for sequencing purposes, nodes are compared according to their keys rather than any part of their associated records.

The major advantage of binary search trees over other data structures is that the related sorting

algorithms and search algorithms such as in-order traversal can be very efficient.

Binary search trees are a fundamental data structure used to construct more abstract data structures such as sets, multisets, and associative arrays.

*** TV(Telescopic Vector)-tree 기법**

*** SS-tree 기법**

SS-트리(Similarity Search tree)는 유사도를 평가하는 척도를 사용하며 질의와 이미지 데이터의 유사성을 비교하여 유사도가 높은 이미지 데이터를 검색하는데 적합하도록 설계된 동적인 색인 구조이다. R-트리 계열의 색인 구조가 데이타 공간을 MBR(Minimum Rectangle Region)로 분할하는 반면, SS-트리는 구 영역(spherical region)으로 데이타 공간을 분할한다. 이는 유사성에 기반한 검색을 용이하게 하는 반면 겹침 영역이 많이 발생함으로써 검색성능을 저하시킨다.

*** X-tree 기법**

컴퓨터학에서, X-tree는 여러차원에 있는 데이터를 저장하기 위하여 사용된 R-tree를 근거로 하는 색인 트리 구조이다. 이것은 R-trees, R+-trees and R*-trees와 다른데, 그 이유는 고차원에서 점점 더 문제가 되는 서로 연결된 박스들에서 overlap의 예방을 강조하기 때문이다. 노드들이 중복을 예방하지 못해서 쪼개질 수 없는 경우에, 쪼개진 노드는 미루어져서 결과적으로 슈퍼 노드가 될 것이다. 극단적인 경우에, 어떤 다른 데이터 구조에서 관찰된 최악의 경우에 발생하는 행위를 방어하기 위하여 그 트리는 선형화될 것이다.

1) **R*-trees** are a variant of R-trees used for indexing spatial information. R*-trees support point and spatial data at the same time with a slightly higher cost than other R-trees.

2) An **R+ tree** is a method for looking up data using a location, often (x, y) coordinates, and often for locations on the surface of the earth. Searching on one number is a solved problem; searching on two or more, and asking for locations that are nearby in both x and y directions, requires craftier algorithms.

Fundamentally, an R+ tree is a tree data structure, a variant of the R tree, used for indexing spatial information.

** 대표적인 검색엔진; **QBIC(Query By Image Content)** 프로젝트, IBM Almaden 연구소의 프로젝트

Content-based image retrieval (CBIR), known as **query by image content** (QBIC) and content-based visual information retrieval (CBVIR) is the application of computer vision techniques to the image retrieval problem, that is, the problem of searching for digital images in large databases. Content-based image retrieval is opposed to concept-based approaches.

"Content-based" means that the search analyzes the contents of the image rather than the metadata such as keywords, tags, or descriptions associated with the image. The term "content" in this context might refer to colors, shapes, textures, or any other information that can be derived from the image itself. CBIR is desirable because most web-based image search engines rely purely on metadata and this produces a lot of garbage in the results. Also having humans manually enter keywords for images in a large database can be inefficient, expensive and may not capture every keyword that describes the image. Thus a system that can filter images based on their content would provide better indexing and return more accurate results.

(오디오 자료의 내용기반색인)

* Query by humming (QbH): 1995년 코넬대학의 Query By Humming(QBH)

a music retrieval system that branches off the original classification systems of title, artist, composer, and genre. It normally applies to songs or other music with a distinct single theme or melody. The system involves taking a user-hummed melody (input query) and comparing it to an existing database. The system then returns a ranked list of music closest to the input query.

* MELDEX: 1997년 뉴질랜드의 Waikato 대학의 MELody in Dex(MELDEX)

the New Zealand Digital Library's Web-based melody index. the MELDEX system, designed to retrieve melodies from a database on the basis of a few notes sung into a microphone.

p. cxxxi

* (자체적으로는 인용색인을 생성하지 않고 "Cited By" 기능을 통해 인용색인 데이터베이스에 링크하는 데이터베이스의 예):

ScienceDirect is website operated by the Anglo-Dutch publisher Elsevier containing (as of 2013) about 11 million articles from 2,500 journals and over 25,000 e-books, reference works, book series and handbooks. The articles are grouped in four main sections: Physical Sciences and Engineering, Life Sciences, Health Sciences, and Social Sciences and Humanities. For most articles on the website, abstracts are freely available; access to the full text of the article (in PDF, and also HTML for newer publications) generally requires a subscription or pay-per-view purchase.

SAGE is a leading international publisher of journals, books, and electronic media for academic, educational, and professional markets. Since 1965, SAGE has helped inform and educate a global community of scholars, practitioners, researchers, and students spanning a wide range of subject areas including business, humanities, social sciences, and science, technology, and medicine.

PMC is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

* Bibliographic Database

서지 데이터베이스란 journal and newspaper articles, conference proceedings, reports, government and legal publications, patents, books, etc.를 포함하여 출판된 문헌에 대한 조직화된 디지털 콜렉션의 레퍼런스들인 서지 레코드의 데이터베이스이다. 도서관 목록 엔트리와는 대조적으로, 서지 데이터베이스에 있는 서지 레코드의 대부분은 완전한 단행본보다는 기사, 회의자료 등을 묘사하며 그것들은 keywords, subject classification terms, or abstracts과 같은 매우 풍부한 주제 묘사를 일반적으로 포함하고 있다.

서지 데이터베이스는 범위가 일반적이거나 특수한 학문적 분야를 다룰 수 있다. 많은 수의 서지 데이터베이스는 아직까지 벤더로부터 또는 그것들을 만든 색인 및 초록 서비스 회사로부터 직접적으로 라이센스를 얻어야만 사용할 수 있는 전매품들이다.

많은 서지 데이터베이스는 디지털 도서관으롤 진화하여 색인된 콘텐트의 풀-텍스트를 제공하고 있다. 나머지들은 비-서지적 학술 데이터베이스로 전환하여 Chemical Abstracts or Entrez와 같은 보다 완벽한 학술탐색 엔지 시스템을 만들고 있다.

1) The **Entrez** Global Query Cross-Database Search System is a powerful federated search engine, or web portal that allows users to search many discrete health sciences databases at the National Center for Biotechnology Information (NCBI) website. The NCBI is a part of the National Library of Medicine (NLM), which is itself a department of the National Institutes of Health (NIH), which in turn is a part of the United States Department of Health and Human Services. The name "Entrez" (a greeting meaning "Come in!" in French) was chosen to reflect the spirit of welcoming the public to search the content available from the NLM.

* Document-oriented database

도큐멘트-지향적 데이터베이스는 semi-structured data로 잘 알려진 도큐멘트-지향적 정보를 저장, 검색, 관리하도록 디자인된 컴퓨터 프로그램이다. 이 데이터베이스는 소위 NoSQL 데이터베이스의 주요 카테고리의 하나이며, "document-oriented database" (or "document store")라는 용어는 NoSQL을 사용하면서 그 인기가 커졌다. 관계형 데이터베이스와 이것들의 “Relations” (또는 “Tables”에 대한 개념과는 대조적으로, 이들 시스템은 “Document”의 추상적 개념을 가지고 디자인한 것이다.

** Documents?

도큐멘트-지향적 데이터베이스의 핵심 개념은 도큐멘트라는 개념이다. 각 도큐멘트-지향적 데이터베이스의 설치가 이것의 정의에 대한 내역에 따라 다르지만, 일반적으로 이것들 모두는 도큐멘트들이 어떤 표준 포맷이나 암호화기법으로 데이(또는 정보)를 감싸서 암호화한다는 것을 가정하고 있다. 이때 사용하는 암호화기법으로는 XML, YAML, JSON, and BSON 뿐만 아니라 PDF and Microsoft Office documents (MS Word, Excel, and so on)와 같은 이진 폼들도 있다.

1) **YAML** (/ˈjæməl/, rhymes with camel) is a human-readable data serialization format that takes concepts from programming languages such as C, Perl, and Python, and ideas from XML and the data format of electronic mail (RFC 2822).

YAML is a recursive acronym for "YAML Ain't Markup Language". Early in its development, YAML was said to mean "Yet Another Markup Language", but it was then reinterpreted (backronyming the original acronym) to distinguish its purpose as data-oriented, rather than document markup.

2) **JSON** (/ˈdʒeɪsɒn/ JAY-soun, /ˈdʒeɪsən/ JAY-son), or JavaScript Object Notation, is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs. It is used primarily to transmit data between a server and web application, as an alternative to XML.

3) **BSON** /ˈbiːsɒn/ is a computer data interchange format used mainly as a data storage and network transfer format in the MongoDB database. It is a binary form for representing simple data structures and associative arrays (called objects or documents in MongoDB). The name "BSON" is based on the term JSON and stands for "Binary JSON".

도큐멘트-지향적 데이터베이스에 있는 도큐멘트들은 관계형 데이터베이스의 레코드나 로우와 몇가지에서 비슷하지만, 엄격성이 좀 떨어진다. 이것들은 하나의 표준 스킴바에 집착하지 않으며, 또한 이것들 모두가 동일한 sections, slots, parts, or keys를 갖진 않는다. 다음은 한 도큐멘트에 대한 예이다:

```
{
   FirstName: "Bob",
   Address: "5 Oak St.",
   Hobby: "sailing"
}
```

A second document might be:

```
{
  FirstName: "Jonathan",
  Address: "15 Wanamassa Point Road",
  Children: [
        {Name: "Michael", Age: 10},
        {Name: "Jennifer", Age: 8},
        {Name: "Samantha", Age: 5},
        {Name: "Elena", Age: 2}
  ]
  }
```

이들 두가지 도큐멘트들은 서로서로 몇가지 구조적 요소를 공유하고 있지만, 각각은 또한 유일한 요소들을 가지고 있다. 모든 레코드가 사용하지 않은 필드를 빈칸으로 남겨 놓으면서 동일한 필드를 갖는 관계형 데이터베이스와 달리, 위의 예에서 어떤 도큐멘트(레코드)에서도 빈 필드는 존재하지 않는다. 이러한 방법은 새로운 정보가 그 데이터베이스에있는 모든 다른 레코드가 동일한 구조를 공유할 것을 요구하지 않고 어떤 레코드에 추가되는 것을 가능하게 한다.

** Keys

Documents are addressed in the database via a unique key that represents that document. This key is often a simple string, a URI, or a path. The key can be used to retrieve the document from the database. Typically, the database retains an index on the key to speed up document retrieval.

** Retrieval

Another defining characteristic of a document-oriented database is that, beyond the simple key-document (or key-value) lookup that can be used to retrieve a document, the database offers an API or query language that allows the user to retrieve documents based on their content. For example, you may want a query that retrieves all the documents with a certain field set to a certain value. The set of query APIs or query language features available, as well as the expected performance of the queries, varies significantly from one implementation to the next.

* Citation Index

인용색인은 이용자로 하여금 어떤 나중의 문서가 어떤 앞선 문서를 인용했는지를 쉽게 알 수 있는 출판물간의 인용 색인인 일종의 서지 데이터베이스이다.

A form of citation index is first found in 12th-century Hebrew religious literature. Legal citation indexes are found in the 18th century and were made popular by citators such as Shepard's Citations (1873). In 1960, Eugene Garfield's Institute for Scientific Information (ISI) introduced the first citation index for papers published in academic journals, first the **Science Citation Index (SCI)**, and later the **Social Sciences Citation Index (SSCI)** and **the Arts and Humanities Citation Index (AHCI)**. The first automated citation indexing was done by **CiteSeer** in 1997. Other sources for such data include **Google Scholar.**

** Major citation indexing services; General-purpose academic citation indexes include:

ISI (now part of Thomson Reuters) publishes the ISI citation indexes in print

and compact disc. They are now generally accessed through the Web under the name Web of Science, which is in turn part of the group of databases in the Web of Knowledge.

Elsevier publishes Scopus, available online only, which similarly combines subject searching with citation browsing and tracking in the sciences and social sciences.

Indian Citation Index is an online citation data which covers peer reviewed journals published from India. It covers major subject areas such as scientific, technical, medical, and social sciences and includes arts and humanities. The citation database is the first of its kind in India.

the ISI databases and Scopus are available by subscription (generally to libraries). In addition, CiteSeer and Google Scholar are freely available online.

*** Impact factor

The impact factor (IF) of an academic journal is a measure reflecting the average number of citations to recent articles published in the journal. It is frequently used as a proxy for the relative importance of a journal within its field, with journals with higher impact factors deemed to be more important than those with lower ones. The impact factor was devised by Eugene Garfield, the founder of the Institute for Scientific Information. Impact factors are calculated yearly starting from 1975 for those journals that are indexed in the Journal Citation Reports.

*** Citation impact; Citation impact can be measured in various ways.

An obvious measure is citation count, which quantifies both the usage and impact of the cited work. This is called citation analysis or bibliometrics. Among the measures that have emerged from citation analysis are the citation counts for:

an individual article (how often it was cited);
an author (total citations, or average citation count per article);
a journal (average citation count for the articles in the journal).

Many measures have been proposed, beyond simple citation counts, to better quantify an individual scholar's citation impact. The best-known measures include the h-index and the g-index. Each measure has advantages and disadvantages, spanning from bias to discipline-dependence and limitations of the citation data source.

1) The **h-index** is an index that attempts to measure both the productivity and impact of the published work of a scientist or scholar. The index is based on the set of the scientist's most cited papers and the number of citations that they have received in other publications. The index can also be applied to the productivity and impact of a group of scientists, such as a department or university or country, as

well as a scholarly journal. The index was suggested by Jorge E. Hirsch, a physicist at UCSD, as a tool for determining theoretical physicists' relative quality and is sometimes called the Hirsch index or Hirsch number.

2) The **g-index** is an index for quantifying scientific productivity based on publication record. It was suggested in 2006 by Leo Egghe. The index is calculated based on the distribution of citations received by a given researcher's publications:

Given a set of articles ranked in decreasing order of the number of citations that they received, the g-index is the (unique) largest number such that the top g articles received (together) at least g2 citations.

Just as with the h-index, the g-index is a number which is the same for two different quantities:

g is (1) the number of highly cited articles, such that each of them has brought (2) on average g citations.

** Eigenfactor

The Eigenfactor score, developed by Jevin West and Carl Bergstrom at the University of Washington, is a rating of the total importance of a scientific journal. Journals are rated according to the number of incoming citations, with citations from highly ranked journals weighted to make a larger contribution to the eigenfactor than those from poorly ranked journals. As a measure of importance, the Eigenfactor score scales with the total impact of a journal. All else equal, journals generating higher impact to the field have larger Eigenfactor scores.

Eigenfactor scores and Article Influence scores are calculated by eigenfactor.org, where they can be freely viewed. The Eigenfactor score is intended to measure the importance of a journal to the scientific community, by considering the origin of the incoming citations, and is thought to reflect how frequently an average researcher would access content from that journal. However, the Eigenfactor score is influenced by the size of the journal, so that the score doubles when the journal doubles in size (measured as published articles per year). The Article Influence score measures the average influence of articles in the journal, and is therefore comparable to the ISI impact factor.

Eigenfactor scores are measures of a journal's importance. It can be used in combination with H-index to evaluate the work of individual scientists.

* Journal Citation Reports

Journal Citation Reports (JCR) is an annual publication by the Science and Scholarly Research division of Thomson Reuters. It has been integrated with the Web of Science and is accessed from the Web of Science-Core Collections. It provides information about academic journals in the sciences and social sciences, including impact factors. The JCR was originally published as a part of Science Citation Index. Currently, the JCR, as a distinct service, is based on citations compiled from the Science Citation Index Expanded (SCIE) and the Social Science Citation Index (SSCI).

** Basic journal information
The information given for each journal includes:

1) the basic bibliographic information of publisher, title abbreviation, language, ISSN.
2) the subject categories (there are 171 such categories in the sciences and 54 in the social sciences)

** Citation information
Basic citation data:
1) the number of articles published during that year and
2) the number of times the articles in the journal were cited during the year by later articles in itself and other journals,

* Coercive citation

Coercive citation is an academic publishing practice in which an editor of a scientific or academic journal forces an author to add spurious citations to an article before the journal will agree to publish it. This is done to inflate the journal's impact factor, thus artificially boosting the journal's scientific reputation. Manipulation of impact factors and self-citation has long been frowned upon in academic circles; however, the results of a 2012 survey indicate that about 20% of academics working in economics, sociology, psychology, and multiple business disciplines have experienced coercive citation. Individual cases have also been reported in other disciplines.

* SCImago Journal Rank

SCImago Journal Rank (SJR indicator) is a measure of scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where such citations come from. The SJR indicator is a variant of the eigenvector centrality measure used in network theory. Such measures establish the importance of a node in a network based on the principle that connections to high-scoring nodes contribute more to the score of the node. The SJR indicator, which is inspired by the PageRank algorithm, has been developed to be used in extremely large and heterogeneous journal citation networks. It is a size-independent indicator and its values order journals by their "average prestige per article" and can be used for journal.

1) **PageRank** is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring

the importance of website pages. According to Google:

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

— Facts about Google and Competition

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known. Google uses an automated web spider called Googlebot to actually count links and gather other information on web pages.

*** Acknowledgement index**

An acknowledgment index is a method for indexing and analyzing acknowledgments in the scientific literature and, thus, quantifies the impact of acknowledgments. Typically, a scholarly article has a section where the authors acknowledge entities such as funding, technical staff, colleagues, etc. that have contributed materials or knowledge or have influenced or inspired their work.

p. clxvii

*** 분류자질 선정의 대표적인 기법**

- 빈도기법(단어빈도, 문헌빈도, 역문헌빈도)
- 상호정보량(mutual information)
- 정보획득량(inf. gain)
- 카이제곱 통계량(X^2)
- 공기어로 알려진 동시출현단어(Co-occurrence word or co-word)

p. clxviii

*** 문헌 범주화에 사용되는 분류기**

나이브 베이즈(naive Bayes) 분류기(classifier)

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

의사결정나무(Decision Tree) 분류기

In computational complexity and communication complexity theories the **decision tree model** is the model of computation or communication in which an algorithm or

communication process is considered to be basically a decision tree, i.e., a sequence of branching operations based on comparisons of some quantities, the comparisons being assigned the unit computational cost.

kNN(k-nearest neighbors) 분류기

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

1) **Nonparametric regression** is a form of regression analysis in which the predictor does not take a predetermined form but is constructed according to information derived from the data. Nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates.

Nearest neighbor search (NNS)

also known as proximity search, similarity search or closest point search, is an optimization problem for finding closest (or most similar) points. Closeness is typically expressed in terms of a dissimilarity function: The less similar are the objects the larger are the function values.

SVM(Support Vector Machine) 분류기

In machine learning, **support vector machines (SVMs**, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.

신경망(Neural Network) 분류기

In computer science and related fields, artificial neural networks are computational models inspired by animals' central nervous systems (in particular the brain) that are capable of machine learning and pattern recognition. They are usually presented as systems of interconnected "neurons" that can compute values from inputs by feeding information through the network.

p. clxxi

*** 유사도 계수(유사계수)**

거리계수(distance coefficient)

In statistics, the **Bhattacharyya distance** measures the similarity of two discrete or continuous probability distributions. It is closely related to the Bhattacharyya coefficient which is a measure of the amount of overlap between two statistical samples or populations. Both measures are named after A. Bhattacharya, a

statistician who worked in the 1930s at the Indian Statistical Institute.[1] The coefficient can be used to determine the relative closeness of the two samples being considered. It is used to measure the separability of classes in classification and it is considered to be more reliable than the Mahalanobis distance, as the Mahalanobis distance is a particular case of the Bhattacharyya distance when the standard deviations of the two classes are the same. Therefore, when two classes have similar means but different standard deviations, the Mahalanobis distance would tend to zero, however, the Bhattacharyya distance would grow depending on the difference between the standard deviations.

유클리드 거리(Euclidean distance),

In mathematics, the **Euclidean distance** or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, and is given by the Pythagorean formula. By using this formula as distance, Euclidean space (or even any inner product space) becomes a metric space. The associated norm is called the Euclidean norm. Older literature refers to the metric as Pythagorean metric.

민코스키 매트릭스(Minincowski metrics)

The **Minkowski distance** is a metric on Euclidean space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.

시티 블록 거리(city block distance)

Taxicab geometry, considered by Hermann Minkowski in 19th century Germany, is a form of geometry in which the usual distance function or metric of Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their Cartesian coordinates. The taxicab metric is also known as rectilinear distance, L1 distance or norm (see Lp space), **city block distance**, Manhattan distance, or Manhattan length, with corresponding variations in the name of the geometry. The latter names allude to the grid layout of most streets on the island of Manhattan, which causes the shortest path a car could take between two intersections in the borough to have length equal to the intersections' distance in taxicab geometry.

연관계수(association coefficient)

코사인(Cosine) 계수: 문헌이나 용어 클러스터링에서 가장 많이 사용되는 유용한 척도.

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of

0° is 1, and it is less than 1 for any other angle. It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1].

Note that these bounds apply for any number of dimensions, and Cosine similarity is most commonly used in high-dimensional positive spaces. For example, in Information Retrieval and text mining, each term is notionally assigned a different dimension and a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter.
The technique is also used to measure cohesion within clusters in the field of data mining.

Cosine distance is a term often used for the complement in positive space, that is: $D_C(A,B) = 1 - S_C(A,B)$. It is important to note, however, that this is not a proper distance metric as it does not have the triangle inequality property and it violates the coincidence axiom; to repair the triangle inequality property whilst maintaining the same ordering, it is necessary to convert to Angular distance (see below.)

One of the reasons for the popularity of Cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.

자카드(Jaccard) 계수: 문헌이나 용어 클러스터링에서 가장 많이 사용되는 유용한 척도

The Jaccard index, also known as the **Jaccard similarity coefficient** (originally coined coefficient de communauté by Paul Jaccard), is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

(If A and B are both empty, we define $J(A,B)=1$.) Clearly,

$$0 \leq J(A,B) \leq 1.$$

다이스 계수(Dice coef.)

The Sørensen–Dice index, also known by other names (see Names, below), is a statistic used for comparing the similarity of two samples. It was independently developed by the botanists Thorvald Sørensen[1] and Lee Raymond Dice,[2] who

published in 1948 and 1945 respectively.

The index is known by several other names, usually Sørensen index or Dice's coefficient. Both names also see "similarity coefficient", "index", and other such variations. Common alternate spellings for Sørensen are Sorenson, Soerenson index and Sörenson index, and all three can also be seen with the -sen ending.

해만 계수(Hamann coef.)???

In information theory, the **Hamming distance** between two strings of equal length is the number of positions at which the corresponding symbols are different. In another way, it measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other.

<Examples>
The Hamming distance between:

"toned" and "roses" is 3.
10111**0**1 and 10**0**10**0**1 is 2.
2**1**73**8**96 and 2**23**37**9**6 is 3.

* 상관계수(correlation coefficient)
# 피어슨 적률(Pearson product moment) 상관계수

In statistics, the **Pearson product-moment correlation coefficient** (/ˈpɪərsɨn/) (sometimes referred to as the PPMCC or PCC,[1] or Pearson's r) is a measure of the linear correlation (dependence) between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is widely used in the sciences as a measure of the degree of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name.

* 내적 계수(inner product coefficient)

In linear algebra, an **inner product space** is a vector space with an additional structure called an inner product. This additional structure associates each pair of

vectors in the space with a scalar quantity known as the inner product of the vectors. Inner products allow the rigorous introduction of intuitive geometrical notions such as the length of a vector or the angle between two vectors. They also provide the means of defining orthogonality between vectors (zero inner product). Inner product spaces generalize Euclidean spaces (in which the inner product is the dot product, also known as the scalar product) to vector spaces of any (possibly infinite) dimension, and are studied in functional analysis.

An **inner product** naturally induces an associated norm, thus an inner product space is also a normed vector space. A complete space with an inner product is called a Hilbert space. An incomplete space with an inner product is called a pre-Hilbert space, since its completion with respect to the norm induced by the inner product becomes a Hilbert space. Inner product spaces over the field of complex numbers are sometimes referred to as unitary spaces.

p. clxxiv

* Cluster analysis or clustering

is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Besides the term clustering, there are a number of terms with similar meanings, including automatic classification, numerical taxonomy, botryology (from Greek βότρυς "grape") and typological analysis. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest. This often leads to misunderstandings between researchers coming from the fields of data mining and machine learning, since they use the same terms and often the same algorithms, but have different goals.

Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939and famously used by Cattell beginning in 1943 for trait theory classification in personality psychology.

Clustering algorithms can be categorized based on their cluster model, as listed above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms. Not all provide models for their clusters and can thus not easily be categorized. An overview of algorithms explained in Wikipedia can be found in the list of statistics algorithms.

There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder." The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically different kind of model. For example, k-means cannot find non-convex clusters.

클러스트링 알고리즘의 계층적 기법: Connectivity based clustering (**hierarchical clustering**)

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the

x-axis such that the clusters don't mix.

Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as "chaining phenomenon", in particular with single-linkage clustering). In the general case, the complexity is , which makes them too slow for large data sets. For some special cases, optimal efficient methods (of complexity) are known: SLINK[5] for single-linkage and CLINK[6] for complete-linkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did however provide inspiration for many later methods such as density based clustering.

단일 연결(single linkage)
완전(complete) 연결
그룹 평균(group average) 연결
워드 기법(Word's method)

클러스트링 알고리즘의 비계층적 기법
single pass 기법

Single-linkage clustering is one of several methods of agglomerative hierarchical clustering. In the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined. The definition of 'shortest distance' is what differentiates between the different agglomerative clustering methods. In single-linkage clustering, the link between two clusters is made by a single element pair, namely those two elements (one in each cluster) that are closest to each other. The shortest of these links that remains at any step causes

the fusion of the two clusters whose elements are involved. The method is also known as nearest neighbour clustering. The result of the clustering can be visualized as a dendrogram, which shows the sequence of cluster fusion and the distance at which each fusion took place.

K-means 기법

centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, **k-means clustering** gives a formal definition as an optimization problem: find the cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd's algorithm, often actually referred to as "k-means algorithm". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

Most k-means-type algorithms require the number of clusters - - to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters (which is not surprising, as the algorithm optimized cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. On the one hand, it partitions the data space into a structure known as a Voronoi diagram. On the other hand, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based classification, and Lloyd's algorithm as a variation of the Expectation-maximization algorithm for this model discussed below.

EM(expectation maximization) 알고리즘

In statistics, an **expectation-maximization (EM)** algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step,

which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

p.clxxvi

*** 웹 문헌 클러스터링**

단어기반(term-based) 클러스터링: 단어의 유사도에 기반, 텍스트가 많은 웹 문헌.
링크기반(link-based) 클러스터링: 이미지가 많은 웹 문헌.
    ## intra-document link
    ## inter-document link
    ## out-link
    ## in-link
혼합형(hybrid) 클러스터링

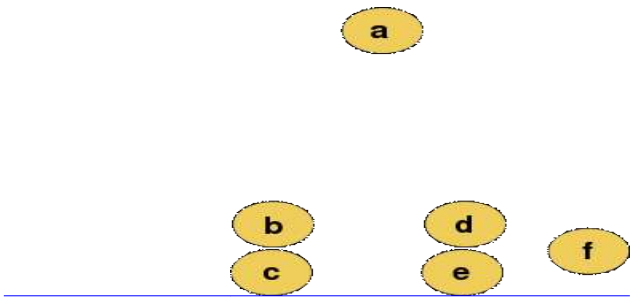
p. clxxx

*** 덴드로그램(dendrogram):** 클러스터 생성 과정을 표현하는 한 방법

A **dendrogram** (from Greek *dendron* "tree" and *gramma* "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples.

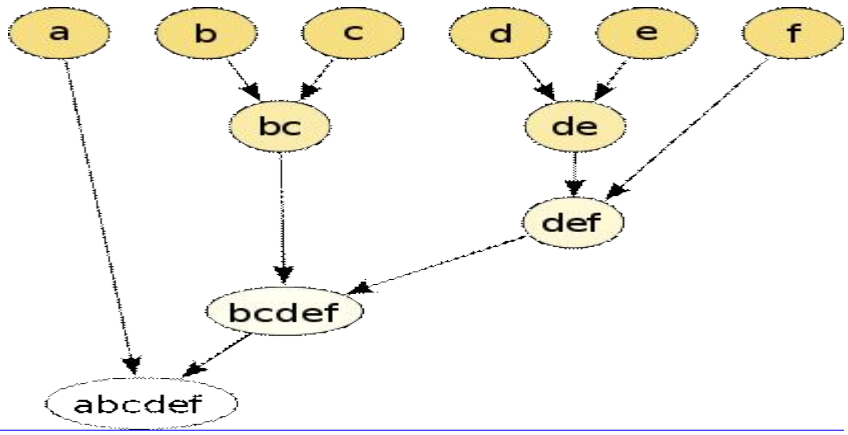
** Clustering Example

For a clustering example, suppose this data is to be clustered using Euclidean distance as the distance metric.



Raw data

The hierarchical clustering dendrogram would be as such:



Traditional representation

The top row of nodes represent data (individual observations), and the remaining nodes represent the clusters to which the data belong, with the arrows representing the distance (dissimilarity).

The distance between merged clusters is monotone increasing with the level of the merger: the height of each node in the plot is proportional to the value of the intergroup dissimilarity between its two daughters (the top nodes representing individual observations are all plotted at zero height).

<제 7장 정보검색 언어>

* Web 1.0, 2.0, 3.0

Web 1.0

That Geocities & Hotmail era was all about read-only content and static HTML websites. People preferred navigating the web through link directories of Yahoo! .

-the mostly read only web
-45million global users(1996)
-focused on companies
-home pages
-owning content
-Britannica Online
-HTML. ports
-web forms
-directories(taxonomy)

-Netscape
-page views
-advertising

Web 2.0

This is about user-generated content and the read-write web. People are consuming as well as contributing information through blogs or sites like Flickr, YouTube, Digg, etc. The line dividing a consumer and content publisher is increasingly getting blurred in the Web 2.0 era.

-the wildly read-write web
-1 billion + golobal users(2006)
-focused on communities
-blogs
-sharing content
-Wikipedia
-XML, RSS
-web applications
-tagging("folksonomy")
-Google
-cost per click
-word of mouth

Web 3.0

This will be about semantic web (or the meaning of data), personalization (e.g. iGoogle), intelligent search and behavioral advertising among other things.

-the portable personal web
-focused on the individual
-lifestream
-consolidating dynamic content
-the semantic web
-widgets, drag & drop mashups

a **mashup** is a Web page or application that uses and combines data, presentation or functionality from two or more sources to create new services. The term implies easy, fast integration, frequently using open APIs and data sources to produce enriched results that were not necessarily the original reason for producing the raw source data.

The main characteristics of the mashup are combination, visualization, and aggregation. It is important to make existing data more useful, moreover for personal and professional use. To be able to permanently access the data of other services, mashups are generally client applications or hosted online.

-user behavior("me-onomy")
-iGoogle, NetVibes
-user engagement
-advertainment

p. clxxxvi

* 토픽맵(Topic Maps) : See Chptr 11.

* 시맨틱 웹 : See Chptr 11.

* 온톨로지: See Chptr 11

* RDF

RDF는 W3C 스펙의 한 종류이며 원래는 메타데이터 데이터 모델로 디자인되었다. 이것은 다양한 syntax notations과 data serialization formats을 사용하여 웹 자원에서 실행되는 정보의 개념적 기술이나 모델링을 위한 일반적인 방법으롤 사용되고 있다.

** Overview

RDF 데이터 모델은 객체-관계 또는 클래스 다이어그램과 같은 고전적 개념적 모델링 기법과 비슷하다. 그 이유는 주체-술어-객체 표현의 형태로 자원(특히 웹자원에서)에 대하여 표현한다는 아이디어를 근거로 하기 때문이다. 이러한 표현식은 RDF 용어에서 triples로서 알려져 있다. 주체는 자원을, 술어는 자원의 속성이나 모습을 나타내며 주체와 객체 사이의 관계를 표현한다.

예를 들어, RDF로 “The sky has the color blue”라는 관념을 표현하는 한 가지 방법은 다음과 같은 트리플: 주체는 “the sky”, 술어는 “has”, 그리고 객체는 “the color blue”로 나타내는 것이다. 그러므로 RDF는 객체를 객체-지향형 디자인에서 엔티티-속성-값의 고전적 개념에 사용될 수 있는 주체로 바꾼다; 객체(sky), 속성(color), 값(blue). RDF는 여러 가지 연속 포맷(다시 말해서 파일 포맷)으로 된 하나의 추상적 모델이다. 그래서 하나의 자원이나 트리플을 코드화하는 특별한 방법은 포맷마다 다르다.

자원을 기술하는 이 메카니즘은 자동화된 소프트웨어가 웹을 통해 분산되는 기계가독형 정보를 저장, 교환, 이용할 수 있도록 하여 이용자로 하여금 보다 커다란 효율성과 확실성을 가지고 정보를 다룰 수 있도록 하는 웹의 진화적 단계인 W3C의 시멘틱 웹 활동에서 중요한 요소이다. RDF의 간단한 데이터 모델과 서로 다른 추상적 개념을 모델화하는 능력은 시멘틱 웹의 활동과 관련이 없은 지식관리 어플에서도 사용의 증가를 가져왔다.

한 무리의 RDF 서술문은 본질적으로 표식이 있고 통제된 멀티 그래프를 나타낸다. 마찬가지로, RDF-의존형 데이터 모델은 관계형 모델이나 기타 온톨로지 모델보다 어떤 종류의 지식을 표현하는데 있어서 보다 자연스럽다. 그렇지만, 실재로, RDF 데이터는 종종 관계

형 데이터베이스나 Triplestore라고 부르는 native representation, 또는 만일 콘텍스트(다시 말해서 명명된 그래프) 역시 각 RDF 트리플용으로 존속한다면 Quad stores 에 존속하고 있다.

1) A **triplestore** is a purpose-built database for the storage and retrieval of triples, a triple being a data entity composed of subject-predicate-object, like "Bob is 35" or "Bob knows Fred".

Much like a relational database, one stores information in a triplestore and retrieves it via a query language. Unlike a relational database, a triplestore is optimized for the storage and retrieval of triples. In addition to queries, triples can usually be imported/exported using Resource Description Framework (RDF) and other formats.

Some triplestores can store billions of triples.

2) **Named graphs** are a key concept of Semantic Web architecture in which a set of Resource Description Framework statements (a graph) are identified using a URI, allowing descriptions to be made of that set of statements such as context, provenance information or other such metadata.

Named graphs are a simple extension of the RDF data model through which graphs can be created but the model lacks an effective means of distinguishing between them once published on the Web at large.

ShEX 또는 Shape Expression은 RDF 그래프의 문제점을 표현하는 언어이다. 여기에는 OSLC Resource Shapes와 Dublin Core Description Set Profiles 뿐만 아니라 분리와 다형성을 위한 논리적 관계에 발생하는 cardinality 문제점도 포함하고 있다. RDFS와 OWL에서 보여준 것처럼 누구나 RDF로 추가적인 온톨로지 언어를 구축할 수 있다.

1) **RDF Schema** (Resource Description Framework Schema, variously abbreviated as RDFS, RDF(S), RDF-S, or RDF/S) is a set of classes with certain properties using the RDF extensible knowledge representation language, providing basic elements for the description of ontologies, otherwise called RDF vocabularies, intended to structure RDF resources. These resources can be saved in a triplestore to reach them with the query language SPARQL

2) The **Web Ontology Language (OWL)** is a family of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases. The languages are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web. OWL is endorsed by the World Wide Web Consortium (W3C) and has attracted academic, medical and commercial interest.

The OWL family contains many species, serializations, syntaxes and specifications with similar names. OWL and OWL2 are used to refer to the 2004 and 2009 specifications, respectively. Full species names will be used, including specification version (for example, OWL2 EL). When referring more generally, OWL Family will be used.

** RDF topics
*** RDF vocabulary
The vocabulary defined by the RDF specification is as follows:

Classes
***** rdf

rdf:XMLLiteral - the class of XML literal values
rdf:Property - the class of properties
rdf:Statement - the class of RDF statements
rdf:Alt, rdf:Bag, rdf:Seq - containers of alternatives, unordered containers, and ordered containers (rdfs:Container is a super-class of the three)
rdf:List - the class of RDF Lists
rdf:nil - an instance of rdf:List representing the empty list

***** rdfs
# rdfs:Resource - the class resource, everything
# rdfs:Literal - the class of literal values, e.g. strings and integers
# rdfs:Class - the class of classes
# rdfs:Datatype - the class of RDF datatypes
# rdfs:Container - the class of RDF containers
# rdfs:ContainerMembershipProperty - the class of container membership properties, rdf:_1, rdf:_2, ..., all of which are sub-properties of rdfs:member

Properties
***** rdf
# rdf:type - an instance of rdf:Property used to state that a resource is an instance of a class
# rdf:first - the first item in the subject RDF list
# rdf:rest - the rest of the subject RDF list after rdf:first
# rdf:value - idiomatic property used for structured values
# rdf:subject - the subject of the subject RDF statement
# rdf:predicate - the predicate of the subject RDF statement
# rdf:object - the object of the subject RDF statement

rdf:Statement, rdf:subject, rdf:predicate, rdf:object are used for reification (see

below).

***** rdfs
# rdfs:subClassOf - the subject is a subclass of a class
# rdfs:subPropertyOf - the subject is a subproperty of a property
# rdfs:domain - a domain of the subject property
# rdfs:range - a range of the subject property
# rdfs:label - a human-readable name for the subject
# rdfs:comment - a description of the subject resource
# rdfs:member - a member of the subject resource
# rdfs:seeAlso - further information about the subject resource
# rdfs:isDefinedBy - the definition of the subject resource

This vocabulary is used as a foundation for RDF Schema where it is extended.

** Serialization formats
Several common serialization formats are in use, including:

Turtle; a compact, human-friendly format.
N-Triples; a very simple, easy-to-parse, line-based format that is not as compact as Turtle.
N-Quads; a superset of N-Triples, for serializing multiple RDF graphs.
JSON-LD; a JSON-based serialization.
N3 or Notation 3; a non-standard serialization that is very similar to Turtle, but has some additional features, such as the ability to define inference rules.
RDF/XML; an XML-based syntax that was the first standard format for serializing RDF.

RDF/XML은 때때로 간단하게 RDF라 부르는 실수를 범한다. 그 이유는 그것이 RDF를 정의하고 있는 다른 W3C 스펙들에서 소개되었기 때문이며 역사적으로 첫 번째 W3C 표준 RDF 연속화 포맷이기 때문이다. 그렇지만, RDF/XML 포맷과 추상적인 RDF 모델 그 자체를 구분하는 것은 중요하다. 비로 RDF/XML 포맷이 아직까지 사용중이라 하더라도, 다른 RDF 연속화는 이제 많은 RDF 사용자에 의해 선호되고 있다. 왜냐하면 이것들은 인간친화적이고, XML QNames의 구분법에 있는 제한으로 인하여 어떤 RDF 그래프는 RDF/XML에서는 표현될 수 없기 때문이다.

RDF triples may be stored in a type of database called a triplestore.
RDF 트리플은 트리플스토어라 부르는 일종의 데이터베이스에 저장될 수 있다.

** Resource identification
RDF 서술문의 주체는 URI 이거나 blank node이다. 둘 다 자원을 나타낸다. blank

node가 가르키는 자원은 익명의 자원이라 부른다. 이것들은 RDF 진술문으로부터 직접적으로 식별할 수 없다. 그것의 술어는 또한 관계를 표현하는 자원을 나타내는 URI이다. 객체는 URI, blank node 또는 Unicode 스트링 리터럴이다.

1) A **string literal** is the representation of a string value within the source code of a computer program. Most often in modern languages this is a quoted sequence of characters (formally "bracketed delimiters"), as in x = "foo", where "foo" is a string literal with value *foo* – the quotes are not part of the value, and one must use escape characters to allow the delimiters themselves to be embedded in the string. However, there are numerous alternate notations for specifying string literals, particularly more complicated cases, and the exact notation depends on the individual programming language in question. Nevertheless, there are some general guidelines that most modern programming languages follow.

시멘틱 웹 어플과 RSS와 FOAF와 같은 RDF의 비교적 인기있는 어플들에서, 자원들은 고의적으로 URIs에 의해 표현되는 경향이 있으며 웹의 실재적인 데이터에 접근하는데 사용할 수 있다. 그러나 RDF는 일반적으로 인터넷-의존형 자원의 표기로 제한하진 않는다. 사실상, 자원의 이름인 URI는 결코 derefernceable되어져서는 안된다. 예를 들어, “http:”로 시작하여 RDF 서술문의 주체로서 사용되는 URI는 반드시 http를 통해 접근할 수 있는 자원만을 표현할 필요는 없으며, 또한 실체적이고 네트워크로 접근가능한 자원을 표현할 필요도 없다 - 그러한 URI는 무조건 어떤 것이든 표현할 수 있다. 그렇지만, 광범위하게 동의하는 것은 HTTP GET에서 사용될 때 300가지의 암호화된 응답을 리턴하는 a bare URI(하나의 # 심복이 없는)는 그것이 접근하는데 성공하는 인터넷 자원을 나타내는 것처럼 처리되어야 한다.

1) **RSS (Rich Site Summary)**; originally RDF Site Summary; often dubbed Really Simple Syndication, uses a family of standard web feed formats[2] to publish frequently updated information: blog entries, news headlines, audio, video. An RSS document (called "feed", "web feed",[3] or "channel") includes full or summarized text, and metadata, like publishing date and author's name.

RSS feeds enable publishers to syndicate data automatically. A standard XML file format ensures compatibility with many different machines/programs. RSS feeds also benefit users who want to receive timely updates from favourite websites or to aggregate data from many sites.

Subscribing to a website RSS removes the need for the user to manually check the web site for new content. Instead, their browser constantly monitors the site and informs the user of any updates. The browser can also be commanded to automatically download the new data for the user.

Software termed "RSS reader", "aggregator", or "feed reader", which can be web-based, desktop-based, or mobile-device-based, present RSS feed data to users. Users subscribe to feeds either by entering a feed's URI into the reader or by clicking on the browser's feed icon. The RSS reader checks the user's feeds regularly for new information and can automatically download it, if that function is enabled. The reader also provides a user interface.

2) **FOAF** (an acronym of Friend of a friend) is a machine-readable ontology describing persons, their activities and their relations to other people and objects. Anyone can use FOAF to describe him- or herself. FOAF allows groups of people to describe social networks without the need for a centralised database.

FOAF is a descriptive vocabulary expressed using the Resource Description Framework (RDF) and the Web Ontology Language (OWL). Computers may use these FOAF profiles to find, for example, all people living in Europe, or to list all people both you and a friend of yours know.[1][2] This is accomplished by defining relationships between people. Each profile has a unique identifier (such as the person's e-mail addresses, a Jabber ID, or a URI of the homepage or weblog of the person), which is

used when defining these relationships.

The FOAF project, which defines and extends the vocabulary of a FOAF profile, was started in 2000 by Libby Miller and Dan Brickley. It can be considered the first Social Semantic Web application, in that it combines RDF technology with 'Social Web' concerns.

Tim Berners-Lee, in a 2007 essay,[3] redefined the Semantic web concept into the Giant Global Graph, where relationships transcend networks and documents. He considers the GGG to be on equal ground with the Internet and the World Wide Web, stating that "I express my network in a FOAF file, and that is a start of the revolution."

그러므로, RDF 서술문의 생산자와 소비자들은 자원 식별자의 어의에 동의하여야 한다. 그러한 동의는 비록 RDF에서 사용하기 위하여 URI space에 부분적으로 포함되는 Dublin Core Metadat와 같이 일반적 용도의 몇 가지 통제어휘가 있다하더라도 RDF 그 자체에 본질적은 것은 아니다. 웹에서 RDF-의존형 온톨로지를 출판하려는 의도는 RDF에서 데이터를 표현하기 위하여 사용된 자원 식별자의 의도된 의미를 종종 제한하거나 수립한다.

For example, the URI:

http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine#Merlot

예의 URI는 양조인에 의해 모든 Merlot 적 포도주의 급을 언급하기 위하여 그것의 소유자가 의도적으로 만든 것이다. 다시 말해서, 위의 URI의 예는 단일 양조인에 의해 생산된 모든 포도주의 등급을 표현하고 있다. 정의가 발생한 OWL 온톨로지 – 그 자체가 하나의 RDF 도큐멘트인 – 에 의해 표현된 정의. 그 정의에 대한 주의 깊은 분석이 없다면 누구나 위의 URI의 예는 포도주의 종류 대신에 물리적인 어떤 것이라는 잘못된 결론을 내릴 수도 있다.

이것은 'bare' 자원 식별자가 아니라 그것보다는 '#' 문자를 포함하고 있고 조각 식별자로 마감하고 있는 URI reference 라는 것을 주목하라.

1) A **URI reference** may take the form of a full URI, or just the scheme-specific portion of one, or even some trailing component thereof – even the empty string. An optional fragment identifier, preceded by #, may be present at the end of a URI reference. The part of the reference before the # indirectly identifies a resource, and the fragment identifier identifies some portion of that resource.

To derive a URI from a URI reference, software converts the URI reference to 'absolute' form by merging it with an absolute 'base' URI according to a fixed algorithm. The system treats the URI reference as relative to the base URI, although in the case of an absolute reference, the base has no relevance. The base URI typically identifies the document containing the URI reference, although this can be overridden by declarations made within the document or as part of an external data transmission protocol. If the base URI includes a fragment identifier, it is ignored during the merging process. If a fragment identifier is present in the URI reference, it is preserved during the merging process.

Web document markup languages frequently use URI references to point to other resources, such as external documents or specific portions of the same logical document.

2) In computer hypertext, a **fragment identifier** is a short string of characters that refers to a resource that is subordinate to another, primary resource. The primary resource is identified by a Uniform Resource Identifier (URI), and the fragment identifier points to the subordinate resource.

The fragment identifier introduced by a hash mark # is the optional last part of a URL for a

document. It is typically used to identify a portion of that document. The generic syntax is specified in RFC 3986. The hash mark separator in URIs does not belong to the fragment identifier

** Statement reification and context

한 무리의 진술서에 의해 모델화된 지식의 실체는 각각의 진술(즉 각 트리플인 주체-술어-객체 모두 함께)이 URI를 할당 받아서, 추가적 진술이 이루어질 수 있는 하나의 자원으로 취급되는 reification에 따라야 할 것이다: 예를 들어 “Jane says that John is the author of document X”. Reification은 각 서술에 대하여 확신의 수준이나 유용성의 정도를 파악하기 위하여 때때로 중요하다.

1) **Reification** in knowledge representation involves the representation of factual assertions, that are referred to by other assertions; which might then be manipulated in some way. e.g., to compare logical assertions from different witnesses in order to determine their credibility.

The message "John is six feet tall" is an assertion involving truth, that commits the speaker to its factuality, whereas the reified statement, "Mary reports that John is six feet tall" defers such commitment to Mary. In this way, the statements can be incompatible without creating contradictions in reasoning. For example the statements "John is six feet tall" and "John is five feet tall" are mutually exclusive (thus, incompatible); but, the statements "Mary reports that John is six feet tall," and "Paul reports that John is five feet tall," are not incompatible, as they both are governed by a conclusive rationale, that either Mary or Paul (or both) is, in fact, incorrect.

** Query and inference languages

RDF 그래프용으로 우수한 쿼리 언어는 SPARQL이다 SPARQL은 SQL-유형의 언어이며 W3C에서 추천하고 있다.

가상의 온톨로지를 사용하여 아프리카에 있는 국가수도를 나타내는 SPARQL 쿼리의 예:

```
PREFIX abc: <nul://sparql/exampleOntology#> .
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y.
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Africa.
}
```

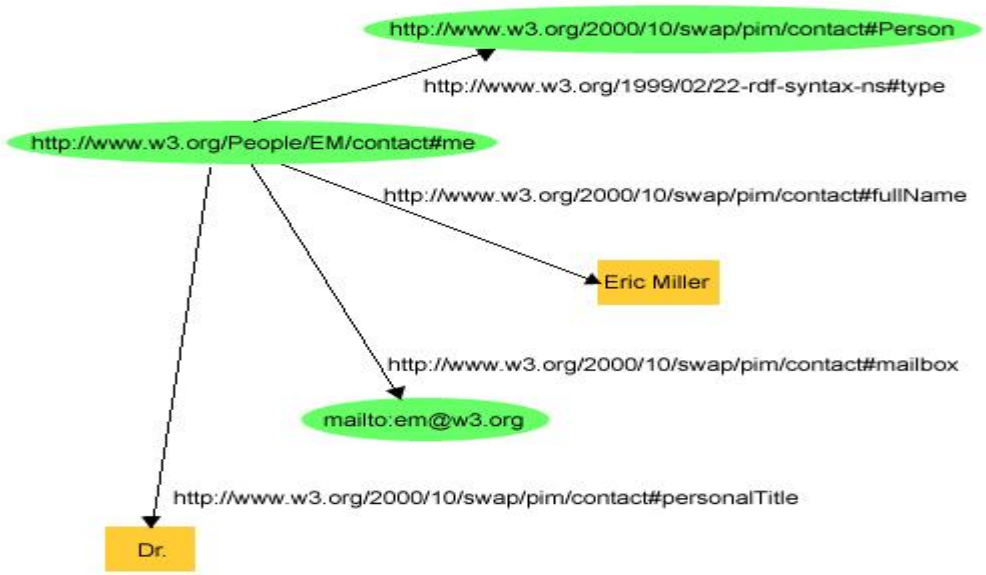
Other non-standard ways to query RDF graphs include:

RDQL, precursor to SPARQL, SQL-like
Versa, compact syntax (non-SQL-like), solely implemented in 4Suite (Python)
RQL, one of the first declarative languages for uniformly querying RDF schemas and resource descriptions, implemented in RDFSuite.[24]
SeRQL, part of Sesame
XUL has a template element in which to declare rules for matching data in RDF.

XUL uses RDF extensively for databinding.

** Examples
# Example 1: RDF Description of a person named Eric Miller

The following example is taken from the W3C website describing a resource with statements "there is a **Person** identified by **http://www.w3.org/People/EM/contact#me**, whose name is **Eric Miller**, whose email address is **em@w3.org**, and whose title is **Dr.**



An RDF Graph Describing Eric Miller

The resource "**http://www.w3.org/People/EM/contact#me**" is the subject.

The objects are:

"Eric Miller" (with a predicate "whose name is"),
mailto:em@w3.org (with a predicate "whose email address is"), and
"Dr." (with a predicate "whose title is").

The subject is a URI.

The predicates also have URIs. For example, the URI for each predicate:

"whose name is" is **http://www.w3.org/2000/10/swap/pim/contact#fullName,**
"whose email address is" is **http://www.w3.org/2000/10/swap/pim/contact#mailbox,**

"whose title is" is http://www.w3.org/2000/10/swap/pim/contact#personalTitle.

In addition, the subject has a type (with URI http://www.w3.org/1999/02/22-rdf-syntax-ns#type), which is person (with URI http://www.w3.org/2000/10/swap/pim/contact#Person).

Therefore, the following "subject, predicate, object" RDF triples can be expressed:

```
# http://www.w3.org/People/EM/contact#me, http://www.w3.org/2000/10/swap/pim/contact#fullName, "Eric Miller"
# http://www.w3.org/People/EM/contact#me, http://www.w3.org/2000/10/swap/pim/contact#mailbox, mailto:em@w3.org
# http://www.w3.org/People/EM/contact#me,  http://www.w3.org/2000/10/swap/pim/contact#personalTitle, "Dr."
# http://www.w3.org/People/EM/contact#me, http://www.w3.org/1999/02/22-rdf-syntax-ns#type, http://www.w3.org/2000/10/swap/pim/contact#Person
```

In standard N-Triples format, this RDF can be written as:

```
<http://www.w3.org/People/EM/contact#me>  <http://www.w3.org/2000/10/swap/pim/contact#fullName> "Eric Miller" .
<http://www.w3.org/People/EM/contact#me>  <http://www.w3.org/2000/10/swap/pim/contact#mailbox> <mailto:em@w3.org> .
<http://www.w3.org/People/EM/contact#me>  <http://www.w3.org/2000/10/swap/pim/contact#personalTitle> "Dr." .
<http://www.w3.org/People/EM/contact#me>  <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://www.w3.org/2000/10/swap/pim/contact#Person> .
```

Equivalently, it can be written in standard Turtle (syntax) format as:

```
@prefix eric:    <http://www.w3.org/People/EM/contact#> .
@prefix contact: <http://www.w3.org/2000/10/swap/pim/contact#> .
@prefix rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

eric:me contact:fullName "Eric Miller" .
eric:me contact:mailbox <mailto:em@w3.org> .
eric:me contact:personalTitle "Dr." .
eric:me rdf:type contact:Person .
```

Or, it can be written in RDF/XML format as:

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF  xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#"
xmlns:eric="http://www.w3.org/People/EM/contact#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
```

```
    <contact:fullName>Eric Miller</contact:fullName>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:mailbox rdf:resource="mailto:em@w3.org"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:personalTitle>Dr.</contact:personalTitle>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.w3.org/People/EM/contact#me">
    <rdf:type rdf:resource="http://www.w3.org/2000/10/swap/pim/contact#Person"/>
  </rdf:Description>
</rdf:RDF>
```

p. ccvii

<제 8장 정보검색의 실제>

* Marcia J. Bates의 Berrypicking Model

고전적인 정보검색모델과는 대조적으로, 딸기줍기 모델은 이용자가 문서집단(아마도 무수한 서로 다른 쿼리로 서로 다른 탐색)을 탐색할 때 그들은 맛있는 것(유용한 참고자료와 현실적 정보)만 수집하고 저장한다는 것을 가정한다. 중요성의 차이는 그것의 목표가 검색된 최종문헌을 생산하는 것이 아니라 그것보다는 하나씩하나씩 이삭줍기를 통하여 정보란 "딸기"를 수집하는 것이다.

딸기줍기를 소개한 논문에서, Marcia Bates는 또한 진화적 탐색의 개념을 소개하였다. 정보검색의 고전 모델에서, 이용자는 쿼리로 표현하고자 하는 변하기 않는 정보요구를 가지고 있다. 대조적으로, 진화적 탐색에서 이용자는 쿼리를 수행하여, 최종결과로 얻은 도큐멘트의 어떤 것은 취하고나면, 실재로 자신들의 정보요구를 약간 변경해야한다는 것을 배울 수도 있다. 그런 다음에 그들은 자신들의 정보요구를 보다 잘 표현하기 위해서뿐만 아니라 정보요구 그 자체가 지금 변했기 때문에 자신들의 쿼리를 다시 조정한다.

이러한 상황에서, 여러분은 "도큐멘트, 정보"의 딸기들이 보다 대규모의 연속적인 탐색을 통해서 서로 다른 시기에 수집되는 것을 알 수 있다. 정보검색의 고전적 모델에서, 누구나 기대할 수 있는 것은 탐색 과정의 말미에 최종 쿼리에 의해 나타난 도큐멘트의 세트가 이용자가 갖고자하는 모든 것을 포함할 때까지 이용자는 자신의 탐색을 간단하게 재조정할 수 있다는 것이다.

고전적 정보검색 모델과의 두 가지 차이는 다음과 같다:

1. 정보요구는 단편적이고 불변적인 것이 아니라 탐색과정을 통해 진화한다.

2. 탐색과정의 결과는 최종 쿼리에 의해 검색된 도큐멘트의 세트가 아니라 그 과정을 거치면서 이삭줍기식으로 검색된 도큐멘트, 참고자료, 정보이다.

* What is the JISC IE(Joint Information Systems Commitee Information Environment?

The Information Environment (IE)는 사람들로 하여금 자신들의 학습, 교수, 또는 연구와 관련해서 효율적이고 효과적으로 정보를 발견하고 관리할 수 있는 서비스를 개발하고 제공하는 JISC의 업무를 말하는 용어이다. 사람들이 요구하는 정보자원은 매우 다양하며 - books, journals, research papers, teaching resources, videos, maps 등 - , 그것들이 어떤 포맷으로 되어 있더라도 점차적으로 디지털화되고 있다.

1) **Jisc** (formerly the Joint Information Systems Committee, and still commonly referred to as JISC) is a United Kingdom non-departmental public body whose role is to support post-16 and higher education, and research, by providing leadership in the use of information and communications technology (ICT) in learning, teaching, research and administration. It is funded by all the UK post-16 and higher education funding councils.

** What does the Information Environment mean in practice?

national resource discovery tools such as the Archives Hub which provides convenient access to information about unique research collections distributed across the UK
software protocols such as SWORD (Simple Web Service Offering Repository Deposit) which enables files to be easily deposited in digital repositories from within other applications
'technical' infrastructure such as the OpenURL router service at EDINA which enables linking between bibliographic records and the electronic or other copy of the item referenced to which a user's home institution has access

1) **EDINA** is a UK-based data centre (funded by the Joint Information Systems Committee - JISC), which provides data applications delivered over the Internet and aimed primarily at Higher Education staff and students in the United Kingdom. (In this context, a "data centre" is an organisation that provides a set of specific datasets which can either be downloaded, or accessed and manipulated directly over the Internet. The two other main UK-based data centres are MIMAS and the UKDA.)

It also conducts research and development (R&D) projects into the delivery of data across networks.

Although funded at a national level, EDINA operates through the University of Edinburgh, where it is a division of Information Services.

centres of expertise such as the Digital Curation Centre
practical guidance such as a methodology for the analysis and costing of the lifecycle of digital objects

** What is being developed as part of the Information Environment at the moment?

Current programme activities include:

exploring how digital repositories of research outputs can be made easier to use
putting digital preservation into practice
increasing understanding of how metadata for digital resources can be created automatically.

<제 9장 검색 인터페이스>

* ORBIT - Questel-Orbit

Questel-Orbit는 지적재산권에 관한 특별한 제공자이다. 이것은 특허 데이터베이스 장서, 상표 데이터베이스, 그리고 CAS, COMPENDEX(engineering), INSPEC(scientific and technical literature), PASCAL(covers the core scientific literature in science, technology and medicine with special emphasis on European literature)에 있는 비-특허 문헌 집단를 제공한다. 이것의 콘텐트는 예를 들어, Dialog에 의해 데이터 장서와 함께 부분적으로 중복되어 제공된다.

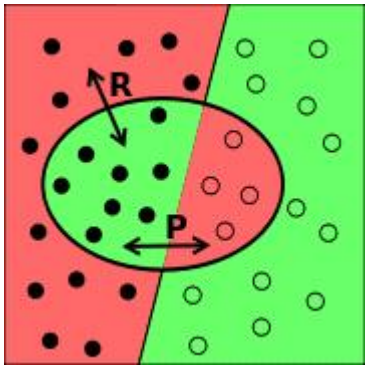
* Dialog and the invention of online information services

1972년 설립된 세계 최초의 상업적 온라인 탐색 서비스이다. Dialog는 최고급 콘텐트 정보원의 선도적 제공자로 긴 그리고 자랑스런 역사를 가지고 있다.

Now, as part of ProQuest, we are focused on providing content and search capabilities to meet the Information needs of professional users. For researchers in corporate, business and government settings worldwide, ProQuest Dialog delivers authoritative answers to support critical decision making, build competitive advantage and drive innovation. and STN International(an online database service that provides global access to published research, journal literature, patents, structures, sequences, properties, and other data), but Questel-Orbit''s strength definitely and exclusively lies with patent and trademark information. The power of Questel-Orbit''s retrieval language is comparable to that of Dialog und STN International. Some of Questel-Orbit''s offers, including patent full-texts, are -- free of charge -- also available at national and international patent offices (e.g. the European Patent Office).

<제 10장 정보검색 시스템의 평가>

* Precision and recall



Recall-precision

위의 그림에서 적합한 아이템들은 직선의 왼쪽에 있는 반면에 검색된 아이템들은 타원형 한에 있다. 붉은 지역은 에러를 표현한다. 왼쪽편에는 검색되지 않은 적합한 아이템(false negatives)가 있으며, 반면에 오른쪽에는 적합하지 않은 검색된 아이템들(false positives)이 있다. 패턴 인식과 정보검색에서, 정확률(또는 positive predictive value라고도 부름)은 적합해서 검색된 경우의 단편인 반면에 재현율(또는 sensitivity로 알려져 있음)은 검색된 적합한 경우의 단편이다. 그러므로 정확률과 재현율 둘 다 적합성의 이해와 척도의 근거이다.

여러 장면에 있는 개들을 인식하는 프로그램이 9마리의 개와 몇 마리의 고양이가 있는 한 장면에서 개 7마리를 찾았다고 가정해 보자. 찾아낸 것 중에서 4마리는 정확하지만 3마리는 실재로는 고양이라면, 이 프로그램의 정확률은 4/7인 반면에 재현율은 4/9이다. 탐색엔진이 단지 20페이지만 적합한 30페이지를 제공하고 추가적으로 적절한 40페이지를 제공하는데 실패한다면, 그것의 정확률은 20/30 = 2/3 이지만, 그것의 재현율은 20/60 = 1/3 이다.

간단히 말해서, 높은 재현율은 그 알고리즘이 대부분이 적합한 결과를 제공하지만 높은 정확률은 그 알고리즘이 본질적으로 부적합한 것보다는 더 많은 적합한 결과를 제공한다는 것을 의미한다.

*** Relevence**

적합성의 개념은 인지과학, 논리학, 문헌정보학과 같은 많은 분야에서 연구되고 있다. 가장 기본적으로 이것은 인식론(지식의 이론)에서 연구되고 있다. 지식에 대한 다양한 이론들은 무엇이 적합한가에 대하여 다양한 주장을 펴고 있으며 이러한 기본적인 견해들은 모든 다른 분야와도 연결되어 있다.

** Definition

"Something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T." (Hjørland & Sejer Christensen,2002).

** Library and information science

This field has considered when documents (or document representations) retrieved from databases are relevant or non-relevant. Given a conception of relevance, two measures have been applied: Precision and recall:

Recall = a : (a + c) X 100%, where a = number of retrieved, relevant documents, c = number of non-retrieved, relevant documents (sometimes termed "silence"). Recall is thus an expression of how exhaustive a search for documents is.

Precision = a : (a + b) X 100%, where a = number of retrieved, relevant documents, b = number of retrieved, non-relevant documents (often termed "noise").

Precision is thus a measure of the amount of noise in document-retrieval.

Relevance itself has in the literature often been based on what is termed "the system's view" and "the user's view". Hjørland (2010) criticize these two views and defends a "subject knowledge view of relevance".

p. cclxxxiv

*** 확률모델**

주어진 쿼리로 각 도큐멘트가 해당 쿼리에 적합할 확률을 베이지언 룰을 활용하여 계산하는데, 독립가정을 전제로 베이지언 룰을 이용하여, 비연관문서 집단에서 쿼의가 포함될 확률에 대한 연관 집단에 포함될 확률을 계산하여 문서를 찾는 모델링이다.

장점 : 문서들이 쿼의에 대하여 적합할 확률의 순서에 내림차순으로 랭크된다.
단점 : 비연관 문서와 연관 문서 집단의 초기 결과 집단을 가정해야만 한다.
불린 모델과 같이 가중치가 없어서 색인어의 빈도수에 대한 가중치를 부여할 수가 없다. (오카피 모델에서 적용)
색인어들에 대한 상호 독립 가정을 전제로 한다.

1) **Bayesian probability** is one of the different interpretations of the concept of probability and belongs to the category of evidential probabilities. The Bayesian interpretation of probability can be seen as an extension of propositional logic that enables reasoning with propositions whose truth or falsity is uncertain. To evaluate the probability of a hypothesis, the Bayesian probabilist specifies some prior probability, which is then updated in the light of new, relevant data.

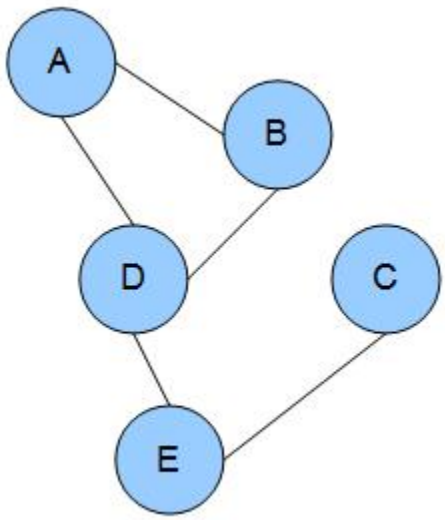
The Bayesian interpretation provides a standard set of procedures and formulae to perform this calculation. In contrast to interpreting probability as the "frequency" or "propensity" of some phenomenon, Bayesian probability is "a quantity that we assign theoretically, for the purpose of representing a state of knowledge".

* Markov random model

쿼리의 위치에서 시작을 해서 A 혹은 B로 도착할 때까지 매 번 랜덤하게 임의의 포인트로 이동하는 방식이다. 즉, 두 문서간의 유사성을 판단할 때 문서간의 직선거리를 생각하는 대신에 쿼리가 문서에 도착할 확률로 보는 모델이다.

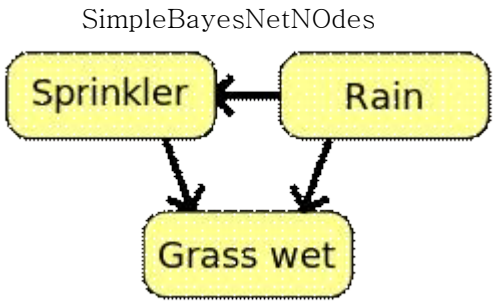
In the domain of physics and probability, a **Markov random field** (often abbreviated as MRF), Markov network or undirected graphical model is a set of random variables having a Markov property described by an undirected graph. A Markov random field is similar to a Bayesian network in its representation of dependencies; the differences being that Bayesian networks are directed and acyclic, whereas Markov networks are undirected and may be cyclic. Thus, a Markov network can represent certain dependencies that a Bayesian network cannot (such as cyclic dependencies); on the other hand, it can't represent certain dependencies that a Bayesian network can (such as induced dependencies).

Markov random field example



* **A Bayesian network**, Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model

a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic(비주기적) graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.



<제 11장 차세대 정보검색>

*** 시멘틱 웹**

시멘틱 웹은 W3C와 같은 국제적 표준 기궁에 의해 발전하고 있는 협업적 움직임이다. 이 표준은 웹 상에서 공동의 데이터 포맷을 발전시키고 있다. 웹페이지에 어의적 콘텐트를 포함시키도록 함으로써, 시멘틱 웹은 비정형화되고 유사-정형화된 도큐멘트를 "web of data"로 바꾼는 것을 목표로 하고 있다. 시멘틱 웹 구조는 W3c의 RDF를 근간으로 한다.
be processed by machines.

W3C에 따르면, "시멘틱 웹은 application, enterprise, and community boundaries 간에 데이터를 공유하고 재사용하도록 하는 공동의 프레임워크를 제공한다." 이 용어는 컴퓨터에 의해 처리될 수 있는 a web of data 용으로 Tim Berners-Lee에 의해 만들어졌다.

** Purpose

시멘틱 웹의 주요 목적은 이용자로 하여금 더욱 더 쉽게 정보를 찾고, 공유하고, 결합할 수 있도록 함으로써 현재의 웹을 발전시키는 것이다. 인간은 웹을 사용하여 "twelve months"의 에스토니아 번역본을 찾고, 도서관 책을 예약하고, 가장 값싼 DVD를 찾는 것과 같은 업무를 수행할 수 있다. 그렇지만, 컴퓨터는 인간의 지시없이는 모든 이러한 일을 할 수 없다. 왜냐하면 웹페이지들은 사람이 읽을 수 있도록 디자인 되었지 컴퓨터가 아니다. 시멘틱 웹은 컴퓨터에 의해 쉽게 해석될 수 있는 정보에 대한 비젼이다. 그러므로 컴퓨터는 컴퓨터는 웹에서 정보를 찾고, 결합하고, 관련된 행동을 하는 것을 포함하여 보다 많은 지루한 일을 할 수 있다.

1) The **Twelve Months** is a Greek fairy tale collected by Georgios A. Megas in Folktales of Greece.
A young and beautiful girl is sent into the cold forest in the winter to perform impossible tasks. She must get violets and apples in midwinter. She meets the 12 months personified who help her. The step mother and sister take the items, without a word of thanks. When the evil stepsister comes and is rude, they disappear, taking their fire, and leaving the stepsister cold and hungry.

시멘틱 웹은 원래 생각되었던 것처럼 컴퓨터가 의미를 근거로 복잡한 인간의 리퀘스트들을 이해하고 반응하는 하나의 시스템이다. 그러한 "이해"는 어의적으로 정형화된 적합한 정보원을 필요로 한다.

Tim Berners-Lee originally expressed the vision of the Semantic Web as follows:

I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize.

시멘틱 웹은 서로 다른 콘텐트, 정보 어플 및 시스템들의 통합체로 여겨진다. 이것은 출판, 브로깅, 그리고 많은 기타 분야에 대한 어플을 가지고 있다.

종종 "semantics", "metadata", "ontologies" and "Semantic Web"은 서로 다르게 사용된다. 특히, 이들 용어들은 연구자나 실무자에 의해 서로 다른 다양한 분야와 개념과 어플분야에 걸쳐서 일상의 용어처럼 사용된다. 따라서 시멘티 웹을 현실화할 수 있는 기술들의 현 상태와 관련해서 혼란이 존재한다.

** Limitations of HTML

전형적으로 컴퓨터의 많은 파일들은 크게 봐서 human readable documents와 machine readable data로 나눌 수 있다. 메일 메시지, 리포트, 브로쉬어 같은 도큐멘트는 인간에 의해 읽혀지며, 칼렌다, 주소책, 플레이리스트, 스프레드쉬트와 같은 데이터는 그것들을 볼 수 있고, 탐색할 수 있고 연결할 수 있는 응용 프로그램을 사용하여 제공된다.

현재, 웹은 이미지와 쌍방형 폼과 같은 멀티미디어 객체에 산재해 있는 텍스트를 코딩하기 위해 사용하는 마크업 규칙인 HTML로 작성된 도큐멘트를 주 근거로 삼고 있다. 메타데이터 태그들은 컴퓨터가 웹 페이지의 콘텐트를 범주화할 수 있는 방법을 제공한다.

예:

```
<meta name="keywords" content="computing, computer studies, computer" />
<meta name="description" content="Cheap widgets for sale" />
<meta name="author" content="John Doe" />
```

(웹브라우저 소프트웨어, 기타 유저 에이전트와 같은 데서 제공하는 HTML과 도구를 가지고 누구나 판매용 아이템을 리스트하고 있는 페이지를 만들어 제시할 수 있다. 이러한 카탈로그 페이지의 HTML은 “this document's title is 'Widget Superstore'"와 같은 간단한 document-level assertions을 만들 수 있지만, 그 HTML 자체는 예를 들어, 아이템 번호 X586172는 도매가격 199 유로인 Acme Gizmo라고 명백하게 주장하거나 그것이 소지자 제품이라고 밝힐 능력을 가지고 있지 않다. 그 보다는 HTML은 단지 텍스트 ”X586712“의 spam(범위)은 "Acme Gizmo" 그리고 "€199", etc. 근처에 위치해야하는 어떤 것이라고는 말할 수 있다. 그러므로 ”this is a catalog“라고 말할 수 있거나 또는 심지어 ”Acme Gizmo“는 타이틀의 일종이라거나 ”€199“는 가격이라는 것을 표현할 방법이 전혀 없다. 또한 이러한 여러 가지 정보를 함께 묶어서 그 페이지에 리스트 되어 있는 다른 아이템과 확실하게 구분해서 그 아이템을 설명할 방법도 전혀 없다.

어의적 HTML은 직접적으로 레이아우트의 내역을 세밀하게 지정하는 것보다는 의미를

수반하는 마크업의 전통적인 HTML 실무를 말한다. 예를 들어, 의 사용은 이탤릭체를 지정하는 <i>보다 “emphasis”를 나타낸다. 레이아우트 내역은 Cascading Style Sheets와 결합함으로써 브라우저에 적용된다. 그러나 이 실무는 판매용 아이템이나 가격과 같은 객체의 어의를 지정하는 데는 단점이 있다.

마이크로포맷은 HTML 구문식을 확대하여 사람, 조직, 사건, 제품과 같은 것에 대하여 기계가독형 어의적 마크업을 만들 수 있도록 하고 있다. 유사한 계획에는 RDFa, Microdata and Schema.org.가 포함되어 있다.

1) A **microformat** (sometimes abbreviated µF) is a web-based approach to semantic markup which seeks to re-use existing HTML/XHTML tags to convey metadata and other attributes in web pages and other contexts that support (X)HTML, such as RSS. This approach allows software to process information intended for end-users (such as contact information, geographic coordinates, calendar events, and similar information) automatically.

As of 2010, microformats allow the encoding and extraction of events, contact information, social relationships and so on. Established microformats such as hCard are published on the web more than alternatives like schema (microdata) and RDFa.

2) **RDFa** (or Resource Description Framework in Attributes) is a W3C Recommendation that adds a set of attribute-level extensions to HTML, XHTML and various XML-based document types for embedding rich metadata within Web documents. The RDF data-model mapping enables its use for embedding RDF subject-predicate-object expressions within XHTML documents. It also enables the extraction of RDF model triples by compliant user agents.

The RDFa community runs a wiki website to host tools, examples, and tutorials.

3) **Microdata** is a WHATWG HTML specification used to nest metadata within existing content on web pages. Search engines, web crawlers, and browsers can extract and process Microdata from a web page and use it to provide a richer browsing experience for users. Search engines benefit greatly from direct access to this structured data because it allows search engines to understand the information on web pages and provide more relevant results to users. Microdata uses a supporting vocabulary to describe an item and name-value pairs to assign values to its properties. Microdata is an attempt to provide a simpler[citation needed] way of annotating HTML elements with machine-readable tags than the similar approaches of using RDFa and microformats.

4) **Schema.org** is an initiative launched on 2 June 2011 by Bing, Google and Yahoo! (the operators of the then world's largest search engines) to “create and support a common set of schemas for structured data markup on web pages.” On 1 November Yandex (whose search engine is the largest one in Russia) joined the initiative. They propose using their ontology and Microdata in HTML5 to mark up website content with metadata about itself. Such markup can be recognized by search engine spiders and other parsers, thus gaining access to the meaning of the sites (see Semantic Web). The initiative started with a small number of formats, but the long term goal is to support a wider range of schemas. The initiative also describes an extension mechanism for adding additional properties. A mailing list is provided, for discussion of the initiative.

** Semantic Web solutions

시멘틱 웹은 더욱이 솔루션을 가지고 있다. 여기에는 특별하게 데이터를 디자인하기 위한 언어로 출판하는 것이 포함된다: Resource Description Framework (RDF), Web Ontology Language (OWL), and Extensible Markup Language (XML). HTML은 이들 간

의 도큐멘트와 링크들을 묘사한다. RDF, OWL 그리고 XML은 대조적으로 사람, 회의 또는 비행기 부품과 같은 임의의 사물을 묘사할 수 있다.

이런 기술들은 웹 도큐멘트의 콘텐트를 보완하거나 대체하는 descriptions를 제공하기 위하여 결합된다. 따라서, 콘텐트는 웹으로 접근이 가능한 데이터베이스에 저장된 기술 데이터처럼 또는 XML이 점재되어 있는 특히 XTHML에 있는, 또는 특히, XML이 점재되어 있는 XTHML로 된, 또는 별도로 저장된 cues를 제공하거나 레이아우트을 갖춘 XML로 된 도큐멘트 속의 markup처럼 manifest될 수도 있다. 기계가독형 descriptions는 content managers로 하여금 그 콘텐트에 의미를 다시 말해서, 그 콘텐트에 대하여 우리가 알고 있는 지식의 구조를 묘사할 수 있도록 추가할 수 있도록 한다. 이런 방식으로, 컴퓨터는 텍스트 대신에 지식을 처리하는데 인간의 연역법과 귀납법과 비슷한 처리과정을 사용함으로써 의미있는 결과를 얻을 수 있고 그럼으로써 컴퓨터를 사용하여 자동정보수집과 연구조사를 수행할 수 있다.

An example of a tag that would be used in a non-semantic web page:

```
<item>blog</item>
```

Encoding similar information in a semantic web page might look like this:

```
<item rdf:about="http://example.org/semantic-web/">Semantic Web</item>
```

Tim Berners-Lee는 Linked Data의 결과로 발생한 네트워크를 the the HTML-based World Wide Web과 대비하여 the Giant Global Graph라고 불렀다. Berners-Lee는 만일 과거가 document sharing이었다면, 미래는 data sharing이라고 주장하였다. “how”라는 질문에 대한 그의 답은 3가지였다;

1) a URL should point to the data.
2) anyone accessing the URL should get data back.
3) relationships in the data should point to additional URLs with data.

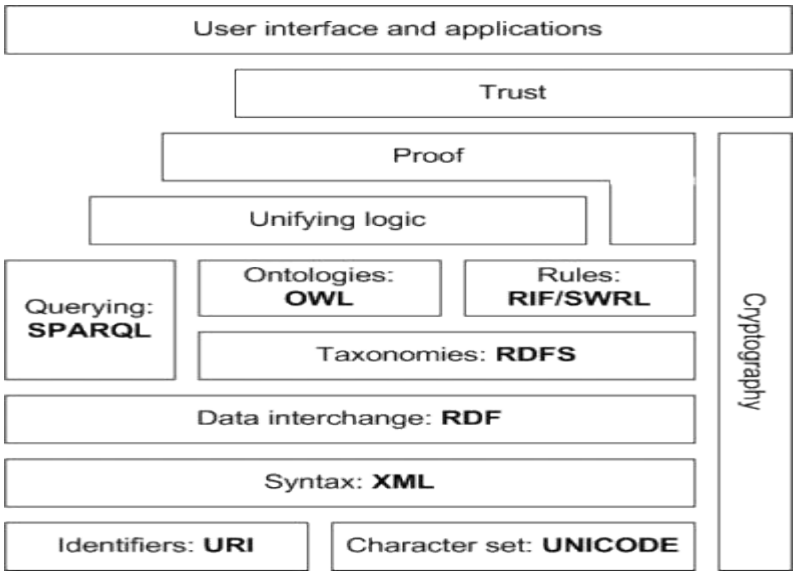
"Semantic Web" is sometimes used as a synonym for "Web 3.0", though each term's definition varies.

** Components

The term "Semantic Web" is often used more specifically to refer to the formats and technologies that enable it. The collection, structuring and recovery of linked data are enabled by technologies that provide a formal description of concepts, terms, and relationships within a given knowledge domain. These technologies are specified as W3C standards and include:

Resource Description Framework (RDF), a general method for describing information

RDF Schema (RDFS)
Simple Knowledge Organization System (SKOS)
SPARQL, an RDF query language
Notation3 (N3), designed with human-readability in mind
N-Triples, a format for storing and transmitting data
Turtle (Terse RDF Triple Language)
Web Ontology Language (OWL), a family of knowledge representation languages
Rule Interchange Format (RIF), a framework of web rule language dialects supporting rule interchange on the Web



Semantic-web-stack

The Semantic Web Stack illustrates the architecture of the Semantic Web. The functions and relationships of the components can be summarized as follows:

XML provides an elemental syntax for content structure within documents, yet associates no semantics with the meaning of the content contained within. XML is not at present a necessary component of Semantic Web technologies in most cases, as alternative syntaxes exists, such as Turtle. Turtle is a de facto standard, but has not been through a formal standardization process.

XML Schema is a language for providing and restricting the structure and content of elements contained within XML documents.

RDF is a simple language for expressing data models, which refer to objects ("web resources") and their relationships. An RDF-based model can be represented

in a variety of syntaxes, e.g., RDF/XML, N3, Turtle, and RDFa. RDF is a fundamental standard of the Semantic Web.

RDF Schema extends RDF and is a vocabulary for describing properties and classes of RDF-based resources, with semantics for generalized-hierarchies of such properties and classes.

OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes.

SPARQL is a protocol and query language for semantic web data sources.

RIF is the W3C Rule Interchange Format. It's an XML language for expressing Web rules which computers can execute. RIF provides multiple versions, called dialects. It includes a RIF Basic Logic Dialect (RIF-BLD) and RIF Production Rules Dialect (RIF PRD).

Current state of standardization; Well-established standards:

Unicode
Uniform Resource Identifier
XML
RDF
RDFS
SPARQL
Web Ontology Language (OWL)
Rule Interchange Format (RIF)

Not yet fully realized:

Unifying Logic and Proof layers

**** 온톨로지**

온톨로지란 공식적으로 도메인에 들어 있는 한 세트의 개념에 대한 유형, 성질, 상호연관성을 나타내기 위하여 공유된 어휘를 사용하여 이러한 개념들을 지식으로 표현한다.

온톨로지는 정보를 조직하는 정형화된 프레임워크이며, artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture에서 자신들 분야

에 대한 지식 표현의 형태처럼 사용되고 있다. 도메인 온톨로지의 제작은 또한 기업조직의 프레임워크에 대한 정의 및 이용에 있어 기본적인 것이다.

** Components

현대의 온톨로지는 많은 구조적 유상성을 공유하고 있다, 그것을 표현하는 언어와 상관없이. 대부분의 온톨로지는 individuals (instances), classes (concepts), attributes, and relations로 묘사된다.

*** Common components of ontologies include:

Individuals: instances or objects (the basic or "ground level" objects)
Classes: sets, collections, concepts, classes in programming, types of objects, or kinds of things
Attributes: aspects, properties, features, characteristics, or parameters that objects (and classes) can have
Relations: ways in which classes and individuals can be related to one another
Function terms: complex structures formed from certain relations that can be used in place of an individual term in a statement
Restrictions: formally stated descriptions of what must be true in order for some assertion to be accepted as input
Rules: statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form
Axioms: assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic. In those disciplines, axioms include only statements asserted as a priori knowledge. As used here, "axioms" also include the theory derived from axiomatic statements
Events: the changing of attributes or relations

온톨로지는 일반적으로 온톨로지 언어(예: OWL)을 사용하여 코드화된다.

** Types of ontologies

Domain ontology

도메인 온톨로지 또는 도메인-지정 온톨로지는 해당 분야의 일부분을 대표하는 특수한 도메인을 모델화 한다. 그 같은 도메인에 적용된 용어들의 특별한 의미들은 도메인 온톨로지에 의해 제공된다. 예를 들어, card 단어는 많은 서로 다른 의미를 가지고 있다. poker 도메인에 대한 온톨로지에서 그 단어에서 “playing card”라는 의미로 모델화될 수 있다. 반면에 컴퓨터 하드웨어 도메인에 대한 온톨로지에서는 “punched card” 그리고 “video card” 의미로 모델화 될 수 있다.

도메인 온톨로지가 매우 특별하고 종종 절충적인(eclectic) 방법으로 개념을 표현하므로, 이것들은 종종 호환성이 없기도 한다. 도메인 온톨로지에 의존하는 시스템들이 늘어남으로써, 이것들은 종종 더 만은 일반적인 표현으로 도메인 온톨로지를 통합할 필요가 있다. 이것이 온톨로지 디자이너에게는 하나의 도전이다. 동일한 도메인에서 서로 다른 온톨로지들은 서로 다른 언어, 서로 다른 의도, 그리고 그 도메인에 대한 서로 다른 인식(문화적 배경, 교육, 이데올로기, 등을 근거로) 으로 인하여 발생한다.

현재에, 일반 foundation ontology에서 발전하지 못한 온톨로지의 통합은 주로 수작업으로 이루어지며, 시간 소모적이고 비용이 많이 든다. 도메인 온톨로지의 요소들에 대한 의미를 특정하는 한 세트의 기본적인 요소들을 제공하기 위하여 동일한 foundation ontology를 사용하는 도메인 온톨로지들은 자동으로 통합될 수 있다. 온톨로지 통합에 대한 일반화된 기술들에 대한 연구가 이루어지고 있지만, 이 분야의 연구는 아직까진 주로 이론적이다.

1) In information science, an **upper ontology (also known as a top-level ontology or foundation ontology)** is an ontology (in the sense used in information science) which describes very general concepts that are the same across all knowledge domains. An important function of an upper ontology is to support very broad semantic interoperability between a large number of ontologies which are accessible ranking "under" this upper ontology. As the rank metaphor suggests, it is usually a hierarchy of entities and associated rules (both theorems and regulations) that attempts to describe those general entities that do not belong to a specific problem domain.

Library classification systems predate these upper ontology systems. Though library classifications organize and categorize knowledge using general concepts that are the same across all knowledge domains, neither system is a replacement for the other.

Upper ontology

upper ontology (or foundation ontology)는 다양한 범위의 도메인 온톨로지 간에 적용할 수 있는 일반 객체들에 대한 모델이다. 이것은 다양하고 적합한 도메인 세트들에서 사용되는 용어들과 연관된 객체 묘사를 포함하고 있는 core glossary를 사용한다.

사용이 가능한 여러 가지의 표준화된 upper ontologies가 존재한다: 예; BFO, Dublin Core, GFO, OpenCyc/ResearchCyc, SUMO, and DOLCE. 어떤 사람들에게는 upper ontology라고 여겨지는 WordNet는 엄격하게 말해서 온톨로지가 아니다. 그렇지만, 이것은 도메인 온톨로지를 배우기 위한 언어적 도구로 사용되고 있다.

1) The **Basic Formal Ontology (BFO)** is a formal ontological framework developed by Barry Smith and his associates that consists in a series of sub-ontologies at different levels of granularity. The ontologies are divided into two varieties: continuant (or snapshot) ontologies, comprehending continuant entities such as three-dimensional enduring objects, and occurrent ontologies, comprehending processes conceived as extended through (or as spanning) time. BFO thus incorporates both three-dimensionalist and four-dimensionalist perspectives on reality within a single framework. Interrelations are defined between the two types of ontologies in a way which gives BFO the facility to deal with both static/spatial and dynamic/temporal features of reality. Each continuant ontology is an inventory of all entities existing at a time. Each occurrent ontology is an inventory (processory) of all the processes unfolding through a given interval of time. Both types of ontology serve as basis for a series of sub-ontologies, each of which can be conceived as a window on a certain portion of reality at a given level of granularity.

2) The **general formal ontology (GFO)** is an upper ontology integrating processes and objects. GFO has

been developed by Heinrich Herre, Barbara Heller and collaborators (research group Onto-Med) in Leipzig. Although GFO provides one taxonomic tree, different axiom systems may be chosen for its modules. In this sense, GFO provides a framework for building custom, domain-specific ontologies. GFO exhibits a three-layered meta-ontological architecture consisting of an abstract top level, an abstract core level, and a basic level.

3) **Cyc** is an artificial intelligence project that attempts to assemble a comprehensive ontology and knowledge base of everyday common sense knowledge, with the goal of enabling AI applications to perform human-like reasoning.

#OpenCyc

The latest version of OpenCyc, 4.0, was released in June 2012. OpenCyc 4.0 includes the entire Cyc ontology containing hundreds of thousands of terms, along with millions of assertions relating the terms to each other; however, these are mainly taxonomic assertions, not the complex rules available in Cyc. The knowledge base contains 239,000 concepts and 2,093,000 facts and can be browsed on the OpenCyc website.

ResearchCyc

In addition to the taxonomic information contained in OpenCyc, ResearchCyc includes significantly more semantic knowledge (i.e., additional facts) about the concepts in its knowledge base, and includes a large lexicon, English parsing and generation tools, and Java based interfaces for knowledge editing and querying.

4) The **Suggested Upper Merged Ontology or SUMO** is an upper ontology intended as a foundation ontology for a variety of computer information processing systems.

SUMO originally concerned itself with meta-level concepts (general entities that do not belong to a specific problem domain), and thereby would lead naturally to a categorization scheme for encyclopedias. It has now been considerably expanded to include a mid-level ontology and dozens of domain ontologies.

5) **DOLCE and DnS**

Developed by Nicola Guarino and his associates at the Laboratory for Applied Ontology (LOA), the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) is the first module of the WonderWeb foundational ontologies library. As implied by its acronym, DOLCE has a clear cognitive bias, in that it aims at capturing the ontological categories underlying natural language and human common sense.

DnS (Descriptions and Situations), developed by Aldo Gangemi (STLab, Rome), is a constructivist ontology that pushes DOLCE's descriptive stance even further. DnS does not put restrictions on the type of entities and relations that one may want to postulate, either as a domain specification, or as an upper ontology, and it allows for context-sensitive 'redescriptions' of the types and relations postulated by other given ontologies (or 'ground' vocabularies). The current OWL encoding of DnS assumes DOLCE as a ground top-level vocabulary. DnS and related modules also exploit 'CPs' (Content ontology design Patterns), which provide a framework to annotate 'focused' fragments of a reference ontology (i.e., the parts of an ontology containing the types and relations that underlie 'expert reasoning' in given fields or communities). The combination of DOLCE and DnS has been used to build a planning ontology known as DDPO (DOLCE+DnS Plan Ontology).

Both DOLCE and DnS are particularly devoted to the treatment of social entities, such as e.g. organizations, collectives, plans, norms, and information objects. It has also been used to study and create domain ontologies for sovereign states, geopolitical boundaries, and the agentivity of social entities. The DOLCE-2.1-Lite-Plus OWL version, including a number of DnS-based modules, has been and is being applied to several ontology projects.

6) **WordNet**, a freely available database originally designed as a semantic network based on psycholinguistic principles, was expanded by addition of definitions and is now also viewed as a dictionary. It qualifies as an upper ontology by including the most general concepts as well as more specialized concepts, related to each other not only by the subsumption relations, but by other semantic relations as well, such as part-of and cause. However, unlike Cyc, it has not been formally axiomatized so as to make the logical relations between the concepts precise. It has been widely used in Natural language processing research.

7) **Unified Foundation Ontology (UFO)**

The Unified Foundational Ontology (UFO), developed by Giancarlo Guizzardi and associates, incorporating developments from GFO, DOLCE and the Ontology of Universals underlying OntoClean in a single coherent foundational ontology. The core categories of UFO (UFO-A) have been completely formally characterized in Giancarlo Guizzardis Ph.D. thesis and further extended at the Ontology and Conceptual Modelling Research Group (NEMO) in Brazil with cooperators from Brandenburg University of Technology (Gerd Wagner) and Laboratory for Applied Ontology (LOA). UFO-A has been employed to analyze structural conceptual modeling constructs such as object types and taxonomic relations, associations and relations between associations, roles, properties, datatypes and weak entities, and parthood relations among objects.

8) **IDEAS**

The upper ontology developed by the IDEAS Group is higher-order, extensional and 4D. It was developed using the BORO Method. The IDEAS ontology is not intended for reasoning and inference purposes; its purpose is to be a precise model of business.

9) **UMBEL**

Upper Mapping and Binding Exchange Layer (UMBEL) is an ontology of 28,000 reference concepts that maps to a simplified subset of the OpenCyc ontology, that is intended to provide a way of linking the precise OpenCyc ontology with less formal ontologies. It also has formal mappings to Wikipedia, DBpedia, PROTON and GeoNames. It has been developed an maintained as open source by Structured Dynamics.

Hybrid ontology

The Gellish ontology is an example of a combination of an upper and a domain ontology.

1) **Gellish** is a formal language that is natural language independent, although its concepts have 'names' and definitions in various natural languages. Any natural language variant, such as Gellish Formal English is a controlled natural language. Information and knowledge can be expressed in such a way that it is computer-interpretable, as well as system-independent and natural language independent. Each natural language variant is a structured subset of that natural language and is suitable for information modeling and knowledge representation in that particular language. All expressions, concepts and individual things are represented in Gellish by (numeric) Unique Identifiers (Gellish UID's). This enables a software to automatically generate expressions that are created in one formal natural language into any other formal natural language. From a data modeling perspective, Gellish is a universal and extendable conceptual data model that also includes domain-specific terminology and definitions. Therefore, it can also be called a semantic data model. The accompanying Gellish modeling methodology thus belongs to the family of semantic modeling methodologies.

** Ontologies as a specification mechanism

공식적으로 표현된 지식은 개념화를 근거로 한다: 관심대상분야 및 이것들 간에 유지되는 관계 속에 존재한다고 추정하는 객체, 개념, 그리고 기타 엔티티. 개념이란 우리가 어떤 목적을 위해 표현하고자 하는 세상의 추상적이고 단순화된 견해이다. 모든 지식 베이스, 지식-베이스드 시스템, 또는 지식-수준 에이전트는 노골적이든 아니든 어떤 개념화 작업을 하고 있다.

온톨로지는 개념화에 대한 명확한 스펙이다. 이 용어는 온톨로지는 존재(Existence)에 대한 체계적인 근거인 철학에서부터 유래되었다. 예를 들어, AI 시스템에서 “exists”하는 것은 표현될 수 있는 무엇이다. 어떤 도메인의 지식을 선언적 형식론(declarative formalism)으로 표현할 때, 표현할 수 있는 객체의 세트를 universe of discourse라고 부른다. 이러한 객체의 세트와 그것들간에 묘사 가능한 관계는 지식-의존 프로그램에 지식을 표현하는데 사용하는 표현 어휘에 영향을 끼친다. 그러므로 AI 환경에서, 우리는 한 세트의 표현 가능한 용어들을 정의함으로써어떤 프로그램의 온톨로지를 묘사할 수 있다. 그 같은 온톨로지에서, 정의들은 universe of discourse (e.g., classes, relations, functions, or other objects)에 있으면서 그 이름이 의미하는 것을 묘사하고 인간이 읽을 수 있는 텍스트와 이러한 용어들의 해석과 잘 짜여진 용도를 제한하는 공식적인 원리를 엔티티의 이름과 결합시킨다. 공식적으로 온톨로지란 논리적 이론의 공술(statament)이다.

우리는 한 세트의 에이전트를 위하여 온톨로지 약정(commitments)을 묘사하기 위하여 일반 온톨로지를 사용한다. 왜냐하면 그것들이 전세계적으로 공유된 이론에 따라 반드시 운영할 필요없이 a domain of discourse에 대해 통신할 수 있기 때문이다. 우리가 말하는 것은 어떤 에이전트는 만일 관찰할 수 있는 행동들이 온톨로지에 있는 정의와 일치한다면 온톨로지에 헌신(commit to)하고 있다는 것이다. 온톨로지 약정의 아이디어는 Knowledge-Level 견해에 근거하고 있다. Knowledge-Level은 그 에이전트에 의해 내부적으로 사용된 symbol-level representation과는 따로 그 에이전트의 지식을 묘사하는 수준이다. 지식은 그것들의 행동을 관찰함으로써 에이전트의 것으로 추정한다; 만일 그것이 마치 정보를 가지고 그것의 목표를 달성하기 위하여 합리적으로 행동하고 있다면 그 에이전트는 어떤 것을 “knows” 한다. 에이전트들의 “actions”는 - 지식의존 서버와 지식의존형 시스템을 포함하여 - logical assertions (tell), 그리고 posing queries (ask)에 의해 클라이언트와 에이전트가 상호작용하는 a tell and ask functional interface를 통해 알 수 있다.

실용적으로 말해서, 일반 온톨로지는 에이전트들 간에 queries와 assertions가 상호 교환되는 어휘를 정의한다. 온톨로지 약정은 명확하게 일치된 방식으로 공유된 어휘를 사용하는 협정문이다. 어휘를 공유하는 에이전트들은 시식 베이스를 공유할 필요가 없다; 각각은 남이 모르는 어떤 일을 안다. 그리고 온톨로지에 헌신하는 에이전트는 공유하는 어휘에 공식화될 수 있는 모든 쿼리에 대답할 것을 요구 받지 않는다.

간단하게 말해서, 공동 온톨로지에 대한 책임은 온톨로지에서 정의한 어휘를 사용하는 queries and assertions(a predicate: a true-false statement)와 관련해서 완벽하지는 않지만, 일관성을 보증하는 것이다.

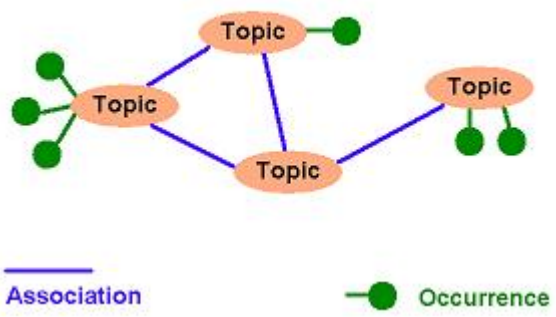
*** 토픽맵(Topic Maps)**

토픽 맵은 지식을 표현하고 상호교환할 수 있는 표준이며 정보의 발견성(findability)을 강조한다. 토픽 맵은 서로 다른 정보원의 복수 색인들을 통합하기 위하여,

back-of-the-book index structures를 표현하는 방법으로 1990년대 말에 최초로 개발되었다. 그렇지만, 개발자들은 신속하게 약간의 추가적인 일반화와 함께 잠재적으로 널리 응용될 수 있는 메타-모델을 창조할 수 있다는 것을 깨달았다. 이것은 공식적으로 ISO/IEC 13250:2003 표준이다.

토픽 맥에서 정보를 표현하는 방법:

topics, representing any concept, from people, countries, and organizations to software modules, individual files, and events,
associations, representing hypergraph relationships between topics, and
occurrences representing information resources relevant to a particular topic.



Topic Map Key Concepts

토픽 맵은 많은 부분에서 concept maps 및 mind maps과 비슷하다. 비로 토픽 맵만이 표준이지만. 토픽 맵은 시멘틱 웹 기술의 한 형태이며 어떤 작업은 W3C's RDF/OWL/SPARQL family of semantic web standards 와 ISO's family of Topic Maps standards 간에 상호 교환적으로 이루어지고 있다.

1) A **concept map** is a diagram that depicts suggested relationships between concepts. It is a graphical tool that designers, engineers, technical writers, and others use to organize and structure knowledge.

A concept map typically represents ideas and information as boxes or circles, which it connects with labeled arrows in a downward-branching hierarchical structure. The relationship between concepts can be articulated in linking phrases such as causes, requires, or contributes to.

2) A **mind map** is a diagram used to visually outline information. A mind map is often created around a single word or text, placed in the center, to which associated ideas, words and concepts are added. Major categories radiate from a central node, and lesser categories are sub-branches of larger branches. Categories can represent words, ideas, tasks, or other items related to a central key word or idea.

3) **Topincs** is a software for rapid development of web databases and web applications. It is based on LAMP and the semantic technology Topic Maps. A Topincs web database makes information accessible through browsing very much like a Wiki. Editing a page on a subject is done through forms rather than markup editing. A web database can be tailored into a web application to provide specific user groups a contextualized approach to the data.

많은 방식에서 토픽 맵의 어의적 표현성은 RDF의 것과 유사하지만 중요한 차이들은 토픽 맵이 첫째, a template of topics, associations and occurrences를 제공하는데 있어서 보다 높은 차원의 어의적 추론을 제공하지만 RDF는 하나의 관계에 의해 링크된 두 가지의 arguments에 대한 a template만을 제공하며, 둘째, 토픽 맵은 어떠한 수의 노드간에도 n-ary 관계(hypergraphs)를 허용하지만 RDF는 triplets로 제한한다.

** Ontology and merging

Topics, associations, and occurrences은 모두가 타입을 갖출 수 있는데, 그 type이란 토픽 맵의 하나이상의 제작자에 의해 정의되어야 한다. 허용된 타입의 정의는 그 토픽 맵의 온톨로지로 알려지게 된다. 토픽 맵은 분명하게 말해서 복수의 토픽 또는 토픽 맵 들간의 아이덴티티의 통합 개념을 지원한다. 더구나, 온톨로지가 토픽 맵 그 자체이기 때문에, 그것들은 또한 다양한 소스에서 나온 정보들을 연관된 새로운 토픽 맵으로 자동적으로 통합할 수 있다. subject identifiers (URIs given to topics) 와 PSIs (Published Subject Indicators)와 같은 특성들은 사로 다른 분류방법들 간에 통합을 조정하기 위하여 사용된다. Scoping on names(이름에 대한 범위지정)은 서로 다른 정보원에 의해 특별한 토픽이 주어진 여러 가지 이름을 조직하는 방법을 제공한다.

** Current standard

The most recent work standardizing Topic Maps (ISO/IEC 13250) is taking place under the umbrella of the ISO/IEC JTC1/SC34/WG3 committee (ISO/IEC Joint Technical Committee 1, Subcommittee 34, Working Group 3 - Document description and processing languages - Information Association).

The Topic Maps (ISO/IEC 13250) reference model and data model standards are defined in a way that is independent of any specific serialization or syntax.

TMRM Topic Maps - Reference Model
TMDM Topic Maps - Data Model***

*** 토픽맵의 정의:국립중앙도서관 도서관연구소 웹진 15호 도서관용어해설**

토픽맵(Topic Maps)은 차세대 웹인 시맨틱 웹 구현을 위한 등장한 개념체계 인 온톨로지를 표현하기 위한 전용 언어 중 하나이다. 온톨로지를 구축하려 면 개념화 구조를 정확하게 표현해야 하는데, XML은 개념의 특성이나 상호관계를 표현하는 데는 미흡하므로 이를 대신하여 RDF/RDFS, DAML, OWL, 토픽맵 등 온톨로지 전용 언어가 개발되었다.

그 중 토픽맵은 국제표준화기구(ISO)에서 제정된 온톨로지 생성 언어로 W3C의 OWL과 상호 보완 및 경쟁 관계에 있으며 원래 용어집, 시소러스, 색인집등 용어의 의미적 구조를 다루는 목적으로 되었다. 그러나 현재는 정보자원을 의미적 관계를 표현하고 의미적 검색이 가능하도록 하는 시맨팀 웹의 핵심기술로 인정받고 있다.

토픽맵 관련 표준으로 ISO/IEC 13250 국제 표준이 있다. 처음에는 토픽맵 표준 규격으로 SGML(Standard Generalized Markup Language)구조와 HyTM 언어였으나 2001년 TopicMaps.org에서 개발한 XTM(XML TopicMaps)으로 통합되면서 현재는 XTM 1.0이 표준 규격이 되었다. 토픽맵 관련 표준은 아래 표와 같다(박여삼 외 2008).

- Fin -